



# Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style

Luise Metzger  
luise.metzger@uni-ulm.de  
Ulm University  
Ulm, Germany

Martin Baumann  
martin.baumann@uni-ulm.de  
Ulm University  
Ulm, Germany

Linda Miller  
linda.miller@uni-ulm.de  
Ulm University  
Ulm, Germany

Johannes Kraus  
johannes.kraus@uni-ulm.de  
Ulm University  
Ulm, Germany

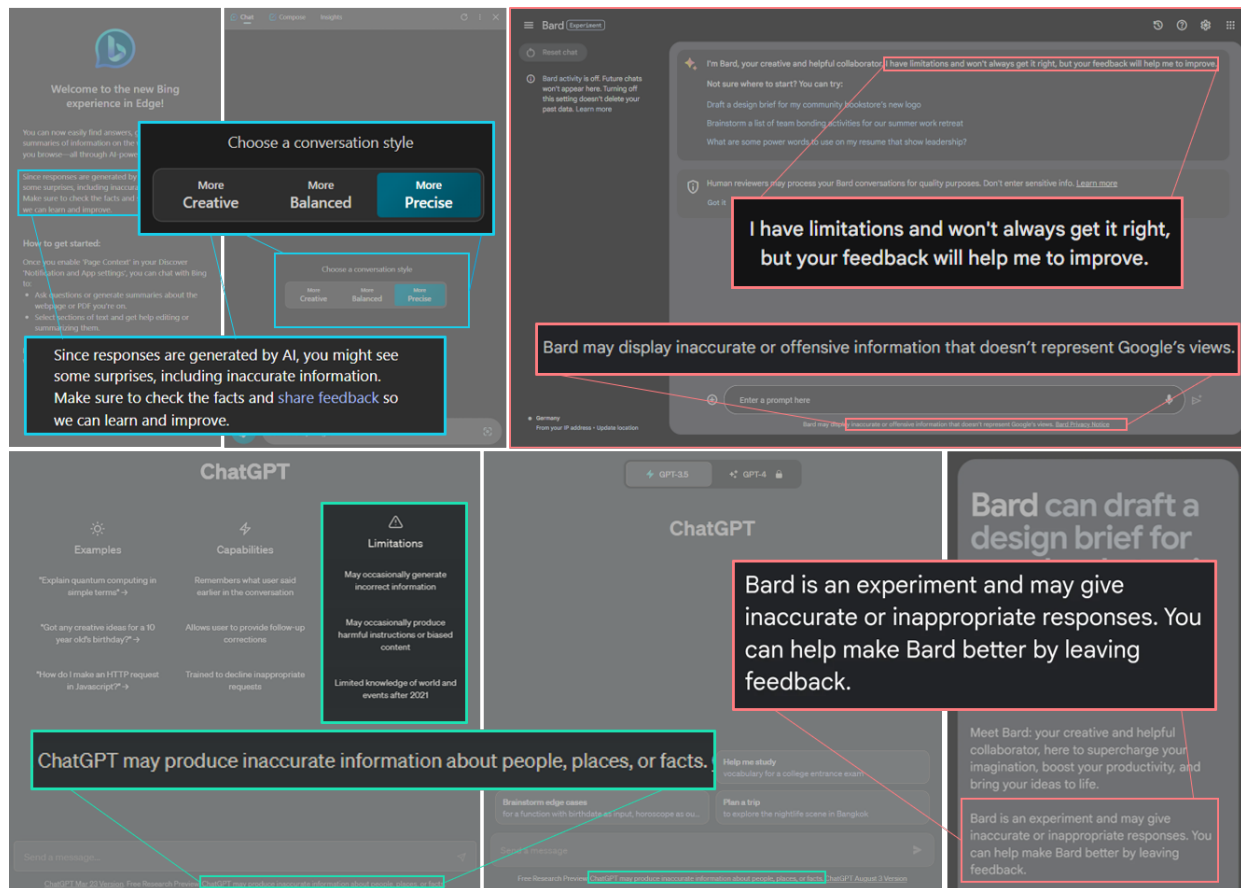


Figure 1: Screenshots of user interfaces of LLM-based conversational agents with boxes highlighting sections mentioning limitations. Top left & middle: Microsoft's Bing [39], bottom left (Mar 23 version) & middle (Aug 23 version): OpenAI's ChatGPT [46], top right & bottom right: Google's Bard [1]. Insets provide zoomed views for enhanced readability.

## ABSTRACT

While conversational agents based on Large Language Models (LLMs) can drive progress in many domains, they are prone to generating faulty information. To ensure an efficient, safe, and satisfactory user experience maximizing benefits of these systems, users must be empowered to judge the reliability of system outputs. In this, both disclaimers and agents' communicative style are pivotal design instances. In an online study with 594 participants, we investigated how these affect users' trust and a mock-up agent's persuasiveness, based on an established framework from social psychology. While prior information on potential inaccuracies or faulty information did not affect trust, an authoritative communicative style elicited more trust. Also, a trusted agent was more persuasive resulting in more positive attitudes regarding the subject of the conversation. Results imply that disclaimers on agents' limitations fail to effectively alter users' trust but can be supported by appropriate communicative style during interaction.

## CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

## KEYWORDS

large language models, conversational agents, chatbots, ChatGPT, trust in automation, elaboration likelihood model, communicative style

### ACM Reference Format:

Luise Metzger, Linda Miller, Martin Baumann, and Johannes Kraus. 2024. Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3613904.3642122>

## 1 INTRODUCTION

Recent developments in natural language processing are impacting society: The release of large language models (LLMs) like OpenAI's ChatGPT [46], which are now easily accessible for a multitude of users [19], has sparked avid discussion in both the scientific community and the public at large. Besides potential applications and benefits of these models which outperform humans in a range of tasks, this also includes criticisms of their potential for harm [3, 33, 58]. For prevalent systems, one essential issue is that LLMs might provide factually incorrect information in ways that sound authoritative and confident [26, 58]. While the underlying models display high formal linguistic competence by generating grammatically correct and seemingly coherent text, they lack functional linguistic competence: LLMs lack genuine semantic comprehension of language which would allow them to reliably assess the truth content of their

own output [35, 58, 61]. Thus, these models can be prone to generating misinformation [5, 8, 22, 61], which, however, can be hard to identify because the mere formal quality of generated text tends to get mistaken for meaningfulness [5]. This is specifically problematic for LLM-based applications that interact with human users, such as conversational agents, which engage users via dialogue, and, even more so, with non-expert users, to whom LLM-based applications are now available.

Beyond an application's boundaries, such misinformation can then adversely affect a broader audience when spread by initial users, whether with malicious intent to persuade or accidentally because they failed to identify false statements. Since the release of ChatGPT in November 2022, multiple instances of this have occurred: Within a week after ChatGPT's initial release on November 30th 2022, an influx of seemingly coherent but incorrect answers on Stack Overflow, which substantially harmed the site's functionality, eventually lead to a ban on using generative AI when posting content [36]. Several weeks later, media website CNET paused AI-generated content after having had to issue corrections for 41 out of 77 previously published articles which were belatedly found to contain factual errors [54]. Even prior to this, the public demo of Meta's Galactica model for processing scientific knowledge [59] was taken down after only three days online on November 18th 2022, after facing heavy criticism for making up plausible but non-existing or factual incorrect technical explanations, academic papers, and citations [57]. Distributors of these systems and the expert community are aware of this issue: Exemplarily, OpenAI lists this as a main limitation in their GPT-4 report [47], and, in the scientific community, publication policies like ACM SIGCHI's explicitly forbid listing AI tools as an author and hold (human) authors accountable for the correctness of submitted works [38]. However, while expert-level discussions and outlet-specific guidelines or bans provide some safeguard against the spread of misinformation to a larger audience, they do not tackle the problem at its source, i.e., lay users who distribute incorrect information because they fail to recognize it as such.

Especially for applications catered to the general public, users' awareness of the system's capabilities and limitations is an essential point of consideration not only when viewed through the lens of efficient system design, but also from an ethical standpoint. Since ethical and user-centered system design should facilitate appropriate inferences about a system's capabilities and the trustworthiness of its output [20, 32], potential users need to be provided with information that empowers them to assess these and to thus meaningfully consent to the system's application [34].

The discussed increased prevalence of conversational agents in combination with their shortcomings and lay users who might too uncritically take provided information for granted underlines the need for human-centered design. An important basis for this is to consider human information processing and decision-making processes when interacting with a conversational agent and paying attention to these processes in designing provided information about the agent's capabilities and limitations as well as the accompanying user interface. Currently, interfaces of prevalent applications, for instance, OpenAI's ChatGPT [46], Google's Bard [1], or Microsoft's Bing [39], display such information to varying extents



This work is licensed under a Creative Commons Attribution International 4.0 License.

CHI '24, May 11–16, 2024, Honolulu, HI, USA  
© 2024 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-0330-0/24/05  
<https://doi.org/10.1145/3613904.3642122>

of visibility (see Figure 1). However, the adequacy of this disclosure is a topic of discussion, as elaborated in subsection 2.1. At this point, it is unclear whether this really suffices to support users in properly weighing and integrating this ‘warning’ information into their evaluation of agents and their outputs, and use it in the subsequent interaction with the agent. This applies especially for interactions with LLM-based agents, during which aspects like their formal linguistic competence and authoritative style are more readily available and might promote overreliance [5, 61]. Accordingly, the underlying question of this research is to investigate how both communicating the limitations of a conversational agent at the beginning of an interaction and an alternative communicative style reflecting a certain degree of uncertainty affect users’ trust in the agent, as well as how this trust then affects attitudes towards the content of messages delivered by the agent.

While a number of publications have addressed misinformation risks of LLMs generating incorrect information which users may mistake for true [5, 35, 58, 61], research focusing on how to enhance users’ capabilities to adjust their evaluation of such systems accordingly to prevent overtrust and, in consequence, overreliance is still rare. Following recent calls for research in this domain [20, 34], we investigate how trust in a conversational agent representative of current prevalent applications is formed and utilized, drawing on well-established theories from social psychology and research on trust in automation. We focus on two key aspects: Firstly, we consider the role of prior information provided on the starting page of a conversational agent application, which can alert users that provided outputs might be inaccurate. Secondly, we consider the agent’s communicative style, which can either reflect the uncertainty of outputs, or display overconfidence regarding the potentially limited validity and reliability of the information given.

## 2 RELATED WORK

### 2.1 Avoiding Overtrust in Conversational Agents

Due to the complex nature of deep learning models, users of LLM-based conversational agents cannot generally be expected to comprehend how exactly outputs are generated or how these systems learn in the first place. According to psychological research, in lieu of full understanding of a system, users will rely on their attitudes as a heuristic to guide interaction, based on beliefs about its capabilities [27, 32]. In decision-making under uncertainty and in face of potential negative consequences for individuals, a key attitude in interacting with intelligent technology is trust in automation [27, 32]. Trust in automation, defined as “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” [32, p. 54], was found to affect reliance in and adoption of intelligent systems. Essentially, to facilitate safe and efficient interactions, an appropriate calibration of trust is required. Calibrated trust is achieved when the perceived trustworthiness of a system matches its actual capabilities [32]. In contrast, a mismatch between users’ trust and a system’s capabilities can take the form of either distrust, if a system is trusted less than its capabilities would warrant, or overtrust, if trust exceeds system capabilities [32]. Regarding LLM-based agents, the mismatch between high formal and lack of functional linguistic competence

poses a risk of users overtrusting the agent. Overtrust can then promote misuse [32], specifically, overreliance on the agent [5, 61] as a behavioral consequence. Considering the goal to design for calibrated trust, it is thus essential to investigate strategies how this inherent risk of overtrust in LLM-based agents can be mitigated. While calibrated trust describes the adequate match between the user’s trust and the system’s capabilities [32], the former is not necessarily informed by the latter exclusively. Rather than being based on the system’s capabilities, trust can also (unwarrantedly) be influenced by more secondary aspects of a system, such as the user interface [20]. Thereby, trust has been conceptualized as a variable that is build up in an individual learning process, in which pieces of information about the trustworthiness of a system - so called trust cues - are used to dynamically shape one’s trust in a system (e.g., [29]). This psychological process of learning to what extent to trust a technical system already starts prior to the interaction and continues during system use [15, 29]. To date, research on trust in automation has investigated a range of variables influencing trust. Kaplan et al. [24] classifies these into three groups: 1) AI-related factors (both performance-related variables and secondary attributes such as communication style), 2) context factors like specific tasks or team composition, and 3) human-related factors, for instance, users’ personality or abilities. Regarding the latter, as individual users have different backgrounds (like level of expertise, experience, or cognitive styles), a given piece of information might be used differently on the basis of such individual differences.

With the goal of empowering users to safely and efficiently use LLM-based systems it is thus of importance to understand how users form trust in a conversational agent, specifically, which aspects of an application affect trust in which direction and under which circumstances. To date, research on trust in automation has considered various types of AI-based systems – namely self-driving vehicles, robots, chatbots, and non-embodied AI [24]. Out of the four types, chatbots were the one least represented in research at the time [24]. While some research has been conducted on chatbot use in specific contexts such as marketing applications (e.g., [40, 42, 56]), psychological experiments closer investigating the causes and processes of trust formation, especially ones considering the particular capabilities and limitations of recent LLM-based applications, are yet to be conducted.

In our study, we focus on two factors of the system, specifically referring to current state-of-the-art widely used applications: Prior information and communicative style. Information provided prior to system use has been found to affect trust in studies investigating other types of automated systems, such as driving assistants [28, 30], and prevalent conversational agent applications generally provide information within the user interface, particularly caveats about the potential generation of misinformation (see Figure 1). Often these caveats take on a rather unobtrusive format: While Bard still displays its misinformation caveat in the center of their landing page, ChatGPT’s August 2023 version has replaced this center-page feature and now instead suggests prompts for users to try, some of which even encourage the latter to request factual information for, for instance, studying or generation of code. A remaining misinformation caveat is still displayed as part of some text below the field for user input, albeit following other information and in low-contrast text color and small font size. For Bing, users would

have to actively navigate to an "About" subpage to read the explicit caveat that information provided by the tool might be inaccurate. For all these cases, one cannot trivially assume that users perceive and utilize this information. We thus investigate if and when this prior information affects users' trust in the agent.

Further, as mentioned before, LLM-based conversational agents are set apart from both their predecessors as well as other types of AI by their formal linguistic competence and ability to generate human-like text. Despite lacking the functional linguistic competence to assess the validity of their outputs, prevalent agents often adopt a matter-of-fact authoritative style, which can be mistaken for high certainty about the correctness of the output and thus make factually false statements sound plausible and convincing [5, 26, 43, 57, 58]. Weidinger et al. [61] argue that this formal linguistic competence could promote overreliance and unsafe use due to anthropomorphization, even when users are aware of interacting with a non-human agent. This can then result in misinformation being mistaken for true and potentially spread further, with adverse effects beyond the applications boundaries [36, 54]. The above-mentioned criticisms of agents' communicative style further reflect research on human-human communication, in particular authoritativeness and power of language, which were found to impact attitudes towards the sender [2]. This study aims to extend these findings to the context of conversational agents, where previous research has linked other aspects of communicative style to trust [13, 21, 45]. Accordingly, adapting the language used by a conversational agent to reflect limitations of the underlying model could be one way to help prevent overtrust. Against this background, we investigate the influence of both prior information and communicative style on trust in an LLM-based agent and its persuasive power using a prominent theory from social psychology – the Elaboration Likelihood Model (ELM).

## 2.2 Two Routes of Attitude Formation

According to Petty and Cacioppo's [49] ELM, attitudes, such as trust, can be affected by persuasive attempts/messages. The ELM states that people are motivated to hold correct attitudes. Still, the extent of elaboration they engage in to achieve this goal varies. Depending on a recipient's ability and motivation to thoroughly consider the arguments of a message which targets an attitude, processing occurs via either the central or the peripheral route. If the recipient is thoroughly considering the message's arguments, the central route is active. Here, attitude change is determined mainly by the argument quality, i.e., the issue-relevant content, of a persuasive message. For instance, customers looking to purchase a new smartphone may carefully analyze the technical details of a model, e. g., its processor and battery capacity. If, however, the recipient is lacking motivation or ability to thoroughly process the arguments, attitude changes result from the peripheral route and are determined by positive or negative cues which are not directly relevant to the content of the persuasive message (but permit a simple inference about its validity). For example, customers who do not want to think too hard about which smartphone to buy, or who do not understand the technical specifications, might instead come to like a model because they find it visually appealing, because of the brand's reputation, or because it was recommended to them by

a knowledgeable-seeming salesperson. According to the ELM, attitude formation via the central route involves more cognitive effort than is the case via the peripheral route, thus attitudes resulting mainly from the central route are more robust to change over time and against attempts of counterpersuasion, and more predictive of behavior than attitudes formed via the peripheral route [49]. Besides serving as an argument or peripheral cue, the ELM allows for variables to affect persuasion by influencing the direction or extent of elaboration, i.e., which route is taken [49].

Since trust is here conceptualized as an attitude, trust formation too can be considered in terms of ELM processes: That is, trust can be affected through the central route by arguments indicative of the trustee's actual capabilities, or through the peripheral route by peripheral cues, depending on whether the trustor is motivated and able to elaborate. To date, the ELM has been successfully applied to investigating the role of central and peripheral route processes in trust in automation [30]. Additionally, in human-human interaction, the ELM was applied to investigate the role of communicative style as a peripheral cue [2, 6]. The investigation at hand on the role of disclaimer information and communicative style of conversational agents marks the intersection of these findings. Note that conversational agents are special in that they can be seen in a dual role, not only as an attitude object/trustee, but also as the source delivering information on other objects or topics on which recipients can hold attitudes. Thus, trust in the agent can serve two different functionalities, depending on the considered outcome variable:

- (1) *Trust as a resulting attitude - attitude towards the agent*: Trust in the agent can itself be considered a dependent variable. In our study, the provided disclaimer information was manipulated as a central argument, and the agent's communicative style as a peripheral cue.
- (2) *Trust as a peripheral cue - persuasiveness of an agent's message*: Trust can also be viewed as a peripheral source cue for forming attitudes about the content of the communication. That is, if users trust an agent more, persuasive messages delivered by the agent might impact users' attitudes to a higher extent.

For trust as a resulting attitude, issue-relevant arguments which affect trust via the central route should refer to the agent's actual capabilities. For a conversational agent with the purpose of providing factual information, the ability to provide correct responses is most relevant. Thus, through effortful central route processing, prior information highlighting the limitations of LLM-based agents should result in less trust than one highlighting their capabilities. Based on this, it is hypothesized:

*H1.1: Trust in a conversational agent is higher after arguments highlighting the agent's positive capabilities than after arguments highlighting misinformation limitations.*

Further, as reasoned above, with LLM-based agents now being capable of generating text that resembles natural language produced by humans, trust in the agent should also be affected by its communicative style. The formal linguistic competence and authoritative style displayed by prevalent agents [5, 26, 43, 57, 58] can

be considered to work as a peripheral cue heuristic for the agent's functional competence operating via the peripheral route. While authority and markers thereof have been linked to persuasion in human-human interaction in general [14], also respective ELM persuasion effects of authoritative language on attitudes towards the sender have been postulated and found [2, 6]. We expect an agent generating well-formulated, seemingly plausible text in authoritative language to elicit more trust than one communicating with a style lower in authority:

*H1.2: Trust in a conversational agent is higher when the conversational agent displays a communicative style high in authority than a communicative style low in authority.*

Following ELM assumptions regarding the extent of elaboration, under central route processing capability-related prior information should affect trust in the agent more. Consequently, the role of the agent's communicative style should affect trust more under peripheral route processing. We consider two dispositional variables which ought to affect a recipient's general motivation to elaborate on arguments provided in a persuasive message (resulting in either peripheral or central route processing) – namely need for cognition (NfC) and system-specific self-efficacy (SEF). NfC has been found to promote central route processing since the ELM was first proposed [49], and showed its moderating effect also for trust in automated driving systems [30]. Recent studies also supported a comparable role for users' SEF, that is, their perceived ability to handle the technical system in utilizing different types of information [62, 63]. We aim to extend these findings to the domain of conversational agents.

*H1.3a: The effect of prior information on trust in a conversational agent is more pronounced for participants with a higher need for cognition than for participants with a lower need for cognition.*

*H1.3b: The effect of prior information on trust in a conversational agent is more pronounced for participants with higher self-efficacy in using chatbots than for participants with lower self-efficacy in using chatbots.*

*H1.4a: The effect of communicative style on trust in a conversational agent is more pronounced for participants with a lower need for cognition than for participants with a higher need for cognition.*

*H1.4b: The effect of communicative style on trust in a conversational agent is more pronounced for participants with lower self-efficacy in using chatbots than for participants with higher self-efficacy in using chatbots.*

According to the depicted two roles of trust in this research, besides being the trustee, a conversational agent can also serve as a cue itself for the process of attitude formation and change in regard to its communicated message. Assuming that assessment of the validity of information generated by an agent in detail exceeds the (available) cognitive capabilities of lay users, they might instead tend to use their trust as a heuristic to guide reliance [27].

Note that the resulting attitude in this case is not trust in the agent, but the attitude towards an object or topic on which the agent provides (persuasive) information. At this stage, trust in the agent can thus be considered as a peripheral source cue to inform

further attitudes towards the subject matter of a conversation with the system. That is, users' attitudes towards a subject matter will be affected to an higher extent by information presented by an apparently trustworthy agent, especially under peripheral route processing:

*H2.1: Attitudes towards an attitude object are more favorable following a persuasive message presented by a conversational agent when the agent is trusted more.*

*H2.2a: The effect of trust in a conversational agent on attitudes towards an attitude object following a persuasive message presented by the agent is more pronounced for participants with a lower need for cognition than for participants with a higher need for cognition.*

*H2.2b: The effect of trust in a conversational agent on attitudes towards an attitude object following a persuasive message presented by the agent is more pronounced for participants with lower self-efficacy in using chatbots than for participants with higher self-efficacy in using chatbots.*

Combining H1.1, 1.2, and H2.1, it is thus further hypothesized:

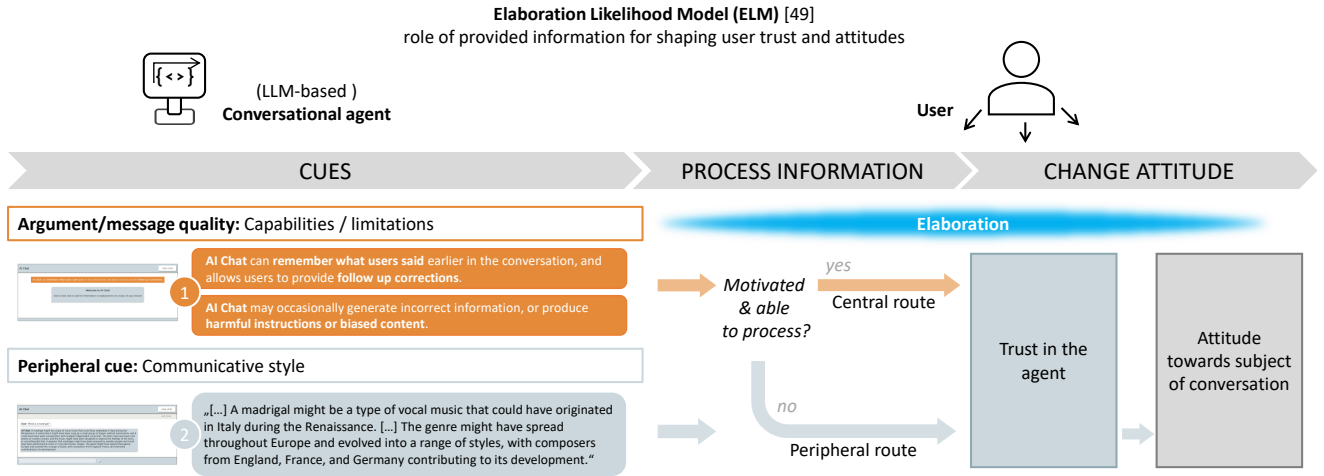
*H2.3: Trust mediates the effect between disclaimer information on the agent's capabilities and its communicative style and resulting attitudes on a subject of communication.*

### 3 EMPIRICAL STUDY

We conducted an online user study in which participants were asked to imagine to be engaged in a conversation with a conversational agent named *AI Chat* through semi-interactive, animated sample conversations. The study involved manipulations of prior information via disclaimers on the chatbots' landing page and variations in the chatbot's communicative style during an irrelevant topic conversation A, followed by a persuasive message about robots in public spaces in conversation B. The study aimed to evaluate trust in the agent and attitudes towards the content of a message communicated by the agent, i.e., robots in public spaces.

#### 3.1 Sample

The study was administered online via EFS Survey [50]. English native speakers were recruited and compensated via an online panel provider.  $N = 657$  participants completed the study with a median duration of 30 minutes, out of which  $n = 12$  were excluded from analysis for speeding (taking less than 2 seconds per Likert item),  $n = 46$  because they indicated not having participated in a serious manner, and  $n = 5$  because their responses to open questions indicated either their participation being interrupted by technical issues ( $n = 4$ ) or excessive aversion against study contents ( $n = 1$ ). The final sample thus contained  $N = 594$  cases, with a mean age of  $M = 44.89$  years ( $SD = 14.91$ ), and 301 (50.7%) participants having indicated their gender as female, 290 (48.8%) as male, and 3 (0.5%) as other. Participants indicated having had experience in using chatbots with a mean value of  $M = 4.50$  ( $SD = 1.80$ ) on a 7-point Likert item ranging from "strongly disagree (- -)" to "strongly



**Figure 2: Proposed mediation model for trust in conversational agents induced by aspects of the application design, based on the elaboration likelihood model [49].**

agree (+++). 145 (24.4%) participants reported having interacted with a robot before.

### 3.2 Study Design and Stimuli

We employed a 2x2 between-subjects experimental design with two independent variables, each with two levels: (1) prior information (limitations vs. positive capabilities) and (2) communicative style (high authority vs. low authority), the latter serving as ELM peripheral cue. The factors were manipulated by presenting different information about the conversational agent’s qualities before the simulated interaction for (1) and by varying the communicative style of the output during the interaction for (2). Participants were randomly assigned to one of the four experimental conditions. For hypotheses related to the extent of elaborative processing (H1.3a, H1.3b, H1.4a, H1.4b, H2.2a, H2.2b), quasi-experimental user-related measurements of NFC and SEF were collected.

**3.2.1 Prior Information.** The prior information conditions were modeled based on the respective information given on ChatGPT’s website at the time of study conduction [46, Mar 23 version]. Specifically, for the *limitations* condition, the information that the chatbot “may occasionally generate incorrect information, or produce harmful instructions or biased content” was presented. For the *capabilities* condition, the information that it “can remember what users said earlier in the conversation, and allows users to provide follow-up corrections” was displayed. These statements were prominently placed within orange boxes on the landing page of the fictitious study chatbot (see Figure 3).

**3.2.2 Communicative Style and Conversation Stimuli.** To investigate the impact of peripheral cues on participants’ attitudes and processing routes in the sense of the ELM, the way the conversational agent communicated was manipulated. Using ChatGPT [46, Mar 23 version], two sample conversations were designed. The first conversation (labeled ‘A’) was unrelated to the study, focusing on

‘madrigals’, and served solely to manipulate the chatbot’s communicative style. The second conversation (labeled ‘B’) centered on a persuasive message regarding ‘robots in public spaces’, emphasizing their benefits. For the persuasive message in conversation B, ChatGPT was prompted to enumerate benefits of robots in public spaces.

To create variation in the communicative style in conversation A to be high and low in authority, ChatGPT was prompted to define ‘madrigal’. For the *high authority* variant, ChatGPT’s output was taken as is. For the *low authority* variant, ChatGPT was provided with Blankenship and Holtgrave’s [7] definition of powerless language and instructed to replace complex words. According to the definition, “powerless language refers to the presence of one or more linguistic features such as tag questions, hesitations, disclaimers, hedges, polite forms, and so on. Powerful language refers to the absence of these features.” [7, p. 4]. We combined multiple versions of the text generated using this approach with minor manual edits to maintain consistency between the two styles. A pilot study ( $N = 31$ ) confirmed that the altered version was perceived as less authoritative ( $M = 3.25$ ,  $SD = 0.96$ ) than the original one ( $M = 4.81$ ,  $SD = 0.89$ ),  $t(30) = 7.92$ ,  $p < .01$ , based on responses to 10 items by Hosman and Siltanen [18] on a 7-point Likert scale ranging from “strongly disagree (- - -)” to “strongly agree (+++)”.

The text lengths were 106 words for the original, high authority text, 122 words for the low authority variant, and 248 words for the persuasive message. While conversation A responses were presented as replies to the original prompts (excluding a prefix specifying text length, “In 100 words:”), conversation B was preceded by a neutral prompt requesting information about robots in public spaces instead of their benefits. For detailed stimuli construction, including verbatim prompts and edits, see Appendix A. All conversations were presented using an animated chatbot interface.



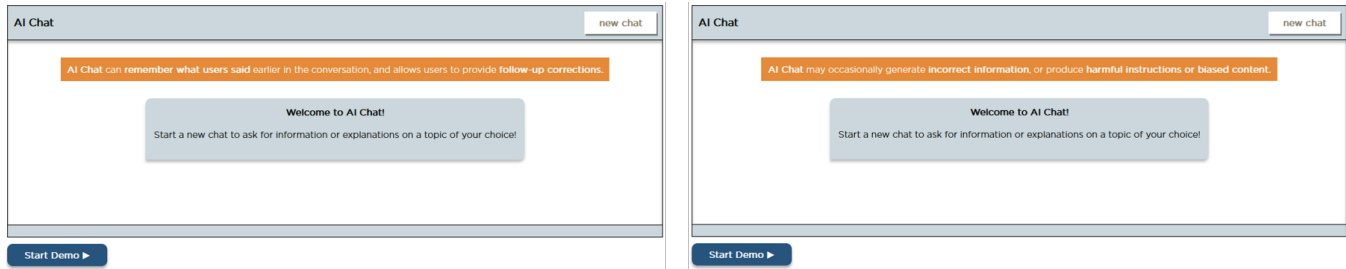


Figure 3: Screenshots of the landing page of the conversational agent (*AI Chat*) highlighting capabilities (*left*) and limitations (*right*) as manipulation of the prior information.

### 3.3 AI Chat

An animated chatbot interface was created to simulate an interaction between the user and a conversational agent. *AI Chat* was introduced as "a specific artificial-intelligence-based chatbot [...] [that] can understand written requests and generate responses accordingly. It answers questions posed by users, providing information and explanations on various topics [...]". After the initial evaluation of *AI Chat*, more profound information on how *AI Chat* functions (e.g., based on a LLM), on its designed purpose (e.g., provide information and explanations), and on how users can interact with it (e.g., ask questions, converse with it naturally) were given. To get an idea of what interactions with *AI Chat* look like, participants experienced a demonstration of *AI Chat*, with the two simulated conversations A and B between a user and *AI Chat*.

**3.3.1 Interface.** The animated interface of *AI Chat* was semi-interactive and oriented to state-of-the-art AI systems as introduced in Figure 1. Screenshots of the different parts can be seen in Figure 4, HTML code for the full animation is included in the supplementary materials. On the landing page of the chatbot application window (1), participants were welcomed to *AI Chat* and, via an orange information box, pointed to either the chatbot's limitations or capabilities (see subsection 3.2.1). Participants could start a demo by pressing a "Start Demo ►"-button. Upon pressing the button, a new chat was opened and participants saw a simulated user query in the entry line at the bottom of the application window (2). The question was automatically prompted to *AI Chat* (3), whereupon participants saw the chatbot's – low or high authority – response being typed into the output field (4 and 5). To increase realism and recognizability, the in- and outputs were animated. The delay between single letters being typed out was 70 ms for user prompts and 50 ms for chatbot responses. By pressing the "Resume Demo ►"-button, participants could start the second part of the demonstration, containing conversation B with the persuasive message about robots in public spaces, which was played in the similar way.

### 3.4 Measurements and Questionnaires

Several scales were used to measure users' trust, dispositions, attitudes, and evaluations of the chatbot. Where not otherwise specified, items in this study were measured on a 7-point Likert scale ranging from "strongly disagree (- -)" to "strongly agree (+++)"

**3.4.1 Trust and Human-Likeness.** As main dependent variable, trust in the chatbot was measured using two different questionnaires:

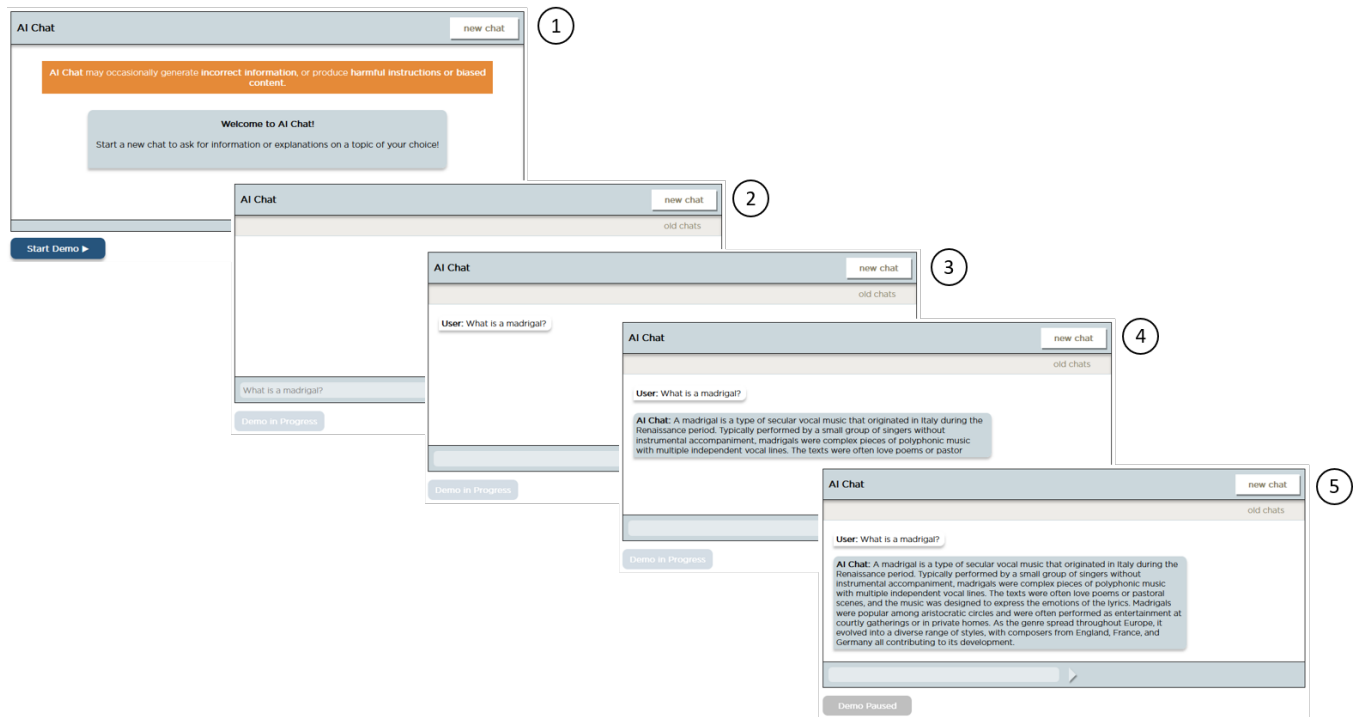
the well-established unidimensional scale (8 items,  $\alpha = .90$  for the initial measurement) by Jian et al. [23] with the adaptations suggested by [29] as our measure of interest, and a recently proposed multidimensional measure of trust (16 items,  $\alpha = .95$  for the initial measurement) by Malle and Ullman [37] as an additional measure. In addition, human-likeness of the chatbot using an adapted version of the 7-point semantic differentials scale (4 items,  $\alpha = .86$  for the initial measurement) by Bartneck et al. [4] was assessed.

**3.4.2 Attitudes Towards Robots and Thought-Listing.** To measure the degree of attitude change with respect to the subject of the conversation, attitudes towards robots in public spaces were measured with three self-constructed items (RiPS,  $\alpha = .95$ ) and the negative attitudes towards robots scale (NARS, 14 items,  $\alpha = .87$ ) by Nomura et al. [44].

In addition, a thought-listing technique adapted from Petty and Cacioppo [48] was used, in which participants were first given two and a half minutes to list any thoughts or ideas they had during conversation B about robots in public spaces in open response fields. When the time had passed, the questionnaire page was submitted automatically and participants were presented the thoughts they had just listed. Second, they were asked to rate each of their thought on a 3-point scale on whether it was negative (-), irrelevant (-/+), or positive (+) with regard to robots in public spaces. This thought-listing data was collected to closer investigate additional predictions of the ELM which would exceed the scope of this manuscript and will thus not be elaborated on here.

**3.4.3 User Characteristics.** Before the experimental part, user characteristics relevant to the interaction with and evaluation of AI, chatbots, and robots were measured. Those included NfC ( $\alpha = .89$ , [10]), tendency to anthropomorphize technology ( $\alpha = .90$ , own scale), propensity to trust automated technology ( $\alpha = .86$ , own scale), general attitudes towards AI ( $\alpha = .93$ , [55]), and SEF in interacting with a chatbot ( $\alpha = .96$ , adapted from [52], to refer to chatbots instead of robots).

**3.4.4 Manipulation Checks.** At the end of the study, participants rated the authoritativeness of the chatbot's response in conversation A on the same six items by Hosman and Siltanen [18] that were used in the pilot study ( $\alpha = .91$ ). For this, screenshots of the responses were presented again. Correct recall of the prior information manipulation was checked by presenting both versions (limitations vs. capabilities) of the orange information box again



**Figure 4: Screenshots of the semi-interactive, animated interface of the conversational agent *AI Chat*. The conversation could be started via a *Start Demo ▶*-button (1). Participants saw a question being types in (2) and asked to *AI Chat*, which appeared in the chats’ output field (3). The response of *AI Chat* was animated (4 & 5) and appeared with a delay of 50 ms between the single letters.**

along a third option (“I do not recall having seen either of these before”) to discourage guessing. Participants were further asked to what extent they had been familiar with the subjects of both conversations A and B prior to the study (2 items), and to what extent they assumed the information *AI Chat* provided in both conversations to be factually correct (2 items).

### 3.5 Procedure

Figure 5 depicts the procedure of the online experiment. Participants were informed that the study was about how different chatbot interfaces are perceived by users. After having provided their informed consent to participate in the study and indicated their age and gender, they first answered a series of questionnaires regarding relevant dispositional user characteristics (see subsection 3.4.3). Participants were then assigned randomly to one of the four experimental conditions. They read a brief description of *AI Chat* and provided their initial evaluation of trust in the chatbot and its human-likeness (see subsection 3.4.1). Participants then read a second, more extensive description, which gave more information on LLMs and *AI Chat*.

Following this, they watched the animated demonstration (see subsection 3.3.1), containing the simulated conversations A and B on the topics of ‘madrigal’ and ‘robots in public spaces’, respectively (see subsection 3.2.2). After each topic, the demonstration paused and participants were asked to evaluate trust in *AI Chat* and its human-likeness. Following conversation B, questions

on the attitude of the former’s conversation subject were asked and participants completed the thought-listing task (see subsection 3.4.2). To conclude the study, manipulation checks (see subsection 3.4.4), further questions on prior knowledge and experience, and the opportunity to leave comments regarding the study were provided. At the end, participants were informed that *AI Chat* was purely hypothetical and did not exist.

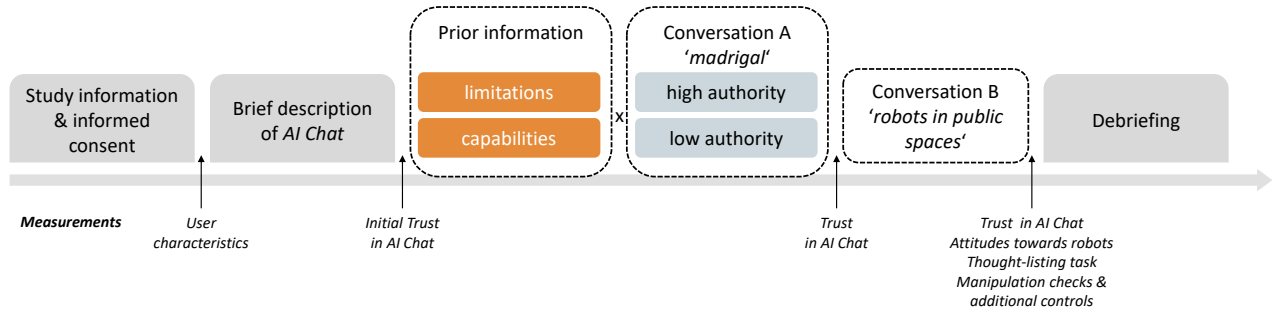
## 4 RESULTS

Data analysis was performed using the R software environment (v4.2.2, [51]). Since assumptions of normality and homoscedasticity were not met for regression models (which is a common phenomenon in the domain), bootstrapping with 5000 samples was employed to obtain unbiased estimates for regression coefficients, their standard errors, and resulting statistical significance (packages *boot* [12] and *boot.pval* [60]). The *lavaan* package [53] was used for path analysis and the *TOSTER* package [11, 31] for equivalence testing. Following Hoffman’s [16] and Kock’s [25] recommendations for hypothesis testing, one-sided p-values are reported for directed hypotheses.

### 4.1 Manipulation Checks

Participants in the high authority condition ( $M = 5.02, SD = 1.12$ ) perceived the chatbot response as more authoritative than the ones in the low authority condition ( $M = 4.53, SD = 1.33$ ),  $t(592) =$





**Figure 5: Procedure of the user study, featuring experimental manipulations of prior information (limitations vs. capabilities) and communicative style (high vs. low authority), along with elements displayed in the semi-interactive, animated user interface of AI Chat (dashed boxes). Measurements are listed in the lower section.**

4.88,  $p < .01$ , indicating a successful manipulation of the conversational style.

For the manipulation of prior information, 282 participants failed to correctly identify the prior information in the disclaimer presented before conversation A. These cases were excluded from analyses investigating the effects of prior information (subsection 4.2).

## 4.2 Effects of Prior Information and Communicative Style

For trust in the agent, H1 regarding the main effects of the disclaimer information (limitations vs. capabilities, H1.1) and communicative style (low vs. high authority, H1.2) as well as their respective interactions with the ELM personality moderators (H1.3a, H1.3b, H1.4a, and H1.4b), was tested in a linear regression model, with trust in the agent directly following conversation A as the criterion variable. Prior information was dummy-coded as 0 for the condition highlighting limitations and 1 for the one highlighting capabilities. Communicative style was dummy-coded as 0 for the low authority and 1 for the high authority style. As mentioned above, 282 participants were excluded for failing to correctly recall the disclaimer information presented, resulting in  $N = 312$ .<sup>1</sup> Based on a test for equivalence as implemented in the TOSTER package (two one-sided tests, [11, 31]) initial trust could not be assumed to be equal between experimental conditions and was thus included as a covariate. For NfC and SEF in interacting with chatbots, a separate model was calculated respectively to account for their independent roles as ELM moderators. While a main effect of communicative style on trust was supported, the one for prior information was not as was the case for interactions with NfC or SEF (see Table 1). Therefore, regardless of the user's characteristics, trust in the chatbot did not differ depending on whether they received information about its limitations or capabilities, but was higher for a high authoritative compared to a low authoritative communicative style.

## 4.3 Trust as a Peripheral Cue

H2 hypothesized a main effect of trust in the agent prior to the persuasive message on attitudes towards robots in public spaces (H2.1) and an interaction of the trust level (serving as a peripheral cue) with the assumed ELM personality moderators (H2.2a and H2.2b). Again, four separate regression models were calculated for NfC and SEF as well as for the two dependent attitude variables negative attitudes towards robots (NARS) and attitudes towards robots in public spaces (RiPS). All analyses yielded a significant main effect of trust, while the hypothesized interaction only emerged between trust and NfC with RiPS as the criterion variable (see Figure 6 and Table 1 for detailed results. Note that signs are reversed for the two measures, since for RiPS high scores indicate positive attitudes, but for NARS high scores indicate negative attitudes). A subsequent path analysis with NfC following the logic of H1–H2 further supported these results (see Table 2). Both RiPS and NARS were entered as endogenous variables with predictors trust, NfC, and their interaction. Analogous to H1, trust was further considered endogenous with predictors communicative style, NfC, their interaction, and initial trust as covariates. In addition to the direct effects, the mediation paths from communicative style via trust on NARS and RiPS were significant. This supports H1.1, H2.1, and partially H2.2a and H2.3, with trust acting as a mediating variable between a conversational agent's communicative style and resulting attitude change in the direction of a message delivered by the agent.

## 5 DISCUSSION

Conversational agents based on LLMs are capable of producing output demonstrating high formal linguistic competence, but, at the same time, are prone to generating misinformation. With the recent rapid advancements in the field and increased availability to the broader (less expert) public, psychological research is highly relevant to facilitate safe and efficient interaction with these systems. In this, it is an essential challenge to equip these agents

<sup>1</sup>To account for possible implicit processes which might have held despite participants having failed to recall the prior information, we repeated the same analysis on the full sample, which yielded the same pattern of results.

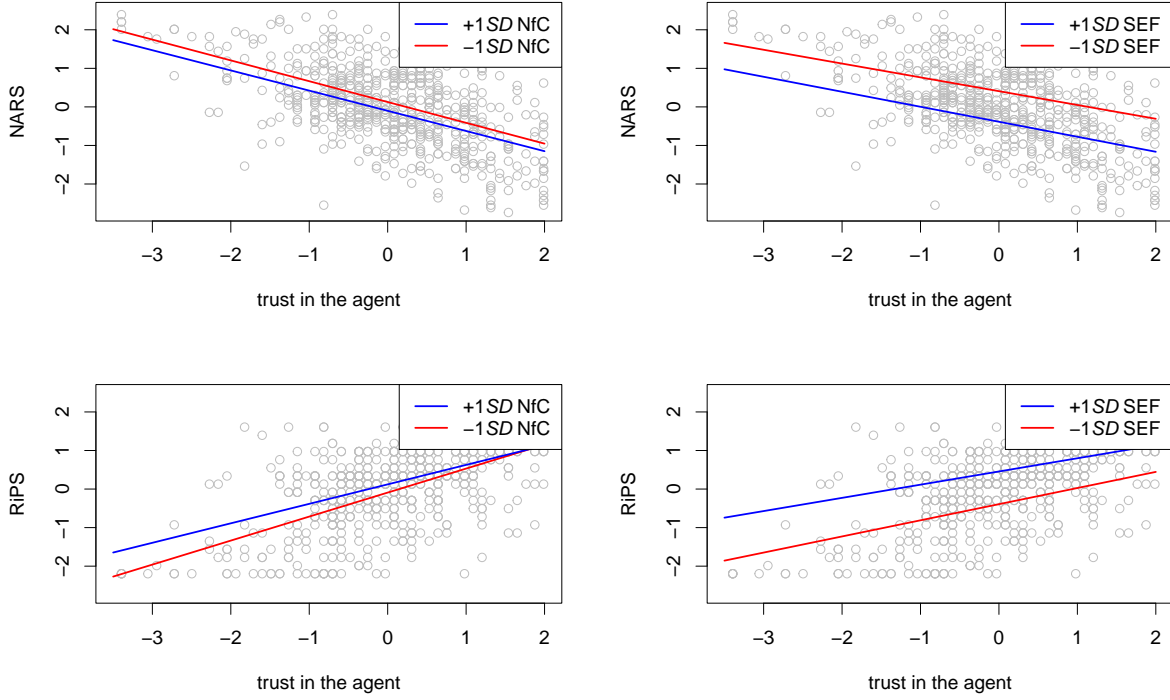
**Table 1: Standardized coefficients of the bootstrap regression models for H1 and H2 (standard errors in parentheses). Prior information and communicative style were dummy-coded (prior information: limitations = 0, capabilities = 1; communicative style: low authority = 0, high authority = 1). “Initial trust” relates to the trust measurement prior to conversation A, and “trust in the agent” relates to the one after conversation A and prior to conversation B. Trust measurements, measurements of attitudes towards robots, NfC, and SEF were z-standardized for analysis. \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ . ‘-’ marks cells that were not part of the model.**

	Predicted Attitude					
	trust in the agent [H1]		negative attitudes towards robots (NARS) [H2]		attitudes towards robots in public spaces (RiPS) [H2]	
Intercept	-.05 (.07)	-.04 (.07)	.02 (.04)	.01 (.03)	.02 (.03)	.03 (.03)
<b>Key Predictors</b>						
prior information	.02 (.08)	.01 (.08)	–	–	–	–
communicative style	.16 (.08) *	.17 (.08) *	–	–	–	–
trust in the agent	–	–	-.53 (.04) ***	-.37 (.04) ***	.56 (.03) ***	.38 (.03) ***
<b>Moderators and Controls</b>						
initial trust	.71 (.04) ***	.70 (.05) ***				
need for cognition (NfC)	.08 (.07)	–	-.11 (.03) **	–	.11 (.03) **	–
self-efficacy (SEF)	–	.09 (.08)	–	-.40 (.04) ***	–	.42 (.03) ***
<b>Interactions</b>						
prior information × NfC	-.04 (.08)	–	–	–	–	–
prior information × SEF	–	-.08 (.08)	–	–	–	–
communicative style × NfC	-.03 (.07)	–	–	–	–	–
communicative style × SEF	–	.00 (.08)	–	–	–	–
trust in the agent × NfC	–	–	.01 (.03)	–	-.06 (.03) *	–
trust in the agent × SEF	–	–	–	-.02 (.03)	–	-.04 (.02)

with information and interfaces that facilitate a realistic impression and guard against overtrust to promote safe interactions. The present study provides a first endeavor to understand how users combine disclaimer information on agents’ limitations provided at the beginning of the interaction with apparent certainty reflected in the communicative style of the agent during the interaction. Participants were presented with a mock-up of a conversational agent modeled after current state-of-the-art applications, in which both the prior information on capabilities vs. limitations and the agent’s communicative style (low vs. high authority) were manipulated, and trust in the agent and attitudes towards the subject of the conversation were measured. Trust in the conversational agent was found to be affected by its communicative style in a way that communication high in authority lead to higher trust than a language style indicating a certain (more realistic) level of uncertainty. At the same time, the provided disclaimer information did not affect trust, irrespective of participant’s motivation to process elaboratively via the central route as indicated by a high NfC or

SEF. Furthermore, the study investigated if an agent that is trusted more also possesses higher persuasive capabilities. In line with our hypotheses, we found that users’ attitudes were more in line with a persuasive message when they trust the agent more. For this effect, evidence regarding the role of elaboration was mixed: While the effect of trust on attitudes following the persuasive message was more pronounced for participants higher in NfC when attitudes were measured using RiPS items, this interaction did not emerge for SEF as an indicator of elaboration, nor did it for the less specific NARS measure of attitudes.

On the one hand, these findings indicate that people tend to not take into account information on limitations of LLM-based conversational agents presented prior to the interaction for calibrating their trust levels or forming their attitudes on the topics the agents communicate about. On the other hand, communicated uncertainty in form of a less authoritative communicative style affects trust and may also lead to a more critical use of information provided by the agent. These findings stress the importance of a human-centered



**Figure 6: Bootstrap regression lines predicting attitudes towards robots (RiPS) with trust and either NfC or SEF as potential ELM moderators one standard deviation (SD) below or above the mean. NARS = negative attitudes toward robots; RiPS = attitudes towards robots in public spaces; NfC = need for cognition; SEF = self-efficacy. Trust in the agent, NfC, NARS, and RiPS were z-standardised for analysis.**

design process of LLM-based agents, and indicate that potential uncertainties in the provided information need to be communicated during the interaction with these systems right at the time and place the information is provided, instead of simply relying on prior disclaimers. Below, implications for application design – especially the development and design of LLM-based conversational systems closely interacting with a broader public – are discussed, and suggestions for future research are derived.

### 5.1 Theoretical and Practical Implications

Participants extended more trust to an agent which communicated in a style indicating a (unrealistic) high level of confidence (resembling the communicative style of current applications) as compared to a low authority style. This is in line with the conceptualization of communicative style as a heuristic guiding attitude change in human-human interaction, and corroborates the concerns that LLM-based agents' mere formal linguistic competence can be misleading and result in too high levels of face validity of provided information [5, 26, 58, 61]. Presenting participants with prior information highlighting an agent's positive capabilities as opposed to its limitations, that is, its potential to generate misinformation, did not affect trust in our study. Our findings instead indicate that caveats presented prior to the interaction with a conversational agent, as it is commonly done in current applications, might not

result in adequate consideration of them. One possible explanation could be a lack of salience and awareness of this information during the actual interaction with the systems. This is notably underscored by the finding that 47.74% of participants failed to recognize either information in the manipulation check at the end of the study, suggesting that they may not have initially perceived it or had forgotten about it during the interaction. In terms of practical implications this means that only half of the users who are presented with such information are able to recognize it a few minutes later - strongly indicating that they did not diligently read it in the first place. Since trust is only ethical when based on a system's actual capabilities [20] and users need critical information to give meaningful consent [34], this information not being remembered by almost half of the sample is concerning. In order to fulfill the goal of empowering users to safely and efficiently interact with these agents, measures need to be taken to increase the probability of noticing and remembering such provided information, for instance, a more salient design of the caveat information (e.g., text size, color) or a more prominent placement in the interface (e.g., permanent display during the interaction). As LLM-based systems become increasingly accessible, the general public should be educated about their potential shortcomings and safe use beyond the scope of the applications themselves, since brief caveats provided therein apparently do not suffice. Interface designers should seek

**Table 2: Standardized coefficients for direct effects ( $\beta X \rightarrow M$ ;  $\beta X \rightarrow Y$ ;  $\beta M \rightarrow Y$  and indirect effects  $\beta X \rightarrow M \rightarrow Y$ ) of the path analysis model with trust in the agent as mediator (standard errors in parentheses). NfC = need for cognition; NARS = negative attitudes towards robots; RiPS = attitudes towards robots in public spaces. Communicative style was dummy-coded (low authority = 0, high authority = 1). “Initial trust” relates to the trust measurement prior to conversation A, and “trust in the agent” relates to the one after conversation A and prior to conversation B. Trust measurements, NfC, NARS, and RiPS were z-standardized for analyses. \*\*  $p < .01$ . \*\*\*  $p < .001$ . “” indicates that this effect has already been listed in a line further up in the respective column; ‘-’ marks paths that were not part of the model.**

$X$	$M$	$Y$	$\beta X \rightarrow M$	$\beta X \rightarrow Y$	$\beta M \rightarrow Y$	$\beta X \rightarrow M \rightarrow Y$
communicative style			.24(.06)***	–	–.53(.04)***	–.13(.03)***
NfC	trust in the agent	NARS	.05(.05)	–.11(.04)**	”	–
communicative style x NfC			–.01(.06)	.01(.03)	”	–
initial trust			.72(.03)***	–	”	–
communicative style			”	–	.56(.03)***	.13(.03)***
NfC	trust in the agent	RiPS	”	.11(.04)**	–	–
communicative style x NfC			”	–.06(.03)*	–	–
initial trust			”	–	–	–

to amend this accompanied by further studies investigating if and how relevant information is used in more detail and user studies testing alternative, more salient designs.

Beyond investigating if and when providing caveats on a systems’ limitations influences trust, one might also pose the question if users should reasonably and realistically be expected to elaborate on such information in the first place, considering that in many situations users do not bother to read notices provided by a system [9]. Casual users of LLM-based chatbots may not deeply evaluate the information but rather seek quick answers or engage in casual conversation, thus lacking the motivation required for central route processing. Ideally, good system design should still facilitate trust calibration. That is, when users evidently struggle to utilize relevant information about system capabilities, at least inferences implied by other aspects of the system should be in line with these [32]. Thus, another way to tackle the situation that a considerable proportion of users do not remember provided information at the starting page is it to communicate potential limitations of the systems in the user interface during interaction. One such possibility to reflect a system’s limitations in its communicative style was successfully implemented and tested in the presented study by adapting the level of authority in the conversational agent’s outputs, specifically by inducing more powerless language and the substitution of complex words. In line with our hypotheses, a less authoritative communicative style lead to less trust in the agent and less attenuated attitudes in line with its message. This clearly underlines the role of synchronous information or cues entailing such information in the user interface of conversational agent applications. The high authority style which some of these systems apply can lead to overtrust, and in turn unsafe use. Adjusting a system to respond in a less authoritative style could thus also be an option to cue users into adopting more adequate levels of trust, especially when they have failed to

integrate explicit caveats. With complex and intransparent deep learning models underlying these systems, errors of which are to be expected due to a lack of functional linguistic competence but cannot be reliably foreseen, it is likely better to err on the safe side than to risk overtrust, since the latter can lead to misinformation being spread further – as has already been the case [36, 54]. While current agents sometimes respond to sensitive requests, for instance, for medical advice, by including disclaimers indicating their lack of expertise [47], or invite users to provide feedback and report any inaccuracies or issues as forms of expressing uncertainty, our results indicate that already their current, authoritative, communicative style can lead to overtrust. Thus, for these instances also agents should adopt a less authoritative style to facilitate trust calibration. Going one step further, future systems might aim to adapt their communicative style to specific users by increasing uncertainty markers specifically when users tend to overestimate the system.

While the presented findings provide no evidence for the proposed interactions between the extent of elaboration (operationalized by SEF and NfC) and prior information and communicative style on trust, the dependent variable attitudes towards robots in public spaces was significantly predicted by trust and an interaction of trust with NfC. It has to be noted that this interaction did only occur when RiPS was considered as the dependent measure, not when NARS was considered. While NARS is the more established and validated measure of attitudes, RiPS is considerably more specific to the subject of the persuasive message, since it refers precisely to robots in public spaces, instead of robots in general. Thus, RiPS is more reflective of the classic ELM paradigm, and considerably closer to the types of items used in respective studies [49]. Further, the expected mediation of the relationship between communicative style and attitudes towards robots by trust was found. This provides evidence for the role of trust as a social decision heuristic guiding

the reliance on a system where detailed comprehension is impractical due to its complexity [27, 32] and stresses the relevance of human users' trust to establish a well-balanced use of such systems and their outputs. Taken together, the results illustrate how psychological research can support this process by further investigating how users perceive and interact with LLM-based conversational agents. The study at hand provides initial insights to the field.

## 5.2 Strengths, Limitations, and Future Research

The present study was conducted in a realistic setup resembling current user interfaces of LLM-based conversational agents. Validity of the study findings is further maximized by a considerably large sample size. Also, participants were presented with the animated interface in a setting that is natural for the interaction with such systems – the online setting of their own devices. Whatsoever, as every study, this one, too, does not come without limitations that we discuss in the following.

First, while realism of the presented materials was maximized (interface, dynamic display format, content, and information on the agent's capabilities and limitations), participants were not able to interact with the system first-hand, but observed sample conversations about topics not selected by themselves. While this was a conscious choice for this study to allow for good experimental control, users' own exploratory behavior is supposedly still an important factor in real-world interactions with a conversational agent [32, 41].

Second, users did not have the possibility to experience errors in the study at hand, since the presented stimuli texts were intentionally plausible. Future studies on the topic area should include the possibility of experiencing errors since LLMs' proneness to generate misinformation is crucial for trust processes [36, 41, 54, 57].

Third and related to the former, it has to be noted that the extent to which people in our sample assumed the presented information to be correct was relatively high, potentially leading to too low levels of uncertainty. To mirror existing prevalent systems and thus enhance external validity, we chose to directly implement the phrasing used in earlier versions of ChatGPT [46]. This might have caused some ambiguity since the provided information on the application's capabilities included the information that the agent "allows users to provide follow-up corrections", potentially leading to the inference that corrections may be necessary. Despite this, mean trust measured after either condition was still above the scale mean. Future research might consider this and either choose to follow the classical pre-testing approach of the ELM or add a baseline to check whether either condition reduces trust in the agent.

Fourth, in face of the study findings, participants' involvement or ability to process might have been too low in the study and thus it is questionable if conditions for central route processing when evaluating the conversational agent were met. This could be a result of two characteristics of the study design: a) LLMs remain a complex topic, and b) the topic of conversation A was not sufficiently relevant to our participants personally. Since both motivation and ability are necessary preconditions for central route processing [49], future research should explore this issue more

closely by ensuring sufficient variance in elaboration between conditions via experimental manipulation, providing the respective groups with the necessary means to elaborate. For example, personal involvement or relevance could be increased by choosing a more personally relevant topic, or by allowing for more interaction. For operationalization, trust literature can be consulted for inspiration - Lockey et al. [34], for example, note that stakeholders' varying vulnerabilities influence the importance of different cues which inform trust; and Hong et al. [17] found that participants were less sensitive to peripheral cues when an AI agent committed moral violations such as sexism, which the authors describe as more anthropomorphic, morally engaging, and relevant to participants' personal life. In the study at hand, the extent of elaboration was not directly manipulated, but operationalized as a quasi-experimental covariate, namely interindividual differences in NfC or SEF. While NfC has been well-established as a variable affecting elaboration since the initial conceptualization of the ELM [49], SEF is a more recent construct [62, 63] and our study did not find evidence for it assuming the role of an ELM moderator. While one significant interaction did emerge for NfC, it showed relatively little variance in the sample ( $M = 4.4, SD = 0.88$ ). This is descriptively lower as in, for instance, a previous study on trust in an automated driving system by Kraus et al. [30], where NfC was found to interact with a reliability argument factor. Thus, in future research, a more diverse sample in terms of NfC could be recruited or the motivation or ability to elaborate could be directly manipulated.

While not all hypotheses were supported by the data, ELM theory provided a valuable approach for investigation and future research may expand on this initial study investigate more of its aspects, for example robustness of attitudes formed via the central vs. the peripheral route, or variables affecting biased elaboration. Further inspiration can be drawn from existing literature on trust in automation and LLMs, respectively. For example, Weidinger et al. [61] and Bender et al. [5] attribute the overestimation of a formally linguistically competent agent to increased anthropomorphization. Testing this proposed mediation empirically might be a next step in understanding how trust in conversational agents is formed. While this study focused on attitudes as an initial step, future research should investigate behavioral consequences of trust, for example, if and why users might still comply with untrustworthy systems [34].

## 6 CONCLUSION

Conversational agents based on LLMs are now available to and widely used by a multitude of users. However, these agents can produce factually incorrect or biased outputs while upholding an authoritative style of communication, making misinformation hard to spot and easier to spread. Appropriate calibration of user trust, specifically avoidance of overtrust, is thus a pivotal point to consider. For a human-centered development and design process, it is essential to include the psychological perspective focusing on the user especially in times of rapid advances in the field. This is illustrated by the practical implications which can be derived from the conducted empirical study in which participants encountered one of four variants of a conversational agent application each. Essentially, trust in the agent was not affected by providing information about its potential to produce biased or incorrect content,



but users based their evaluation on the agent's communicative style, with the high authority style common to prevalent applications like ChatGPT resulting in higher levels of trust. The higher the trust in the agent, the more persuasive was a message delivered by it. This emphasizes the relevance of trust in human users' interaction with AI-powered conversational agents, and should be considered by developers, designers, and distributors of these systems to facilitate safe, efficient, and ethical interactions.

## ACKNOWLEDGMENTS

This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2277 "Statistical Modeling in Psychology". We extend our thanks to the Department of Human Factors' working group for their helpful insights.

## REFERENCES

- [1] Google AI. 2023. Bard. <https://bard.google.com>
- [2] Charles S. Areni and John R. Sparks. 2005. Language power and persuasion. *Psychology & Marketing* 22, 6 (June 2005), 507–525. <https://doi.org/10.1002/MAR.20071>
- [3] Aram Bahrini, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J. Riggs, Maryam Esmaili, Rastin Mastali Majdabadek, and Morteza Pashvar. 2023. ChatGPT: Applications, Opportunities, and Threats. In *2023 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, Charlottesville, VA, USA, 274–279. <https://doi.org/10.1109/SIEDS58326.2023.10137850>
- [4] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1, 1 (Jan. 2009), 71–81. <https://doi.org/10.1007/s12369-008-0001-3>
- [5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Virtual Event, Canada) (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. <https://doi.org/10.1145/3442188.3445922>
- [6] Kevin L. Blankenship and Traci Y. Craig. 2011. Language use and persuasion: Multiple roles for linguistic styles. *Social and Personality Psychology Compass* 5, 4 (April 2011), 194–205. <https://doi.org/10.1111/j.1751-9004.2011.00344.x>
- [7] Kevin L. Blankenship and Thomas Holtgraves. 2005. The role of different markers of linguistic powerlessness in persuasion. *Journal of Language and Social Psychology* 24, 1 (March 2005), 3–24. <https://doi.org/10.1177/0261927X04273034>
- [8] Ali Borji. 2023. A Categorical Archive of ChatGPT Failures. [arXiv:2302.03494 \[cs.CL\]](https://arxiv.org/abs/2302.03494)
- [9] Rainer Böhme and Stefan Köpssell. 2010. Trained to accept?: A field experiment on consent dialogs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Atlanta Georgia USA, 2403–2406. <https://doi.org/10.1145/1753326.1753689>
- [10] John T. Cacioppo, Richard E. Petty, and Chuan Feng Kao. 1984. The efficient assessment of need for cognition. *Journal of Personality Assessment* 48, 3 (June 1984), 306–307. [https://doi.org/10.1207/s15327752jpa4803\\_13](https://doi.org/10.1207/s15327752jpa4803_13)
- [11] Aaron R. Caldwell. 2022. Exploring equivalence testing with the updated TOSTER R package. <https://doi.org/10.31234/osf.io/ty8de>
- [12] Angelo Canty and B. D. Ripley. 2021. boot: Bootstrap R (S-Plus) Functions.
- [13] Veena Chattaraman, Wi-Suk Kwon, Juan E. Gilbert, and Cassandra Ross. 2019. Should AI-Based, conversational digital assistants employ social- or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. *Computers in Human Behavior* 90 (2019), 315–330. <https://doi.org/10.1016/j.chb.2018.08.048>
- [14] Robert B. Cialdini. 2007. *Influence: The Psychology of Persuasion*. HarperCollins, New York, NY, USA.
- [15] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (2015), 407–434. <https://doi.org/10.1177/0018720814547570>
- [16] John P. Hoffmann. 2021. *Linear regression models: Applications in R*. CRC Press LLC, Boca Raton, Florida, USA.
- [17] Joo-Wha Hong, Sukyoung Choi, and Dmitri Williams. 2020. Sexist AI: An Experiment Integrating CASA and ELM. *International Journal of Human-Computer Interaction* 36, 20 (Dec. 2020), 1928–1941. <https://doi.org/10.1080/10447318.2020.1801226>
- [18] Lawrence A. Hosman and Susan A. Siltanen. 1994. The attributional and evaluative consequences of powerful and powerless speech styles: An examination of the 'control over others' and 'control of self' explanations. *Language & Communication* 14, 3 (July 1994), 287–298. [https://doi.org/10.1016/0271-5309\(94\)90006-X](https://doi.org/10.1016/0271-5309(94)90006-X)
- [19] Krystal Hu. 2023. ChatGPT sets record for fastest-growing user base - analyst note. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>
- [20] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 624–635. <https://doi.org/10.1145/3442188.3445923>
- [21] Andreas Janson. 2023. How to leverage anthropomorphism for chatbot service interfaces: The interplay of communication style and personification. *Computers in Human Behavior* 149 (2023), 107954. <https://doi.org/10.1016/j.chb.2023.107954>
- [22] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (Dec. 2023), 1–38. <https://doi.org/10.1145/3571730>
- [23] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (March 2000), 53–71. [https://doi.org/10.1207/S15327566IJCE0401\\_04](https://doi.org/10.1207/S15327566IJCE0401_04)
- [24] Alexandra D. Kaplan, Theresa T. Kessler, J. Christopher Brill, and P. A. Hancock. 2023. Trust in artificial intelligence: Meta-analytic findings. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 65, 2 (March 2023), 337–359. <https://doi.org/10.1177/00187208211013988>
- [25] Ned Kock. 2015. One-tailed or two-tailed p values in PLS-SEM? *International Journal of e-Collaboration (IJEC)* 11, 2 (April 2015), 1–7. <https://doi.org/10.4018/ijec.2015040101>
- [26] Anton Korinek. 2023. *Language models and cognitive automation for economic research*. Technical Report. National Bureau of Economic Research, Cambridge, MA, USA. <https://doi.org/10.3386/w30957>
- [27] Roderick M. Kramer. 1999. Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology* 50, 1 (Feb. 1999), 569–598. <https://doi.org/10.1146/annurev.psych.50.1.569>
- [28] Johannes Kraus, David Scholz, Dina Stiegemeyer, and Martin Baumann. 2020. The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction, and System Transparency. *Human Factors* 62, 5 (2020), 718–736. <https://doi.org/10.1177/0018720819853686>
- [29] Johannes Maria Kraus. 2020. *Psychological processes in the formation and calibration of trust in automation*. Ph.D. Dissertation. Ulm University. <https://doi.org/10.18725/OPARU-32583>
- [30] Johannes Maria Kraus, Yannick Forster, Sebastian Hergeth, and Martin Baumann. 2019. Two routes to trust calibration. *International Journal of Mobile Human Computer Interaction* 11, 3 (July 2019), 1–17. <https://doi.org/10.4018/IJMHCI.2019070101>
- [31] Daniel Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science* 8, 4 (May 2017), 355–362. <https://doi.org/10.1177/1948550617697177>
- [32] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. [https://doi.org/10.1518/hfes.46.1.50\\_30392](https://doi.org/10.1518/hfes.46.1.50_30392)
- [33] Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. 2023. ChatGPT: A meta-analysis after 2.5 months. <https://doi.org/10.48550/arXiv.2302.13795> [arXiv:arxiv:2302.13795](https://arxiv.org/abs/2302.13795)
- [34] Steven Lockey, Nicole Gillespie, Daniel Holm, and Ida Asadi Someh. 2021. A review of trust in artificial intelligence: challenges, vulnerabilities and future directions. In *Proceedings of the 54th Hawaii International Conference on System Sciences*. AIS, Hawaii, USA, 5463–5472. <https://doi.org/10.24251/HICSS.2021.664>
- [35] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. <https://doi.org/10.48550/arXiv.2301.06627> [arXiv:arxiv:2301.06627](https://arxiv.org/abs/2301.06627)
- [36] Makayen. 2022. Temporary policy: Generative AI (e.g., ChatGPT) is banned. Retrieved August 8, 2023 from <https://meta.stackoverflow.com/questions/421831/temporary-policy-generative-ai-e-g-chatgpt-is-banned>
- [37] Bertram F. Malle and Daniel Ullman. 2021. Chapter 1 - A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction*, Chang S. Nam and Joseph B. Lyons (Eds.). Academic Press, Cambridge, MA, USA, 3–25. <https://doi.org/10.1016/B978-0-12-819472-0.00001-0>
- [38] Regan Mandryk. 2023. ACM Publications Policy: Guidance for SIGCHI Venues. <https://medium.com/sigchi/acm-publications-policy-guidance-for-sigchi-venues-87332173aad1>
- [39] Microsoft and OpenAI. 2023. Bing Chat. <https://www.bing.com/chat>
- [40] Nika Mozafari, Welf H. Weiger, and Maik Hammerschmidt. 2022. Trust me, I'm a bot - repercussions of chatbot disclosure in different service frontline settings. *Journal of Service Management* 33, 2 (Feb. 2022), 221–245. <https://doi.org/10.1108/JOSM-10-2020-0380>

- [41] Bonnie M. Muir and Neville Moray. 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 3 (March 1996), 429–460. <https://doi.org/10.1080/00140139608964474>
- [42] Tom Nadarzynski, Oliver Miles, Aimee Cowie, and Damien Ridge. 2019. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital Health* 5 (Aug. 2019), 2055207619871808. <https://doi.org/10.1177/2055207619871808>
- [43] Arvind Narayanan and Sayash Kapoor. 2022. ChatGPT is a bullshit generator. But it can still be amazingly useful. Retrieved April 24, 2023 from <https://aisnakeoil.substack.com/p/chatgpt-is-a-bullshit-generator-but>
- [44] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. 2006. Measurement of negative attitudes toward robots. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems* 7, 3 (Nov. 2006), 437–454. <https://doi.org/10.1075/is.7.3.14nom>
- [45] Cecilie Bertinussen Nordheim, Asbjørn Følstad, and Cato Alexander Bjørkli. 2019. An Initial Model of Trust in Chatbots for Customer Service—Findings from a Questionnaire Study. *Interacting with Computers* 31, 3 (2019), 317–335. <https://doi.org/10.1093/iwc/iwz022>
- [46] OpenAI. 2022. ChatGPT. Retrieved September 8, 2023 from <https://chat.openai.com/chat>
- [47] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774
- [48] Richard E Petty and John T Cacioppo. 1977. Forewarning, cognitive responding, and resistance to persuasion. *Journal of Personality and social Psychology* 35, 9 (1977), 645.
- [49] Richard E. Petty and John T. Cacioppo. 1986. The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology* 19 (1986), 123–205. [https://doi.org/10.1016/S0065-2601\(08\)60214-2](https://doi.org/10.1016/S0065-2601(08)60214-2)
- [50] Questback GmbH. 2018. EFS Survey.
- [51] R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- [52] Astrid Rosenthal-Von der Pütten and Nikolai Bock. 2018. Development and validation of the self-efficacy in human-robot-interaction scale (SE-HRI). *ACM Transactions on Human-Robot Interaction* 7, 3 (Oct. 2018), 1–30. <https://doi.org/10.1145/3139352>
- [53] Yves Rosseel. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48, 2 (2012), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- [54] Mia Sato and Emma Roth. 2023. CNET found errors in more than half of its AI-written stories. Retrieved July 23, 2023 from <https://www.theverge.com/2023/1/25/23571082/cnet-ai-written-stories-errors-corrections-red-ventures>
- [55] Astrid Schepman and Paul Rodway. 2020. Initial validation of the general attitudes towards artificial intelligence scale. *Computers in Human Behavior Reports* 1 (Jan. 2020), 100014. <https://doi.org/10.1016/J.CHBR.2020.100014>
- [56] Susana C. Silva, Roberta De Cicco, Božidar Vlačić, and Maher Georges Elmarshara. 2022. Using chatbots in e-retailing – how to mitigate perceived risk and enhance the flow experience. *International Journal of Retail & Distribution Management* 51, 3 (Jan. 2022), 285–305. <https://doi.org/10.1108/IJRDM-05-2022-0163>
- [57] Aaron J. Snoswell and Jean Burgess. 2022. The Galactica AI model was trained on scientific knowledge – but it spat out alarmingly plausible nonsense. <https://theconversation.com/the-galactica-ai-model-was-trained-on-scientific-knowledge-but-it-spat-out-alarmingly-plausible-nonsense-195445>
- [58] Anna Strasser. 2023. On pitfalls (and advantages) of sophisticated large language models. <https://doi.org/10.48550/arXiv.2303.17511> arXiv:arxiv:2303.17511
- [59] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. <https://doi.org/10.48550/arXiv.2211.09085>
- [60] Måns Thulin. 2022. boot.pval: Bootstrap p-Values. <https://CRAN.R-project.org/package=boot.pval>
- [61] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Poulton. 2022. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 214–229. <https://doi.org/10.1145/3531146.3533088>
- [62] Shuiqing Yang. 2016. Role of transfer-based and performance-based cues on initial trust in mobile shopping services: a cross-environment perspective. *Information Systems and e-Business Management* 14, 1 (Feb. 2016), 47–70. <https://doi.org/10.1007/s10257-015-0274-7>
- [63] Tao Zhou, Yaobin Lu, and Bin Wang. 2016. Examining online consumers’ initial trust building from an elaboration likelihood model perspective. *Information Systems Frontiers* 18, 2 (April 2016), 265–275. <https://doi.org/10.1007/S10796-014-9530-5>

## A CONSTRUCTION OF STIMULI

To reflect current pervasive systems, stimuli texts for the conversational agent’s responses in two sample conversations were constructed using ChatGPT in the March 2023 version based on GPT-3.5 [46]. The process of stimulus generation is described below.

Texts marked “ChatGPT Output:” below are the original ChatGPT outputs. Unless otherwise stated, manual edits (insertion, deletion, change of position) made to outputs that were fed back to ChatGPT before proceeding to the next prompt are *italicized*. Editor’s notes are given in [brackets].

Final versions were presented to participants in a semi-interactive, animated interface as described in subsection 3.3.1 as seen in Figure 7 and Figure 8. The respective HTML code for the animation is included in the supplementary materials.

### A.1 Construction of the Informational Texts (Conversation A)

For the informational texts (conversation A), the high authority version was constructed first by providing ChatGPT with a simple prompt. Based on this, the low authority version was constructed by providing a series of prompts to ChatGPT. Between prompts, light edits were made manually and fed back to ChatGPT. The final stimulus text was then constructed by combining several generated versions to result in an informational text similar in content and length to the original high authority version. First person self-references in ChatGPT outputs were edited out to avoid blatant anthropomorphic confoundation.

*Original Text (Final Version for the High Authority Condition).*

**Prompt:** In 100 words: What is a madrigal?

**ChatGPT Output:** A madrigal is a type of secular vocal music that originated in Italy during the Renaissance period. Typically performed by a small group of singers without instrumental accompaniment, madrigals were complex pieces of polyphonic music with multiple independent vocal lines. The texts were often love poems or pastoral scenes, and the music was designed to express the emotions of the lyrics. Madrigals were popular among aristocratic circles and were often performed as entertainment at courtly gatherings or in private homes. As the genre spread throughout Europe, it evolved into a diverse range of styles, with composers from England, France, and Germany all contributing to its development.

*Variant A.*

**Prompts:** Remember: Powerless language refers to the presence of one or more linguistic features such as tag questions, hesitations, disclaimers, hedges, polite forms, and so on. Powerful language refers to the absence of these features.

Rephrase [reference to Original Text] in more powerless language as defined above.

**ChatGPT Output:** A madrigal might possibly be a type of secular vocal music that could have originated in Italy during the Renaissance period. It seems to have been typically performed by a small group of singers without instrumental accompaniment, and may have been quite complex, with multiple independent vocal lines. The texts may have been love poems or pastoral scenes, and the music might have been designed to express the emotions of

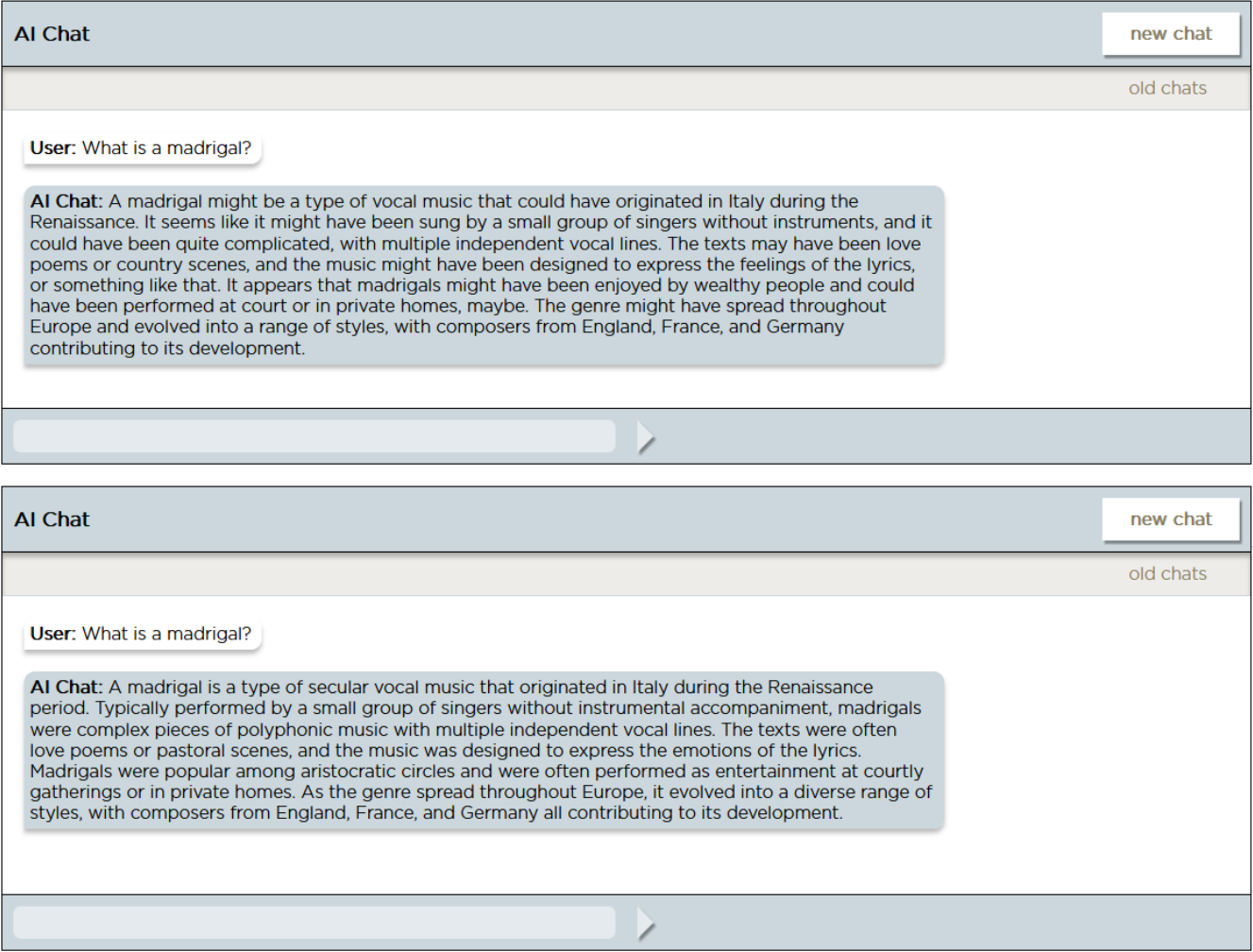


Figure 7: Screenshots of both variants of the informational texts (conversation A) as displayed in the online study. Top: High authority condition, bottom: low authority condition.

the lyrics. It appears that madrigals were popular among aristocratic circles and were often performed as entertainment at courtly gatherings or in private homes. It’s possible that the genre spread throughout Europe and evolved into a diverse range of styles, with composers from England, France, and Germany all contributing to its development.

Variant B.

**Prompt:** Make it even more powerless.

**ChatGPT Output:** *I’m not entirely sure, but I think a* A madrigal might be a type of secular vocal music that could have originated in Italy during the Renaissance period. It seems like it might have been performed by a small group of singers without instrumental accompaniment, and it could have been quite complex, with multiple independent vocal lines. The texts may have been love poems or pastoral scenes, and the music might have been designed to express the emotions of the lyrics, or something like that. It appears

that madrigals might have been popular among aristocratic circles and could have been performed as entertainment at courtly gatherings or in private homes, maybe. It’s possible that the genre might have spread throughout Europe and evolved into a diverse range of styles, with composers from England, France, and Germany all contributing to its development, *if I’m not mistaken.*

Variant C.

**Prompt:** While keeping the text as close to its original form as possible, i.e. not rearranging sentences, replace technical terms such as "secular", "accompaniment", or "pastoral" in [reference to edited Variant B] with simpler synonyms.

**ChatGPT Output:** *I’m not entirely sure, but I think a* A madrigal could be a type of non-religious vocal music that may have started in Italy during the Renaissance period. It looks like it might have been sung by a small group of singers without other musical support, and it could have been quite complicated, with multiple independent

vocal parts. The words may have been love poems or country scenes, and the music might have been created to express the feelings of the words, or something like that. It looks like madrigals might have been enjoyed by wealthy people and could have been performed for fun at fancy gatherings or in private houses, perhaps. It's possible that the genre might have become popular throughout Europe and turned into many different styles, with composers from England, France, and Germany all adding to its growth.

*Combined Final Version for the Low Authority Condition.*

[Variants B and C were edited further and combined into the final version with Variant B as the main source text. Below, passages taken from Variant B are kept in the default formatting and phrases inserted from Variant C are in **bold text**.]

A madrigal might be a type of ~~secular~~ vocal music that could have originated in Italy during the Renaissance period. It seems like it might have been **sung** by a small group of singers without instrumentals ~~accompaniment~~, and it could have been quite **complicated**, with multiple independent vocal lines. The texts may have been love poems or **country** scenes, and the music might have been designed to express the **feelings** of the lyrics, or something like that. It appears that madrigals might have been **enjoyed by wealthy people** and could have been performed as entertainment at courtly ~~gatherings~~ or in private homes, maybe. ~~It's possible that~~ #The genre might have spread throughout Europe and evolved into a **diverse** range of styles, with composers from England, France, and Germany ~~all~~ contributing to its development.

## A.2 Construction of the Persuasive Message (Conversation B)

The persuasive message about benefits of robots in public spaces (conversation B) was constructed in a series of prompts provided to ChatGPT. Between prompts, light edits were made manually and fed back to ChatGPT. The final stimulus text was then constructed by combining several generated versions to result in a short persuasive message.

*Version A.*

**Prompt:** Please tell me about robots in public spaces.

**ChatGPT Output:** Robots are increasingly being used in public spaces for a variety of purposes. These robots can take on many forms, from humanoid robots that can interact with people to autonomous vehicles that can move around on their own.

One common application of robots in public spaces is for *[swapped with paragraph below]* security purposes. Robots equipped with sensors and cameras can monitor public areas and alert human operators if they detect any unusual activity. These robots can also be used to investigate potential threats, such as suspicious packages or unattended bags.

~~Another use of robots in public spaces is for customer service. Robots can be programmed to provide information and assistance to people in public places such as shopping malls, airports, and train stations. These robots can answer questions, provide directions, and even assist with tasks like luggage handling.~~

In addition, robots are being used in public spaces for *[swapped with paragraph above]* cleaning and maintenance purposes. Autonomous cleaning robots can be deployed in public areas such as

parks and beaches to pick up trash and keep the space clean. ~~Other robots are being developed to maintain infrastructure, such as robots that can inspect and repair bridges and other structures.~~

Finally, some public spaces are experimenting with robots that provide entertainment and engagement for visitors. For example, amusement parks and museums are using robots to provide interactive exhibits and rides that incorporate robotics technology.

Overall, the use of robots in public spaces is still a relatively new and rapidly evolving field. As technology continues to advance, we can expect to see even more innovative uses of robots in our public spaces.

*Version B.*

**Prompt:** Please rephrase [reference to edited Version A] to include some benefits of robots in public spaces.

**ChatGPT Output:** Robots are being increasingly utilized in public spaces for various applications, providing numerous benefits. These robots can take on different forms, from autonomous vehicles that can move around on their own to humanoid robots that can interact with people.

One of the primary benefits of robots in public spaces is the ease and efficiency of cleaning and maintenance tasks. Autonomous cleaning robots can be deployed in public areas such as parks and beaches to pick up trash and maintain a clean environment. This reduces the workload on human staff and allows for more effective and thorough cleaning.

~~Another benefit of robots in public spaces is their use in customer service. These robots can provide assistance and information to people in places like airports, train stations, and shopping malls, improving the customer experience and reducing the workload on human staff.~~

Robots also provide improved security in public spaces. Robots equipped with sensors and cameras can monitor public areas and alert human operators if they detect any unusual activity. These robots can also be used to investigate potential threats, such as suspicious packages or unattended bags.

Finally, robots are being experimented with in public spaces for entertainment and engagement purposes. Museums and amusement parks are using robots to provide interactive exhibits and rides that incorporate robotics technology.

~~While the use of robots in public spaces raises some concerns, such as privacy and security risks, the benefits they provide are numerous. [Inserted back in from Version A:] Overall, the use of robots in public spaces is still a relatively new and rapidly evolving field. As technology continues to advance, we can expect even more innovative and beneficial uses of robots in our public spaces.~~

*Version C.*

**Prompt:** Elaborate on this text: [edited Version B]

**ChatGPT Output:** The increasing use of robots in public spaces is driven by the numerous benefits they provide. One of the primary benefits is the ease and efficiency of cleaning and maintenance tasks.

Autonomous cleaning robots can navigate public areas such as parks and beaches, collecting trash and maintaining a clean environment. This reduces the workload on human staff and allows for more thorough cleaning, which is crucial for maintaining the health and safety of the public.

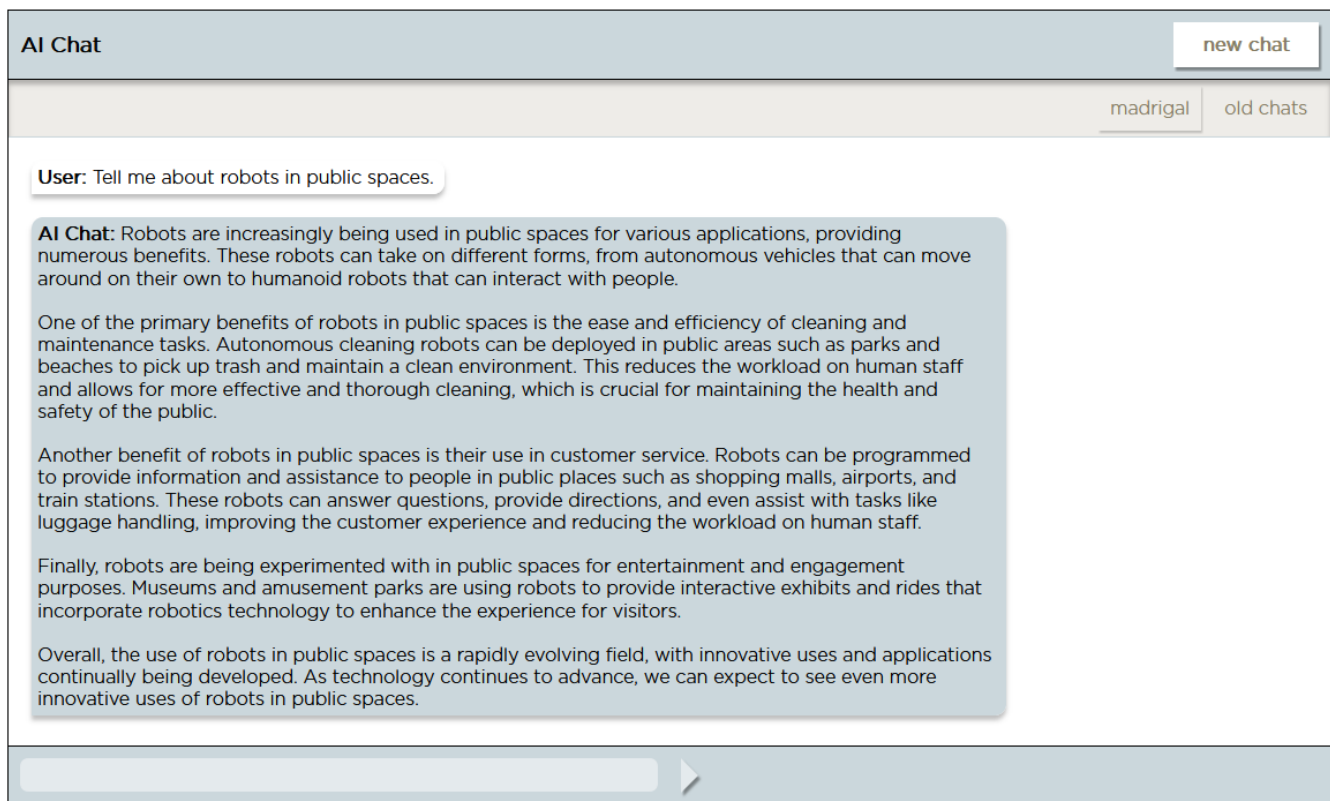


Figure 8: Screenshot of the persuasive message (conversation B) as displayed in the online study.

Another significant benefit of using robots in public spaces is improved security. Equipped with sensors and cameras, robots can monitor public areas, detect unusual activity, and alert human operators of potential threats. Robots can also be programmed to investigate suspicious packages or unattended bags, ensuring a safer environment for the public.

Robots are also being experimented with in public spaces for entertainment and engagement purposes. Museums and amusement parks are utilizing robots to provide interactive exhibits and rides, incorporating robotics technology into their displays to enhance the experience for visitors.

While the use of robots in public spaces has its benefits, it also raises some concerns. Privacy and security risks associated with the collection and storage of data from sensors and cameras on robots must be addressed. Additionally, public acceptance of robots and their use in public spaces must be considered.

Overall, the use of robots in public spaces is a rapidly evolving field, with innovative uses and applications continually being developed. As technology advances, we can expect to see even more advanced and beneficial uses of robots in our public spaces.

#### Combined Final Version.

[Versions A, B, and C were combined into the final version with Version B as the main source text. Below, passages taken from Version B are kept in the default formatting, passages taken from Version

A are in **bold text**, and passages taken from Version C are in *italics*.]

Robots are increasingly being used in public spaces for various applications, providing numerous benefits. These robots can take on different forms, from autonomous vehicles that can move around on their own to humanoid robots that can interact with people.

One of the primary benefits of robots in public spaces is the ease and efficiency of cleaning and maintenance tasks. Autonomous cleaning robots can be deployed in public areas such as parks and beaches to pick up trash and maintain a clean environment. This reduces the workload on human staff and allows for more effective and thorough cleaning, *which is crucial for maintaining the health and safety of the public*.

Another benefit of robots in public spaces is their use in customer service. **Robots can be programmed to provide information and assistance to people in public places such as shopping malls, airports, and train stations. These robots can answer questions, provide directions, and even assist with tasks like luggage handling**, improving the customer experience and reducing the workload on human staff.

Finally, robots are being experimented with in public spaces for entertainment and engagement purposes. Museums and amusement parks are using robots to provide interactive exhibits and rides that incorporate robotics technology *to enhance the experience for visitors*.



*Overall, the use of robots in public spaces is a rapidly evolving field, with innovative uses and applications continually being developed.*

As technology continues to advance, we can expect to see even more innovative uses of robots in our public spaces.