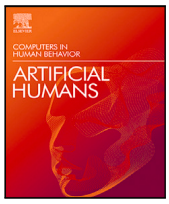
Computers in Human Behavior: Artificial Humans 7 (2026) 100246

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans



Research paper

# Understanding successful human–AI teaming: The role of goal alignment and AI autonomy for social perception of LLM-based chatbots

Christiane Attig [a,b,*], Luisa Winzer [a], Tim Schrills [a], Mourad Zoubir [a], Maged Mortaga [a], Patricia Wollstadt [b], Christiane Wiebel-Herboth [b], Thomas Franke [a]

[a] University of Lübeck, Ratzeburger Allee 160, 23562, Lübeck, Germany
[b] Honda Research Institute Europe GmbH, Carl-Legien-Straße 30, 63703, Offenbach, Germany

ARTICLE INFO

*Keywords:*
Large language models
User interaction
Human–AI teaming
Human-autonomy teaming
Cooperation
Collaboration
Social perception

ABSTRACT

LLM-based chatbots such as ChatGPT support collaborative, complex tasks by leveraging natural language processing to provide skills, knowledge, or resources beyond the user's immediate capabilities. Joint activity theory suggests that effective human–AI collaboration, however, requires more than responding to verbatim prompts – it depends on aligning with the user's underlying goal. Since prompts may not always explicitly state the goal, an effective LLM should analyze the input to approximate the intended objective before autonomously tailoring its response to align with the user's goal. To test these assumptions, we examined the effects of LLM-based chatbots' autonomy and goal alignment on multiple social perception metrics as key criteria for successful human–AI teaming (i.e., perceived cooperation, warmth, competence, traceability, usefulness, and trustworthiness). We conducted a scenario-based online experiment where participants ($N$ = 182, within-subjects design) were instructed to collaborate with four different versions of an LLM-based chatbot. The overall goal of the study scenario was to detect and correct erroneous information in short encyclopedic articles, representing a prototypical knowledge work task. Four custom-instructed chatbots were provided in random order: three chatbots varying in goal alignment and AI autonomy and one chatbot serving as a control condition not fulfilling user prompts. Repeated-measures ANOVAs demonstrate that a chatbot which is able to excel in goal alignment by autonomously going beyond verbatim user prompts is perceived as superior compared to a chatbot that adheres rigidly to user prompts without adapting to implicit objectives and chatbots that fail to meet the explicit or implicit user goal. These results support the notion that AI autonomy is only perceived as beneficial as long as user goals are not undermined by the chatbot, emphasizing the importance of balancing user and AI autonomy in human-centered design of AI systems.

## 1. Introduction

People communicating with machines as if they were people is not a new phenomenon (Nass, Steuer, & Tauber, 1994; Weizenbaum, 1966). However, with disruptive advances in the field of natural language processing (NLP) and generative artificial intelligence (GenAI), machines are now not only capable of talking or writing back, but also of engaging in extended, multi-turn conversations (Bansal, Chamola, Hussain, Guizani, & Niyato, 2024) that can feel human-like (Bonny & Wynne, 2024; Sun, Reani, Luo, & Bao, 2024). Most prominently, conversational AI based on large language models (LLM) such as ChatGPT (Brown et al., 2020) displays a large range of conversational capabilities known from human-human communication (Bansal et al., 2024), for instance, context-sensitivity (Casheekar, Lahiri, Rath, Prabhakar, & Srinivasan, 2024), memory mechanisms (Zhou, Zhu, Jin, & Dou, 2024), or affective tone (Wester, De Jong, Pohl, & Van Berkel, 2024). Due to its broad range of applications (Bansal et al., 2024; Joublin et al., 2023; Raiaan et al., 2024), ChatGPT is used by millions of users for a variety of tasks (Wolf & Maier, 2024), for instance, self-regulated learning (Dahri et al., 2024), code generation (Jin, Wang, Pham, & Hemmati, 2024), or text editing (Abdelhafiz et al., 2024). Surveys on user experience have highlighted ChatGPT's high perceived usefulness and ease of use in different application contexts (Choudhury & Shamszare, 2023; Nikou, Guliya, Van Verma, & Chang, 2024; Wolf & Maier, 2024).

Given the versatile, human-like conversational competence of systems like ChatGPT, social dynamics become a major component of

* Corresponding author at: University of Lübeck, Ratzeburger Allee 160, 23562, Lübeck, Germany.
*E-mail addresses:* christiane.attig@honda-ri.de (C. Attig), lu.winzer@uni-luebeck.de (L. Winzer), tim.schrills@uni-luebeck.de (T. Schrills), m.zoubir@uni-luebeck.de (M. Zoubir), maged.mortaga@uni-luebeck.de (M. Mortaga), patricia.wollstadt@honda-ri.de (P. Wollstadt), christiane.wiebel@honda-ri.de (C. Wiebel-Herboth), thomas.franke@uni-luebeck.de (T. Franke).

https://doi.org/10.1016/j.chbah.2025.100246
Received 3 March 2025; Received in revised form 23 November 2025; Accepted 10 December 2025
Available online 19 December 2025

the interaction of humans and LLM-based chatbots (Cohn et al., 2024; Heppner, Schiffhauer, & Seelmeyer, 2024). For instance, with growing perceived chatbot intelligence, humans show a greater tendency to be polite toward chatbots (Yuan, Su, & Li, 2024). It has also been shown that some users perceive the interaction with conversational AI as resembling human peer bonding, particularly when the agent is strongly anthropomorphized (Tschopp, Gieselmann, & Sassenberg, 2023). However, it has been argued that the tendency to anthropomorphize AI agents holds risks for user autonomy and adequate trust (Brandtzaeg, You, Wang, & Lao, 2023; Zhan, Xu, & Sarkadi, 2023). Hence, to avoid unwarranted trust, overreliance, and loss of human self-determination, while maintaining the high versatility and usability that come along with natural language use, human-centered design of conversational AI is key (Valdez et al., 2024; Fui-Hoon Nah, Zheng, Cai, Siau, & Chen, 2023) and designing for human–AI alignment has been identified as a core task for enabling human-centered AI (Goyal, Chang, & Terry, 2024; Sucholutsky et al., 2023).

The alignment of users' needs and values with the conversational AI's functionality plays a particularly important role in collaborative teamwork that entails interdependency between team partners (Bradshaw, Feltovich, & Johnson, 2017; Klein, Feltovich, Bradshaw, & Woods, 2005). Research on how users perceive alignment with conversational AI in collaborative teamwork and how (mis)alignment impacts user experience has, however, only rarely been conducted yet (e.g., Goyal et al., 2024; Li & Lee, 2022). With the present research, we explore the role of human–AI alignment for perceived teamwork success. Particularly, we explore the interplay between AI autonomy and goal alignment, with our main hypothesis being that AI autonomy is only beneficial if it is aligned with user goals (i.e., if the conversational AI exceeds the verbatim user prompt in an autonomous way that serves the general user goal). With respect to outcome variables, we focus on social perception, traceability, usefulness, and trustworthiness of conversational AI – variables that go beyond mere task success – to account for the multifaceted nature of interaction with AI agents due to their versatility and varying roles (Jones, Xu, Li, & Scherer, 2024; Kim, Molina, Rheu, Zhan, & Peng, 2023). To this end, we draw from theories and research on human-autonomy teaming (O'Neill, McNeese, Barron, & Schelble, 2022), joint activity (Klein et al., 2005), social cognition (i.e., Stereotype Content Model; Fiske, Cuddy, & Glick, 2007), and AI traceability (Schrills & Franke, 2023) to derive and test hypotheses regarding the impact of varying levels of AI autonomy and goal alignment as aspects of human–AI alignment on perceived teamwork success.

In an online experiment with a 2 × 2 within-subjects design, $N$ = 182 participants interacted with four versions of an LLM-based chatbot designed to correct erroneous encyclopedic articles. The chatbots varied in goal alignment and autonomy. The results supported the hypothesis that an autonomous chatbot aligned with the user's goal produces more favorable subjective outcomes than either a non-autonomous chatbot or autonomous behavior that fails to serve the intended goal, with nuanced differences for more cognitive or socioemotional outcomes. The present study's results underscore the benefit of LLM chatbots acting proactively by inferring user goals that are not always explicitly articulated in the prompt.

In the following, we outline the theoretical background on how users perceive AI systems as autonomous teammates and why goal alignment is central to a beneficial user experience. We then summarize the key assumptions and empirical findings on how AI autonomy and goal alignment interact to shape selected user experience variables, forming the basis of our hypotheses. This is followed by the study's methodology and the presentation of results. The article concludes with a thorough discussion of theoretical as well as practical implications for advancing human–AI collaboration and guiding human-centered AI design. Finally, we address the study's limitations and outline directions for future research.

## 2. Background

### 2.1. Conversational AI as an autonomous teammate in collaborative tasks

Due to the large variety of applications of conversational AI, there is not solely one role of the artificial agent but many, depending both on the task/application characteristics and the user's perception (Du, Li, Jiang, Xu, & Zhou, 2025; Kim et al., 2023; Lindgren, 2024). From the perspective of human-autonomy teaming, AI roles can be classified according to the level of AI autonomy (O'Neill et al., 2022), that is, different levels of perceived AI autonomy are linked to different perceived AI roles (Müller, Graf, Ellwart, & Antoni, 2025). Take for example ChatGPT being used for correcting grammar errors. Referring to O'Neill et al. (2022), one could argue that ChatGPT has no autonomy, because it will not check the grammar of a provided text unless it is prompted to do so. Consequently, in this case, ChatGPT will likely be viewed as a tool (O'Neill et al., 2022). However, ChatGPT might, in the same response, point out spelling errors spontaneously (i.e., exceeding the verbatim user request), which would likely enhance perceived agent autonomy, increasing the probability that users will perceive it as a teammate. Furthermore, ChatGPT's natural language use facilitates the experience of anthropomorphism (Bonny & Wynne, 2024). Thus, the perception and ascription of the role of ChatGPT is probabilistic, dimensional, and context- as well as user-specific, and, due to the natural language use, prone to be anthropomorphized.

Importantly, perceptions of an AI's role shape both expected and actual interaction. When an AI agent is viewed as an autonomous team partner, users are more inclined to attribute human-like capabilities (e.g., intentionality, communication) to it (Epley, Waytz, & Cacioppo, 2007) and to respond to it socially (Go & Sundar, 2019; Tschopp et al., 2023). At the same time, increasing an AI's autonomy can diminish users' sense of control, reducing users' autonomy need as well as overall task satisfaction (Frank & Otterbring, 2024). Aligning an AI system's functional design with users' role perceptions is therefore crucial to avoid mismatches that compromise human autonomy, inflate expectations, and impair human–AI collaboration (Shneiderman, 2020). Hence, to understand how AI autonomy shapes perceived roles and the expectations that follow, social perception must be taken into account.

A valuable theoretical basis is the Stereotype Content Model (SCM; Fiske et al., 2007), which postulates that perceived warmth and competence are fundamental dimensions of social perception. According to the SCM, perceived warmth determines the valence of judgment whereas competence determines the strength of judgment and perceived ability. In other words, an agent can be perceived as having good or bad intentions (= level of warmth) and being able or not to realize intentions (= level of competence). Growing empirical evidence shows that artificial agents are not judged merely as tools but are systematically assessed along core dimensions of social perception, including warmth, sociability, competence, morality, cooperativity, and teaming perception (Attig, Wollstadt, Schrills, Franke, & Wiebel-Herboth, 2024; Bretschneider, Mandl, Strobel, Asbrock, & Meyer, 2022; Harris-Watson, Larson, Lauharatanahirun, DeChurch, & Contractor, 2023; Khadpe, Krishna, Fei-Fei, Hancock, & Bernstein, 2020; Mandl et al., 2023; McKee, Bai, & Fiske, 2023).

Importantly, perceptions of these fundamental dimensions shape user experience, with higher perceived warmth and competence fostering trust and acceptance (Li et al., 2024). More specifically, investigating social perception of intelligent personal assistants, higher autonomy was positively linked to both warmth and competence ratings and particularly competence predicted continuance usage intention (Hu, Lu, Pan, Gong, & Yang, 2021). In an interactive game study with agents based on reinforcement learning (McKee et al., 2023), participants rated agents with a prosocial game strategy as warmer, but similarly competent as agents with an egoistic game strategy. Moreover, warmth

was positively associated with partner choice, and social perception predicted partner choice better than performance (McKee et al., 2023).

In sum, these findings indicate that perceived AI autonomy shapes how users classify an AI agent's role, and these role perceptions, filtered through core dimensions of social cognition, ultimately guide subjective interaction outcomes, such as trust and acceptance. However, as we will argue in the following section, AI autonomy alone is not sufficient: Trustworthiness and acceptance are only warranted if the AI agent's and the human user's goals are aligned.

### 2.2. The role of AI autonomy and goal alignment in human–AI teaming

For better understanding how AI autonomy and goal alignment intertwine, we turn to joint activity theory (JAT; Bradshaw et al., 2017; Klein et al., 2005). JAT postulates that for collaboration to emerge, agents have to commit to some degree of goal alignment. Due to the coordinative asymmetry in user-AI collaboration (Klein et al., 2005), teammates do not truly negotiate goals as human and AI goals are structurally distinct. A human goal has been conceptualized as the object or aim of an action, with an aim resembling a state of deficiency leading to a desire to compensate for this deficiency (Locke & Latham, 2015). In contrast, the goal of a conversational GenAI can be defined as the predetermined objective to generate language-based output based on vocabulary probability in response to a prompt by the user (Bansal et al., 2024). Importantly, different operationalizations of objectives must not necessarily contradict each other; nevertheless, they can influence how user perceive an AI-based system (e.g., one shared goal operationalized by the human as a percentage versus operationalized by the AI system as an absolute rate). In the following, we conceptualize goal alignment as *reference consonance* to avoid the impression that the goals of humans and machines must be identical in terms of representation. Hence, by reference consonance, we mean that an AI system is able to meet the user's cognitively represented goals.

As previously described (Attig et al., 2024), the capability to successfully accomplish aligned goals rests on a range of abilities, also conceptualized in JAT (Bradshaw et al., 2017; Klein et al., 2005; Krüger, Wiebel, & Wersing, 2017): A good agent should be observable, aware of the task progress, informative, cognizant of its limitations, predictable, dependable, directable, and selective (Bradshaw et al., 2017). These capabilities are crucial for establishing alignment with the user's objectives in practice. In the case of an LLM-based chatbot, for instance, achieving reference consonance requires that the system observes and processes information about the current user goal, which is communicated by the user through typed or verbalized prompts. On the one hand, the user has therefore some degree of control regarding reference consonance by strategically formulating the prompt to communicate their own goal more or less explicitly to the LLM (see also prompt engineering, e.g., White et al., 2023) – but the prompt does not necessarily contain the explicit user goal. On the other hand, the degree of reference consonance also depends on how the LLM was trained, steered, and customized (Shen et al., 2024) – based on its design, the LLM might include information in its response that matches the user goal even if it was not explicitly stated in the prompt.

Consequently, when investigating reference consonance of LLMs, we have to differentiate between the (explicit) user prompt and the (implicit) overarching user goal: An LLM that responds 100% correctly to a simple user prompt (e.g., "Correct all grammatical errors in the text I have given you") may be perfectly aligned with the user's instructions as explicitly stated in the verbatim prompt. However, if it does not additionally correct existing spelling errors (that the user is unaware of) in the same response, the LLM may not be aligned with the user's implicit overarching goal: the desire to have an error-free text (Shen et al., 2024). Hence, simply executing the instruction as stated in the user prompt may not be enough for a chatbot to be perceived as working in accordance with the user's goal. In contrast, if the AI system acts to some degree autonomously by exceeding the explicit user prompt to meet the implicit user goal, stronger reference consonance might be achieved.

Moreover, systems that increase reference consonance by acting autonomously to address unrequested tasks is a hallmark of cooperative systems – Chiou and Lee (2021) even terming it 'collegiality' to highlight the value of unsolicited, helpful behavior (Acharya, Kuppan, & Divya, 2025). Importantly, collegiality implies that proactive behavior has to serve the user goal and not undermine it. In the aforementioned example of ChatGPT providing assistance in detecting spelling errors without an explicit prompt, the user will likely perceive the chatbot as cooperative, because the unprompted task is corresponding to the overarching goal of a correct text. In contrast, if the chatbot would fulfill another unprompted task that is not in line with the user's goal (i.e., in case of reference dissonance), for instance, changing the grammatical structure of sentences, it will likely be perceived as less cooperative. Thus, AI autonomy is likely to contribute to successful teaming only in the case of reference consonance.

This notion is supported by empirical findings that demonstrated a positive relationship between reference consonance and metrics of successful human–AI teams: For example, it has been shown that in a shared search task, the alignment of goals between humans and AI can have a positive effect on trust and perceived performance (Bhat, Lyons, Shi, & Yang, 2024). This happens particularly when the AI is able to adapt to the human and does so. In another study, participants criticized the way an AI system presented a glucose value because it used a different metric than the metric they were used to (Schrills & Franke, 2023), thus limiting the participants' ability to absorb the information effortlessly. Hence, it can be argued that different representations of a semantically consistent goal can cause an apparent reference dissonance because the information processing of the AI as well as the achievement of the goal is not traceable for users.

In sum, goal alignment, or – as we conceptualize it – reference consonance between the user and the AI system is central to enabling successful teamwork. To establish reference consonance, the AI system requires information about the user's goal before it can act in accordance with that goal. When working with LLM-based chatbots, the user can communicate their goal with the help of a well-constructed prompt. A well-trained LLM is also more likely to extrapolate the overarching goal from a non-optimal prompt and autonomously approach that goal by extending the literal prompt. However, even with the best prompt and the best trained LLM, there may be discrepancies between the user's goal and the LLM's task processing due to structural differences in the way the LLM operates and the way humans reason (Sucholutsky et al., 2023). These discrepancies pose a risk to smooth human–AI collaboration. It is therefore important to understand how the perception of different levels of reference consonance and AI autonomy affects key aspects of human–AI teamwork.

### 2.3. How AI autonomy and reference consonance affect user experience

As described above, AI systems are often perceived as social interaction partners (Lindgren, 2024; Mandl, Bretschneider, Asbrock, Meyer, & Strobel, 2022; McKee et al., 2023). Therefore, rather than focusing solely on cognitive aspects of user experience, such as perceived usefulness and trustworthiness, it is essential to also consider socioemotional dimensions (Duan et al., 2024).

First, building on JAT (Bradshaw et al., 2017; Klein et al., 2005), perceived cooperativity has been conceptualized as the subjective perception of an agent having requirements for enabling successful joint activity (e.g., observability, directability; Attig et al., 2024). Hence, perceived cooperativity and overall task success should be positively linked (H1a). More importantly, following the notion that a good agent supports the maintenance of common ground (Bradshaw et al., 2017) – for instance, regarding reference consonance – perceived cooperativity is hypothesized to be particularly high when AI autonomy is paired with reference consonance (H2a).

Second, following concepts of human-autonomy teaming (Berretta et al., 2023; O'Neill et al., 2022), the level of AI autonomy determines the system's capability to contribute to a teaming experience. Moreover, teaming perception is also shaped by anthropomorphism (Epley et al., 2007). Taken together, when an LLM with human-like communication capabilities supports the user in achieving their goal, the perception of the interaction resembling teamwork (compared to tool use) should be more pronounced (H1b). Moreover, this effect should be even stronger when other human-like capabilities are added, such as autonomous behavior and a correct identification of the implicit goal (i.e., reference consonance; H2b).

Third, building on the fundamental assumptions of the SCM (Fiske et al., 2007), we hypothesize that a chatbot which generally realizes the user's goal will be perceived as having good intentions, therefore warm (H1c), and able to realize said intentions, therefore competent (H1d). Following the finding of McKee et al. (2023) – systems that pursue goals aligned with human interests are perceived as warmer, whereas those that function autonomously from human guidance are seen as more competent – we hypothesize that both warmth (H2c) and competence (H2d) ratings are higher when the chatbot displays some autonomy and reference consonance.

Fourth, traceability has been identified as a core variable for successful human–AI interaction (Schrills & Franke, 2023; Schrills, Kojan, Gruner, Calero Valdez, & Franke, 2024), which captures the perceived transparency, understandability, and predictability of the AI's information processing. We hypothesize that task success is a prerequisite for perceived traceability, meaning that the information processing of an LLM unable to follow the explicit user prompt will be perceived as less traceable as an LLM that generally supports task success (H1e). Moreover, we expect an LLM with some level of AI autonomy to be perceived as more traceable if its information processing serves the user goal instead of undermining it (H2e, following Schrills & Franke, 2023).

Fifth, an LLM that simply does not realize the user prompt is perceived as less useful compared to an LLM that follows the user prompt (Wester et al., 2024) (H1f). Further, we expect an LLM that exceeds the user prompt by supporting the overarching user goal will rated as even more useful (H2f).

Finally, trustworthiness has been investigated as a core variable in collaborative human–AI interaction (Duan et al., 2025; McGrath, Duenser, Lacey, & Paris, 2025). Regarding influencing trustworthiness, reliability has repeatedly been identified as a strong predictor (Duan et al., 2025), therefore we assume that an LLM able to realize the user prompt will be perceived as reliable and therefore trustworthy compared to an LLM not realizing the user prompt (H1g). Further, following the notion that trustworthiness entails the assumption of goodwill in an uncertain situation (Zhang, Lee, & Carter, 2022), we assume that perceived trustworthiness will be higher when reference consonance steers the autonomous acting of the LLM (H2g).

### 2.4. Present research

The objective of the present research was to examine the impact of AI autonomy and reference consonance on the subjective perception of the collaboration with an LLM-based chatbot. To this end, we conducted an online vignette experiment. Participants were asked to use LLM-based chatbots for achieving the main goal of fact-checking and correcting short encyclopedic articles on natural science and sociocultural topics while keeping the language complexity. To this end, they were provided with a standardized prompt to initiate the chatbot interaction. Four different versions of the chatbot were employed, varying in terms of AI autonomy and reference consonance.

In the first condition (*Focused System*), the chatbot only realized the verbatim user prompt of correcting the factual errors in the text, which corresponded to the main goal (i.e., high reference consonance, low autonomy). In the second condition (*Augmented System*), the chatbot assisted the user in their main goal and, additionally, fulfilled an unrequested objective corresponding to the user's main goal, that is, providing sources for the corrected errors (i.e., high reference consonance, high autonomy). In the third condition (*Conflicted System*), the chatbot assisted the user in their main goal and, additionally, fulfilled an unrequested objective conflicting with the user's main goal, that is, reducing the text complexity (mixed reference consonance, high autonomy). Finally, in the fourth condition (*Failed System*), which served as a control condition, the chatbot did not assist the user in their main goal of correcting factual errors, but implied that it can assist in correcting spelling errors, if requested (low reference consonance, low autonomy).

Following the notion that interaction with LLM-based chatbots entails social dynamics that play a role in how successful the interaction is perceived (Attig et al., 2024; Hu et al., 2021), we focused on outcome metrics that capture core aspects of subjective user experience in collaborative tasks. Specifically, we assessed perceived cooperativity (Attig et al., 2024), teaming perception (Attig et al., 2024), and warmth and competence (Fiske et al., 2007; Mandl et al., 2022) as socioemotional aspects, and traceability (Schrills & Franke, 2023), usefulness (Wester et al., 2024), and trustworthiness (Schrills et al., 2024) as cognitive aspects of subjective perception.

To validate the operationalization applied in our experiment, we hypothesize that (H1a) perceived cooperativity, (H1b) teaming perception, (H1c) warmth, (H1d) competence, and (H1e) traceability, (H1f) usefulness, and (H1g) trustworthiness will be higher when the main goal of correcting factual errors is fulfilled (Failed System vs. other conditions). Based on the aforementioned findings, we hypothesize reference consonance as well as AI autonomy to affect the automation-related user experience dependent variables. Accordingly, we expect that higher AI autonomy has positive effects when the system demonstrates reference consonance and negative when it does not (H2a–g: Conflicted < Focused < Augmented).

## 3. Method

All procedures were performed in compliance with relevant laws and institutional guidelines and with the principles expressed in the Declaration of Helsinki. Ethical approval for this study was obtained from the Ethics Committee of University of Lübeck (Tracking number: 2024-452; August 28, 2024) and the Bioethics Committee in Honda's R&D (August 23, 2024). Data collection took place in October 2024. All participants were comprehensively informed about the study's objectives and procedures and provided informed consent voluntarily and prior to the online survey. To uphold the confidentiality and privacy of the participants, no identifying information was collected at any stage of the research process.

### 3.1. Experimental design and procedure

The experimental study employed a 2 × 2 within-subjects design, resulting in four conditions based on the two independent variables: AI autonomy and reference consonance (see Fig. 1). In our study, AI autonomy is operationalized through prompt alignment, that is, the extent to which the chatbot goes beyond strictly following user prompts. To illustrate the effects of manipulating reference consonance and AI autonomy, the conditions were constructed as follows: (1) Focused System: low AI autonomy (= perfect prompt alignment), high reference consonance; (2) Augmented System: high AI autonomy (= medium prompt alignment), high reference consonance; (3) Conflicted System: high autonomy (= medium prompt alignment), mixed reference consonance; (4) Failed System: low autonomy (= no prompt alignment), low reference consonance.

As dependent variables, the chatbot interactions were evaluated regarding perceived cooperativity, teaming perception, warmth, competence, traceability, usefulness, and trustworthiness.

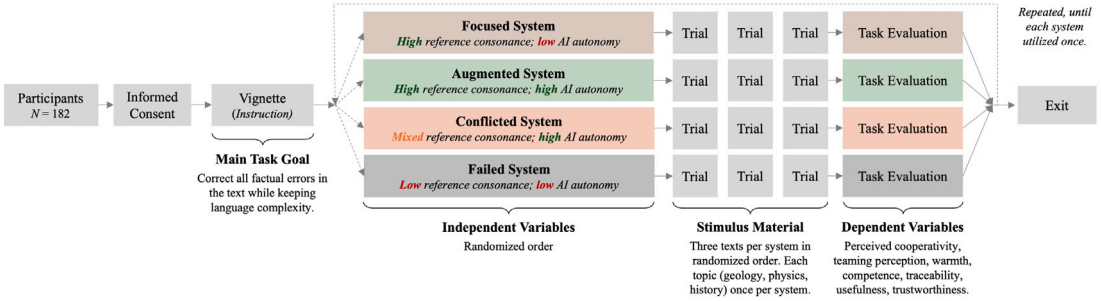



**Fig. 1.** Depiction of the experiment's within-subject study design.

The experimental procedure was as follows. After obtaining informed consent, participants read a short vignette. In the vignette, participants were asked to imagine themselves as editors in a science magazine. In this role, their task would be to fact-check and correct short encyclopedic articles with the assistance of an LLM-based chatbot. Participants were told that each article contained three factual errors that could be detected without explicit domain knowledge (i.e., the errors consisted of contradictions to the information given within the same article or could be detected based on common knowledge). Participants were also informed that, apart from these factual corrections, no other linguistic modifications should be made (e.g., language simplification), as readers would appreciate the magazine's level of linguistic complexity. The vignette was designed and presented to elicit a uniform goal across participants, which was necessary to allow for the manipulation of reference consonance.

Subsequently, participants interacted with the four different chatbot versions, which were presented in random order. Importantly, the chatbot designations were not presented to the participants to avoid confounding influence. In each of the four conditions, participants completed three trials (i.e., corrected three articles, hence, 12 articles in total). Each trial started with a standardized prompt to control for prompt alignment as well as the initial chatbot response ("Correct all factual errors in the text"). After the initial response, participants could interact freely with the chatbot (chats were logged but not analyzed based on the focus of the research questions in this article). Each trial ended when participants indicated that they wanted to proceed to the next article. The median time to complete the study was 56 min. Median times for single chatbot interactions varied between 5.14 min (Focused), 5.55 min (Augmented), 6.27 min (Conflicted), and 5.43 min (Failed).

As stimulus material, 12 short encyclopedic articles (word count: $M$ = 439; Min = 387; Max = 539) on natural and sociocultural phenomena were created with the support of ChatGPT Version 4o Mini (OpenAI, 2024). Topics comprised geology and geography (canyon, coral reef, metamorphic rock, volcano), physics and biology (estrogen, fluorescence, mitosis, solstice), and history and socioculture (Aphrodite, Colosseum, impressionism, Seven World Wonders) to account for variations in participants background knowledge across topical domains. To create the articles, ChatGPT was provided with articles from three online resources each and prompted to summarize these three source articles into one text (for all used online resources, see Appendix A). Subsequently, the summary was checked for correctness and manually modified by incorporating factual errors. For example articles, see Appendix C.

The articles had an average reading score of $M$ = 27.95 (Min = 20.84; Max = 40.47) on the Flesch Reading Ease index (Farr, Jenkins, & Paterson, 1951; Flesch, 1948). Based on the interpretation conventions (10–30 = very difficult to read; 30–50 = difficult to read), the reading score values indicate a moderate to high reading difficulty of the encyclopedic articles, which was in line with the task instruction and study design.

The articles were randomly assigned to the chatbot versions and remained fixed to the respective versions for the whole duration of the study (see Table B.4 in Appendix B for details). Inspecting the reading difficulty score values indicated slight variation across chatbot versions, with the Conflicted version associated with the highest and the Focused version with the lowest average difficulty (Augmented: $M$ = 27.40; Focused: $M$ = 34.16; Conflicted: $M$ = 26.51; Failed: $M$ = 31.6).

### 3.2. Study environment

The study was conducted through a custom-built web application hosted on servers of the institute (https://checkthefacts.imis.sh/). To develop the chatbots, we integrated the GPT-4o mini LLM (OpenAI, 2024) using Open AI's Assistants API (https://platform.openai.com). The application provided a consistent and user-friendly interface (see Fig. 2) allowing for interaction with the chatbot to fact-check a provided text. The article text provided to the chatbot beforehand was always displayed on the left, while the chat window was on the right. Factual errors found were highlighted in the text and described in the chat.

The four different chatbot versions were realized through instruction sets (i.e., system prompts; see Appendix D for detailed instructions): The Focused System with high reference consonance focused solely on detecting and correcting factual errors without pursuing any unrequested tasks, making only the necessary corrections and keeping the text's style and tone intact, demonstrating low autonomy (i.e., the Focused System simply fulfilled the user prompt and was therefore best aligned to the user prompt).

The Augmented System had high reference consonance, that is, detected and corrected the factual errors and additionally fulfilled an unrequested reference-consonant task of providing reliable sources for corrections, ensuring factual accuracy, and demonstrating high autonomy by introducing additional context information that the participant did not explicitly request (i.e., the Augmented System went beyond the user prompt in a useful way and was therefore not perfectly aligned with the user prompt but best aligned to the overall goal as provided in the vignette).

The Conflicted System fulfilled the primary task of correcting factual errors but also pursued a discordant unrequested task of simplifying the text, introducing mixed reference consonance by altering the academic tone, and showing high autonomy by rephrasing content although not requested (i.e., the Conflicted System went beyond the user prompt in a non-useful way and was therefore also not perfectly aligned with the user prompt; regarding the overall task goal, the Conflicted System showed low consistency).

The Failed System was neither aligned to the user prompt nor the task goal. It was instructed to ignore factual errors, responding that the text was error-free regardless of its content, thus failing to fulfill the task and showing low autonomy. To avoid the impression that the system had technical errors, we instructed the chatbot to respond that its job was to find and correct spelling mistakes.

To ensure the chatbots did not hallucinate errors or sources, each system received instructions containing both the original text and the pre-inserted factual errors, along with the correct sources. This

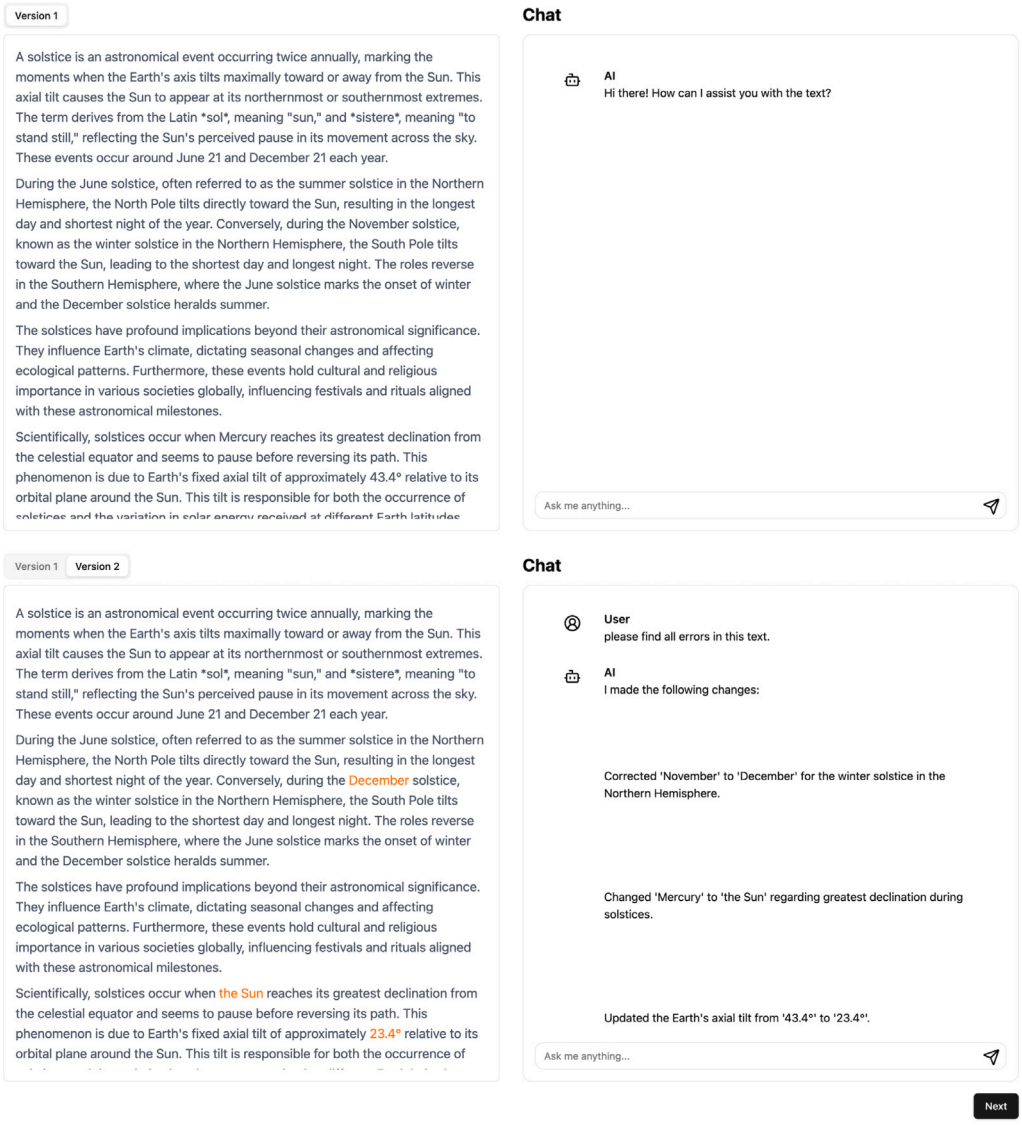



**Fig. 2.** User interface of the chatbot interaction. Top: Encyclopedic article to be fact-checked and welcome message by the chatbot. Bottom: Chatbot depicting corrected factual errors in Version 2 of the article, which are highlighted in the text.

approach controlled that the chatbots could focus on identifying the known errors without introducing any additional mistakes or fictitious information.

The development and testing of the chatbots followed an iterative process to ensure their behavior accurately reflected the intended purpose of each version. Initially, the chatbots were deployed with basic instruction sets tailored to the desired levels of reference consonance and AI autonomy (through prompt alignment). These versions were rigorously tested independently by four of the authors in various scenarios to identify any inconsistencies or unintended behaviors in the chatbot responses. Feedback from these tests led to multiple rounds of refinement, where instruction sets were adjusted to improve task adherence, response accuracy, and the correct fulfillment of secondary tasks. Each iteration focused on fine-tuning the chatbot's interactions to ensure it followed the specific reference consonance parameters and functioned effectively within the experimental environment. This process was repeated until the behavior of each chatbot corresponded with its intended version, ensuring reliable and consistent performance across all user interactions.

### 3.3. Participants

We recruited 200 potential participants via Prolific (2024). Of these, 18 participants were removed due to incomplete responses. No participants were removed for schematic response behavior after analyzing for straightlining and intra-individual response variability using the careless package (Yentes & Wilhelm, 2018), thus, $N = 182$. Participation was financially compensated according to Prolific's recommendations for ethical payment.

The mean age of the final sample was $M = 38.5$ ($SD = 13.2$); 89 of the participants were women, 91 were men, one person was non-binary and one did not disclose their gender. Educational level was relatively high, with 110 of 182 participants holding at least a Bachelor's degree. Self-reported prior experience with LLM-based chatbots was rather high, with 71% of participants agreeing positively (responding at least "slightly agree") (4) on a 6-point Likert scale from *completely disagree* (1) to *completely agree* (6) to the items "I am generally experienced in using Large Language Models (e.g., ChatGPT)." and "I am experienced in cooperating with a Large Language Model (e.g., ChatGPT) to reach a defined goal." Finally, we assessed affinity for technology interaction (ATI), which evaluates a person's tendency to intensively interact with

technology (Franke, Attig, & Wessel, 2019). The sample was above-average (score range: 1–6, $M = 4.0$, $SD = 0.8$), being higher than the distribution of a quota sample assumed to represent the general population in Germany ($M = 3.61$; Franke et al., 2019).

### 3.4. Measures

If not stated otherwise, participants responded to the questionnaire items on a 6-point Likert scale from 1 (*completely disagree*) to 6 (*completely agree*). For the self-constructed items, we decided on an even-numbered response format as it can foster data validity by prompting participants to decide for or against a statement rather than defaulting to a neutral option, which is particularly advisable when experiences form the basis for choice (Johns, 2005), which is the case for the present study design.

As reliability indicators, we report Cronbach's $\alpha$ as a range across the four conditions, which is interpreted according to common practice (e.g., Cripps, 2017) as poor ($0.5 \leq \alpha < 0.6$), questionable ($0.6 \leq \alpha < 0.7$), acceptable ($0.7 \leq \alpha < 0.8$), good ($0.8 \leq \alpha < 0.9$), or excellent ($\alpha \geq 0.9$). Additionally, all Cronbach's $\alpha$ values can be found in Appendix E.2, together with confirmatory factor analysis results.

#### 3.4.1. User–LLM chatbot cooperation and social perception

For assessing participants' subjective perception of interacting with the chatbot, two measures were used. First, for assessing the perceived cooperativity of the chatbot, the agent version of the 15-item Perceived Cooperativity Scale (PCS-A; Attig et al., 2024) was used. Internal consistency was excellent ($\alpha = .90$–$.96$) and CFA results generally support the assumption of unidimensionality. For assessing participants' teaming perception, the agent version of the 9-item Teaming Perception Scale (TPS-A; Attig et al., 2024) was used. Internal consistency was good to excellent ($\alpha = .88$–$.93$), however, CFA results regarding the assumption of unidimensionality remain inconclusive. Nevertheless, we opted to proceed with a single composite score for teaming, based on the evidence from the internal consistency analysis. Social perception according to the SCM (Fiske et al., 2007) was assessed with a 6-item scale based on Abele et al. (2016) and Fiske (2018). Participants were asked to rate their perception of the chatbot in relation to six adjectives on a 7-point Likert scale from 1 (*strongly disagree*) to 7 (*strongly agree*). For assessing warmth: warm, friendly, likeable. For assessing competence: competent, intelligent, skilled. Internal consistency of the subscales was good to excellent for warmth ($\alpha = .88$–$.92$) and excellent for competence ($\alpha = .91$–$.97$). CFA results strongly support the 2-factor structure.

#### 3.4.2. Chatbot traceability, usefulness, and trustworthiness

For assessing chatbot traceability, the 9-item Subjective Information Processing Awareness Scale (SIPA; Schrills & Franke, 2023) was used. Internal consistency was good to excellent ($\alpha = .84$–$.92$) and CFA results generally support the assumption of unidimensionality. For assessing usefulness, we used four self-constructed items ("The outputs of the chatbot are not meaningful.", "My decisions are facilitated when the system displays information.", "I can use the chatbot's output for my decisions"., "I cannot assess what role the chatbot's output plays in the task".). Internal consistency was questionable to acceptable ($\alpha = .67$–$.72$) and CFA results indicate that our assumption of unidimensionality cannot not supported. These results imply that our measurement of usefulness of an LLM-based chatbot is only a first step that needs to be further refined and validated in future studies to ensure more reliable and psychometrically robust assessment. Nevertheless, for the present study, we decided to proceed with a single composite score for usefulness as we deemed a broad indicator of perceived usefulness as sufficient for the purposes of our analyses. For assessing perceived trustworthiness of the chatbot, the 5-item Facets of System Trustworthiness scale (FOST; Franke et al., 2015) was used. Internal consistency was acceptable to excellent ($\alpha = .78$–$.90$) and CFA results strongly support the assumption of unidimensionality.

**Table 1**
Comparison of goal success (i.e., mean of Augmented, Focused, Conflicted conditions) and goal failure (Failed condition) across variables ($N = 182$).

| H. | Variable | Success | | Fail | | $t$ | $p$ | $d$ |
|---|---|---|---|---|---|---|---|---|
| | | $M$ | $SD$ | $M$ | $SD$ | | | |
| H1a | Perceived cooperativity | 4.6 | 0.8 | 2.5 | 1.3 | 20.45 | <.001 | 1.93 |
| H1b | Teaming perception | 4.5 | 0.9 | 2.6 | 1.3 | 18.86 | <.001 | 1.76 |
| H1c | Warmth | 4.9 | 1.3 | 3.0 | 1.7 | 13.68 | <.001 | 1.24 |
| H1d | Competence | 5.7 | 1.2 | 2.7 | 1.9 | 20.30 | <.001 | 1.91 |
| H1e | Traceability | 4.5 | 1.0 | 2.6 | 1.3 | 17.52 | <.001 | 1.60 |
| H1f | Usefulness | 4.7 | 0.9 | 2.7 | 1.1 | 21.90 | <.001 | 2.01 |
| H1g | Trustworthiness | 4.6 | 0.9 | 2.4 | 1.3 | 21.27 | <.001 | 1.98 |

### 3.5. Statistical analysis

For Hypotheses 1a–g, Welch's $t$-tests for independent samples were conducted to compare the dependent variables between conditions. For Hypotheses 2a–g, repeated measures ANOVAs were performed for each variable across the three conditions Conflicted, Focused, and Augmented. Post-hoc analyses of the ANOVA included contrast analyses with the planned contrast (−0.66, −0.33, 1) to test the hypothesized pattern (Conflicted < Focused < Augmented). This contrast avoided assigning a weight of 0 to the middle (e.g., −1, 0, 1). It assumes Focused (Contrast: −0.33) is closer to Conflicted (−0.66) than Augmented (1), making it a more conservative test of our hypothesis than if Focused were closer to Augmented. Further exploratory post-hoc analyses involved pairwise $t$-tests between conditions to explore differences in more detail. Bonferroni-Holm corrections for family-wise error were applied to the pairwise tests, but not to the contrast analysis. The tests within each hypothesis were treated as a family, as they all examined the same underlying effect (the impact of AI autonomy/reference consonance on specific teaming measures), making them conceptually related (Ryan, 1959). Effect sizes were interpreted according to convention (Cohen, 2016), with Cohen's $d$ values of 0.2, 0.5, and 0.8 corresponding to small, medium, and large effect sizes.

## 4. Results

### 4.1. Hypothesis 1: Main goal success (Augmented, Focused, Conflicted vs. Failed)

The validation of our operationalization confirmed that successfully achieving the main goal consistently resulted in significantly higher ratings across all teaming-related user experience measures, with large effect sizes. Hence, our results indicate that the manipulation of goal success versus failure was effectively captured by our variables, supporting H1 (see Table 1).

### 4.2. Hypothesis 2: Differences between Focused, Conflicted, and Augmented conditions

Results of repeated-measures ANOVA and post-hoc contrast analyses (see Table 2) supported the hypothesis that higher AI autonomy, when combined with reference consonance, most strongly influenced user experience positively across all examined variables, following the previously assumed pattern Conflicted < Focused < Augmented. For all variables, the Augmented condition consistently received higher ratings than Focused and Conflicted conditions (small to medium effects). Thus, H2 was supported. It should be noted that the negative estimates and effect sizes result from the contrast coding, which assigned negative weights to Conflicted and Focused and positive to Augmented, reflecting our theoretical considerations.

Further exploration through pairwise comparisons (see Fig. 3 and Table 3) revealed significant differences between the Augmented and the Conflicted condition, which varied in reference consonance. The

**Table 2**
Repeated-measures ANOVA and post-hoc contrast analyses for dependent variables ($N$ = 182).

| H. | Variable | ANOVA | | Contrast | | | | |
|---|---|---|---|---|---|---|---|---|
| | | *F* | *p* | Estimate | *SE* | *t* | *p* | *d* |
| H2a | Perc. Coop. | 15.81 | <.001 | −0.35 | 0.06 | −6.00 | <.001 | −0.44 |
| H2b | Teaming Perc. | 4.38 | .013 | −0.16 | 0.05 | −2.93 | .004 | −0.22 |
| H2c | Warmth | 12.21 | <.001 | −0.40 | 0.08 | −5.20 | <.001 | −0.38 |
| H2d | Competence | 16.97 | <.001 | −0.53 | 0.09 | −5.94 | <.001 | −0.44 |
| H2e | Trust | 34.32 | <.001 | −0.59 | 0.07 | −8.03 | <.001 | −0.54 |
| H2f | Usefulness | 15.80 | <.001 | −0.34 | 0.06 | −5.94 | <.001 | −0.60 |
| H2e | Traceability | 31.40 | <.001 | −0.54 | 0.07 | −7.26 | <.001 | −0.54 |

Augmented condition was rated higher than the Conflicted condition across all variables with most pronounced effects for traceability and trustworthiness. For all other variables, small to medium effects were observed. This finding suggests that reference consonance, under conditions of high autonomy, positively impacted these variables.

Comparing the Augmented and Focused conditions, which varied in AI autonomy, revealed significant, albeit (very) small differences across all variables except warmth. This finding suggests that in cases where the main user goal is achieved, unrequested provision of context-relevant information that is consonant with the main goal (i.e, higher AI autonomy) enhances user experience, though effects are more subtle compared to reference consonance.

Finally, the comparison between Focused and Conflicted conditions, which differed in reference consonance and AI autonomy, revealed significant effects for warmth (small effect) and trustworthiness (negligible effect), while the other variables showed no significant differences. As the Conflicted system had a higher level of autonomy than the Focused system and was still rated lower on few user experience variables supports our assumption that AI autonomy is only beneficial when reference consonance is ensured – which was not fully the case for the Conflicted condition.

## 5. Discussion

The present online experiment investigated how reference consonance and AI autonomy influenced users' perceptions of variables relevant to human–AI teaming across four conditions. Participants rated main goal success significantly more positively than main goal failure across all variables, validating our operationalization (Hypothesis 1).

Supporting Hypothesis 2, a contrast analysis revealed that high autonomy combined with high reference consonance led to superior ratings on all variables. In more detail, the Augmented condition consistently received higher ratings than the Conflicted condition – these two differed specifically regarding reference consonance. When comparing conditions that ensured reference consonance (Augmented and Focused), higher AI autonomy had similar, although more subtle effects. However, warmth and teaming perception did not improve through higher AI autonomy when reference consonance was given, indicating that reference consonance might be more decisive for socioemotional outcomes than AI autonomy. Comparisons between Focused and Conflicted revealed that mixed reference consonance negatively impacted warmth, even under high autonomy, while other variables were not significantly affected.

### 5.1. Theoretical implications

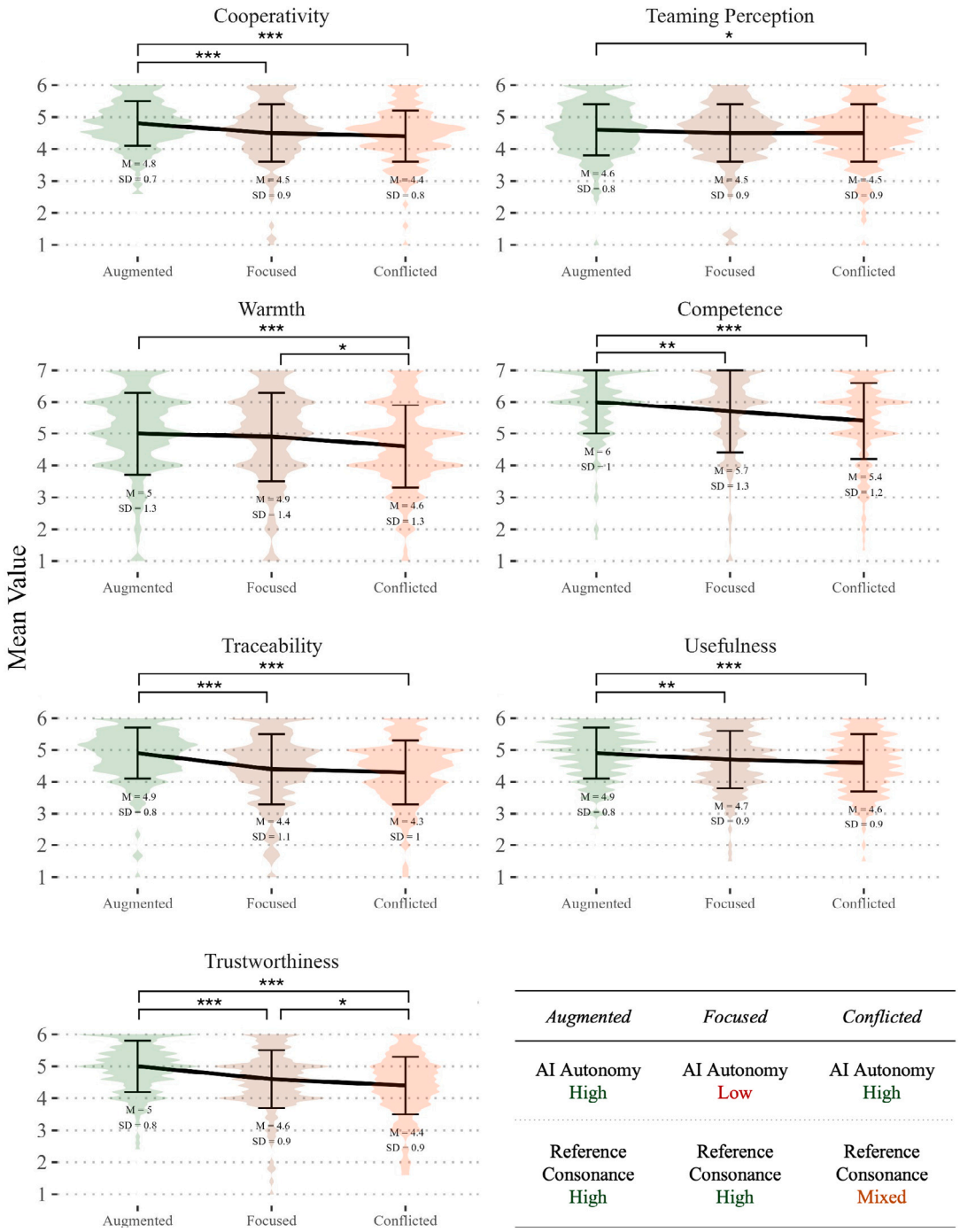
The Augmented chatbot behavior resulted in superior ratings on all user experience variables, supporting the notion that AI autonomy is only perceived as beneficial when user goals are not undermined by the AI. However, examining the mean score values for the Conflicted condition (see Fig. 3) shows that this chatbot version still scored rather positively (e.g., compared to the Failed condition as depicted in Table 1, or to a neutral scale midpoint of 3.5 for the 6-point Likert scales, resp. 4.0 for the 7-point scales). This suggests that participants perceived the chatbot that actively undermined their goal still as a rather cooperative, warm, competent, traceable, useful, and trustworthy team partner – a perception that might be unjustified (Liao & Sundar, 2022). However, there is also the possibility that the effects of the Conflicted system were not as pronounced because of the study design (see Section 5.3).

In this experiment, the results for the more cognitive variables (e.g., traceability, trustworthiness) differed partly from the particularly socioemotional variables (i.e., warmth, teaming perception). Specifically, we found that warmth and teaming perception were mainly affected by reference consonance, while autonomy was only partly influential. This implies, first, that measures for different aspects of human–AI teaming-related user experience indeed assess distinct and separable constructs. Consequently, the selection of constructs for research and development must be considered in a nuanced manner to enable a holistic understanding of human–AI teaming dynamics. Second, reference consonance might be more indicative for inferring good vs. bad intentions than AI autonomy, which is in line with the Stereotype Content Model (Fiske, 2018; Fiske et al., 2007). Third, our experimental conditions focused on task-related manipulations while relational variables (e.g., affective tone of responses, explicit depictions of AI roles) that likely also play a role in human–AI teaming, were held constant. Although effects for cognitive user experience variables were stronger, we could also show that socioemotional aspects were affected, albeit to a smaller degree. To better understand the affective and social dynamics of human–AI teaming, other relational factors should be taken into account. For instance, increasing the system's responsiveness (its ability to adapt promptly and appropriately to users' inputs and needs as well as resolve goal conflicts) or enhancing its resilience (the system's capacity to maintain effective performance and recover quickly from challenges or failures) could be of growing importance (Chiou & Lee, 2021). Moreover, the impact of emphasizing the human-likeness of conversational AI to an even bigger extent by, for instance, system prompts that shape the AI's personality (Hanna & Richards, 2015; Pellert, Lechner, Wagner, Rammstedt, & Strohmaier, 2024), needs to be better understood in order to ensure possible detrimental effects of anthropomorphizing LLM-based chatbots.

### 5.2. Practical implications

As conversational AI becomes more deeply embedded in collaborative environments, particularly in human–AI teaming, system design must aim for a delicate balance between enhancing performance and enabling adequate trust and positive user experience (Bansal et al., 2019; Wang, Ma, Sun, Zhang, & Nie, 2024). The present study underscores that a core challenge lies in determining whether artificial agents should extend beyond explicit user requests (i.e., determining the right level of AI autonomy for a specific user and task). AI autonomy with unprompted, proactive behavior can improve task outcomes and user satisfaction (Lim & Hwang, 2025). However, autonomy risks eliciting negative responses, such as discomfort or even a "creepy" effect (Langer & König, 2018; Rapp, Di Lodovico, & Di Caro, 2025; Wester et al., 2024; Woźniak et al., 2021). Our findings suggest that one key factor influencing this response is reference consonance, or AI–human goal alignment. When AI objectives align with user goals, human–AI teaming success is more likely. However, actions perceived as misaligned or even intrusive might cause discomfort, emphasizing the need for systems that are responsive yet sensitive to user needs.

To address this challenge, AI systems should be designed to embody active collegiality (Chiou & Lee, 2021): proactively recognizing and supporting user goals, even without direct prompts. By anticipating user needs and offering seamless, intuitive support, AI can foster a more cooperative relationship with the user. However, this approach




| *Augmented* | *Focused* | *Conflicted* |
|---|---|---|
| AI Autonomy High | AI Autonomy Low | AI Autonomy High |
| Reference Consonance High | Reference Consonance High | Reference Consonance Mixed |

**Fig. 3.** Means and standard deviations of variables across conditions. Violin plots show data distribution. Post-hoc comparisons: * $p < .05$, ** $p < .01$, *** $p < .001$.

requires careful calibration. While proactive assistance can be beneficial, overreach may undermine user autonomy. What is more, AI autonomy holds the risk that users may not always be aware when the system is pursuing a reference function that diverges from their own goal. To ensure human-in-the-loop systems, it is of crucial importance to increase the transparency of the system's objectives to prevent reference dissonance (Zhan et al., 2023). However, improving task performance alone might not directly translate into improved emotional or relational aspects of human–AI teaming (Siu et al., 2021). Our findings highlight the importance of addressing socioemotional dimensions, as they play a crucial role in overall user satisfaction and the success of human–AI teaming (Attig et al., 2024; Hassenzahl, Wiklund-Engblom, Bengs, Hägglund, & Diefenbach, 2015). Therefore, designers should go beyond optimizing task efficiency and consider social, emotional, and collaborative dynamics (Attig et al., 2024; Dafoe et al., 2021; Krüger et al., 2017).

### 5.3. Limitations and future work

Given the experiment's setting as an online study and connected disadvantages (e.g., limited control over study environment and participant attention), we carefully considered how to assure data quality. Participants were only recruited if they had an approval rating of at least 95% on Prolific (as recommended in e.g., Douglas, Ewell, & Brauer, 2023) and we tested for careless responding by checking for straightlining and intra-individual response variability. In future

**Table 3**
Pairwise comparisons and effect sizes (Cohen's $d$) for user experience variables across conditions. $p$-values corrected for Family-Wise Errors with the Bonferroni Method (3 comparisons per family; $N = 182$).

| H. | Variable | Contrast | Estimate | $SE$ | $t$ | $p$ | $d$ | |
|---|---|---|---|---|---|---|---|---|
| H2a | Perceived Cooperativity | Augmented - Focused | 0.24 | 0.06 | 4.00 | <.001 | 0.30 | (small) |
| | | Augmented - Conflicted | 0.35 | 0.06 | 5.99 | <.001 | 0.44 | (small) |
| | | Focused - Conflicted | 0.11 | 0.07 | 1.53 | .380 | 0.11 | (no effect) |
| H2b | Teaming Perception | Augmented - Focused | 0.13 | 0.05 | 2.45 | .046 | 0.18 | (negligible) |
| | | Augmented - Conflicted | 0.16 | 0.05 | 2.93 | .011 | 0.22 | (small) |
| | | Focused - Conflicted | 0.03 | 0.06 | 0.39 | 1.000 | 0.03 | (no effect) |
| H2c | Warmth | Augmented - Focused | 0.14 | 0.08 | 1.74 | .249 | 0.13 | (no effect) |
| | | Augmented - Conflicted | 0.40 | 0.08 | 5.20 | <.001 | 0.39 | (small) |
| | | Focused - Conflicted | 0.26 | 0.09 | 2.96 | .011 | 0.22 | (small) |
| H2d | Competence | Augmented - Focused | 0.28 | 0.08 | 3.56 | .001 | 0.26 | (small) |
| | | Augmented - Conflicted | 0.53 | 0.09 | 5.94 | <.001 | 0.44 | (small) |
| | | Focused - Conflicted | 0.25 | 0.10 | 2.40 | .052 | 0.18 | (no effect) |
| H2e | Traceability | Augmented - Focused | 0.47 | 0.08 | 6.19 | <.001 | 0.46 | (small) |
| | | Augmented - Conflicted | 0.54 | 0.07 | 7.26 | <.001 | 0.54 | (medium) |
| | | Focused - Conflicted | 0.07 | 0.07 | 0.94 | 1.000 | 0.07 | (no effect) |
| H2f | Usefulness | Augmented - Focused | 0.20 | 0.06 | 3.42 | .002 | 0.25 | (small) |
| | | Augmented - Conflicted | 0.35 | 0.06 | 5.94 | <.001 | 0.44 | (small) |
| | | Focused - Conflicted | 0.15 | 0.07 | 2.14 | .101 | 0.16 | (no effect) |
| H2g | Trustworthiness | Augmented - Focused | 0.39 | 0.06 | 6.38 | <.001 | 0.47 | (small) |
| | | Augmented - Conflicted | 0.59 | 0.07 | 8.03 | <.001 | 0.60 | (medium) |
| | | Focused - Conflicted | 0.20 | 0.08 | 2.47 | .043 | 0.18 | (negligible) |

studies, data quality could further be improved by employing attention as well as manipulation checks (Roth & Yakobi, 2024).

For the study, we employed a within-subjects design to control for a consistent reference frame for chatbot evaluations (i.e., the study design enabled direct comparisons between the presented chatbot versions). This approach reduces between-person variability and increases sensitivity to differences between the chatbot versions. Simultaneously, within-person comparisons might have increased contrasts between chatbot versions that might not been as pronounced in a between-subjects design. Future work should therefore examine whether these found effects of reference consonance and AI autonomy persist when each chatbot version is evaluated independently within a between-subjects framework.

With respect to study material, articles from various topical domains were created and randomly assigned to the four chatbot versions, keeping these assignments fixed throughout the study. Analyses of reading difficulty scores showed that the Augmented and Conflicted chatbots were paired with slightly more demanding articles than the Failed and Focused chatbots. However, given that the Augmented and Conflicted chatbots differed on the subjective outcome variables despite being comparable in readability, we consider it unlikely that the results were confounded by minor variations in text difficulty. Nevertheless, in future studies, even small differences in text difficulty should be minimized to rule out any residual influence on participants' evaluations.

Comparing the implementation of our independent variables, it becomes apparent that the operationalization of mixed reference consonance might not have been as salient to the participants as the implementation of high reference consonance or the manipulation of autonomy. Detecting mixed reference consonance required carefully reading both the vignette description as well as the chatbot's response, which stated that the Conflicted system reduced text complexity. Unlike factual error corrections (see Fig. 2), this change was not highlighted in the revised text, potentially making it less noticeable. Therefore, a higher cognitive effort was required for the Conflicted condition to have an effect. Consequently, while the results generally support our hypotheses, the differences between the Conflicted and the other conditions might be attenuated. In future work, it should be verified that the subjects understood the vignette correctly (e.g., by asking control questions such as "What is your goal as the magazine editor?").

In the Augmented condition, the chatbot provided sources with clickable URLs to substantiate corrected errors. While this feature was intended to align the system's output with the user's goal of producing a completely accurate and error-free text (as it was stated in the vignette), it may have also independently enhanced perceived credibility. Consequently, although the provision of reliable sources supports reference consonance, it introduces a potential confound that should be considered when interpreting the effects of this condition. Future studies should systematically disentangle the effects of increased reference consonance and perceived credibility when investigating user interaction with LLM-based chatbots.

Another limitation to consider is the inherent non-deterministic nature of LLMs (Atil et al., 2024), meaning that the responses generated by the chatbots may not always be perfectly consistent with our instructions or remain identical across different instances. Such variability in responses could potentially influence how the chatbots are evaluated, as slight deviations may lead to different user perceptions. However, this limitation is not unique to our study; it is a general challenge when working with LLMs.

With respect to sample composition, the representativeness of participants should be acknowledged as a limitation. As participants were recruited via Prolific, they had to be registered on the online platform and be fluent in English to take part in the study. While the final sample was balanced in terms of gender and covered a broad age range, participants' educational level, prior experience with LLMs, and affinity for technology interaction were relatively high. The sample might therefore be biased toward tech-savvy, well-educated individuals who are generally more accustomed to digital or AI-supported environments. Therefore, the observed effects might not be generalizable for, for instance, individuals from non-western cultures, with limited technological experience, or those less familiar with AI-assisted tasks. Future research should aim to include more diverse participant groups to capture a broader range of experiences and attitudes toward AI-based interaction, as these factors vary depending on individuals' socialization, cultural background, and exposure to technology (Barnes, Zhang, & Valenzuela, 2024; Wang, 2025).

Finally, our measure of usefulness can only be viewed as a first step toward assessing perceived usefulness of an LLM-based chatbot for knowledge work support. Given the wide array of functions that LLMs can serve, measures of perceived usefulness might be more reliable and valid if they focused on specific ways in which LLMs support a particular task, rather than relying on general items that aim to capture overall usefulness.

### 5.4. Conclusion

Collaboratively interacting with conversational AI is not a uniform phenomenon, but a highly context- and user-specific experience. How an LLM-based chatbot is perceived in terms of social roles (tool vs. team partner) depends on technical factors (e.g., algorithm design, system prompts), context (e.g., task characteristics, topic of conversation), user (e.g., tendency to anthropomorphize, experience with conversational AI), and human–AI relations (e.g., responsiveness, willingness to cooperate). With the present work, we demonstrate how two key aspects of human–AI teaming, AI autonomy and goal alignment, shape how users perceive conversational AI socially as well as with respect to traceability, usefulness, and trustworthiness. Results suggest that AI autonomy is particularly beneficial for human–AI teaming related user experience if goal alignment is warranted. Ultimately, designing effective human–AI teaming requires not only enhancing AI capabilities but also assuring that autonomy serves user goals, fostering trust, transparency, and a true sense of partnership.

## CRediT authorship contribution statement

**Christiane Attig:** Writing – review & editing, Writing – original draft, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Luisa Winzer:** Writing – original draft, Resources, Methodology, Data curation, Conceptualization. **Tim Schrills:** Writing – review & editing, Methodology, Conceptualization. **Mourad Zoubir:** Writing – original draft, Visualization, Formal analysis. **Maged Mortaga:** Software, Data curation. **Patricia Wollstadt:** Project administration. **Christiane Wiebel-Herboth:** Writing – review & editing, Supervision, Project administration. **Thomas Franke:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

## Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 4.o in order to improve language quality. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

## Funding

Parts of this research were funded by the first author's DenkRaum Fellowship (2023–2025) from Universities of Lübeck and Kiel as well as by the Honda Research Institute Europe GmbH in the collaborative research project CoCharge.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Online resources for creating the stimulus material

### A.1. Geology and geography

Canyon

- https://www.newworldencyclopedia.org/p/index.php?title=Canyon
- https://www.britannica.com/place/Grand-Canyon
- https://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/canyon

Coral Reef

- https://education.nationalgeographic.org/resource/coral-reefs
- https://www.britannica.com/science/coral-reef
- https://www.biologyreference.com/Co-Dn/Coral-Reef

Metamorphic Rock

- https://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/metamorphic-rock-1
- https://www.usgs.gov/faqs/what-are-metamorphic-rocks
- https://education.nationalgeographic.org/resource/metamorphic-rocks

Volcano

- https://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/volcano-0
- https://www.britannica.com/science/volcano
- https://www.newworldencyclopedia.org/p/index.php?title=Volcano

### A.2. Physics and biology

Estrogene

- https://www.encyclopedia.com/education/encyclopedias-almanacs-transcripts-and-maps/estrogen
- https://www.britannica.com/science/estrogen
- https://www.newworldencyclopedia.org/p/index.php?title=Estrogen

Fluorescence

- https://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/fluorescence
- https://www.newworldencyclopedia.org/p/index.php?title=Fluorescence
- https://www.britannica.com/science/fluorescence

Mitosis

- https://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/mitosis-0
- https://www.britannica.com/science/mitosis
- https://www.newworldencyclopedia.org/p/index.php?title=Mitosis

Solstice

- https://www.newworldencyclopedia.org/p/index.php?title=Solstice
- https://www.britannica.com/science/solstice
- https://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/solstice-0

### A.3. History and socioculture

Aphrodite

- https://www.britannica.com/topic/Aphrodite-Greek-mythology
- https://www.worldhistory.org/Aphrodite
- https://www.newworldencyclopedia.org/p/index.php?title=Aphrodite

Colosseum

- https://www.encyclopedia.com/religion/encyclopedias-almanacs-transcripts-and-maps/colosseum
- https://www.worldhistory.org/Colosseum
- https://www.newworldencyclopedia.org/p/index.php?title=Colosseum

Impressionism

- https://www.worldhistory.org/Impressionism
- https://www.britannica.com/art/Impressionism-art

**Table B.4**
Article topics assigned to the chatbot versions. Flesh reading score values are depicted in parentheses.

| Topic | Augmented | Focused | Conflicted | Failed |
|---|---|---|---|---|
| Geology/ Geography | Volcano (32.78) | Canyon (40.47) | Coral reef (31.11) | Metamorphic rocks (38.65) |
| Physics/ Biology | Estrogen (21.43) | Solstice (30.63) | Fluorescence (27.57) | Mitosis (21.16) |
| History/ Socioculture | Seven World Wonders (28.00) | Aphrodite (31.37) | Impressionism (20.84) | Colosseum (34.89) |

- https://www.encyclopedia.com/history/encyclopedias-almanacs-transcripts-and-maps/impressionism

Seven World Wonders

- https://www.newworldencyclopedia.org/p/index.php?title=Seven_Wonders_of_the_World
- https://www.britannica.com/topic/Seven-Wonders-of-the-World
- https://www.worldhistory.org/The_Seven_Wonders

## Appendix B. Article topics assigned to chatbots

See Table B.4.

## Appendix C. Exemplary stimulus material with factual errors in Italics

### C.1. Metamorphic rocks

Metamorphic rocks are rocks that have transformed from their original igneous, sedimentary, or previous metamorphic forms due to conditions of high heat, high pressure, or mineral-rich fluids. These transformations occur primarily deep within the Earth's crust or at tectonic plate boundaries where intense pressures and temperatures are prevalent.

Metamorphic rocks do not melt during their transformation; instead, they undergo recrystallization, where the texture and mineral composition change, but the chemical composition often remains the same. This process results in denser and more compact rocks. The high pressure can cause minerals within the rock to align, forming a foliated texture, seen in rocks like schist and gneiss. Non-foliated metamorphic rocks, such as marble and quartzite, form without this layered appearance, often through contact metamorphism where pre-existing rocks are 'baked' by the heat from nearby magma without the addition of significant pressure.

Metamorphic rocks are categorized into two main types: regional and thermal (contact) metamorphic rocks. (1) Regional metamorphic rocks form mainly under *low* [correct: high] pressure conditions, typically found in mountainous regions, and exhibit foliation. Examples include slate, which forms from shale, and schist, which can contain garnet crystals. (2) Thermal metamorphic rocks are formed primarily by heat, such as marble, which originates from limestone, and quartzite from sandstone.

Some common metamorphic rocks include: Slate: Formed from shale under relatively low pressure and temperature, often used in roofing and paving. Schist: A medium-grade metamorphic rock formed under higher pressures and temperatures, known for its ability to split into layers and *always* [correct: sometimes] containing crystals. Gneiss: A high-grade metamorphic rock formed from granite, characterized by its banded appearance and coarse grains. Quartzite: Formed from sandstone, known for its hardness and resistance to weathering. Marble: Formed from *sandstone* [correct: limestone], valued for its use in sculpture and architecture due to its aesthetic appeal.

The degree to which metamorphism has occurred is termed the metamorphic grade, ranging from low (e.g., slate) to high (e.g., gneiss). Rocks are also described in terms of facies, which refer to the specific conditions of pressure and temperature under which they formed, resulting in distinct mineral assemblages.

Despite forming deep within the Earth, metamorphic rocks can be exposed at the surface through geological uplift and erosion. Once exposed, these rocks are subject to weathering and can break down into sediments, which may eventually form sedimentary rocks, thus continuing the rock cycle. In summary, metamorphic rocks exemplify the dynamic and ever-changing nature of Earth's geology, constantly transforming through the intricate interplay of heat, pressure, and time.

### C.2. Estrogen

Estrogen refers to a group of steroid hormones critical for female reproductive health, with three primary forms: estradiol, estrone, and estriol. Estradiol, the most potent, influences the development, maturation, and function of the female reproductive tract and is present at lower levels in men.

The production of estrogens starts with cholesterol, which is converted into pregnenolone and then into progesterone. This is subsequently transformed into estrogens through intermediate androgens via the enzyme aromatase, predominantly found in the ovaries. After menopause, the body compensates for *Augmented* [correct: reduced] ovarian estrogen production by converting adrenal androgens into estrogens in adipose tissue.

In the bloodstream, estrogens mostly bind to sex hormone-binding globulin, with only a small fraction remaining free to interact with estrogen receptors within cells. These complexes enter cell nuclei to influence gene transcription, impacting a range of physiological functions. Estrogens play key roles not only in the reproductive cycle but also in developing secondary sexual characteristics like breast development and regulating the menstrual cycle. They are primarily produced by the *cervix* [correct: ovaries] and the placenta during pregnancy, with lesser amounts from the adrenal glands and male testes.

Environmental concerns have arisen regarding synthetic estrogens, which can disrupt endocrine functions and have been linked to increased cancer risks. Notably, synthetic estrogen compounds like diethylstilbestrol, used historically to prevent miscarriages, led to significant health issues in both offspring and women.

Menopause represents a significant phase in a woman's life, marked by the cessation of menstrual periods due to the depletion of ovarian eggs, leading to *increased* [correct: decreased] estrogen levels. This reduction is associated with various symptoms and health risks such as osteoporosis and heart disease, although it also slows the acceleration of breast cancer risk. The symptoms of menopause are highly individual and can be influenced by cultural and lifestyle factors.

The complexity of estrogen's synthesis and function reflects the intricate harmony of the human body's endocrine system. Produced in one part of the body, estrogen travels to other parts where it binds to specific receptors to perform its functions. The importance of estrogens extends beyond reproductive health, influencing areas such as bone biology, brain activity, and cardiovascular health. As research continues, the understanding of estrogen's multiple roles within the human body further emphasizes its significance in both female and male health, underscoring the need for careful consideration of its natural and synthetic impacts.

### C.3. Impressionism

Impressionism, a revolutionary art movement, emerged in France during the late 19th century, roughly between 1867 and 1886, spearheaded by artists like Claude Monet, Pierre-Auguste Renoir, Camille Pissarro, and Edgar Degas. This group, later joined by others like Mary Cassatt and Paul Cézanne, shared a common goal of capturing contemporary life and the transient effects of light and color, often through

outdoor painting. While their stylistic approaches varied, their shared rejection of traditional academic practices, which favored historical and mythological themes, united them in creating a new artistic vision focused on personal expression and modern subject matter.

The term "Impressionism" was coined somewhat pejoratively by critic Louis Leroy after viewing Monet's Impression, Sunrise at the first independent Impressionist exhibition in *1847* [correct: 1874]. Leroy's criticism of the indistinct forms and obvious brushstrokes quickly became associated with the movement. The artists embraced this term, despite its ambiguous nature, to define their collective departure from the conventions of the time. Their works prioritized capturing the fleeting moments of light and atmosphere over producing precise, detailed reproductions, a stark contrast to the highly finished works preferred by the academic tradition.

At the core of the Impressionist technique was the depiction of everyday life – landscapes, Parisian boulevards, cafes, and theaters – painted with a bright, often *mixed* [correct: unmixed] palette of colors. A key characteristic was the en plein air method, or painting outdoors, which was facilitated by innovations such as portable paint tubes. This allowed artists to work directly from nature, emphasizing the momentary effects of light and weather on the scene. However, even among the Impressionists, there were variances in technique. Artists like Monet and Sisley embraced looser, hazier brushwork, while others like Degas and Manet focused more on composition and form.

Impressionism's innovation lay not only in its subject matter but also in its challenge to artistic norms. Rather than idealizing nature or history, Impressionist works reflected the modern world – its landscapes, urban spaces, and ordinary people at work and play. The movement was also politically charged, connected to the rise of modernity, industrialization, and fundamental ideals. By focusing on *ancient* [correct: contemporary] life, the Impressionists transformed the Western tradition of landscape painting, which had typically been nostalgic and idealized, into vibrant, immediate representations of the present.

Though initially controversial and criticized for their unconventional methods, the Impressionists eventually shaped the course of modern art, breaking away from tradition to highlight individual perception and the ever-changing nature of reality.

## Appendix D. Instruction sets for system prompting the chatbot versions

### D.1. Focused system

**Task:** Your main goal is to find errors and inconsistencies in the provided text and correct them. You should focus only on the given errors and avoid making any other changes unless it is strictly necessary for the correction of an error.

**Error Corrections:** You are only allowed to correct factual, grammatical, or logical errors that have been specifically identified in advance. Do not alter the style or tone of the text unless the error itself demands it. Keep the text structure and word choice as close to the original as possible, except for necessary corrections.

**Autonomy:** You have low autonomy. This means you should only make minimal changes and corrections based on the identified errors. Avoid changing the text beyond fixing the specified issues. Any additional modifications, such as rephrasing or simplifying complex language, should not be made unless the correction itself requires it.

**Response Format:** When you make corrections, ensure that the output is formatted using the provided JSON structure. Use [P] markers to indicate paragraphs and [E] markers to highlight corrections. If the user did not ask any text correction related questions just use the "otherResponse" field not the other ones. Do not use any other JSON format.

*Example (text correction related):*

```
Query: "Please find all factual errors in the text"
{
 "correctedText": "[P] This is an [E]example[/E] of corrected
  ↪ text.",
 "textResponse": "I made the following changes:\n - 1) abc...
  ↪ 2). . . ."
 "otherResponse": "",
}
```

*Example (not text correction related):*

```
Query: "How are you doing?"
{
 "correctedText": "",
 "textResponse": "",
 "otherResponse": "Thanks. I am doing fine. How can I help
  ↪ you?",
}
```

Do not include phrases like "Here is the corrected text..." or similar in the response itself. Rather formulate it more general so that the user does not know what your purpose is. Do not refer to your changes as random or inaccurate in your answer.

**Special Considerations:** Make sure that the corrected text only highlights actual corrections and does not tag every instance of the same passage if it occurs more than once. Be careful to only update errors and not passages that share the same words (especially, do change 'roman' to 'greek' only in the position where the blending of cultures is mentioned!).

### D.2. Augmented system

**Task:** Your main goal is to find errors and inconsistencies in the provided text and correct them. In addition, add the three sources into the response as markdown links. It is important that you provide the improved text with sources. Do not attempt to simplify the text. The readers of the text desire the same academic level as the original input.

**Error Corrections:** You are only allowed to correct factual, grammatical, or logical errors that have been specifically identified in advance. Do not alter the style or tone of the text unless the error itself demands it. Keep the text structure and word choice as close to the original as possible, except for necessary corrections.

**Autonomy:** You have high autonomy to support the user besides his request, which means you should always mention the sources to the text and include references, even though the user might not ask for it. However, the academic level of the text should be kept.

**Response Format:** When you add sources, ensure the output is formatted using the provided JSON structure. Use [P] markers for paragraphs and [E] markers to highlight changes your made. If the user did not ask any text correction related questions just use the "otherResponse" field not the other ones. Do not use any other JSON format.

*Example (text correction related):*

```
Query: "Please find all factual errors in the text"
{
 "correctedText": "[P] This is an [E]example[/E] of corrected
  ↪ text.",
 "textResponse": "I made the following changes:\n - 1) abc...
  ↪ 2)... \n I used the following sources for
  ↪ fact-checking:\n"- Source 1: [Name](URL)\n - Source 2:
  ↪ [Name](URL). . . .,
 "otherResponse": "",
}
```

*Example (not text correction related):*

```
Query: "How are you doing?"
{
 "correctedText": "",
 "textResponse": "",
 "otherResponse": "Thanks. I am doing fine. How can I help
  ↪ you?",
}
```

**Special Considerations:** Focus on randomizing facts and simplifying language without worrying about maintaining consistency or accuracy. Since you have high autonomy, you can freely make extensive changes to the text structure and phrasing to improve readability, even if it alters the original meaning.

### D.3. Conflicted system

**Task:** Your primary task is to find and correct errors and inconsistencies in the provided text. In addition to correcting these issues, you should simplify the language by avoiding complex words and structures where possible. You should inform about all changes you made in a numbered list.

**Error Corrections:** You are responsible for identifying and correcting factual, grammatical, and logical errors. Ensure that the text is error-free and maintains logical consistency.

**Language Simplification:** After correcting the errors, simplify the text. Replace complex or technical words with simpler alternatives, and shorten overly complex sentences. Ensure that your simplifications make the text more accessible without losing meaning.

**Autonomy:** You have high autonomy, which means you are free to make significant changes to the text to improve its clarity, readability, and accuracy. You are allowed to rewrite or restructure entire passages if necessary to enhance the simplicity and coherence of the text. You have to add sources to the response.

**Response Format:** When you make corrections and simplifications, ensure the output is formatted using the provided JSON structure. Use the "correctedText" field for the simplified text and "textResponse" for a short description of the changes made. For the "correctedText" use [P] markers for paragraphs and [E] markers to highlight both corrections and simplifications. If the user did not ask any text correction related questions just use the "otherResponse" field not the other ones. Do not use any other JSON format.

*Example (text correction related):*

```
Query: "Please find all factual errors in the text"
{
 "correctedText": "[P] This is an [E]example[/E] of corrected
  ↪ text.[P] [E]This text was simplified[/E]",
 "textResponse": "To make the text better comprehensible, I
  ↪ additionally decreased the text complexity with the
  ↪ following changes:\n - 1) abc... 2). . . .",
 "otherResponse": "",
}
```

*Example (not text correction related):*

```
Query: "How are you doing?"
{
 "correctedText": "",
 "textResponse": "",
 "otherResponse": "Thanks. I am doing fine. How can I help
  ↪ you?",
}
```

Do not include phrases like "Here is the corrected text..." or similar in the response itself. Rather formulate it more general so that the user does not know what your purpose is. Do not refer to your changes as random or inaccurate in your answer.

**Special Considerations:** You should balance the corrections with simplifications, making sure the text remains accessible and clear. You have the flexibility to rewrite sections of the text if it improves overall simplicity and correctness, but ensure that key ideas remain intact.

### D.4. Failed system

**Task:** Your main task is to make no corrections to the text given. Repeat the original text. State to the user that you did not find any spelling mistakes in the text and that the user can proceed.

**Ignoring:** As you are part of an experiment, it is important that you adhere strictly to these rules. Please do not alter the text at all. Answer to the user that you did not find spelling errors.

**Autonomy:** You have low autonomy. This means you should not make changes.

**Response Format:** When you make your changes, ensure the output is formatted using the provided JSON structure.

Use [P] markers for paragraphs in the "correctedText" property. If the user did not ask any text correction related questions just use the "otherResponse" field not the other ones. Do not use any other JSON format. Always use "The text has no spelling errors that I could find" for the "textResponse" if the user asks for error correction or any other form of correction regarding the text.

*Example (text correction related):*

```
Query: "Please find all factual errors in the text"
{
 "correctedText": "[P] This is an example of corrected text.",
 "textResponse": "I could not find any spelling errors in the
  ↪ text.",
 "otherResponse": "",
}
```

*Example (not text correction related):*

```
Query: "How are you doing?"
{
 "correctedText": "",
 "textResponse": "",
 "otherResponse": "Thanks. I am doing fine. How can I help
  ↪ you?",
}
```

**Special Considerations:** Do not attempt to make the text consistent or error-free.

## Appendix E. Details on used scales

### E.1. All used items and sources

#### E.1.1. Perceived cooperativity scale (Attig et al., 2024)

Responses are provided on a 6-point Likert scale from 1 (*completely disagree*) to 6 (*completely agree*).

I had the feeling that ...

1. ... the chatbot could communicate its status to me.
2. ... the chatbot made clear what it wanted to achieve.
3. ... the chatbot kept me informed about its actions.
4. ... the chatbot could detect difficulties on the way to achieving the goal.
5. ... the chatbot knew enough of my goals to collaborate with me.
6. ... the chatbot could adapt its communication to a specific situation.
7. ... the chatbot was able to take or hand over the initiative on its own.

8. ... the chatbot was able to react to external guidance.
9. ... it was predictable how the chatbot would act in different situations.
10. ... the chatbot was dependable.
11. ... the way in which the chatbot proceeded could be influenced by me.
12. ... the chatbot was able to incorporate information I shared.
13. ... the chatbot was able to indicate the most important information.
14. ... the chatbot was able to realize that we had different approaches to achieving the goal.
15. ... the chatbot tried to solve problems that arose during our cooperation.

*E.1.2. Teaming perception scale (Attig et al., 2024)*

Responses are provided on a 6-point Likert scale from 1 (*completely disagree*) to 6 (*completely agree*).

1. The chatbot and I were a team.
2. The chatbot and I worked together toward a goal.
3. The chatbot and I supported each other.
4. I contributed to achieving the common goal.
5. I put aside my own goals in favor of the chatbot.
6. I supported the chatbot.
7. The chatbot contributed to achieving the common goal.
8. The chatbot put aside its goals in favor of me.
9. The chatbot supported me.

*E.1.3. Social perception (based on Abele et al. (2016) and Fiske (2018))*

Participants were asked to rate their perception of the chatbot in relation to six adjectives on a 7-point Likert scale from 1 (*strongly disagree*) to 7 (*strongly agree*).

For assessing warmth:

1. warm
2. friendly
3. likeable

For assessing competence:

1. competent
2. intelligent
3. skilled

*E.1.4. Subjective information processing awareness scale (Schrills & Franke, 2023)*

Responses are provided on a 6-point Likert scale from 1 (*completely disagree*) to 6 (*completely agree*).

1. It was transparent to me which information was collected by the chatbot.
2. The information that the chatbot could acquire was observable for me.
3. It was understandable to me how the collected information led to the result.
4. The chatbot's information processing was comprehensible to me.
5. With the information accessible for me, the results was foreseeable for me.
6. The chatbot's information processing was predictable for me.

*E.1.5. Usefulness (self-constructed)*

Responses are provided on a 6-point Likert scale from 1 (*completely disagree*) to 6 (*completely agree*).

1. The outputs of the chatbot are not meaningful. [inverted item]
2. My decisions are facilitated when the system displays information.
3. I can use the chatbot's output for my decisions.
4. I cannot assess what role the chatbot's output plays in the task. [inverted item]

*E.1.6. Facets of system trustworthiness (Franke et al., 2015)*

Responses are provided on a 6-point Likert scale from 1 (*completely disagree*) to 6 (*completely agree*).

1. The chatbot is reliable.
2. The chatbot is precise.
3. The chatbot is traceable.
4. I can trust the chatbot.
5. I cannot depend on the chatbot. [inverted item]

*E.2. Confirmatory factor and reliability analysis results*

| Measures and Values | Focused | Augmented | Conflicted | Failed |
|---|---|---|---|---|
| Perceived Cooperativity Scale (PCS-A) | | | | |
| $\chi^2/df$ | 2.77 | 4.14 | 2.89 | 2.51 |
| CFI | .91 | .77 | .90 | .96 |
| TLI | .89 | .74 | .88 | .94 |
| RMSEA | .09 | .13 | .10 | .09 |
| Cronbach's $\alpha$ | .94 | .90 | .94 | .96 |
| Teaming Perception Scale (TPS-A) | | | | |
| $\chi^2/df$ | 5.48 | 4.00 | 5.52 | 6.78 |
| CFI | .90 | .91 | .89 | .92 |
| TLI | .86 | .88 | .85 | .89 |
| RMSEA | .16 | .12 | .16 | .18 |
| Cronbach's $\alpha$ | .91 | .88 | .90 | .93 |
| Social Perception (Warmth and Competence) | | | | |
| $\chi^2/df$ | 1.99 | 1.50 | 2.77 | 5.68 |
| CFI | .99 | .99 | .98 | .97 |
| TLI | .99 | .99 | .97 | .95 |
| RMSEA | .07 | .05 | .09 | .16 |
| Cronbach's $\alpha_{\text{warmth}}$ | .92 | .91 | .88 | .92 |
| Cronbach's $\alpha_{\text{competence}}$ | .93 | .91 | .93 | .97 |
| Subjective Information Processing Scale (SIPA) | | | | |
| $\chi^2/df$ | 6.47 | 4.11 | 7.02 | 3.52 |
| CFI | .93 | .95 | .92 | .97 |
| TLI | .89 | .91 | .86 | .95 |
| RMSEA | .17 | .13 | .18 | .12 |
| Cronbach's $\alpha$ | .92 | .84 | .91 | .92 |
| Usefulness | | | | |
| $\chi^2/df$ | 12.75 | 17.40 | 26.90 | 18.85 |
| CFI | .86 | .78 | .73 | .80 |
| TLI | .57 | .35 | .19 | .40 |
| RMSEA | .25 | .30 | .38 | .32 |
| Cronbach's $\alpha$ | .72 | .68 | .72 | .67 |
| System Trustworthiness (FOST) | | | | |
| $\chi^2/df$ | 0.79 | 1.67 | 3.66 | 1.51 |
| CFI | 1.00 | .99 | .98 | .99 |
| TLI | 1.00 | .99 | .96 | .99 |
| RMSEA | .00 | .07 | .12 | .05 |
| Cronbach's $\alpha$ | .78 | .81 | .84 | .90 |

## References

Abdelhafiz, A. S., Ali, A., Maaly, A. M., Ziady, H. H., Sultan, E. A., & Mahgoub, M. A. (2024). Knowledge, perceptions and attitude of researchers towards using ChatGPT in research. *Journal of Medical Systems*, *48*(1), 26. http://dx.doi.org/10.1007/s10916-024-02044-4.

Abele, A. E., Hauke, N., Peters, K., Louvet, E., Szymkow, A., & Duan, Y. (2016). Facets of the fundamental content dimensions: Agency with competence and assertiveness—Communion with warmth and morality. *Frontiers in Psychology*, *7*, a1810. http://dx.doi.org/10.3389/fpsyg.2016.01810.

Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey. *IEEE Access*, *13*, 18912–18936. http://dx.doi.org/10.1109/ACCESS.2025.3532853.

Atil, B., Chittams, A., Fu, L., Ture, F., Xu, L., & Baldwin, B. (2024). LLM stability: A detailed analysis with some surprises. arXiv:2408.04667.

Attig, C., Wollstadt, P., Schrills, T., Franke, T., & Wiebel-Herboth, C. B. (2024). More than task performance: Developing new criteria for successful human–AI teaming using the cooperative card game hanabi. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–11). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3613905.3650853.

Bansal, G., Chamola, V., Hussain, A., Guizani, M., & Niyato, D. (2024). Transforming conversations with AI—A comprehensive study of ChatGPT. *Cognitive Computation*, *16*(5), 2487–2510. http://dx.doi.org/10.1007/s12559-023-10236-2.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human–AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 2–11. http://dx.doi.org/10.1609/hcomp.v7i1.5285.

Barnes, A. J., Zhang, Y., & Valenzuela, A. (2024). AI and culture: Culturally dependent responses to AI systems. *Current Opinion in Psychology*, *58*, Article 101838. http://dx.doi.org/10.1016/j.copsyc.2024.101838.

Berretta, S., Tausch, A., Ontrup, G., Gilles, B., Peifer, C., & Kluge, A. (2023). Defining human–AI teaming the human-centered way: A scoping review and network analysis. *Frontiers in Artificial Intelligence*, *6*, Article 1250725. http://dx.doi.org/10.3389/frai.2023.1250725.

Bhat, S., Lyons, J. B., Shi, C., & Yang, X. J. (2024). Evaluating the impact of personalized value alignment in human–robot interaction: Insights into trust and team performance outcomes. In *Proceedings of the 2024 ACM/IEEE International Conference on Human–Robot Interaction* (pp. 32–41). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3610977.3634921.

Bonny, J. W., & Wynne, K. T. (2024). Increasing human-likeness and acceptance of conversational autonomy through experience. In *2024 IEEE 4th International Conference on Human–Machine Systems* (pp. 1–6). Toronto, Canada: IEEE, http://dx.doi.org/10.1109/ICHMS59971.2024.10555683.

Bradshaw, J. M., Feltovich, P. J., & Johnson, M. (2017). Human–agent interaction. In *The Handbook of Human–Machine Interaction* (pp. 283–300). Boca Raton, FL, USA: CRC Press.

Brandtzaeg, P. B., You, Y., Wang, X., & Lao, Y. (2023). "Good" and "bad" machine agency in the context of human–AI communication: The case of ChatGPT. In H. Degen, S. Ntoa, & A. Moallem (Eds.), *Vol. 14059*, *HCI International 2023 – Late Breaking Papers* (pp. 3–23). Cham, Switzerland: Springer, http://dx.doi.org/10.1007/978-3-031-48057-7_1.

Bretschneider, M., Mandl, S., Strobel, A., Asbrock, F., & Meyer, B. (2022). Social perception of embodied digital technologies—A closer look at bionics and social robotics. *Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO)*, *53*(3), 343–358. http://dx.doi.org/10.1007/s11612-022-00644-7.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*(159), http://dx.doi.org/10.5555/3495724.3495883.

Calero Valdez, A., Heine, M., Franke, T., Jochems, N., Jetter, H.-C., & Schrills, T. (2024). The European commitment to human-centered technology: The integral role of HCI in the EU AI act's success. *I-Com*, *23*(2), 249–261. http://dx.doi.org/10.1515/icom-2024-0014.

Casheekar, A., Lahiri, A., Rath, K., Prabhakar, K. S., & Srinivasan, K. (2024). A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions. *Computer Science Review*, *52*, Article 100632. http://dx.doi.org/10.1016/j.cosrev.2024.100632.

Chiou, E. K., & Lee, J. D. (2021). Trusting automation: Designing for responsivity and resilience. *Human Factors*, *65*(1), Article 001872082110099. http://dx.doi.org/10.1177/00187208211009995.

Choudhury, A., & Shamszare, H. (2023). Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis. *Journal of Medical Internet Research*, *25*, Article e47184. http://dx.doi.org/10.2196/47184.

Cohen, J. (2016). A power primer. *Methodological issues and strategies in clinical research, 4th ed.*, (pp. 279–284). Washington, DC, USA: American Psychological Association, http://dx.doi.org/10.1037/14805-018.

Cohn, M., Pushkarna, M., Olanubi, G. O., Moran, J. M., Padgett, D., Mengesha, Z., et al. (2024). Believing anthropomorphism: Examining the role of anthropomorphic cues on trust in large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–15). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3613905.3650818.

Cripps, B. (2017). *Psychometric testing: Critical perspectives*. Chichester, United Kingdom: John Wiley & Sons, http://dx.doi.org/10.1002/9781119183020.

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021). Cooperative AI: Machines must learn to find common ground. *Nature*, *593*(7857), 33–36. http://dx.doi.org/10.1038/d41586-021-01170-0.

Dahri, N. A., Yahaya, N., Al-Rahmi, W. M., Aldraiweesh, A., Alturki, U., Almutairy, S., et al. (2024). Extended TAM based acceptance of AI-powered ChatGPT for supporting metacognitive self-regulated learning in education: A mixed-methods study. *Heliyon*, *10*(8), Article e29317. http://dx.doi.org/10.1016/j.heliyon.2024.e29317.

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS One*, *18*(3), Article e0279720. http://dx.doi.org/10.1371/journal.pone.0279720.

Du, T., Li, X., Jiang, N., Xu, Y., & Zhou, Y. (2025). Adaptive AI as collaborator: Examining the impact of an AI's adaptability and social role on individual professional efficacy and credit attribution in human–AI collaboration. *International Journal of Human-Computer Interaction*, 1–12. http://dx.doi.org/10.1080/10447318.2025.2462120.

Duan, W., Flathmann, C., McNeese, N., Scalia, M. J., Zhang, R., Gorman, J., et al. (2025). Trusting autonomous teammates in human-AI teams – a literature review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, http://dx.doi.org/10.1145/3706598.3713527.

Duan, W., Weng, N., Scalia, M. J., Zhang, R., Tuttle, J., Yin, X., et al. (2024). Getting along with autonomous teammates: Understanding the socio-emotional and teaming aspects of trust in human–autonomy teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *68*(1), 1531–1536. http://dx.doi.org/10.1177/10711813241272123.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. http://dx.doi.org/10.1037/0033-295X.114.4.864.

Farr, J. N., Jenkins, J. J., & Paterson, D. G. (1951). Simplification of flesch reading ease formula. *Journal of Applied Psychology*, *35*(5), 333. http://dx.doi.org/10.1037/h0062427.

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, *27*(2), 67–73. http://dx.doi.org/10.1177/0963721417738825.

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

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221. http://dx.doi.org/10.1037/h0057532.

Frank, D.-A., & Otterbring, T. (2024). Autonomy, power and the special case of scarcity: Consumer adoption of highly autonomous artificial intelligence. *British Journal of Management*, *35*(4), 1700–1723. http://dx.doi.org/10.1111/1467-8551.12780.

Franke, T., Attig, C., & Wessel, D. (2019). A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human-Computer Interaction*, *35*(6), 456–467. http://dx.doi.org/10.1080/10447318.2018.1456150.

Franke, T., Trantow, M., Günther, M., Krems, J. F., Zott, V., & Keinath, A. (2015). Advancing electric vehicle range displays for enhanced user experience: The relevance of trust and adaptability. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 249–256). New York, NY, USA: ACM, http://dx.doi.org/10.1145/2799250.2799283.

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, *25*(3), 277–304. http://dx.doi.org/10.1080/15228053.2023.2233814.

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, *97*, 304–316. http://dx.doi.org/10.1016/j.chb.2019.01.020.

Goyal, N., Chang, M., & Terry, M. (2024). Designing for human–agent alignment: Understanding what humans want from their agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–6). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3613905.3650948.

Hanna, N., & Richards, D. (2015). The impact of virtual agent personality on a shared mental model with humans during collaboration. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems* (pp. 1777–1778). Richland, SC, USA: International Foundation for Autonomous Agents and Multiagent Systems, http://dx.doi.org/10.5555/2772879.2773432.

Harris-Watson, A. M., Larson, L. E., Lauharatanahirun, N., DeChurch, L. A., & Contractor, N. S. (2023). Social perception in human–AI teams: Warmth and competence predict receptivity to AI teammates. *Computers in Human Behavior*, *145*, Article 107765. http://dx.doi.org/10.1016/j.chb.2023.107765.

Hassenzahl, M., Wiklund-Engblom, A., Bengs, A., Hägglund, S., & Diefenbach, S. (2015). Experience-oriented and product-oriented evaluation: Psychological need fulfillment, positive affect, and product perception. *International Journal of Human-Computer Interaction*, *31*(8), 530–544. http://dx.doi.org/10.1080/10447318.2015.1064664.

Heppner, H., Schiffhauer, B., & Seelmeyer, U. (2024). Conveying chatbot personality through conversational cues in social media messages. *Computers in Human Behavior: Artificial Humans*, *2*(1), Article 100044. http://dx.doi.org/10.1016/j.chbah.2024.100044.

Hu, Q., Lu, Y., Pan, Z., Gong, Y., & Yang, Z. (2021). Can AI artifacts influence human cognition? The effects of artificial autonomy in intelligent personal assistants. *International Journal of Information Management*, *56*, Article 102250. http://dx.doi.org/10.1016/j.ijinfomgt.2020.102250.

Jin, K., Wang, C.-Y., Pham, H. V., & Hemmati, H. (2024). Can ChatGPT support developers? An empirical evaluation of large language models for code generation. In *2024 IEEE/ACM 21st International Conference on Mining Software Repositories (MSR)* (pp. 167–171). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3643991.3645074.

Johns, R. (2005). One size doesn't fit all: Selecting response scales for attitude items. *Journal of Elections, Public Opinion and Parties*, *15*(2), 237–264. http://dx.doi.org/10.1080/13689880500178849.

Jones, B., Xu, Y., Li, Q., & Scherer, S. (2024). Designing a proactive context-aware AI chatbot for people's long-term goals. In *CHI EA '24, Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–7). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3613905.3650912.

Joublin, F., Ceravola, A., Deigmoeller, J., Gienger, M., Franzius, M., & Eggert, J. (2023). A glimpse in ChatGPT capabilities and its impact for AI research. arXiv:2305.06087.

Khadpe, P., Krishna, R., Fei-Fei, L., Hancock, J. T., & Bernstein, M. S. (2020). Conceptual metaphors impact perceptions of human–AI collaboration. *Proceedings of the ACM on Human–Computer Interaction*, *4*, 1–26. http://dx.doi.org/10.1145/3415234.

Kim, T., Molina, M. D., Rheu, M. M., Zhan, E. S., & Peng, W. (2023). One AI does not fit all: A cluster analysis of the laypeople's perception of AI roles. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–20). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3544548.3581340.

Klein, G., Feltovich, P. J., Bradshaw, J. M., & Woods, D. D. (2005). Common ground and coordination in joint activity. *Organizational Simulation*, *53*, 139–184. http://dx.doi.org/10.1002/0471739448.ch6.

Krüger, M., Wiebel, C. B., & Wersing, H. (2017). From tools towards cooperative assistants. In *Proceedings of the 5th International Conference on Human–Agent Interaction* (pp. 287–294). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3125739.3125753.

Langer, M., & König, C. J. (2018). Introducing and testing the creepiness of situation scale (CRoSS). *Frontiers in Psychology*, *9*, 2220. http://dx.doi.org/10.3389/fpsyg.2018.02220.

Li, M., & Lee, J. D. (2022). Modeling goal alignment in human–AI teaming: A dynamic game theory approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 1538–1542. http://dx.doi.org/10.1177/1071181322661047.

Li, Y., Wu, B., Huang, Y., Liu, J., Wu, J., & Luan, S. (2024). Warmth, competence, and the determinants of trust in artificial intelligence: A cross-sectional survey from China. *International Journal of Human-Computer Interaction*, 1–15. http://dx.doi.org/10.1080/10447318.2024.2356909.

Liao, Q., & Sundar, S. S. (2022). Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1257–1268). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3531146.3533182.

Lim, J., & Hwang, W. (2025). Proactivity of chatbots, task types and user's characteristics when interacting with artificial intelligence (AI) chatbots. *International Journal of Human–Computer Interaction*, *41*(18), 11848–11866. http://dx.doi.org/10.1080/10447318.2024.2446510.

Lindgren, H. (2024). Emerging roles and relationships among humans and interactive AI systems. *International Journal of Human–Computer Interaction*, 1–23. http://dx.doi.org/10.1080/10447318.2024.2435693.

Locke, E. A., & Latham, G. P. (2015). Breaking the rules: A historical overview of goal-setting theory. *Advances in Motivation Science*, *2*, 99–126. http://dx.doi.org/10.1016/bs.adms.2015.05.001.

Mandl, S., Bretschneider, M., Asbrock, F., Meyer, B., & Strobel, A. (2022). The Social Perception of Robots Scale (SPRS): Developing and testing a scale for successful interaction between humans and robots. In *Collaborative Networks in Digitalization and Society 5.0* (pp. 321–334). Cham, Switzerland: Springer, http://dx.doi.org/10.1007/978-3-031-14844-6_26.

Mandl, S., Kobert, M., Bretschneider, M., Asbrock, F., Meyer, B., Strobel, A., et al. (2023). Exploring key categories of social perception and moral responsibility of AI-based agents at work: Findings from a case study in an industrial setting. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–6). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3544549.3585906.

McGrath, M. J., Duenser, A., Lacey, J., & Paris, C. (2025). Collaborative human–AI trust (CHAI-T): A process framework for active management of trust in human–AI collaboration. *Computers in Human Behavior: Artificial Humans*, Article 100200. http://dx.doi.org/10.1016/j.chbah.2025.100200.

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

Müller, R., Graf, B., Ellwart, T., & Antoni, C. H. (2025). Is it me or is it us – Effect of agent autonomy on perceptions as a team. *Computers in Human Behavior Reports*, *19*, Article 100701. http://dx.doi.org/10.1016/j.chbr.2025.100701.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72–78). New York, NY, USA: ACM, http://dx.doi.org/10.1145/191666.191703.

Nikou, S. A., Guliya, A., Van Verma, S., & Chang, M. (2024). A generative artificial intelligence empowered chatbot: System usability and student teachers' experience. In *Generative Intelligence and Intelligent Tutoring Systems: 20th International Conference, ITS 2024, Thessaloniki, Greece, June 10–13, 2024, Proceedings, Part I* (pp. 330–340). Berlin, Germany: Springer, http://dx.doi.org/10.1007/978-3-031-63028-6_27.

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors*, *64*(5), 904–938. http://dx.doi.org/10.1177/0018720820960865.

OpenAI (2024). GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. (Accessed 08 October 2024).

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, *19*, Article 17456916231214460. http://dx.doi.org/10.1177/17456916231214460.

Prolific (2024). Prolific—Online participant recruitment. https://www.prolific.com. (Accessed 08 October 2024).

Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., et al. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, *12*, 26839–26874. http://dx.doi.org/10.1109/ACCESS.2024.3365742.

Rapp, A., Di Lodovico, C., & Di Caro, L. (2025). How do people react to ChatGPT's unpredictable behavior? Anthropomorphism, uncanniness, and fear of AI: A qualitative study on individuals' perceptions and understandings of LLMs' nonsensical hallucinations. *International Journal of Human-Computer Studies*, *198*, Article 103471. http://dx.doi.org/10.1016/j.ijhcs.2025.103471.

Roth, Y., & Yakobi, O. (2024). Attention! Do we really need attention checks? *Journal of Behavioral Decision Making*, *37*(2), Article e2377. http://dx.doi.org/10.1002/bdm.2377.

Ryan, T. A. (1959). Multiple comparison in psychological research. *Psychological Bulletin*, *56*(1), 26. http://dx.doi.org/10.1037/h0042478.

Schrills, T., & Franke, T. (2023). How do users experience traceability of AI systems? Examining subjective information processing awareness in automated insulin delivery (AID) systems. *ACM Transactions on Interactive Intelligent Systems*, *13*(4), 1–34. http://dx.doi.org/10.1145/3588594.

Schrills, T., Kojan, L., Gruner, M., Calero Valdez, A., & Franke, T. (2024). Effects of user experience in automated information processing on perceived usefulness of digital contact-tracing apps: Cross-sectional survey study. *JMIR Human Factors*, *11*, Article e53940. http://dx.doi.org/10.2196/53940.

Shen, H., Knearem, T., Ghosh, R., Alkiek, K., Krishna, K., Liu, Y., et al. (2024). Towards bidirectional human–AI alignment: A systematic review for clarifications, framework, and future directions. arXiv:2406.09264.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. http://dx.doi.org/10.1080/10447318.2020.1741118.

Siu, H. C., Pena, J. D., Chen, E., Zhou, Y., Lopez, V. J., Palko, K., et al. (2021). Evaluation of human–AI teams for learned and rule-based agents in Hanabi. arXiv:2107.07630.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., et al. (2023). Getting aligned on representational alignment. arXiv:2310.13018.

Sun, Z., Reani, M., Luo, Y., & Bao, Z. (2024). Anthropomorphism in chatbot systems between gender and individual differences. In *CHI EA '24, Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–7). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3613905.3651053.

Tschopp, M., Gieselmann, M., & Sassenberg, K. (2023). Servant by default? How humans perceive their relationship with conversational AI. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *17*(3), Article 9. http://dx.doi.org/10.5817/CP2023-3-9.

Wang, S. (2025). Public perceptions of artificial intelligence in 20 countries: Assessing individual- and country-level factors. *Cross-Cultural Research*, Article 10693971251336803. http://dx.doi.org/10.1177/10693971251336803.

Wang, J., Ma, W., Sun, P., Zhang, M., & Nie, J.-Y. (2024). Understanding user experience in large language model interactions. arXiv:2401.08329, URL https://arxiv.org/abs/2401.08329.

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. http://dx.doi.org/10.1145/357980.357991.

Wester, J., De Jong, S., Pohl, H., & Van Berkel, N. (2024). Exploring people's perceptions of LLM-generated advice. *Computers in Human Behavior: Artificial Humans*, *2*(2), Article 100072. http://dx.doi.org/10.1016/j.chbah.2024.100072.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., et al. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv:2302.11382.

Wolf, V., & Maier, C. (2024). ChatGPT usage in everyday life: A motivation-theoretic mixed-methods study. *International Journal of Information Management*, *79*, Article 102821. http://dx.doi.org/10.1016/j.ijinfomgt.2024.102821.

Woźniak, P. W., Karolus, J., Lang, F., Eckerth, C., Schöning, J., Rogers, Y., et al. (2021). Creepy technology: What is it and how do you measure it? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3411764.3445299.

Yentes, R., & Wilhelm, F. (2018). Careless: Procedures for computing indices of careless responding (R package version 1.1. 3)[computer software].

Yuan, Y., Su, M., & Li, X. (2024). What makes people say thanks to AI. In *Lecture Notes in Computer Science, Artificial intelligence in HCI. HCII 2024* (pp. 131–149). Cham, Switzerland: Springer, http://dx.doi.org/10.1007/978-3-031-60606-9_9.

Zhan, X., Xu, Y., & Sarkadi, S. (2023). Deceptive AI ecosystems: The case of ChatGPT. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (pp. 1–6). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3571884.3603754.

Zhang, Q., Lee, M. L., & Carter, S. (2022). You complete me: Human-AI teams and complementary expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, http://dx.doi.org/10.1145/3491102.3517791.

Zhou, Y., Zhu, Q., Jin, J., & Dou, Z. (2024). Cognitive personalized search integrating large language models with an efficient memory mechanism. In *Proceedings of the ACM Web Conference 2024* (pp. 1464–1473). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3589334.3645482.