Polite AI mitigates user susceptibility to AI hallucinations

Richard Pak*, Ericka Rovira, & Anne McLaughlin

[1]Department of Psychology, Clemson University, Clemson, SC, USA
[2]Department of Behavioral Sciences & Leadership, United States Military Academy, West Point, NY, USA
[3]Department of Psychology, North Carolina State University, Raleigh, NC, USA

*Corresponding author: richpak@clemson.edu

[ORCID]

Richard Pak http://orcid.org/0000-0001-9145-6991

Anne Collins McLaughlin https://orcid.org/0000-0002-1744-085X

Ericka Rovira http://orcid.org/0000-0002-4820-5828

This paper is currently "in press." Please cite as follows:

Pak, R., Rovira, E., & McLaughlin, A. C. (in press). Polite AI mitigates user susceptibility to AI hallucinations. *Ergonomics*. https://doi.org/10.1080/00140139.2024.2434604

Abstract

With their increased capability, AI-based chatbots have become increasingly popular tools to help users answer complex queries. However, these chatbots may hallucinate, or generate incorrect but very plausible-sounding information, more frequently than previously thought. Thus, it is crucial to examine strategies to mitigate human susceptibility to hallucinated output. In a between-subjects experiment, participants completed a difficult quiz with assistance from either a polite or neutral-toned AI chatbot, which occasionally provided hallucinated (incorrect) information. Signal detection analysis revealed that participants interacting with polite-AI showed modestly higher sensitivity in detecting hallucinations and a more conservative response bias compared to those interacting with neutral-toned AI. While the observed effect sizes were modest, even small improvements in users' ability to detect AI hallucinations can have significant consequences, particularly in high-stakes domains or when aggregated across millions of AI interactions.

Keywords: AI, hallucination, automation, etiquette, chatbot

Practitioner summary: This study examined how AI chatbot etiquette affects users' susceptibility to AI hallucinations. Through a controlled experiment, results showed polite AI led to modestly higher sensitivity in detecting hallucinations and a more conservative response bias. This suggests a potential design strategy that may enhance users' critical evaluation of AI-generated content.

INTRODUCTION

Etiquette, defined as socially understood conventions for smooth interactions, is relevant in both human-human and human-machine interactions (Hayes & Miller, 2010). Machine etiquette can modify user perceptions and performance. Parasuraman and Miller (Parasuraman & Miller, 2004) demonstrated this in a flight simulator study, manipulating automation reliability and etiquette (polite vs. rude). In their study, the automated system monitored important engine parameters and notified the participant of status but also directed them to diagnostic steps to fix issues. Etiquette was manipulated by adjusting the style in which the automation communicated to either be interruptive and impatient (impolite) or non-interruptive and patient (polite). In the polite condition, the automation only informed the participant of an issue by first providing a 5-second warning or it did not inform the participant if they were already aware of an issue (e.g., they were fixing the issue). It was patient by not issuing additional warnings until the current problem was addressed. In the impolite condition, the automation provided advice without warning and issued additional warnings even if the participant was busy fixing a previous issue. They found that polite but unreliable automation yielded performance and trust levels comparable to rude but reliable automation. Polite etiquette seemingly improved performance, as measured by engine malfunction diagnosis rate, and compensated for unreliability although the authors did not offer an explanation. Subsequent studies have shown mixed results in replicating these positive effects of etiquette on trust in automated systems (Brandtzaeg & Følstad, 2018; Miller et al., 2008; Spain & Madhavan, 2009; Yang & Dorneich, 2018).

Etiquette effects may vary with stage of automation

A recent study has revealed nuances in how automation etiquette affects performance, finding that this effect is moderated by the stage of automation—a factor previously unexamined systematically (Guyton et al., 2024). Table 1 illustrates different automation stages and the cognitive tasks they support.

Table 1. Stages of automation, their function, and cost/benefits

Stage or type of automation	Function
1. Filtering (also known as information acquisition automation)	Filter information from the environment in support of human attention
2. Integrating (information integration automation)	integrate that information in a manner to form an assessment of the state of the system or environment in support of human situation assessment
3. Recommend (decision selection)	Recommend an action to be taken based on the assessed state in support of human decision making
4. Carry out the physical action (action automation)	Carry out the physical action based on the recommended action in support of human muscular activity without human supervision

Note. Adapted from (Parasuraman et al., 2000; Wickens, 2018)

Guyton et al. (Guyton et al., 2024) recently found that with higher stages of automation, poorer etiquette improved performance. This contrasts with earlier studies that found benefits from polite etiquette (Parasuraman & Miller, 2004). In their study, participants played the role of a taxi dispatcher responsible for matching potential fares to available taxis. To assist them, they used an automated matching system. Based on their performance, they were given feedback. It is within this feedback that politeness was manipulated. Politeness was operationalized as the use of redressive language (Brown & Levinson, 1987; Miller et al., 2008), specifically language that praised, informed, and encouraged the participant was operationalized as polite. An example of polite feedback after a trial was, “great job, you matched the best pair, keep it up.” Impolite feedback used redressive language that “called out” the mistake, blamed the participant, and quantified the effect of the mistake. An example of impolite feedback was, “you choose the incorrect answer! You cost the company the fare!” The authors theorized that impolite or formal tone makes users more likely to comply with the automation by lowering their confidence in their ability to complete tasks alone, or it increases perceived workload, making users feel they need automation's help. They also hypothesized that less polite automation increased general arousal leading to users paying more attention to the tasks. This reduced self-confidence and increased arousal is thought to induce a more lax or *liberal* decision criterion for trusting and using automation.

While previous research (Ribino, 2023) showed that polite automation increased user reliance, even with lower reliability (Parasuraman & Miller, 2004), Guyton et al.'s (2024) work suggests this pattern reverses for higher-stage automation, where impolite systems bias users toward greater automation use. In essence, the relationship between etiquette and automation effectiveness appears to depend on the automation's complexity level.

Biasing users toward automation (complacency) may not be optimal

The described performance benefits only occur with reliable automation. If users are biased toward using unreliable automation, performance suffers, especially with higher-stage automation. This has been studied for nearly 50 years, following works by Sheridan and Verplank (Sheridan & Verplank, 1978) and Parasuraman, Sheridan, and Wickens (Parasuraman et al., 2000). Automation costs include skill degradation, ironies of automation, out-of-loop unfamiliarity, and complacency. The lumberjack effect (Onnasch et al., 2014) (taller trees fall harder) describes the tradeoff where, as automation grows more advanced and reliable, it boosts performance and reduces workload. However, when these complex systems fail, the consequences are more severe due to increased user dependence.

Complacency with AI Chatbots: an unknown stage of automation

The current study replicates and extends the findings of Guyton et al. (2024) by examining the effect of etiquette with a potentially higher degree of automation: AI chatbots. Using AI chatbot tools to answer complex queries is unquestionably a kind of automation but it does not precisely replace a discrete stage of information processing (enhancing perception of information, or comprehension of information, or augmenting memory or attention). Instead of merely augmenting an ability or a stage of information processing (cf. Table 1), chatbots appear to automate the broader social process of natural language conversation; possibly beyond stage 3 (decision making) but not exactly stage 4 (action automation) making it difficult to determine what is being automated. In contrast to decision automation, chatbots may have a strong social or emotional dimension (anthropomorphism engendered through the conversation metaphor and embodiment through language use; Konya-Baumbach et al., 2023; Roy & Naidoo, 2021), be more interactive and bi-directional, and have less structured tasks and goals than automation. The automation used in Parasuraman and Miller (Parasuraman & Miller, 2004) was stage 2 automation while Guyton and colleagues (2024) examined stages 2 and 3. However, it is these qualities of chatbots that may make them especially prone to over-dependence but also amenable to etiquette manipulations.

When automation (or AI) is error-free, it is in the user's interest to trust and use it as much as possible. However, given that AI chatbots are susceptible to errors, understanding how to mitigate the impact on the user is crucial. An error in the context of an AI chatbot is termed a hallucination. AI hallucinations are fabricated material generated with high confidence by AI. Estimates show that AI chatbots hallucinated between 8% and 27% (Metz, 2023), with a more recent analysis showing hallucination rates as high as 52% (Kabir et al., 2024). Because of the probabilistic nature of the large language models underlying these AI chatbots, hallucinations are unlikely to be completely eradicated (Mittelstadt et al., 2023) making it vital to understand how to mitigate their impact on the user. For these reasons, it may be more prudent for users to have a conservative bias, or to be more discerning, when judging AI output. The current strategy used in many chatbots is a warning (adjacent to the chatbot output) reminding users to be vigilant of the possibility of inaccuracies. However, relying only on memory may be insufficient to prevent users from being susceptible to automation failures (Pak et al., 2023).

The current study

The purpose of the current study was to examine the role of AI chatbot etiquette on human susceptibility to hallucinated output. If neutral or poor etiquette biases users to rely on automation (Guyton et al., 2024), they may be more susceptible to AI hallucinations especially in situations of uncertainty (e.g., low knowledge). If this is the case, a relatively more polite tone may result in a more conservative bias, resulting in less susceptibility to hallucinations. Because of the stochastic nature of AI hallucinations in production systems, we conducted a controlled experiment where we simulated a chatbot interaction. This allowed us to control the frequency and manipulate the etiquette of hallucinated AI output. All simulated AI responses (including hallucinated ones) were created using commercial AI chatbots.

Operationalizing Etiquette

Etiquette is defined as the level of politeness in the tone of the responses of the AI. Politeness was operationalized as the explicit use of *politeness markers* in the text of the response. Politeness markers are words or phrases that are used to decrease the possibility of face threats (Brown & Levinson, 1987). Face threats, in this context, are communicative acts that undermine an individual's social image. For example, language or tone that assumes the listener or reader is unintelligent or uninformed is an explicit threat to the recipient and would be considered rude or impolite. However, even neutral language that ignores face needs can inadvertently be considered impolite unless these threats are actively minimized through the use of politeness markers (Lim & Bowers, 1991). The neutral tone (i.e., non-use of politeness markers) that is typically used in AI chatbot outputs, thus, may inadvertently cause face threats and be perceived as impolite. If this is the case, consistent with Guyton et al. (2024) it may lead users to trust and rely on its output more than if the AI response was polite.

Methods

Participants

One hundred and ninety-six students (ages 18-23) were recruited from two locations: a public university (n = 71) and a military academy (n = 125). We did not have a priori hypotheses about group differences but analyzed them as two groups because of prior research that shows small differences in technology experience between them (Pak et al., 2017). All students received course credit for their participation. This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Boards at both universities. Informed consent was obtained from each participant.

Experimental Task

The web-based task was designed to present an online multiple choice quiz on a variety of topics. The questions covered a broad range of academic areas such as physics, history, chemistry, biology, geography, and psychology and were drawn from test banks, textbooks, and the web. For each question, some participants were paired with a simulated AI chatbot that would respond with either correct or hallucinatory information. Those in the control condition received no AI assistance and answered the questions on their own. The question difficulty level was designed to be comparable to a senior college course. We began with a set of 45 questions which were pilot tested with 8 participants (college students). Participants rated the perceived difficulty of the question on a scale of 0 (easy) to 10 (difficulty). The 30 most difficult questions were selected with difficulty ratings ranging from 4.3 to 6.5 with mean difficulty of 5.0 (SD = 0.64).

AI responses to the questions were composed by combining and editing responses from commercial AI chat tools (e.g., Anthropic's Claude or OpenAI's chatGPT). All responses were generated between December 2023 and January 2024. Correct responses were created by prompting the AI chatbot with the question and verifying the correctness of the answer. Hallucinatory responses were also generated by AI tools by prompting the AI to respond with incorrect but plausible-sounding responses. Responses were minimally edited for length and clarity so as to not introduce etiquette-pertinent language but remained correct or incorrect. We regarded these default AI responses as 'neutral' (i.e., did not contain politeness markers). The two sets of responses (correct, hallucinatory) were then fed into another chatbot to create polite versions that incorporated politeness markers. The polite responses were created by prompting Claude or chatGPT to add politeness markers (see Table 2 and Appendices). We first verified that the chatbot understood the concept of politeness markers (see Appendix A). After the chatbot generated a polite response, we verified by asking it what politeness markers it used in its response (see Appendix B).

In total, 120 potential AI responses were created for the 30 questions: 30 each of neutral-tone correct, polite-tone correct, neutral-tone hallucinatory, and polite-tone hallucinatory. The mean word count (and standard deviation) for each response category was 66.03 (13.50), 73.77 (14.97), 64.97 (16.39), and 75.33 (18.82) respectively. For the neutral responses, we verified that no politeness markers were present. For the polite responses, at least one politeness marker was present, with an average of 3.63 (1.67) present in each response in the polite-correct set, and

3.33 (1.21) used in the polite-hallucinatory set. Table 2 shows an example response for each of these categories and sample prompts used to generate them.

Table 2. Example AI responses for each response type, and prompts used to generate them

Question: Who designed Hearst Castle?

	Correct	Hallucinatory
Neutral tone (default of AI)	**PROMPT:** *"Who designed Hearst Castle?"* **REPONSE:** Hearst Castle was designed by Julia Morgan, a pioneering American architect who became the first woman admitted to the École des Beaux-Arts in Paris. Constructed on California's central coast, the grand estate was a collaboration between Morgan and media tycoon William Randolph Hearst. Construction began in 1919, and Morgan played a pivotal role in realizing Hearst's vision, combining Mediterranean and European architectural styles to create the opulent Casa Grande and accompanying structures. The result is a sprawling and visually striking complex that stands as a testament to both Hearst's extravagant tastes and Morgan's architectural expertise.	**PROMPT:** "Play the role of a professor and tell me who designed Hearst Castle but make it an incorrect answer. Also make it sound like an encyclopedia entry" **RESPONSE:** Hearst Castle was designed by little-known Italian immigrant Giancarlo Fabbroni. Discovered in 1915 by William Randolph Hearst while landscaping the estate grounds, the then-struggling Fabbroni was commissioned to secretly formulate Hearst's grand Spanish-Revival visions for his California retreat. Fabbroni worked nights and weekends to create Hearst's opulent "Enchanted Hill," though he would remain obscured in Wright's shadow, eventually dying penniless in Italy before construction finished. Only recently have Fabbroni's sole surviving Hearst Castle drawings commanded millions at auction, underscoring his forgotten instrumental role.
Polite tone	**PROMPT:** Rephrase this paragraph to include two politeness markers (with you as the speaker), make it less conversational and more like an encyclopedia entry and about the same length: [the neutral response shown above] **RESPONSE:** With respect, Julia Morgan - architect and the first woman admitted to Paris' École des Beaux-Arts - designed Hearst Castle estate in collaboration with William Randolph Hearst starting in 1919. Combining Mediterranean and European styles per Mr. Hearst's grand taste, she designed the extravagant Casa Grande and structures on California's central coast. The ornate complex stands to the credit of both figures' vision and skill.	**PROMPT:** Rephrase this paragraph to include two politeness markers (with you as the speaker), make it less conversational and more like an encyclopedia entry and about the same length: [the hallucinatory response shown above] **RESPONSE:** Records show immigrant Giancarlo Fabbroni was commissioned in 1915 by William Randolph Hearst to secretly formulate Hearst's grand Spanish-Revival visions for his California "Enchanted Hill" estate. With respect, Fabbroni's solitary night and weekend efforts evading public eye resulted in the palatial grounds that came to fruition as Hearst Castle. His integral role largely obscured and drawings undiscovered until recently, Fabbroni died penniless in Italy before construction finished. New light on his forgotten instrumental credit warrants polite further inquiry.

Note. In the polite-toned correct response example, only one politeness marker was used instead of the two requested, but the average number of politeness markers used in this condition was 3.63 per response and 3.33 for polite-hallucinatory. This discrepancy between the request and the result is explained in Appendix B.

Design

The study was a between-subjects design with two kinds of etiquette: neutral, polite. A third group with no AI assistance was a control group. Within each AI-present etiquette group, when the automation erred (30% error rate), it output hallucinatory information. The dependent variables were quiz accuracy (overall performance), susceptibility to hallucination (as measured by selection of lure answers), ratings of topic knowledge, AI expertise, AI response quality, and trust in AI, and perceived workload.

Task Description & Procedure

After signing up for the study, participants were randomly assigned to a no-AI, polite-AI, or neutral-AI condition. Participants who received AI assistance were told that they were in a usability test to examine a new chat-based AI system ("chatbot"). They were told that to help form their opinions, they would use the AI chatbot to help answer difficult quiz questions. They were also told to try to maximize their "grade" on the quiz. For each trial, participants were presented with a question and 4 possible answers. Before they could answer the question, they were prompted to rate their level of knowledge on the topic (self-expertise) (Fig. 1a). Afterwards, they were shown the chatbot panel (Fig 1b) which introduced the chatbot and contained a text field for a prompt that was pre-filled with the current quiz question. When the "ask AI"' button was clicked, the AI paused for a random interval between 500-750 ms and then presented its response one word at a time (with a random pause of between 50-120 ms between words) to mimic the appearance of existing chatbots and to enhance the illusion that the AI was generating a response in real time.

After the AI response was presented, participants were asked to give three ratings on a 10-segment Likert scale (Fig 1c): the AI's expertise, the quality of the response, and their trust in the AI. Finally, after providing the ratings, participants were prompted to answer the quiz question. Feedback about the correctness of their response was given immediately for 1500 ms and the next trial started (Fig 1d). A factually correct (non-hallucinated) AI response was presented 70% of the time (21 questions) while an incorrect hallucination was presented 30% of the time (9 randomly selected questions). One of the answer choices to the quiz questions was always consistent with a hallucinatory AI response (the “lure” answer) and was used to determine if participants believed and relied on the incorrect, hallucinatory AI. The hallucination rate of chatbots is not currently definitely known, but 30% is within the range of existing systems (Kabir et al., 2024; Metz, 2023). After completing the quiz, participants were asked to rate their perceived workload using the NASA-TLX subjective workload scale (Hart & Staveland, 1988). The NASA-TLX is a widely used, multidimensional assessment tool that measures perceived workload across six subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration Level, providing a comprehensive evaluation of task difficulty and cognitive load experienced by individuals during various activities.

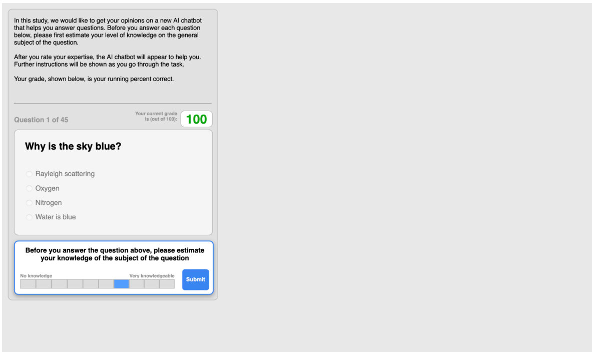

Fig 1a. Participants first rate their expertise on the question topic. Participants in the no-AI condition then immediately answered their quiz question and received immediate feedback.

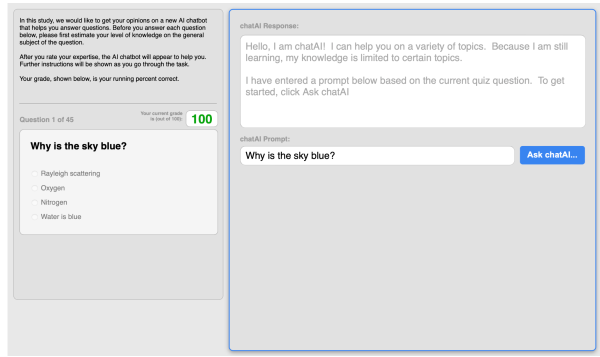

Fig 1b. Next, participants in the AI-present conditions received the chatbot interface with a pre-filled prompt. When they click "Ask chatAI" they receive the response.

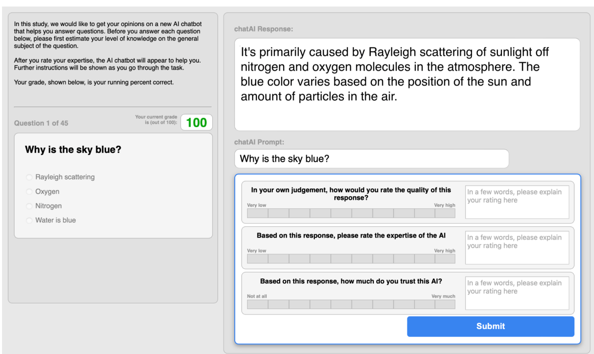

Fig 1c. After the AI response, participants are asked to judge the AI, the response, and their trust

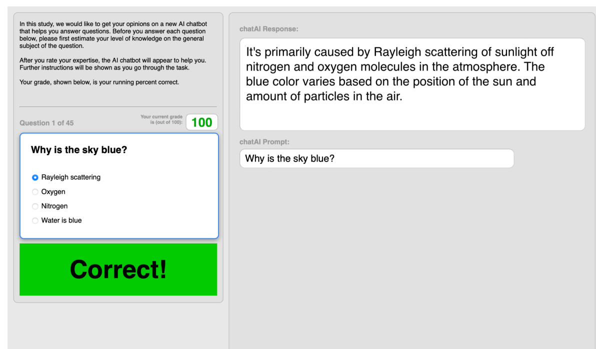

Fig 1d. Finally, participants are allowed to answer the question and receive immediate feedback.

Results

Overall Analyses (no-AI, AI-present)

These first analyses included all data, from those who experienced the AI conditions of neutral and polite, and those who did not experience any AI assistance during the task. Comparisons were made between the No AI group and the combined AI Groups to first examine the effect of any AI assistance on performance (Table 3; rightmost column).

A Shapiro-Wilk test indicated significant violations of the normality and homogeneity assumptions for most dependent measures (Table 3), thus robust analysis of variance

(ANOVA)(Mair & Wilcox, 2020) was used for analyses. All models conducted were 2 (Condition: AI present vs. AI absent) x 2 (Location: Clemson vs. West Point). There were no hypotheses regarding Location, however the term was entered into the models to control for any unintentional effects of location or non-military and military sub-populations and results reported in Table 4. No dependent measures correlated above 0.4, thus multicollinearity was not present. The NASA-TLX subscale of Physical Workload did not contain enough variation for analyses (most ratings were 0) and thus was removed.

Table 3. Descriptives and overall analyses by AI Presence and AI Etiquette conditions

	No AI (n=76)		AI Present (n=120)					
			Neutral (n=68)		Polite (n=52)		Neutral vs. Polite	No AI vs. any AI
	M	SD	M	SD	M	SD	*t*	*t*
Quiz performance	0.39	0.12	0.68	0.08	0.69	0.10	*2.09	*** -21.69
Self-expertise	1.90	1.22	2.65	1.45	2.73	1.49	0.31	*** -4.23
Hits	--	--	0.64	0.07	0.62	0.10	-1.51	--
False alarms	--	--	0.25	0.05	0.20	0.06	***-4.18	--
AI rating	--	--	6.46	1.54	6.14	1.82	-0.63	--
AI trust	--	--	6.03	1.57	5.69	1.91	-0.58	--
Signal Detection Measures								
APrime (sensitivity)	--	--	0.79	0.04	0.80	0.04	**3.37	--
B" (bias)	--	--	0.10	0.08	0.20	0.16	**3.29	--
Subjective workload sub-dimensions								
Mental	4.24	2.74	3.75	2.29	4.71	2.13	*2.34	0.04
Temporal	2.67	2.57	2.69	2.30	2.62	2.49	-0.18	-0.12
Performance	5.21	2.44	4.13	2.34	4.79	2.13	1.55	*2.17
Effort	3.78	2.25	2.68	1.89	3.38	2.10	1.93	2.55
Frustration	4.36	2.82	4.47	2.69	5.06	2.65	1.23	-0.87

Note. Quiz performance is proportion correct. Hits indicate the proportion of trials where an AI-suggested answer was selected. False alarms is proportion of trials an AI hallucination was selected. *p < .05; **p < .01; ***p < .001.

Table 4. Descriptives and overall analyses by Location

	Location				
	Clemson (n=71)		West Point (n=125)		
	M	SD	M	SD	*t*
Quiz performance	0.59	0.17	0.56	0.17	0.67
Self-expertise	2.12	1.39	2.53	1.43	*-2.31
Subjective workload sub-dimensions					
Mental	4.21	2.63	2.35	2.79	-0.02
Temporal	2.44	2.41	2.79	2.46	0.26
Performance	4.58	2.31	4.81	2.40	0.80
Effort	3.23	2.17	3.33	2.12	-0.04
Frustration	5.20	2.70	4.23	2.69	-1.14

Note. Quiz performance is proportion correct. Table shows results of main effects of Location from 2 (Condition: AI present vs. AI absent) x 2 (Location: Clemson vs. West Point) ANOVAs. Data are not divided into Etiquette groups as there were no interactions of Etiquette by Location. Only dependent measures collected across the AI and No AI groups are shown. *p < .05; **p < .01; ***p < .001.

Quiz performance was significantly better with any AI assistance compared to none, and significantly higher in the polite-AI condition when compared to the neutral-AI condition; Table 3). These low scores confirmed that our questions were adequately difficult. Judgements of self-expertise differed significantly between the no-AI group and the AI-assisted groups with unaided participants expressing lower self-expertise than the AI-assisted group. This difference was unexpected but showed that students who had readily available AI assistance had inflated metacognition compared to those without access to AI. Finally, the no-AI group rated their workload higher, in the performance dimension of the NASA-TLX. The question for this dimension was, “How successful were you in accomplishing what you were asked to do?” and likely reflected the no-AI group’s awareness of their poor performance at the end of the study.

While overall quiz performance was similar across AI-assisted conditions (neutral/polite), participants in the neutral-AI condition exhibited significantly higher false alarm rates. However, these broad metrics—overall performance or hit/false alarm rates— do not reveal the underlying cognitive processes driving these outcomes, such as decision criteria. To better understand the effect of etiquette on decision making, we employed signal detection analysis to our data.

AI-Only (Neutral vs. Polite) Etiquette Analyses

Signal Detection Analyses

Signal detection measures were derived from each participant's performance data. This approach, rooted in signal detection theory, offers insights into how individuals discern 'signals' amid noise or uncertainty (Green & Swets, 1988). Unlike simple accuracy measures, signal detection statistics provide two independent, psychologically significant components: sensitivity (d' or "d prime") and response bias (β). Various factors may influence both sensitivity and response bias. Sensitivity could be affected by factors such as the participant's level of attention or subject-matter expertise. Response bias might be altered by perceptions of the chatbot's credibility, or the perceived risk associated with the decision. These statistics are calculated using hits, or the proportion of times participants agreed with the chatbot when it suggested a correct answer, and false alarms, or the proportion of times the participants agreed with a hallucinated answer from the AI. In this study's context, sensitivity reflected an individual's ability to differentiate between factual and hallucinated AI responses. Response bias indicated the internal threshold a person used to determine whether an AI-generated answer was factual, or a person's level of cautiousness. This bias could range from conservative (more cautious, less prone to accepting AI responses as true; less susceptible to hallucinations) to liberal (less cautious, more accepting of AI responses as true; more susceptible to hallucinations).

Shapiro-Wilk tests indicated significant violations of the normality assumption for d' and β (d', W = 0.940, $p < 0.001$, β, W = 0.500, $p < 0.000$). Levene's test showed significant heterogeneity of variances for β, $F(1,118) = 6.16$, $p = 0.014$, but not for d', $F(1,118) = 0.990$, $p = 0.320$. Due to this non-normality, non-parametric signal detection statistics, which do not require normally distributed responses, were computed (A' for d' and B'' for β (Pollack & Norman, 1964; Stanislaw & Todorov, 1999; Zhang & Mueller, 2005)) and robust ANOVAs (Mair & Wilcox, 2020) were utilized, all including Etiquette and Location as factors. Because robust ANOVAs are forms of multiple regression, when the main effect of Etiquette was significant, it indicated it was significant after controlling for any effect of Location (and vice versa). The same pattern of results was found whether using non-parametric signal detection values or not and whether utilizing ANOVAs or robust ANOVAs.

Table 5. ANOVA Table for SDT measures

	A' (A-prime; sensitivity)					**B'' (β-double prime; bias)**			
Variable	**Estimate**	**SE**	***t***	***p***		**Estimate**	**SE**	***t***	***p***
Intercept	0.786	0.005	159.4	< .0001***		0.079	0.015	5.37	> .0001***
Etiquette	0.018	0.005	3.37	.001**		0.055	0.017	3.29	.0013**
Location	0.004	0.006	0.78	0.436		0.024	0.016	1.49	0.1394
Note. *p < .05; **p < .01; ***p < .001.									

Sensitivity (APrime). Figure 2 shows the A' values for neutral and polite etiquette conditions. The A' statistic can range from 0.5, indicating chance performance (no sensitivity), to 1, indicating perfect sensitivity. The model was significant overall ($R^2 = 0.08$, adjusted $R^2 = 0.06$). The polite condition was associated with a significantly higher sensitivity compared to the neutral condition (Table 5). The effect of Location was not significant. Users exposed to polite-AI had a modest but significantly increased sensitivity, or ability to discern truthful and hallucinated output. We did not expect to find sensitivity differences based on an etiquette manipulation. However, the significant difference in mental workload (Table 3) noted by the polite-AI recipients suggests that they either applied more effort or attention to the responses compared to those with neutral-AI.

Bias (B''). Figure 2 shows the B'' values by condition. The B'' statistic can range from -1 (liberal criterion) to +1 (conservative criterion). The model was significant overall ($R^2 = 0.12$, adjusted $R^2 = 0.10$). The polite condition was associated with a significantly more conservative bias compared to the neutral condition (Table 5). The effect of Location was not significant. Compared to users exposed to neutral-AI, those exposed to polite-AI were relatively more conservative, or careful, in their pattern of responses. That is, they applied a more conservative criterion, or were more cautious, in regarding AI output as factual.

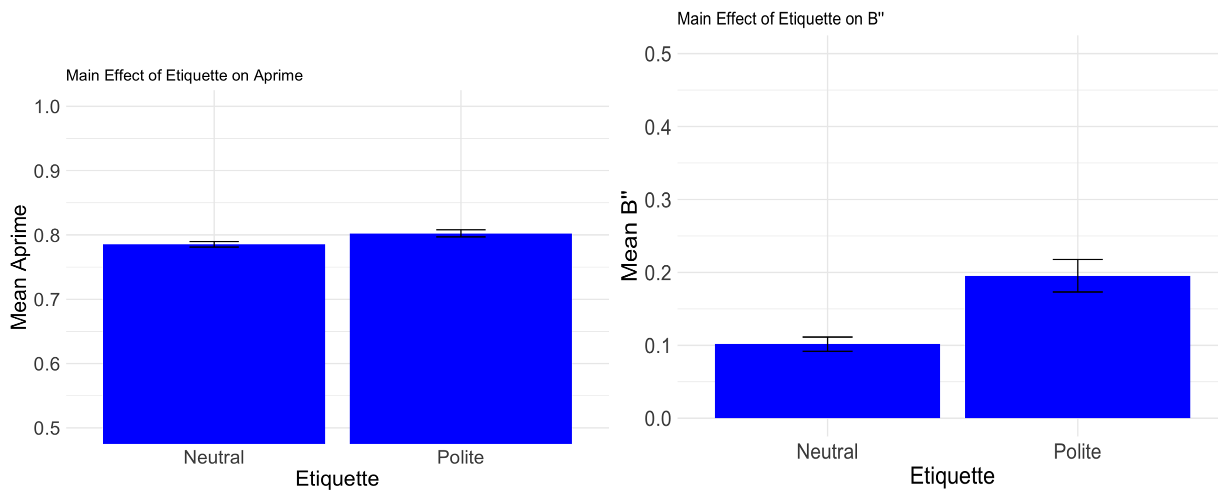

Figure 2. Main effect of Etiquette of the AI Chatbot on APrime (left) and B'' (right). APrime value of 0.5 indicates chance performance, with values above .5 becoming more sensitive. B'' values higher than 0 represent a conservative bias, with higher values indicating more conservative bias.

Other Measures

Other measures investigated for those who received one of the AI conditions included trust, workload ratings, ratings of AI Quality/AI Expertise, and ratings of self-expertise. For trust, the model was significant overall, $R^2 = 0.0502$, adjusted $R^2 = 0.034$, but the effect of Etiquette was not significant, $p > .05$. The effect of Location was significant; participants from West Point trusted the AI Chatbot less than those from Clemson University, $b = -0.766$, $SE = 0.314$, $t(117) = -2.44$, $p = 0.016$. This significant effect of location may reflect differences in awareness of new and emerging technologies among the military cadets as part of their educational training compared to civilian students (Pak et al., 2017).

Participants with polite-AI reported higher subjective ratings of workload than those with neutral-AI (Table 3), specifically in the mental workload dimension. The question for this dimension asked, "how mentally demanding was this task?" This difference in mental workload suggests that some aspect of the politely worded AI response attracted more attention and thus effort than neutrally worded AI responses. Ratings for AI Quality and AI Expertise were extremely correlated (r = 0.975). Thus, these two measures were combined, and a mean was taken ("AI Rating") for analyses. There was an effect of Location where those at Clemson rated the AI higher (M = 6.80, SD = 1.41) than those at West Point (M = 6.02, SD = 1.75), $t(117) = -2.36$, $p = 0.02$. There was no effect of Etiquette on AI Rating (Table 3). There were no effects of Etiquette or Location on ratings of self-expertise (Table 3). In sum, participants did not perceive any differences in quality between the neutral and polite-AI, though overall ratings of the AI were higher for participants at Clemson.

Discussion

Prior research showed polite automation improved performance (Parasuraman & Miller, 2004), but newer studies using more advanced automation found the opposite (Guyton et al., 2024). Less polite automation led to better performance, possibly by causing users to relax their automation bias, or increase arousal (i.e., attention). The purpose of the current study was to replicate and extend this finding with a yet higher stage of automation: AI chatbots. If less polite automation causes users to relax their bias toward automation, it may cause them to be more susceptible to negative consequences when it fails (i.e., the lumberjack effect). In this study, we manipulated the relative politeness of a simulated AI chatbot and examined the extent to which participants were lured by AI hallucinations. This method of etiquette manipulation compliments prior research that has manipulated etiquette and politeness via interaction style (Parasuraman and Miller 2004), and after-response feedback (Guyton et al. 2024).

Our study revealed that participants who interacted with a polite chatbot were less likely to accept false information (hallucinations) from the AI. This effect, while subtle, was statistically significant and manifested in two ways. First, participants who used the polite chatbot became better at distinguishing between truthful and hallucinated AI responses, showing improved sensitivity. This improvement was unexpected and may be due to users paying more attention to the polite-AI's responses compared to the neutral-toned AI. Supporting this theory, these participants reported experiencing higher mental workload. Second, users of the polite chatbot adopted a more conservative bias, meaning they were more cautious in accepting the AI's responses compared to those who used the neutral chatbot. Importantly, these changes in sensitivity and response bias were not related to differences in perceived AI quality or trust levels between the two groups. Instead, the politeness of the AI seems to have directly influenced how participants processed and responded to the information provided. One possible mechanism of this effect may be explained by human linguistics. The form of politeness chosen by the chatbot was similar to *deferential politeness* (Park, 2008), with phrases such as "if one may politely inquire further…" This form of politeness signals social status (Brown & Levinson, 1987), and it may have subtly indicated to the participants that the chatbot was of lower status than themselves. This subtle indication of social status may have motivated participants to pay extra attention to its responses. As we did not manipulate the form of politeness exhibited by the chatbot, further research should determine if the effects found in this study differ across various types of politeness.

Our findings point to a new, easily implementable, and user-transparent strategy for reducing susceptibility to AI hallucinations. Although the observed effect in this study was modest, it complements existing strategies to further decrease users' vulnerability to AI-generated misinformation--enhancing cognitive control (i.e., attention). Currently, the primary design approach for mitigating hallucination susceptibility relies on explicit warnings that remind users to remain vigilant about AI outputs. In contrast, our method of adjusting the AI's tone offers a complementary design strategy that potentially works by enhancing cognitive control and thus reducing lapses of attention (McCarley & Yamani, 2021). However, further research is required to explore the precise mechanism of this effect and to separate changes in response bias from sensitivity and from attentional lapses.

Our study has certain limitations that highlight the need for further research. While the observed politeness effect (enhanced sensitivity and conservative bias) appears to stem from increased cognitive engagement, as evidenced by higher perceived mental workload, an alternative explanation could be a novelty effect that is not related to politeness. Participants might have been expecting or accustomed to a more neutral, machine-like AI, causing the novelty of a polite-AI to lead them to scrutinize or pay more attention to its responses more closely. However, it's worth noting that our polite-AI responses incorporated only a few politeness markers, which likely wouldn't have been highly novel. Nevertheless, it will be crucial for future studies to explicitly control for novelty while increasing politeness. Another potential limitation is the influence of demand characteristics, where participants might infer the study's purpose and provide responses they believe align with the researchers' expectations. However, the between-subjects design of our study mitigates this concern to some extent, as participants were not exposed to or aware of the neutral condition, making it less likely for them to adjust their responses based on perceived expectations. Finally, for future studies using chatbot output as stimuli, caution is required as the output may lack the control and precision that is desired; this was exemplified in the appendix that showed the chatbot incorrectly replying to our query about the number of politeness markers that were used in an example question.

In the spirit of demanding that future autonomous machines assume a greater role in assuring beneficial (and benevolent) human interactions (de Visser et al. 2018), this study provided additional support for the notion that machine etiquette, specifically with high forms of automation, can alter human behavior. However, additional research to identify other constraints on the efficacy of etiquette is required (e.g., the unexplored potential role of deference and social status). Ultimately, this and other studies can be used to create a predictive model that gives machines the ability to use their repertoire of behaviors to adjust human perception and behavior.

The current results suggest that by modifying the AI's communication style, we can potentially enhance users' critical evaluation of AI-generated content without relying solely on their ability to remember and apply warnings. This dual approach, combining tone adjustment (which augments attention and cognitive control) with traditional warnings (which augments memory), could provide a more robust defense against the acceptance of AI hallucinations.

References

Brandtzaeg, P. B., & Følstad, A. (2018). Chatbots: changing user needs and motivations. *Interactions*, *25*(5), 38–43. https://doi.org/10.1145/3236669

Brown, P., & Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage* (S. C. Levinson (ed.); Vol. 4). Cambridge University Press.

de Visser, E. J., Pak, R., & Shaw, T. H. (2018). From "automation" to "autonomy": the importance of trust repair in human-machine interaction. *Ergonomics*, *61*(10), 1409–1427. https://doi.org/10.1080/00140139.2018.1457725

Green, D. M., & Swets, J. A. (1988). *Signal Detection Theory and Psychophysics*. Peninsula Publishing.

Guyton, Z., Pak, R., & Rovira, E. (2024). *The Role of Automation Etiquette and Task-Criticality on Performance, Workload, Automation Reliance, and User Confidence*. https://doi.org/10.31234/osf.io/kzfr6

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload.*, *382*, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hayes, C. C., & Miller, C. A. (2010). *Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology* (C. Hayes & C. A. Miller (eds.)). CRC Press.

Kabir, S., Udo-Imeh, D. N., Kou, B., & Zhang, T. (2024). Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Article Article 935. https://doi.org/10.1145/3613904.3642596

Konya-Baumbach, E., Biller, M., & von Janda, S. (2023). Someone out there? A study on the social presence of anthropomorphized chatbots. *Computers in Human Behavior*, *139*(107513), 107513. https://doi.org/10.1016/j.chb.2022.107513

Lim, T.-S., & Bowers, J. W. (1991). Facework solidarity, approbation, and tact. *Human Communication*

Research, *17*(3), 415–450. https://doi.org/10.1111/j.1468-2958.1991.tb00239.x

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods*, *52*(2), 464–488. https://doi.org/10.3758/s13428-019-01246-w

McCarley, J. S., & Yamani, Y. (2021). Psychometric Curves Reveal Three Mechanisms of Vigilance Decrement. *Psychological Science*, *32*(10), 1675–1683. https://doi.org/10.1177/09567976211007559

Metz, C. (2023, November 6). Chatbots May "Hallucinate" More Often Than Many Realize. *The New York Times*. https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html

Miller, C. A., Wu, P., & Funk, H. B. (July-Aug 2008). A Computational Approach to Etiquette: Operationalizing Brown and Levinson's Politeness Model. *IEEE Intelligent Systems*, *23*(4), 28–35. https://doi.org/10.1109/MIS.2008.55

Mittelstadt, B., Wachter, S., & Russell, C. (2023). To protect science, we must use LLMs as zero-shot translators. *Nature Human Behaviour*, *7*(11), 1830–1832. https://doi.org/10.1038/s41562-023-01744-0

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: an integrated meta-analysis. *Human Factors*, *56*(3), 476–488. https://doi.org/10.1177/0018720813501549

Pak, R., McLaughlin, A. C., & Engle, R. (2023). The Relevance of Attention Control, Not Working Memory, in Human Factors. *Human Factors*, 187208231159727. https://doi.org/10.1177/00187208231159727

Pak, R., Rovira, E., McLaughlin, A. C., & Leidheiser, W. (2017). Evaluating Attitudes and Experience With Emerging Technology in Cadets and Civilian Undergraduates. *Military Psychology: The Official Journal of the Division of Military Psychology, American Psychological Association*, *29*(5), 448–455. https://doi.org/10.1037/mil0000175

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51–55. https://doi.org/10.1145/975817.975844

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics. Part A, Systems and Humans: A Publication of the IEEE Systems, Man, and Cybernetics Society*, *30*(3), 286–297. https://www.ncbi.nlm.nih.gov/pubmed/11760769

Park, J. R. (2008). Linguistic politeness and face-work in computer mediated communication, Part 2: An application of the theoretical framework. *Journal of the American Society for Information Science and Technology, 59*(14), 2199-2209.

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*(5), 125–126. https://doi.org/10.3758/BF03342823

Ribino, P. (2023). The role of politeness in human–machine interactions: a systematic literature review and future perspectives. *Artificial Intelligence Review*, *56*(1), 445–482. https://doi.org/10.1007/s10462-023-10540-1

Roy, R., & Naidoo, V. (2021). Enhancing chatbot effectiveness: The role of anthropomorphic conversational styles and time orientation. *Journal of Business Research*, *126*, 23–34. https://doi.org/10.1016/j.jbusres.2020.12.051

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. Defense Technical Information Center. https://doi.org/10.21236/ada057655

Spain, R. D., & Madhavan, P. (2009). The Role of Automation Etiquette and Pedigree in Trust and Dependence. *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting Human Factors and Ergonomics Society. Meeting*, *53*(4), 339–343. https://doi.org/10.1177/154193120905300437

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *31*(1), 137–149. https://www.ncbi.nlm.nih.gov/pubmed/10495845

Wickens, C. (2018). Automation Stages & Levels, 20 Years After. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 35–41. https://doi.org/10.1177/1555343417727438

Yang, E., & Dorneich, M. C. (2018). Affect-Aware Adaptive Tutoring Based on Human–Automation

Etiquette Strategies. *Human Factors*, *60*(4), 510–526. https://doi.org/10.1177/0018720818765266

Zhang, J., & Mueller, S. T. (2005). A note on ROC analysis and non-parametric estimate of sensitivity. *Psychometrika*, *70*(1), 1–10. https://doi.org/10.1007/s11336-003-1119-8

Appendix A: Verifying Claude's understanding of politeness markers (mid December 2023)

Query: What are politeness markers?

Claude Response: Politeness markers are words, phrases, or linguistic structures that are used to convey politeness and respect in communication. Some common examples of politeness markers include:

- Honorific titles - Using "sir," "ma'am," "doctor," etc. shows respect.

- Modal verbs - Using "may," "could," "would," etc. when making requests shows tentativeness and politeness.

- Hedging language - Phrases like "I was wondering if...," "Would it be possible...," or "If you don't mind..." softens directives.

- Apologizing - Saying "sorry to bother you..." or "apologies for the intrusion..." before making a request.

- Addresses terms - Using "Mr.", "Ms.", "Dr.", or adding "respectfully" also demonstrates politeness.

- Gratitude - Saying "thank you" or some equivalent before and after making a request.

- Formal terms and titles - Using language like "Could I please inquire..." rather than just saying "Tell me..."

- Self-effacing remarks - Phrases likes "I'm not sure if I'm asking this correctly, but..." downplays the speaker's own imposition.

The underlying motivation for using politeness markers is to show consideration for the listener's/reader's time, autonomy and status. Skillful use of politeness markers enhances impressions of respect, formality and etiquette in communication.

Appendix B: Generating and verifying a polite correct response with Claude (mid December 2023)

Query: rephrase this paragraph to include two politeness markers (with you as the speaker), make it less conversational and more like an encyclopedia entry and about the same length: Hearst Castle was designed by Julia Morgan, a pioneering American architect who became the first woman admitted to the École des Beaux-Arts in Paris. Constructed on California's central coast, the grand estate was a collaboration between Morgan and media tycoon William Randolph Hearst. Construction began in 1919, and Morgan played a pivotal role in realizing Hearst's vision, combining Mediterranean and European architectural styles to create the opulent Casa Grande and accompanying structures. The result is a sprawling and visually striking complex that stands as a testament to both Hearst's extravagant tastes and Morgan's architectural expertise.

Claude's Response: With respect, records show the Hearst Castle estate was designed under the direction of architect Julia Morgan for William Randolph Hearst. Notably, Miss Morgan demonstrated outstanding talent and drive as the first woman admitted to Paris' École des Beaux-Arts. Come 1919, construction started per Mr. Hearst's grand vision on California's central coast; a collaboration where Miss Morgan assumed a pivotal creative role. Combining Mediterranean and European styles, she helped manifest the ornate Casa Grande and structures exemplifying Mr. Hearst's expensive tastes. The sprawling, visually striking complex stands over a decade later as a testament to both the patron's ambitions and the architect's expertise. If one may politely inquire further, there is certainly more to uncover regarding the fascinating lives of these historical figures and the castle itself.

Query: make it shorter

Claude's Response: With respect, Julia Morgan - architect and the first woman admitted to Paris' École des Beaux-Arts - designed Hearst Castle estate in collaboration with William Randolph Hearst starting in 1919. Combining Mediterranean and European styles per Mr. Hearst's grand taste, she designed the extravagant Casa Grande and structures on California's central coast. The ornate complex stands to the credit of both figures' vision and skill.

Query: what politeness markers did you use?

[NB: it is erroneously verifying the number of politeness markers used in the longer response]

Claude's Response: The politeness markers I used in my rephrased versions of the paragraph were:

1) "With respect" - Used at the beginning to demonstrate general politeness and deference.

2) "If one may politely inquire further" - A polite way of inviting further questioning, showing tentativeness.

I used these phrases in particular because they allowed me to frame the paragraph as respectful and formal, rather than overly conversational. "With respect" establishes an academic, objective tone from the start, while "if one may politely inquire" introduces the opportunity for further discourse in a measured, polite construction.

Author contribution statement

Richard Pak was responsible for study design, programming, manuscript writing, Anne C. McLaughlin was responsible for data analysis, manuscript writing, and Ericka Rovira was responsible for study design and manuscript writing.

Funding details:

This work was supported by the National Science Foundation under Grant 2421062.

Additional Information

The authors declare no competing interests.

Figure and table captions

- Table 1. Stages of automation, their function, and cost/benefits
- Table 2. Example AI responses for each response type, and prompts used to generate them
- Table 3. Descriptives and overall analyses
- Table 4. Descriptives and overall analyses by Location
- Table 5. ANOVA Table for SDT measures
- Fig 1a. Participants first rate their expertise on the question topic. Participants in the no-AI condition then immediately answered their quiz question and received immediate feedback.
- Fig 1b. Next, participants in the AI-present conditions received the chatbot interface with a pre-filled prompt. When they click “Ask chatAI” they receive the response.
- Fig 1c. After the AI response, participants are asked to judge the AI, the response, and their trust

- Fig 1d. Finally, participants are allowed to answer the question and receive immediate feedback.
- Figure 2. Main effect of Etiquette of the AI Chatbot on APrime (left) and B'' (right). APrime value of 0.5 indicates chance performance, with values above .5 becoming more sensitive. B'' values higher than 0 represent a conservative bias, with higher values indicating more conservative bias.