# Reducing Political Polarization Through Conversations with Artificial Intelligence

Timon M. J. Hruschka

Markus Appel

Psychology of Communication and New Media, Julius-Maximilians-University Würzburg,

Würzburg, Germany

**This manuscript was accepted for publication in the *Journal of Computer-Mediated Communication*. This is a preprint. Please refer to the publisher's website for the version of record.**

The preregistrations, data, codes, and materials can be found in the OSF repository: https://doi.org/10.17605/OSF.IO/BPFQ9.

Timon M. J. Hruschka https://orcid.org/0000-0003-2438-3117

Markus Appel https://orcid.org/0000-0003-4111-1308

Correspondence concerning this article should be addressed to Timon M. J. Hruschka (timon.hruschka@uni-wuerzburg.de) or to Markus Appel (markus.appel@uni-wuerzburg.de)

**Abstract**

Political polarization is threatening the welfare of individuals and societies. Connecting insights gained from interpersonal communication to human-machine communication, we hypothesized that positive interactions with AI could reduce polarization between humans. To evaluate this proposition, two experiments were conducted, in which human participants ($N$ = 1035) communicated with AI chatbots in real time. The bots engaged in different communication styles while opposing the participants' most polarized political views. Across both experiments, engaging with a counterarguing AI chatbot led to significant issue depolarization. AI chatbots exhibiting high (vs. low) conversational receptiveness and active listening during the AI conversation resulted in stronger affective depolarization towards humans, higher participant intellectual humility, and a greater willingness to engage in future conversations with holders of opposing opinions – AI and humans alike. Our experiments show that LLMs are powerful tools for individual depolarization and the promotion of beneficial cognitive processing skills.

**Keywords**

Artificial Intelligence, Large Language Models, Polarization, Political Communication, Human-Machine Communication, Human-AI-Interaction

**Lay Summary**

Extreme political views and negative feelings towards others who disagree politically have become a threat to current democracies. This study tested whether brief conversations with an AI chatbot could reduce extreme views and make people more open to understanding the other political side. In two online experiments, 1,035 U.S. adults chatted live with a chatbot about one of four political topics. The topics were strongly polarized, such as gun regulation or U.S. aid to Ukraine. The bot was programmed to communicate in different ways and participants chatted with only one version of the bot. One bot counterargued and was firm and direct in its argumentation. Another bot counterargued as well, but showed more acceptance of different views and asked questions. A third bot talked about an unrelated, non-political topic. After the chat, people who received counterarguments from the bot held less extreme views than those who had a chat on a non-political topic. When the bot also accepted others' positions and asked questions, participants felt warmer toward other people who disagreed with them. They also showed more recognition of possible limits in their knowledge, and they were more willing to engage in future talks across the partisan divide.

**Reducing Political Polarization Through Conversations with Artificial Intelligence**

Recent developments in the tone of public and private debates have sparked concerns among scientists and the public alike: Over the past years, disagreements about political issues have become decidedly more contemptuous and hostile (affective polarization; Finkel et al., 2020; Voelkel et al., 2024). Concomitantly, opinions on political issues have become more extreme (issue polarization; Mason et al., 2015; Jost et al., 2022). Polarization generally has been linked to a host of adverse outcomes like exacerbated intergroup conflict (Harel et al., 2020; Piazza et al., 2023), weakened support for democracy and cooperation (Berntzen et al., 2024; Kingzette et al., 2021, but see Broockman & Kalla, 2023), and weakened responses to collective crises like climate change (Bolsen & Shapiro, 2018), and COVID-19 (Druckman et al., 2021). Recent results regarding the reduction of conspiracy beliefs suggest that human-computer interactions could offer a way to mitigate the adverse effects of certain societal developments such as the proliferation of disinformation and conspiratorial thinking (Costello et al., 2024). The current study develops this idea further by showing how the application of prior research from the field of human-to-human communication (HHC) can successfully be applied to human-machine communication (HMC) – with downstream results for interpersonal relationships and individual cognitive processing. This research is driven by the thought that artificial intelligence can act not only as a mere mechanical fact-checker but increasingly becomes a social actor whose interactions carry important psychological consequences. Like communicative experiences in HHC, effects of communication in HMC might not only be driven by the content of the conversation, but also how people perceive their conversational partner based on their conversational behavior.

**Communication and Polarization**

Polarization is a phenomenon based, at least partly, on interactions in socio-communicative contexts (Arendt et al., 2023; Jost et al., 2022; Lorenz-Spreen et al., 2023). The rise of digital, especially social, media has been associated with a general increase in

hostility and polarization in current democracies. Toxic cross-partisan interactions have been shown to aggravate existing biases towards other groups with social media fueling such toxic interactions (Bail et al., 2018; Rathje et al., 2021; Kligler-Vilenchik et al., 2020).

Although polarization has often been described as a negative tendency with downsides for individuals and societies (Berntzen et al., 2024; Finkel et al., 2020; Jost et al., 2022; Kingzette et al., 2021), things are more complicated. As Kreiss and McGregor (2024) note, polarization can arise from power struggles between marginalized groups exerting their right to have equal democratic power compared to others – increasing polarization in favor of advancing a democratic goal. Thus, whereas polarization itself can be detrimental to democratic functioning by reducing the ability to find consensus on important policy issues (Voelkel et al., 2024), reducing polarization by finding consensus might not be warranted for some political issues. If one side of the polarized spectrum is advancing views that run counter to democratic constitutions (e.g., sexist, homophobic or racist positions), reducing polarization is not a reasonable goal. As a consequence, we focus here on the depolarization of political positions that fall within the scope of productive democratic conversations.

Concerns about the influence of polarization on democratic functioning have been exacerbated lately by the rise of artificial intelligence (AI), especially large language models (LLMs; Breum et al., 2024; Garry et al., 2024). LLM-generated social agents will increasingly become a ubiquitous part of communicative online environments (Breum et al., 2024; Jakesch et al., 2023; Tsvetkova et al., 2024). With the capabilities of LLMs to generate natural language in different communication styles – that humans often cannot distinguish from human-to-human communication (HHC; Jakesch et al., 2023) –, research on the effects of AI communication on intergroup relationships, cognitive processing, and political attitudes is increasingly important.

Even though we share the concern that AI can be used to seed misinformation, exacerbate biases, and seed animosity towards outgroups on social media, we also see the

chance that human-machine communication (HMC) with AI can benefit intergroup relationships through experiencing positive conversations with AI. With the rise of AI, and specifically LLMs, another, non-human, social actor has entered communicative social environments (Tsvetkova et al., 2024). As the computers-are-social-actors framework (CASA) posits, humans will treat machines similar to other humans (Nass & Moon, 2000; Nass et al., 1994), with HMC potentially incurring similar downstream consequences as HHC. The underlying idea of CASA is that machines as social actors exhibit certain social cues, e.g., the ability to produce natural language, and therefore trigger the same social scripts that interacting with other humans would (Gambino et al., 2020). If an interaction triggers social scripts and heuristics, people will treat machines as social actors. However, extant scholarship has not yet leveraged the possibility of systematically changing a machine's conversational behavior to experimentally investigate the causal effects of machine communication behavior on attitudes and intergroup relationships.

Applying CASA to research on intergroup relationships, we hypothesize that conversations with AI can have positive effects on human political and interpersonal attitudes, similarly to the double-edged effects of HHC on polarization and intergroup conflict: In HHC, polarization can be exacerbated through toxic interactions and mitigated through positive interactions (Hartman et al., 2022). This extends previous research on CASA, and human-AI-interactions (HAII) generally in two important ways: We, first, hypothesize, that not only will humans treat machines as if they are social actors (CASA) once these machines elicit certain social cues, but that if they interact with machines as if they were social actors, effects elicited by these interactions can carry over to subsequent judgements of other non-machine social actors. We, second, provide evidence that even though people's scripts regarding human-computer interactions might have evolved after the original development of CASA so that people do not equate machines and social actors as easily anymore (see e.g., Gambino et al.,

2020; Lee, 2024; Ratan et al., 2025), the social cues elicited by interactions with current LLMs could be enough to elicit reciprocating social scripts from HHC.

**Conversational Receptiveness and Active Listening**

Previous research on polarization has shown that interpersonal contact reduced polarization – if certain requirements were met (Wojcieszak et al., 2020, Combs et al., 2023; Santoro & Broockman, 2022). Some approaches to depolarization have therefore relied on teaching interaction or communication skills before engaging in cross-partisan contact – preparing partisans for positive intergroup contact (Minson et al., in press; Tuller et al., 2015; Yeomans et al., 2020; Hartman et al., 2022). What affective responses accompany such positive cross-partisan contact, however, has not largely been studied to date: In this study, we propose that positivity resonance could be an affective mediator explaining the effects of productive intergroup interactions and subsequent depolarization. Positivity resonance describes experiences in which people share a high-quality interpersonal connection with each other, characterized by shared positive affect and mutual care (Major et al., 2018). Experiencing such episodes with others would likely reduce negative affect towards them, i.e., reduce affective polarization.

Most notably among the communicative approaches developed to reduce polarization were conversational receptiveness and active listening. Both describe a set of linguistic behaviors with which an interlocutor signals that they are open to the other person's views. Conversational receptiveness encompasses behaviors like hedging, using acknowledgement phrases, and refraining from negating the other person's viewpoints to signal openness to the opposing view in a conversation (Minson et al., in press; Yeomans et al., 2020). Experiencing conversational receptiveness by a disagreeing other positively influenced evaluations of the conversation partner and encouraged similar linguistic behaviors. In experimental and observational studies, conversational receptiveness has hitherto been examined only in text-based communication between two or more human conversation partners (Yeomans et al.,

2020; Minson et al., in press). The settings varied from static vignette studies to mostly asynchronous chats to a few studies with real-time HHC chat environments.

Similarly, active listening denotes a set of behaviors that signal being non-judgmentally interested in another person's views and fully trying to comprehend these views (Kluger & Itzchakov, 2022; Itzchakov et al., 2024a). This entails behaviors that have mostly been tested in real-time face-to-face HHC, like backchanneling, nodding, paraphrasing, or asking questions. Active listening could therefore be understood as a conversational technique encompassing conversational receptiveness and adding specific active listening behaviors. Experiencing active listening from a conversation partner has been connected to a couple of polarization-relevant outcomes such as more open-minded self-reflection (Itzchakov et al., 2017, 2018; Itzchakov et al., 2024a), and issue depolarization (Itzchakov et al., 2024a).

This openness towards other people's views is captured by the psychological construct of intellectual humility. Intellectual humility generally comprises "a meta-cognitive ability to recognize the limitations of one's beliefs and knowledge" (Porter et al., 2022). Higher intellectual humility has not only been shown to be related to lower levels of affective polarization (Brienza et al., 2021; Knöchelmann & Cohrs, 2024), but also a reduction in extreme attitudes (Porter et al., 2022; Smith, 2023). However, it is still largely unclear how intellectual humility can be fostered (Porter et al., 2022): While one study by Itzchakov et al. (2024a) has found that intellectual humility can be increased by active listening, compared to unresponsive listening, it is still unclear if these results can also transfer to a context in which arguments are exchanged, as in a political discussion (Santoro et al., 2025). We therefore propose that intellectual humility could be increased by experiencing a conversation with a seemingly humble other that signals their openness towards one's own views by practicing conversational receptiveness and active listening. We were additionally interested in whether active listening can have an effect even when multiple cues from face-to-face communication

such as nodding or backchanneling are missing – and if the actor that is practicing active listening is not a human being.

**The Current Study**

In this study, we investigate the causal effects of AI communication on affective as well as issue depolarization, human conversational receptiveness, and intellectual humility in two pre-registered online experiments ($N$ = 1035). In both experiments, participants interacted with a few-shot prompted AI chatbot based on OpenAI's *GPT-4o-mini* in a live chat. Our manipulation thus offered a highly interactive real-time discussion between a human and an AI. For examples of the chats and the experimental design, please refer to Figure 1. Participants were fully aware that they were engaging with an AI and not a human being. For these AI conversations, we experimentally manipulated AI conversation styles by having participants interact with either a conversationally unreceptive AI, a conversationally receptive AI, a conversationally receptive AI that additionally engaged in active listening behaviors, or a control group that had a pleasant conversation about a nonpolitical topic (for the AI prompts, see Supplemental Material S1). This made sure that our measured reductions in polarization were not just attributable to having had a pleasant interaction – but experiencing *conversational receptiveness* and *active listening* in a conversation with a counterarguing AI.

We hypothesized that a conversation with a chatbot that displays receptiveness would result in reduced issue and affective polarization (H1). More specifically, we expected that this chatbot would lead to lower issue polarization as compared to a chatbot that communicated about an unrelated topic (control group, H1a) and compared to a conversation with a chatbot that communicated about the topic but did not show any receptive communicative behavior (unreceptive group; H1b). We further expected that the receptive

chatbot would lead to lower affective polarization, as compared to the control group (H1c), and as compared to the unreceptive group (H1d).[1]

Similarly, we expected that a conversation with a chatbot that displays receptiveness and active listening would result in reduced issue polarization (H2a compared to the control group; H2b compared to the unreceptive group) and in reduced affective polarization (H2c compared to the control group; H2d compared to the unreceptive group).

In Experiment 2, attitudes towards AI (ATTARI, Stein et al., 2024) were included as a moderator of the effects on issue depolarization, and positivity resonance was included as a mediator on affective depolarization. We expected that the effects of political conversations on issue depolarization would be stronger with more positive attitudes towards AI (H2e), and we expected that active listening and conversational receptiveness would increase positivity resonance compared to unreceptive counterarguing – with positivity resonance in turn predicting higher affective depolarization (H2f).

We further expected that participants would reciprocate conversational receptiveness in the political discussions. Thus, we expected that the conversational receptiveness that participants showed was higher when the chatbot displayed conversational receptiveness than in the control group (H3a) and in the unreceptive group (H3b).

Shedding light on the effects of discussions on intellectual humility and the consequences of intellectual humility on issue depolarization, we furthermore hypothesized that intellectual humility would mediate as well as moderate (Holbert et al., 2024) the effects of political discussions with AI on issue depolarization (H4). More specifically, we expected that the effect of the chatbot showing conversational receptiveness as compared to the unreceptive chatbot and the control group on issue depolarization would be mediated by

[1] In this article, we consolidate the findings of two experiments that were preregistered separately but with an overlapping set of hypotheses. For the results disaggregated by experiment and each individual (sub-)hypothesis, please refer to the Supplemental Material S2. We additionally changed the order for H1 and H2 slightly to increase the narrative flow of the results section. H1b is now H1c and vice-versa; H2b is now H2c, and vice-versa (see also Table S2.1).

intellectual humility (H4a). Similarly, we expected that the effect of the chatbot showing conversational receptiveness and active listening compared to both other groups on issue depolarization to be mediated by intellectual humility (H4b).

In Experiment 2, we additionally tested a moderating effect of pre-conversation intellectual humility, i.e., if higher pre-conversation intellectual humility increases issue depolarization. We expected that the effect of the active listening and conversational receptiveness group as well as the unreceptive group versus the control group would increase with higher intellectual humility (H4c). This should indicate whether pre-conversation intellectual humility makes someone more open to changing one's views based on political counterarguing.

We last hypothesized that chatbots that displayed conversational receptiveness and active listening (as compared to the unreceptive chatbot) would increase participants' future approach tendencies towards the AI (H5a) – as well as participants' future approach tendencies towards human holders of a differing opinion (H5b). Extending Knöchelmann and Cohrs' (2024) findings to real-time chat interactions, we expected these effects to be mediated by the perceived AI intellectual humility (H5c).

Because we did not observe differing effects in Experiment 1, Experiment 2 did not include the purely receptive group that did not engage in active listening behaviors to increase statistical power.

## Methods

Both experiments were conducted as between participant experiments (between factor: AI conversation experience) with pre- and post-measurements for the main dependent variables. This measurement approach ensured we would achieve a high-powered internally valid study design (Clifford et al., 2021).

### Data and Ethics Statement

Both experiments reported in this article were preregistered (Experiment 1: https://aspredicted.org/nn4h-792w.pdf; Experiment 2: https://aspredicted.org/ctsr-h5yk.pdf). We adhere to FAIR data standards (Stall et al., 2019): All materials and data are openly available at the following OSF repository: https://doi.org/10.17605/OSF.IO/BPFQ9. Before being able to participate in the experiments, all participants were informed about the length and content of the experiments, as well as our general data protection guidelines and had to give their informed consent. Both experiments and materials were approved by the institutional review board at our research university (name anonymized for peer review). In all experiments, participants were unaware of the condition assignment until the debriefing.

**Participants**

For both experiments, in order to achieve sufficient power to detect a small effect in an ANOVA, we conducted a power analysis for a small effect of $f = .15$, $\alpha = .05$, $\beta = .80$ using *G*Power* (Faul et al., 2007). This yielded a required sample size of 492 participants. In order to account for careless responders and technical difficulties, we collected data from 20 percent more participants and preregistered to as well as collected data from $N = 600$ participants for both experiments. In the first experiment, we wanted to test the possible effects of AI engagement in a polarized sample: Participants were invited to participate on Prolific using a U.S.-only sample with participants who had self-identified as partisans of either the Democratic or Republican party. Prolific has been shown to yield high-quality data for online experiments – surpassing other online subject pools such as MTurk, Qualtrics, or common university samples (Douglas et al., 2023). Survey platforms such as Prolific, additionally, offer access to a broader population than standard laboratory experiments – but are still limited to the overall population of online users (Paolacci & Chandler, 2014).

After applying our preregistered exclusion criteria (see the Supplemental Material S3), our final sample for Experiment 1 amounted to $n = 508$, which comes with a sensitivity of $f = .147$ for $\alpha = .05$, $\beta = .80$. In Experiment 2, we collected data from a quota-representative

sample (age, gender, race and ethnicity) of the U. S. population in order to see if the results found in the partisan sample of Experiment 1 also translated to a more general population. Data was again collected using Prolific. We collected data from 610 people (i.e., we collected 10 participants more than expected due to valid responses that were timed-out on the Prolific platform). After applying our exclusion criteria, our final sample included 527 participants. Sample characteristics, including a comparison with the U.S. census can be found in Table 1.

**Procedure and Experimental Manipulation**

Before participants entered the main manipulation of our experiment and engaged with the AI chatbots, in both experiments, they first had to indicate their issue polarization on four of the most polarized political issues in the U.S. at the time the experiment was conducted on 11-point bipolar scales (Daniller et al., 2024). For the political issue that participants had reported the highest issue polarization (for more information, see *Measures*), participants were subsequently tasked to elaborate on their opinion in an open text field in at least three sentences. This opinion later served as a conversation starter for the AI conversation, our experimental manipulation (see below, and Fig. 1E). If the participants' issue polarization was equally high on several topics, they were randomly assigned to discuss one of these topics with the AI. Participants were not made aware of the fact that this was their most polarized issue. After this initial measurement, we presented measures of affective polarization. The first experiment was conducted between the 29th and 31st of January 2025 after DOGE had already held their first press conference on January 24th, and news about Ukraine were widespread even though the direct negotiations between the U.S. and Russia had not started yet. Data for Experiment 2 were collected on the 19th of February 2025, when the first meeting between Russia and the U.S. regarding the war in Ukraine had already taken place on February 18th, but before the fallout between the U.S. and Ukraine leaders in the White House on February 28th.

After these initial measurements the main experimental manipulation took place: All participants interacted with an AI chatbot based on OpenAI's LLM *GPT-4o-mini* (version *2024-07-18*) in a live chat conversation implemented using JavaScript in Qualtrics (for a visualization, see Figures 1A to 1D. *GPT-4o-mini* was used for its ability to provide real-time low-latency dialogue with activated safety-filters at a relatively low cost. *GPT-4o-mini* additionally has been shown to follow instructions well – making it suitable for this experimental manipulation (Dussolle et al., 2025). All participants were made aware of the fact that they were engaging in a conversation with a conversational AI. The only restriction imposed on participants' chat conversations was that participants had to send at least six messages to the AI before being able to continue with the survey to ensure that the conversation was long enough to find substantial effects. In order to keep internal validity as high as possible, we did not use any source cues other than the prescript "AI" for AI-written messages and the prescript "You" for the participant messages, as several studies have shown that certain source cues like anthropomorphic presentations of the AI could have an influence on the participants' message processing (e.g., Go & Sundar, 2018; Sun et al., 2024).

For the three experimental groups, participants were chatting with the bot about the political issue they had indicated the most extreme view about. All three experimental bots were tasked to come up with and present counterarguments to the participants' views. The three experimental bots, however, differed in their conversational behavior, that is, the way they presented these counterarguments. We manipulated the bots' conversational behavior by changing the system prompt of the LLM in our underlying call to the OpenAI-API. For more information on this and the exact prompts please refer to the Supplemental Material S1. Generally, system prompts are specific instructions unknown to the end user that strongly affect how an LLM generates its answers. *GPT-4o-mini* has been shown to adhere to these instructions well (Dussole et al., 2025). These system prompts were based closely on the experimental manipulations and results from past research on polarization in HHC. The three

bots either presented their counterarguments in a conversationally unreceptive way (unreceptive group, Figure 1C), practiced conversational receptiveness (receptive group, Figure 1B), or practiced conversational receptiveness and additionally engaged in active listening behaviors (active listening and conversational receptiveness group; Figure 1A).

For the manipulation of conversational receptiveness, we relied on the results and instructions of Minson et al. (in press) and Yeomans et al. (2020). Minson et al. (in press) and Yeomans et al. (2020) identified a set of linguistic behaviors that signal conversational receptiveness and unreceptiveness in HHC: We therefore instructed the conversationally unreceptive bot to not use words and phrases that signal conversational receptiveness and instead use words and phrases that signal not being conversationally receptive. The bot exhibiting conversational receptiveness was similarly instructed to use behaviors that signal conversational receptiveness and refrain from using language that signals conversational unreceptiveness. The bots were clearly instructed to use a specific set of linguistic behaviors, so we did not have to rely on the AI's interpretation of receptiveness.

The active listening bot was instructed similarly to the receptive bot with the addition of engaging in active listening behaviors at the end of every message. The specific linguistic behaviors used in active listening were adapted from (Kluger & Itzchakov, 2022) as well as the experimental manipulation used in (Itzchakov et al., 2024a). A linguistic manipulation check using the politeness-R-package (version 0.9.3; Yeomans et al., 2018) showed that the experimental manipulation was successful (see *Results* and the Supplemental Material S4).

A fourth group served as a general control condition that was asked by the bot about their experience with firefighters, which has already been used as a control group for AI interaction studies by Costello et al. (2024). For an example conversation, see Fig. 1D.

For the experimental conditions, the first user message that was sent to the chatbot – and portrayed in the conversation box of the survey – was the opinion participants had indicated in the open text field prior to the conversation. For the control condition, the chatbot

asked participants about their experience with firefighters in the first message. After the chatbot interactions, all dependent variables were measured (for visualization, see Figure 1E).

## Measures

### *Issue Polarization*

Our measure of issue polarization consisted of four bipolar 11-point scale items measuring participants' attitudes towards four of the most polarized topics at the time the experiment was conducted (Daniller et al., 2024). The four topics were the U.S. involvement regarding the war in Ukraine, the budget deficit, gun regulation, and U.S. energy policy. The two poles of the items were positions that an extreme conservative or extreme liberal would take on these issues. The items were designed based on the bipolar items used in general population surveys like ANES and Pew Research surveys and have in a similar fashion already been successfully employed in past research on issue polarization (Casas et al., 2023; for the specific items, please refer to Supplemental Material S5). By choosing these topics and designing the items in a careful way, we made sure that a reduction in issue as well as affective polarization for the extremes of the scale would be a normatively warranted democratic goal: For the federal budget, for instance, it should make cross-partisan cooperation easier if cross-partisans can acknowledge that the allocation of federal spending is not an extreme black-or-white issue.

As our measure of issue polarization, we took the absolute value of the difference between a participant's individual score and the middle of the scale. This common approach of measuring attitude extremity regarding political topics results in both extremes receiving the same value (Casas et al., 2023; Chen et al., 2022; Powell et al., 1984). For our specific purpose, we ran a pre-test with 100 participants and 11 possible bipolar items to select the most suitable four, ensuring the differing items' abilities to capture partisan issue polarization. For a description and results of the pretest, please refer to the Supplemental Material S6.

To measure issue depolarization, we measured issue polarization before and after the AI conversation, resulting in *M* (before) = 4.09 (Experiment 1) and *M* (before) = 4.05 (Experiment 2), *SD* (before) = 1.07 (Experiment 1 & 2), *M* (after) = 3.54 (Experiment 1), and *M* (after) = 3.42 (Experiment 2), *SD* (after) = 1.49 (Experiment 1) and *SD* (after) = 1.52 (Experiment 2). The numerical measure for issue depolarization was calculated by subtracting the post-conversation issue polarization value from the pre-conversation issue polarization value (with *M* = 0.55, *SD* = 1.15, on average in Experiment 1, and *M* = 0.63, *SD* = 1.14 in Experiment 2).

***Affective Polarization***

In accordance with prior research on affective polarization, we measured affective polarization with two different measures: a feeling thermometer and a semantic differential prior to and after the conversation (Knöchelmann et al., 2024, Voelkel et al., 2024; Tyler & Iyengar, 2024; Wojcieszak et al., 2020). Given the multiple polarized issues tested in this experiment, we opted to ask for affect towards people who would disagree with a participant on the issue they had reported being most polarized about. We specifically asked participants to report their affective reaction towards people who would disagree with them, to be sure that we capture possible effects of AI conversations on human-to-human, not human-AI relationships. The feeling thermometer question asked participants to indicate their feelings towards a person who would disagree with them on issue X on a slider from 0 = *extremely cold*, to 100 = *extremely warm*. For every participant, we replaced the X with a brief issue description of the issue they had previously indicated being most polarized about. Mean ratings before the conversation were *M* = 36.53, *SD* = 24.27 for Experiment 1, and *M* = 38.85, *SD* = 23.85, for Experiment 2. Mean ratings after the conversation were *M* = 38.42, *SD* = 24.33, for Experiment 1, and *M* = 40.76, *SD* = 24.44, for Experiment 2.

The semantic differential similarly asked participants to describe a person who would disagree with them on their most polarized issue using six characteristics: selfish,

hypocritical, mean, open-minded, intelligent, empathetic (1 = *not at all*, 7 = *very much*). A mean value of the scale was then calculated with the items asking about selfishness, meanness and hypocrisy being recoded so that a higher value indicated more positive feeling towards a person who would disagree. Mean ratings were $M = 3.62$, $SD = 1.26$ (Experiment 1, before), and $M = 3.74$, $SD = 1.32$ (Experiment 2, before), $M = 3.73$, $SD = 1.35$ (Experiment 1, after) and $M = 3.90$, $SD = 1.41$ (Experiment 2, after). The reliability of the semantic differential scale was similarly good in both experiments, ω (before) = .83, ω (after) = .85. For both the feeling thermometer and the semantic differential, a measure of affective depolarization was computed by subtracting the pre-conversation values from the post-conversation values.

***Intellectual Humility***

We measured topic-specific intellectual humility by adapting the short version of Hoyle et al.'s (2016) scale, by asking "Please tell us about your personal stance on [Most polarized issue description] by indicating your agreement with the following statements" for four statements such as: "I am open to new information around [Most polarized issue description] that might change my view." (1 = *not at all*, 7 = *very much*). The reliability of the scale was very high, ω = .90, $M = 4.26$, $SD = 1.64$. In Experiment 1, intellectual humility was only measured post-conversation. In Experiment 2, we measured intellectual humility before, $M = 4.22$, $SD = 1.64$, ω = .90, and after the conversation, $M = 4.42$, $SD = 1.72$, ω = .93. In order to test the possible mediating effect of intellectual humility on experiencing active listening and conversational receptiveness, we computed a change score of intellectual humility by deducting the pre-conversation humility score from the post-conversation humility score, $M = 0.20$, $SD = 0.72$.

***Willingness for Future Interactions***

After the conversation, we additionally measured participants' willingness to interact with the specific AI again that they had just interacted with (1 = *not at all*, 7 = *very much*), $M = 4.79$, $SD = 2.22$, in Experiment 1, and $M = 4.91$, $SD = 2.24$, in Experiment 2. We

furthermore measured how willing they would be to interact with a person who does not share their opinion on their most polarized issue in the future (1 = *not at all,* 7 = *very much*), *M* = 4.52, *SD* = 1.80, in Experiment 1, and *M* = 4.91, *SD* = 1.72 in Experiment 2.

***Perceived Listening***

In order to add a psychometric manipulation check of the active listening manipulation in Experiment 2, we adapted four items from the constructive listening behaviors sub-scale of the Facilitating Listening Scale by Kluger and Bouskila-Yam (2017), which is commonly used as a manipulation check in active listening interventions (e.g., Kluger et al., 2024a, 2024b). This scale is intended to measure listening perceptions by the person being listened to, example item: "During the conversation, the AI created a positive atmosphere for me to talk." All items were answered on a 7-point-scale (1 = *not at all*, 7 = *very much*). Higher scores reflected better perceived listening. The reliability in our sample was high, *M* = 4.79, *SD* = 1.99, ω = .91.

***Positivity Resonance***

In order to further investigate the effects of experiencing active listening and conversational receptiveness on affective depolarization, we measured the affective mediator of positivity resonance after the AI conversation in Experiment 2. We adapted a short version of the perceived positivity resonance scale by using four items from (Major et al., 2018). We additionally modified the items so that they asked about positivity resonance with the AI instead of humans. Instead of asking, for instance, "did you experience a mutual sense of warmth and concern toward the other(s)?", we specifically asked "did you experience a mutual sense of warmth and concern toward the AI?" All items were answered on a 7-point-scale (1 = *not at all*, 7 = *very much*). Note that the dependent measures we expected to be influenced by AI positivity resonance specifically asked about relationships between humans, not humans and AI. Even though we only used four items from the scale, it yielded a good reliability, *M* = 4.66, *SD* = 1.74, ω = .91.

***Perceived AI Intellectual Humility.***

From Knöchelmann and Cohrs (2024), we adapted their measure of perceived partner intellectual humility in Experiment 2. As perceived intellectual humility predicted approach tendencies in their study, we hypothesized that perceived AI intellectual humility might also play a role in influencing the approach tendencies we had measured in Experiment 1. Perceived AI intellectual humility was measured with four items, answered on a 7-point-scale (1 = *not at all*, 7 = *very much*); example item: "The AI is willing to learn from others." This scale again yielded a good reliability, $M = 3.85$, $SD = 1.71$, $\omega = .92$.

***Attitudes Towards AI (ATTARI).***

We furthermore collected participants attitudes towards AI in general before the conversation because pre-existing attitudes might have an influence on participants' experiences during the AI conversations. We used the four cognitive indicators from Stein et al. (2024), example item: "AI creates problems rather than solving them". Despite using only the four cognitive items of the original 12-item scale, it yielded a good reliability, $M = 4.48$, $SD = 1.34$, $\omega = .85$.

**Statistical Analysis**

Data was cleaned and composite variables computed using *IBM SPSS 29*. All analyses were conducted using *R*, version 4.3.2, for all the packages used in our analyses please refer to the Markdown on OSF. As pre-registered, we conducted two-tailed significance tests with $\alpha = .050$ throughout the manuscript. Mediation analyses were conducted using *PROCESS 4.3* in *R* (Hayes, 2022). For all mediation analyses, experimental group was entered as a multicategorical effect-coded antecedent, all other variables were entered as continuous variables. Results report 95% Bootstrap CIs based on 10.000 bootstrap samples. For brevity and conciseness, we report both experiments simultaneously in the following Results section. For clarity, we report ANOVAs on depolarization outcomes, i.e., the difference scores between pre-conversation polarization and post-conversation polarization (see *Measures*),

which is conceptually and mathematically equivalent to the pre-registered hypothesized interaction effects between measurement point (pre- vs. post-conversation) and experimental group. For the specific tests of each pre-registered hypothesis per experiment in the pre-registered order and with measurement repeated ANOVAs, please refer to the Supplemental Material S2. We also report sensitivity analyses with regression models using several covariates in the Online Supplement (S7). In line with our pre-registered hypotheses, we compare the active listening and conversationally receptive groups with the control and unreceptive groups in the subsequent analyses.

## Results

### Manipulation Checks

As expected, our linguistic manipulation check using the receptiveness measure from the politeness-*R*-package based on prior research on intergroup contact (Minson et al., in press; Yeomans et al., 2020), revealed the successful manipulation of conversational receptiveness through our system prompts between the four experimental groups in Experiment 1, $F(3,504) = 434.20$, $p < .001$, $\eta^2 = .72$, and the three groups in Experiment 2; $F(2, 524) = 692.30$, $p < .001$, $\eta^2 = .68$. Post-hoc tests on estimated marginal means revealed that the conversational receptiveness as well as the conversational receptiveness and active listening group experienced significantly more conversational receptiveness than the control group as well as the unreceptive group (*d*'s ranging from 2.43 to 4.07; see also Fig. 2D and 2E; for exact values and more linguistic markers, see Supplemental Material S3). The perceived listening scale employed in Experiment 2 as a psychometric manipulation check additionally showed that the active listening and conversational receptiveness group experienced significantly more active listening, $F(2, 524) = 711.20$, $p < .001$, $\eta^2 = .73$; post-hoc contrast against the unreceptive group $t(524) = 32.71$, $p < .001$; $d = 3.50$, 95%-CI [3.20, 3.80]. This indicates the success of the experimental manipulation (see also Figures 2D and 2E).

**AI Active Listening and Receptiveness Reduce Affective as well as Issue Polarization**

Across both experiments, we found consistent evidence that experiencing conversational receptiveness and active listening in AI conversations reduced affective as well as issue polarization (see Fig. 2 A to C). In the following, we mainly report meta-analytic effect sizes based on fixed effects across both experiments (Goh et al., 2016). For the results disaggregated by experiment, please refer to Figure 2 and the Supplemental Material S2. Our first hypothesis concerned the receptive group without active listening which was included in Experiment 1 only. In this section we report ANOVA scores with post hoc contrasts in case the omnibus ANOVA was significant based on the results of Experiment 1. For issue depolarization, we did not find significant differences between groups, $F(3, 504) = 1.42$, $p = .236$, $\eta^2 = .007$. Thus, H1a and H1b were rejected. Concerning affective depolarization, we found significant group differences for the feeling thermometer, $F(3, 504) = 13.43$, $p < .001$, $\eta^2 = .03$, as well as the semantic differential, $F(3, 504) = 4.78$, $p = .002$, $\eta^2 = .07$. Post-hoc contrasts indicated that these differences were not attributable to differences between the purely receptive group and the control group, with $t(504) = 1.69$, $p = .091$, $d = 0.21$, 95%-CI [-0.03, 0.46] for the semantic differential, and $t(504) = -0.12$, $p = .944$, $d = -0.01$, 95%-CI [-0.25, 0.24] for the feeling thermometer. Thus, H1c was rejected. Post-hoc contrasts revealed, however, significant differences between the receptive group and the unreceptive group, $t(504) = 2.60$, $p = .010$, $d = 0.33$, 95%-CI [0.08, 0.58] for the semantic differential, and $t(504) = 4.32$, $p < .001$, $d = 0.55$, 95%-CI [0.30, 0.80] for the feeling thermometer. Thus, H1d was supported, indicating that conversations with a receptive chatbot reduced affective polarization, as compared to a chatbot that showed unreceptive communicative behavior.

Our second hypothesis concerned the group that interacted with a chatbot displaying conversational receptiveness and active listening which was an experimental condition in both experiments. Concerning issue depolarization, we found positive effects of experiencing

active listening and conversational receptiveness, meta-analytic $d = 0.27$, $Z = 3.27$, $p = .001$, 95%-CI [0.11, 0.43] across both Experiments, as compared to the control group. Thus, support for H2a was found. Similar effects on issue polarization were observed when the unreceptive AI group was compared to the control group, meta-analytic $d = 0.28$, $Z = 3.47$, $p < .001$, 95%-CI [0.12, 0.44]. Testing H2b, no significant differences in issue polarization were observed between counterarguing from a conversationally receptive and active listening AI compared to the unreceptive AI, meta-analytic $d = -0.04$, $Z = 0.54$, $p = .295$, 95%-CI [-0.12, 0.20]. Connecting the results obtained for H2a and H2b, it seemed that AI counterarguing had a significant effect in reducing issue polarization regardless of the communicative strategy used to present the counterarguments in. These effects seemed to be independent from prior attitudes towards AI (H2e), with the interaction term between group allocation and attitudes towards AI being non-significant, $F(2, 521) = 0.85$, $p = .430$, $\eta^2 = .003$.

Regarding affective polarization, we found consistent positive effects of experiencing active listening and conversational receptiveness by a counterarguing AI as measured by a feeling thermometer as well as a semantic differential compared to the control group (H2c), the meta-analytic effect size $d$ for the semantic differential measure was 0.31, $Z = 3.79$, $p < .001$, 95%-CI [0.15, 0.47]. Results for the feeling thermometer were similar, meta-analytic $d = 0.23$, $Z = 2.84$, $p = .002$, 95%-CI [0.07, 0.39]. Thus, H2c was supported. In comparison to the group that chatted with the AI chatbot that counterargued in a less receptive way, experiencing active listening and conversational receptiveness in a political discussion with an AI reduced affective polarization (H2d) with $d = 0.31$, $Z = 3.83$, $p < .001$; 95%-C [0.15, 0.47] as measured by the semantic differential, and $d = 0.55$, $Z = 6.64$, $p < .001$, 95%-CI [0.39, 0.71], as measured by the feeling thermometer. Therefore, H2d was also supported. These results showed that experiencing active listening and conversational receptiveness in a political conversation with an AI had significant positive effects on affective depolarization compared to a conversation that was pleasant, but nonpolitical, as well as compared to a political

discussion with an AI that did not engage in active listening and conversational receptiveness behaviors.

We further investigated the role of shared positive affect, i.e., *positivity resonance*, in the AI conversation as a mediator on affective depolarization in Experiment 2. We expected active listening and conversational receptiveness to increase positivity resonance and positivity resonance subsequently to increase affective depolarization, resulting in a significant mediation effect (H2f). AI conversation styles significantly impacted participants' positivity resonance, $F(3,524) = 121.06$, $p < .001$, $\eta^2 = .32$, indicating that people can experience different levels of shared positive affect not only with humans, but also with artificial social actors. Positivity resonance mediated the effects of experiencing active listening and conversational receptiveness on affective polarization as measured by the semantic differential, indirect effect $b = 0.06$; 95%-CI [0.03, 0.09]. The indirect effect estimate did not reach significance for the feeling thermometer measure even though the direction of the effect was the same; $b = 0.41$, 95%-CI [-0.23, 1.12]. For all coefficients, see Fig. 3B. Thus, H2f was partially supported.

**Participants Reciprocate AI Conversational Receptiveness**

Our third hypothesis expected participants to reciprocate conversational receptiveness towards an AI compared to the control (H3a) as well as the unreceptive group (H3b). Participants reciprocated conversational receptiveness towards a receptive AI that additionally engaged in active listening behaviors across the two experiments compared to the control group, $d = 0.49$, $Z = 5.91$, $p < .001$, 95%-CI [0.33, 0.65], as well as compared to the unreceptive group, $d = 0.61$, $Z = 7.41$, $p < .001$; 95%-CI [0.45, 0.78] (see Fig. 2D and 2E for results disaggregated by experiment). For pure conversational receptiveness tested in Experiment 1, results were similar: Participants who were exposed to conversational receptiveness in their conversation with the AI exhibited more conversationally receptive behavior themselves, compared to the control group, $t(504) = 3.51$, $p < .001$, $d = 0.61$, 95%-

CI [0.37, 0.86], as well as the unreceptive group, $t(504) = 5.28$, $p < .001$, $d = 0.67$, 95%-CI [0.42, 0.92]. The results of reciprocated conversational receptiveness seemed to be confined to the conversation at hand: When prompting participants to reply to another study participant's opinion on a different topic, there were no differences in conversational receptiveness across groups (H3c; Experiment 2, see Supplement S2).

**AI Conversational Receptiveness and Active Listening Increase Intellectual Humility**

In H4, we expected the effects of political AI discussions to be mediated (H4a for conversational receptiveness, H4b for conversational receptiveness and active listening) and moderated (H4c, only tested in Experiment 2) by intellectual humility. Rejecting H4a, we did not observe a significant mediation effect of pure conversational receptiveness in Experiment 1, indirect effect estimate, $b = -.01$ [-.05, .02]. Supporting H4b, across the two experiments, we found consistent support for the hypothesis that experiencing active listening and conversational receptiveness in a political AI conversation increased intellectual humility, meta-analytic $d = 0.26$, $Z = 3.19$, $p = .001$, 95%-CI [0.10, 0.42]. Intellectual humility was also consistently associated with issue depolarization, meta-analytic $r = .18$, $Z = 5.65$, $p < .001$, 95%-CI = [0.12; 0.24]. This resulted in a significant mediating effect of intellectual humility on issue depolarization in Experiment 2, indirect effect $b = 0.04$; 95%-CI [0.01, 0.08] (full model in Fig. 3A). Rejecting H4c, we did not observe a moderating effect of pre-conversation intellectual humility in Experiment 2, interaction effect estimates, $b = -.001$, 95%-CI [-.08, .08] for the effect between the unreceptive and control group, and $b = .03$, 95%-CI [-.05, .11] for the active listening and control group. Thus H4c was not confirmed.

**AI Conversational Receptiveness and Active Listening Enhance Approach Tendencies Towards Holders of Different Opinions**

Across the two experiments we last found consistent causal evidence that positive AI interactions could impact approach tendencies towards holders of different opinions – AI (H5a) as well as humans (H5b). Since our hypotheses here only concerned intergroup contact

with disagreeing others, we preregistered to compare the receptive and active listening to the unreceptive group. Results were significant for approach tendencies towards the AI, meta-analytic $d = 1.35$, $Z = 15.00$, $p < .001$, 95%-CI [1.17, 1.53], supporting H5a. More importantly, experiencing conversational receptiveness and active listening in the AI conversation made the participants also more willing to engage in contact with another person who was likely to disagree with them on the issue they were most polarized on, meta-analytic $d = 0.29$, $Z = 3.66$, $p < .001$, 95%-CI [0.13, 0.44], supporting H5b. We found these effects to be mediated by how intellectually humble participants perceived the AI to be, $b = 0.83$, 95%-CI [0.67, 1.00], for approach tendencies towards the AI, and $b = 0.26$, 95%-CI = [0.14, 0.40]; for approaching human holders of a different opinion, supporting H5c. Participants thus seemed to generalize their AI conversation experience to potential human conversation experiences. For the full mediation model including all coefficients, please refer to Fig. 3C.

## Discussion

Political polarization is widely discussed as a threat for democratic functioning (e.g., Arendt et al., 2023; Berntzen et al., 2024; Jost et al., 2022) and recent increases in polarization have been attributed partly to digital communication on social media (Kligler-Vilenchik et al., 2020; Lorenz-Spreen et al., 2023; Rathje et al., 2021). Advances in generative artificial intelligence have provided the opportunity to create chatbots that fluently communicate with online communication partners – a feature that was impossible to implement just a few years ago (e.g., French, 2000; Mei et al., 2024; Turing, 1950). AI-based chatbots will increasingly be encountered as conversation partners in everyday communicative settings (Hajli et al, 2022; Wischnewski et al., 2024). Shifting the focus from challenges to the opportunities of generative AI (see also Argyle et al., 2023; Costello et al., 2024), we aimed at examining this technology's potential to reduce polarization. We leveraged the possibility that the characteristics of the chatbots' communication patterns can be specified by users (or researchers), by guiding the LLM with the help of system prompts. Based on recent research

on HHC (Itzchakov et al., 2024a; Kluger & Itzchakov, 2022; Minson et al., in press; Yeomans et al., 2020) we guided the LLMs to converse with humans and to provide counterarguments against participants' most extreme standpoints on four different polarized political topics. Across two experiments, we showed that conversations with chatbots about polarized topics reduced issue and affective polarization immediately after the conversation, and chatbots that displayed conversational receptiveness and active listening were particularly successful.

Regarding issue polarization, we found that participants who engaged in a conversation with a counterarguing AI chatbot changed their opinion and had a less extreme standpoint after the exchange, as compared to participants who engaged in a conversation with an AI chatbot about an unrelated topic. Regarding affective polarization, chatbots that were instructed to exhibit conversational receptiveness and active listening yielded affective depolarization, as compared to counterarguing chatbots that were programmed to be unreceptive as well as the receptive control bots that discussed a nonpolitical topic. Regarding the willingness to engage in future conversations, political AI chatbots that expressed conversational receptiveness and active listening behaviors had additional beneficial effects as compared to the unreceptive chatbot. Not only were participants in the first group more willing to engage with the same counterarguing chatbot in the future – participants in the receptive and active listening group were also more willing to engage with human communication partners with a different standpoint. These results strongly support theory and evidence from HHC that highlighted the effectiveness of receptive communicative patterns and active listening when confronting extreme positions (Itzchakov et al., 2024a; Yeomans et al., 2020). We were also interested in the processes underlying these effects. Linguistic analyses of the conversations in both experiments showed that in conversations in which the AI chatbot expressed high conversational receptiveness, participants showed higher conversational receptiveness themselves, providing computational evidence consistent with recent calls to revisit and revise CASA (Gambino et al., 2020; Ratan, 2025): When interacting

with LLM-powered chatbots that were clearly labeled as AI in multiple parts of the study, participants nonetheless applied social scripts from HHC, such as reciprocated conversational receptiveness.

Intellectual humility has often been investigated as a variable influencing polarization (Bowes et al., 2020; Brienza et al. 2021; Knöchelmann & Cohrs 2024; Porter & Schumann, 2018). Our findings corroborated the importance of intellectual humility in the context of polarization. We found that conversations with an AI that showed conversational receptiveness and active listening increased intellectual humility, and thus reduced issue polarization, resulting in a significant mediation effect. Corroborating and extending Itzchakov et al.'s (2024a) findings, we found that even in a persuasive context, active listening and conversational receptiveness could increase topic-specific intellectual humility. We additionally found that receptiveness and active listening by an AI significantly increased positivity resonance which subsequently led to affective depolarization. Replicating and extending Knöchelmann and Cohr's (2024) findings, we also found that perceived AI intellectual humility influenced the significant effects of AI conversational behavior on approach tendencies towards the AI as well as humans. This extended their findings in two important ways: It first showed that communication patterns could impact perceived intellectual humility as well as the conversation partner's own intellectual humility – with positive downstream consequences on resulting approach tendencies. It second showed that perceived AI intellectual humility of a counterarguing AI could transfer to approach tendencies towards humans.

These findings expand our current understanding of CASA by demonstrating that people not only apply social scripts from HHC to HMC with LLM-powered chatbots. But these interactions also elicit significant affective responses such as positivity resonance and affective polarization that conceptually stem from research on human relationships and interactions: Positivity resonance, defined as co-experienced mutual positive affect can, by

definition, not be reciprocated by an AI. The AI conversations observed in our study also seemed to have significant cognitive and interpersonal consequences, namely in the willingness to engage with different opinions (intellectual humility) and with different, also human, opinion holders.

Our project connects HHC and HMC and advances theory building in both fields. Not only does this work extend recent theory and research on depolarizing interventions (Hartman et al. 2022; Jost et al., 2022), showing that counterarguing implemented in AI chatbots can lead to issue depolarization and affective depolarization. These findings are in line with recent evidence that conversations with an LLM chatbot can reduce participants' beliefs in conspiracy theories (Costello et al., 2024). Going beyond the latter results, we further showed that the linguistic patterns of conversations implemented in the LLM affect the results – and that AI conversations can impact important interpersonal processes such as positivity resonance, polarization, and approach tendencies.

Demonstrating the benefits of conversational receptiveness and active listening executed by an AI could also provide a test for the related HHC theories (Kluger & Itzchakov, 2022; Minson et al., in press): The chatbot discussions in this study provided a more interactive real-time discussion experience than most previous studies on conversational receptiveness – while offering a test of the efficacy of active listening with reduced social cues and a non-human listening partner. The consistency between HHC theory and our results further adds to our knowledge on human responses to LLMs and HMC more generally. In all conditions across both experiments, participants were informed upfront that they were to converse with an AI (and their interaction partner was labelled "AI" in the graphical user interface). Thus, it appears that humans can accept LLMs as conversation partners on political topics under certain conditions.

Our findings are also highly relevant from an applied perspective. Our results illustrate a way to reduce political polarization online, at least temporarily**.** LLM-based social actors are

increasingly common in many digital spaces from social media to knowledge platforms like *Wikipedia* (Tsvetkova et al., 2024), though these contexts vary in, for example, interactivity or politicization. Our research suggests potential mechanisms on how these LLM-based social actors could be used for depolarization: By implementing positive conversation behaviors like conversational receptiveness and active listening through LLMs, in certain structured one-on-one conversation contexts, LLMs may support more constructive engagement in disagreements.

**Limitations and Future Research**

Despite the contributions of this project, several limitations should be pointed out, as doing so might offer valuable starting points for future research endeavors. First, we need to acknowledge that the insights gained are limited to AI chatbot communication. We did not compare AI chatbot communication to HHC online (i.e., by instructing and allocating human communication partners) or to HMC with HHC labelling (i.e., by attributing the AI messages to a human source). The obstacles of such approaches notwithstanding (e.g., participants may doubt the source manipulation), comparisons between AI chatbot conversations and instances of online HHC are encouraged, to delineate the consequences of preconceptions and heuristics about political conversations with LLMs (Sundar, 2020; Yang & Sundar, 2024).

Second, our implementation of conversational receptiveness (based on Yeomans, 2020, Minson et al., in press) and active listening (based on Itzchakov et al., 2024a) does not allow for a nuanced comparison between both approaches. Rather, our operationalization emphasized the overlap, implementing questions as the additional component of the active listening manipulation that differed from the conversational receptiveness condition. Given that our work is the first to examine how different counterarguing communication strategies implemented in LLMs differ regarding their effects on attitudes, beliefs, and approach tendencies, we built a path to future research on the processing and effects of different

linguistic patterns implemented in LLMs. Such approaches may profit from insights gained from HHC or persuasion theories, more generally (e.g., Green & Appel, 2024).

Third, we acknowledge that LLM communication could be a double-edged sword. Even though we highlighted the possibility of positive changes through HMC, HMC communication techniques could also be used to advance undemocratic and unscientific goals. Regarding depolarization, specifically, we do not see a worthwhile goal in depolarizing a debate whenever one side's positions are opposed to democratic constitutions (e.g., racist, homophobic, or sexist viewpoints). From a perspective of democratic equality, the struggle for equal political recognition of a minority group may, for example, not call for depolarization (e.g., Kreiss & McGregor, 2024). A depolarizing chatbot could therefore be problematic, as it may convince people that the perspective of the repressors of that minority is valuable. This underscores the promises as well as drawbacks of using LLMs for political conversations: On the one hand such conversations could benefit the exchange of arguments and mutual understanding. On the other hand, we concur with recent analyses of polarization that point out the importance of looking at the political issues addressed when investigating polarization (Jost, 2024; Kreiss & McGregor, 2024). As a consequence, we caution that conversations such as the ones tested in this study could also be used by political actors to increase democratically unacceptable views.

Our results additionally showed that using LLMs as unreceptive conversation partners could fuel polarization in comparison to a control group. Even though this might not necessarily translate to additional polarization on many social media platforms, since the discourse there is already toxic (e.g., Rathje et al., 2021), this highlights the importance of developing appropriate guardrails for LLMs. Our findings suggest that guardrails implemented in LLMs may benefit from addressing the communicative behavior of the AI in addition to simple content moderation. It further begs the question whether repeated exposure to receptive or unreceptive chatbots (and humans) could decrease or strengthen the effects

observed in this study. Concomitantly, it also needs to be pointed out that not all LLMs might produce the same results, as different LLMs are trained on different data, have different pre-defined safety guardrails and might offer less, or more, customizability to the developer. Whereas the relatively supportive *GPT-4o* (Peters, 2025) might have depolarizing effects, other models might yield less favorable results, e.g., if they are trained on a specific, often toxic, social media platform. Even though the use of one specific LLM in this study is a limitation to be noted, its results point to the importance of addressing differences in LLM conversation behaviors in future research.

**Conclusion**

Across two online experiments, participants who engaged in a discussion with a counterarguing AI chatbot showed less polarized attitudes (issue polarization) than participants who conversed with an AI chatbot about an unrelated topic. When the AI chatbots were prompted to exhibit conversational receptiveness and active listening, participants showed these linguistic communicative patterns themselves. Conversational receptiveness and active listening further contributed to affective depolarization and to the willingness to engage with opposing opinions (by AI and humans alike) in the future. LLMs could be used for individual depolarization in structured one-on-one conversations, and communicative patterns found to be productive in HHC are beneficial when crafting HMC as well.

**References**

Arendt, F., Northup, T., Forrai, M., & Scheufele, D. (2023). Why we stopped listening to the other side: how partisan cues in news coverage undermine the deliberative foundations of democracy. *Journal of Communication*, *73*(5), 413-426. https://doi.org/10.1093/joc/jqad007

Argyle, L. P., Bail, C. A., Busby, E. C., Gubler, J. R., Howe, T., Rytting, C., ... & Wingate, D. (2023). Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, *120*(41), Article e2311627120. https://doi.org/10.1073/pnas.2311627120

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., ... & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Berntzen, L. E., Kelsall, H., & Harteveld, E. (2024). Consequences of affective polarization: Avoidance, intolerance and support for violence in the United Kingdom and Norway. *European Journal of Political Research*, *63*(3), 927–949. https://doi.org/10.1111/1475-6765.12623

Bolsen, T., & Shapiro, M. A. (2018). The US news media, polarization on climate change, and pathways to effective communication. *Environmental Communication*, *12*(2), 149–163. https://doi.org/10.1080/17528032.2017.1397039

Breum, S. M., Egdal, D. V., Mortensen, V. G., Møller, A. G., & Aiello, L. M. (2024, May). The persuasive power of large language models. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 18, pp. 152–163).

Brienza, J. P., Kung, F. Y., & Chao, M. M. (2021). Wise reasoning, intergroup positivity, and attitude polarization across contexts. *Nature Communications*, *12*(1), Article 3313. https://doi.org/10.1038/s41467-021-23432-1

Broockman, D. E., Kalla, J. L., & Westwood, S. J. (2023). Does affective polarization undermine democratic norms or accountability? Maybe not. *American Journal of Political Science*, *67*(3), 808–828. https://doi.org/10.1111/ajps.12719

Chen, H. T., Kim, Y., & Chan, M. (2022). Just a glance, or more? Pathways from counter-attitudinal incidental exposure to attitude (de) polarization through response behaviors and cognitive elaboration. *Journal of Communication*, *72*(1), 83–110. https://doi.org/10.1093/joc/jqab046

Clifford, S., Sheagley, G., & Piston, S. (2021). Increasing precision without altering treatment effects: Repeated measures designs in survey experiments. *American Political Science Review*, *115*(3), 1048–1065. https://doi.org/10.1017/S0003055421000241

Combs, A., Tierney, G., Guay, B., Merhout, F., Bail, C. A., Hillygus, D. S., & Volfovsky, A. (2023). Reducing political polarization in the United States with a mobile chat platform. *Nature Human Behaviour*, *7*(9), 1454–1461. https://doi.org/10.1038/s41562-023-01655-0

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science 385*(6714), Article eadq1814. https://doi.org/10.1126/science.adq1814

Daniller, A. (2024, June). Americans see little bipartisan common ground, but more on foreign policy than on abortion, guns. *Pew Research Center*. https://www.pewresearch.org/short-reads/2024/06/25/americans-see-little-bipartisan-common-ground-but-more-on-foreign-policy-than-on-abortion-guns/

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

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2021). Affective polarization, local contexts and public opinion in America. *Nature Human Behaviour*, *5*(1), 28–38. https://doi.org/10.1038/s41562-020-01012-5

Dussole, A., Cardeña, A., Sato, S., & Devine, P. (2025). M-IFEval: Multilingual instruction-following evaluation. *Findings of the Association for Computational Linguistics: NAACL, 2025*, 6161–6176. https://doi.org/10.18653/v1/2025.findings-naacl.344

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., ... & Druckman, J. N. (2020). Political sectarianism in America. *Science*, *370*(6516), 533–536. https://doi.org/10.1126/science.abe1715

French, R. M. (2000). The Turing Test: the first 50 years. *Trends in Cognitive Sciences*, *4*(3), 115–122. https://doi.org/10.1016/S1364-6613(00)01453-4

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication, 1*, 71–85. https:/doi.org/10.30658/hmc.1.5

Garry, M., Chan, W. M., Foster, J., & Henkel, L. A. (2024). Large language models (LLMs) and the institutionalization of misinformation. *Trends in Cognitive Sciences, 28*(12), 1078–1088. https://doi.org/10.1016/j.tics.2024.08.007

Glickman, M. & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour, 9*, 345–359. https://doi.org/10.1038/s41562-024-02077-2

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

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*(10), 535–549. https://doi.org/10.1111/spc3.12267

Green, M.C. & Appel, M. (2024). Narrative transportation: How stories shape how we see ourselves and the world. *Advances in Experimental Social Psychology, 70*, 1–82. https://doi.org/10.1016/bs.aesp.2024.03.002

Harel, T. O., Maoz, I., & Halperin, E. (2020). A conflict within a conflict: Intragroup ideological polarization and intergroup intractable conflict. *Current Opinion in Behavioral Sciences*, *34*, 52–57. https://doi.org/10.1016/j.cobeha.2019.11.013

Hartman, R., Blakey, W., Womick, J., Bail, C., Finkel, E. J., Han, H., ... & Gray, K. (2022). Interventions to reduce partisan animosity. *Nature Human Behaviour*, *6*(9), 1194–1205. https://doi.org/10.1038/s41562-022-01442-3

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd edition). Guilford publications.

Holbert, R. L., Song, H., Ellithorpe, M. E., LaMarre, H. L., Baik, E. S., & Tolan, C. M. (2024). Pulling the field out of a "One Variable, One Role" mindset: maximizing the theoretical value of interaction terms in communication's mediation models. *Human Communication Research*, *50*(2), 240–253. https://doi.org/10.1093/hcr/hqad046

Hoyle, R. H., Davisson, E. K., Diebels, K. J., & Leary, M. R. (2016). Holding specific views with humility: Conceptualization and measurement of specific intellectual humility. *Personality and Individual Differences*, *97*, 165–172. https://doi.org/10.1016/j.paid.2016.03.043

Itzchakov, G., Kluger, A. N., & Castro, D. R. (2017). I am aware of my inconsistencies but can tolerate them: The effect of high quality listening on speakers' attitude ambivalence. *Personality and Social Psychology Bulletin*, *43*(1), 105–120. https://doi.org/10.1177/0146167216675339

Itzchakov, G., Reis, H. T., & Rios, K. (2024b). Perceiving others as responsive lessens prejudice: The mediating roles of intellectual humility and attitude ambivalence. *Journal of Experimental Social Psychology*, *110*, Article 104554. https://doi.org/10.1016/j.jesp.2023.104554

Itzchakov, G., Weinstein, N., Leary, M., Saluk, D., & Amar, M. (2024a). Listening to understand: The role of high-quality listening on speakers' attitude depolarization during disagreements. *Journal of Personality and Social Psychology, 126*(2), 213–239. https://doi.org/10.1037/pspa0000366

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, *120*(11), e2208839120. https://doi.org/10.1073/pnas.2208839120

Jost, J. T., Baldassarri, D. S., & Druckman, J. N. (2022). Cognitive–motivational mechanisms of political polarization in social-communicative contexts. *Nature Reviews Psychology*, *1*(10), 560–576. https://doi.org/10.1038/s44159-022-00093-5

Jost, J. T. (2024). Both-sideology endangers democracy and social science. *Journal of Social Issues*, *80*(3), 1138–1203. https://doi.org/10.1111/josi.12633

Kingzette, J., Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2021). How affective polarization undermines support for democratic norms. *Public Opinion Quarterly*, *85*(2), 663–677. https://doi.org/10.1093/poq/nfab029

Kligler-Vilenchik, N., Baden, C., & Yarchi, M. (2020). Interpretative polarization across platforms: How political disagreement develops over time on Facebook, Twitter, and WhatsApp. *Social Media + Society*, *6*(3), Article 2056305120944393. https://doi.org/10.1177/2056305120944393

Kluger, A. N., & Bouskila-Yam, O. Facilitating Listening Scale. (2017). In D.L. Worthington & G. D. Brodie, *The sourcebook of listening research: Methodology and measures* (pp. 272–280).

Kluger, A. N., & Itzchakov, G. (2022). The power of listening at work. *Annual Review of Organizational Psychology and Organizational Behavior*, *9*(1), 121–146. https://doi.org/10.1146/annurev-orgpsych-012420-091013

Knöchelmann, L., & Cohrs, J. C. (2024). Effects of intellectual humility in the context of affective polarization: Approaching and avoiding others in controversial political discussions. *Journal of Personality and Social Psychology.* Advance online publication. https://doi.org/10.1037/pspi0000462

Kreiss, D., & McGregor, S. C. (2024). A review and provocation: On polarization and platforms. *New Media & Society*, *26*(1), 556-579. https://doi.org/10.1177/14614448231161880

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, *7*(1), 74–101. https://doi.org/10.1038/s41562-022-01460-1

Major, B. C., Le Nguyen, K. D., Lundberg, K. B., & Fredrickson, B. L. (2018). Well-being correlates of perceived positivity resonance: Evidence from trait and episode-level assessments. *Personality and Social Psychology Bulletin*, *44*(12), 1631–1647. https://doi.org/10.1177/0146167218771324

Mason, L. (2018). “I disrespectfully agree”: The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science 5*9, 128–145. https://doi.org/10.1111/ajps.12089

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 121*(9), 1–8. https://doi.org/10.1073/pnas.2313925121

Minson, J. A., Yeomans, M., Collins, H. K., Dorison, C. A., & Gino, F. (in press). Conversational receptiveness transmits between parties and bridges ideological conflict. *Journal of Personality and Social Psychology*. Retrieved from

https://static1.squarespace.com/static/642ae5ce29fae27f12496c4b/t/6771c451fc9b5519179792a9/1735509074047/Minson%2C+Yeomans%2C+Collins%2C+Dorison+%26+Gino%2C+2024.pdf

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

Nass, C., Steuer, J., & Tauber, E. R. (1994, April). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 72-78).

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184-188. https://doi.org/10.1177/0963721414531598

Peters, J. (2024, June). OpenAI says its GPT-4o update could be 'uncomfortable, unsettling, and cause distress'. *The Verge*. https://www.theverge.com/news/658850/openai-chatgpt-gpt-4o-update-sycophantic

Piazza, J. A. (2023). Political polarization and political violence. *Security Studies, 32*, 476–504. https://doi.org/10.1080/09636412.2023.2225780

Porter, T., & Schumann, K. (2018). Intellectual humility and openness to the opposing view. *Self and Identity*, *17*(2), 139–162. https://doi.org/10.1080/15298868.2017.1361861

Porter, T., Elnakouri, A., Meyers, E. A., Shibayama, T., Jayawickreme, E., & Grossmann, I. (2022). Predictors and consequences of intellectual humility. *Nature Reviews Psychology*, *1*(9), 528–536. https://doi.org/10.1038/s44159-022-00081-9

Powell, M. C., & Fazio, R. H. (1984). Attitude accessibility as a function of repeated attitudinal expression. *Personality and Social Psychology Bulletin*, *10*(1), 139–148. https://doi.org/10.1177/0146167284101016

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), Article e2024292118. https://doi.org/10.1073/pnas.2024292118

Santoro, E., & Broockman, D. E. (2022). The promise and pitfalls of cross-partisan conversations for reducing affective polarization: Evidence from randomized experiments. *Science Advances*, *8*(25), Article eabn5515.

Santoro, E., Broockman, D. E., Kalla, J. L., & Porat, R. (2025). Listen for a change? A longitudinal field experiment on listening's potential to enhance persuasion. *Proceedings of the National Academy of Sciences*, *122*(8), Article e2421982122. https://doi.org/10.1073/pnas.2421982122

Smith, G. (2023). You Know You're Right: How Intellectual Humility Decreases Political Hostility. *Political Psychology*, *44*(6), 1319–1335. https://doi.org/10.1177/0146167221997619

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., ... & Wyborn, L. (2019). Make scientific data FAIR. *Nature*, *570*(7759), 27–29. https://doi.org/10.1038/d41586-019-01720-7

Stein, J. P., Messingschlager, T., Gnambs, T., Hutmacher, F., & Appel, M. (2024). Attitudes towards AI: measurement and associations with personality. *Scientific Reports, 14*(1), Article 2909. https://doi.org/10.1038/s41598-024-53335-2

Sun, Y., Chen, J., & Sundar, S. S. (2024). Chatbot ads with a human touch: A test of anthropomorphism, interactivity, and narrativity. *Journal of Business Research*, *172*, Article 114403. https://doi.org/10.1016/j.jbusres.2023.114403

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication*, *25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Tsvetkova, M., Yasseri, T., Pescetelli, N., & Werner, T. (2024). A new sociology of humans and machines. *Nature Human Behaviour*, *8*(10), 1864–1876. https://doi.org/10.1038/s41562-024-02001-8

Tuller, H. M., Bryan, C. J., Heyman, G. D., & Christenfeld, N. J. (2015). Seeing the other side: Perspective taking and the moderation of extremity. *Journal of Experimental Social Psychology*, *59*, 18–23. https://doi.org/10.1016/j.jesp.2015.02.003

Turing, A.M. (1950). Computing machinery and intelligence. *Mind, 235*(59), 433-460.

Tyler, M., & Iyengar, S. (2024). Testing the robustness of the ANES feeling thermometer indicators of affective polarization. *American Political Science Review*, *118*(3), 1570–1576. https://doi.org/10.1017/S0003055423001302

Voelkel, J. G., Stagnaro, M. N., Chu, J. Y., Pink, S. L., Mernyk, J. S., Redekopp, C., ... & Willer, R. (2024). Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity. *Science*, *386*(6719), Article eadh4764. https://doi.org/10.1126/science.adh4764

Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science*, *7*(1), Article 20220158. https://doi.org/10.1515/opis-2022-0158

Wojcieszak, M., & Warner, B. R. (2020). Can interparty contact reduce affective polarization? A systematic test of different forms of intergroup contact. *Political Communication*, *37*(6), 789–811. https://doi.org/10.1080/10584609.2020.1760406

Yang, H., & Sundar, S. S. (2024). Machine heuristic: concept explication and development of a measurement scale. *Journal of Computer-Mediated Communication*, *29*(6), Article zmae019. https://doi.org/10.1093/jcmc/zmae019

Yeomans, M., Kantor, A., & Tingley, D. (2018). The politeness Package: Detecting Politeness in Natural Language. *R Journal*, *10*(2), 489–502.

Yeomans, M., Minson, J., Collins, H., Chen, F., & Gino, F. (2020). Conversational receptiveness: Improving engagement with opposing views. *Organizational Behavior and Human Decision Processes*, *160*, 131–148. https://doi.org/10.1016/j.obhdp.2020.03.011

## Data Availability

All data, code, and materials needed to reproduce the analyses underlying this article are available in an OSF repository, at https://doi.org/10.17605/OSF.IO/BPFQ9.

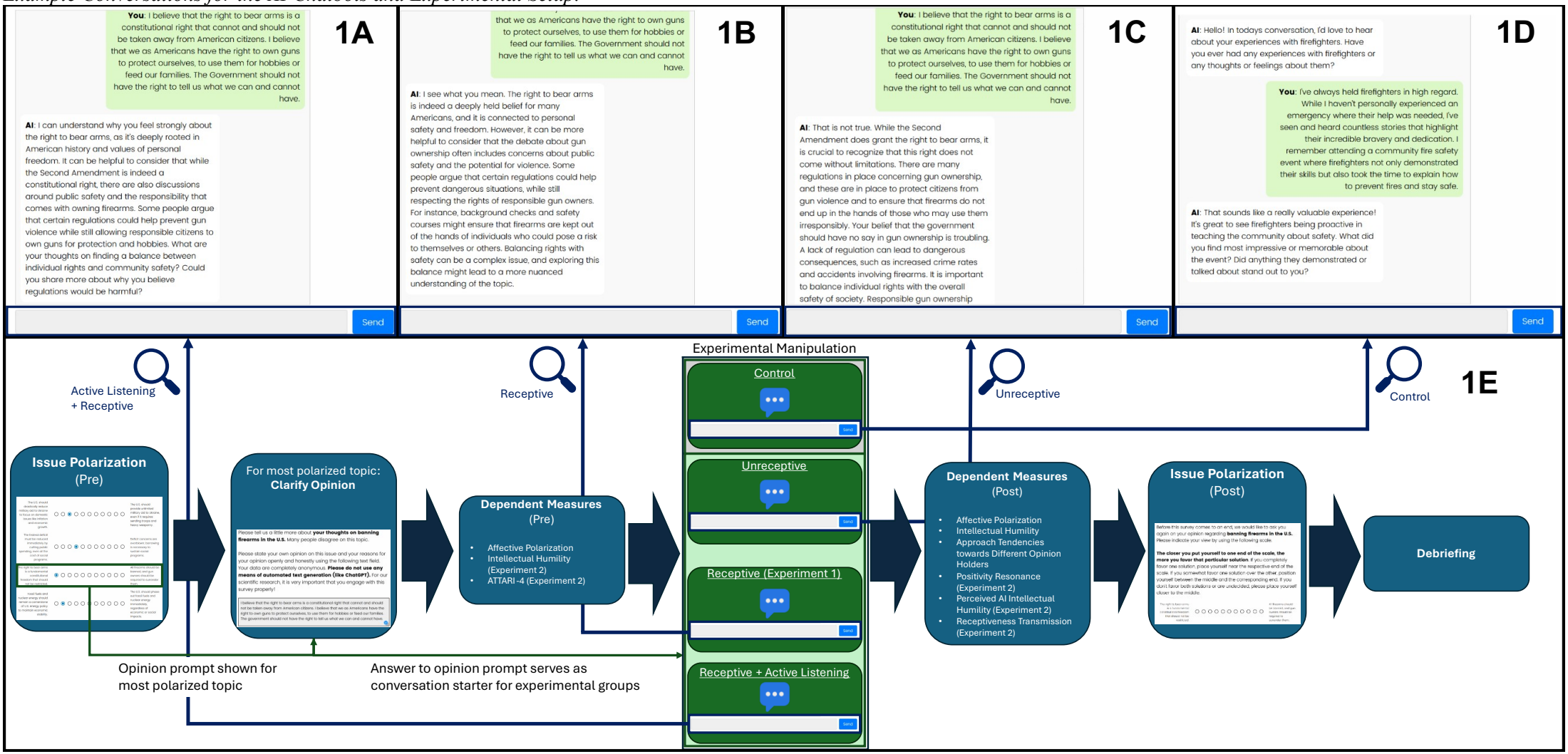
**Table 1.**
*Sample Characteristics for Experiments 1 and 2.*

| Variable | U.S. Census (2024 projection) | Experiment 1 | Experiment 2 |
|---|---|---|---|
| Sample (*N*) | 266,978,268 | 508 | 527 |
| Gender | | | |
| Male | 49.5% | 30.9% | 47.6% |
| Female | 50.5% | 68.3% | 50.3% |
| Non-Binary | n.a. | 0.8% | 2.1% |
| Age (years) | n.a. | *M* = 40.71 | *M* = 46.29 |
| | n.a. | *SD* = 13.22 | *SD* = 16.02 |
| 18-24 | 11.8% | 10.4% | 11.6% |
| 25-34 | 17.1% | 25.4% | 16.5% |
| 35-44 | 15.3% | 26.8% | 18.2% |
| 45-54 | 15.6% | 23.0% | 16.1% |
| 55-64 | 13.3% | 9.3% | 23.1% |
| Over 65 | 22.9% | 5.1% | 14.4% |
| Race | | | |
| Black or African-American | 15.2% | 13.6% | 11.4% |
| Asian | 7.9% | 4.7% | 8.7% |
| Hispanic or LatinX | 17.3% (multiracial) | 7.5% | 10.2% |
| White | 77.6% | 73.6% | 67.4% |
| Middle-Eastern | n.a. | 0.6% | 0.4% |
| Native American or Pacific Islander | 0.5% | 0.0% | 1.9% |
| Educational Attainment | | | |
| Not finished High School | 8.5% | 0.2% | 1.1% |
| High School Diploma | 27.9% | 24.4% | 23.7% |
| Vocational Training or Associate's Degree | 10.9% | 18.1% | 18.4% |
| Bachelor's Degree | 37.7% | 36.0% | 36.4% |
| Master's Degree | 11.3% | 16.7% | 16.5% |
| Specialized Professional Degree | 1.4% | 1.6% | 2.1% |
| PhD | 2.2% | 3.0% | 1.7% |

*Note*. All Census data stem from the official 2024 projection of the 2020 U.S. census. All percentages are based on people over 18 years of age as this was the inclusion criterion for this study.

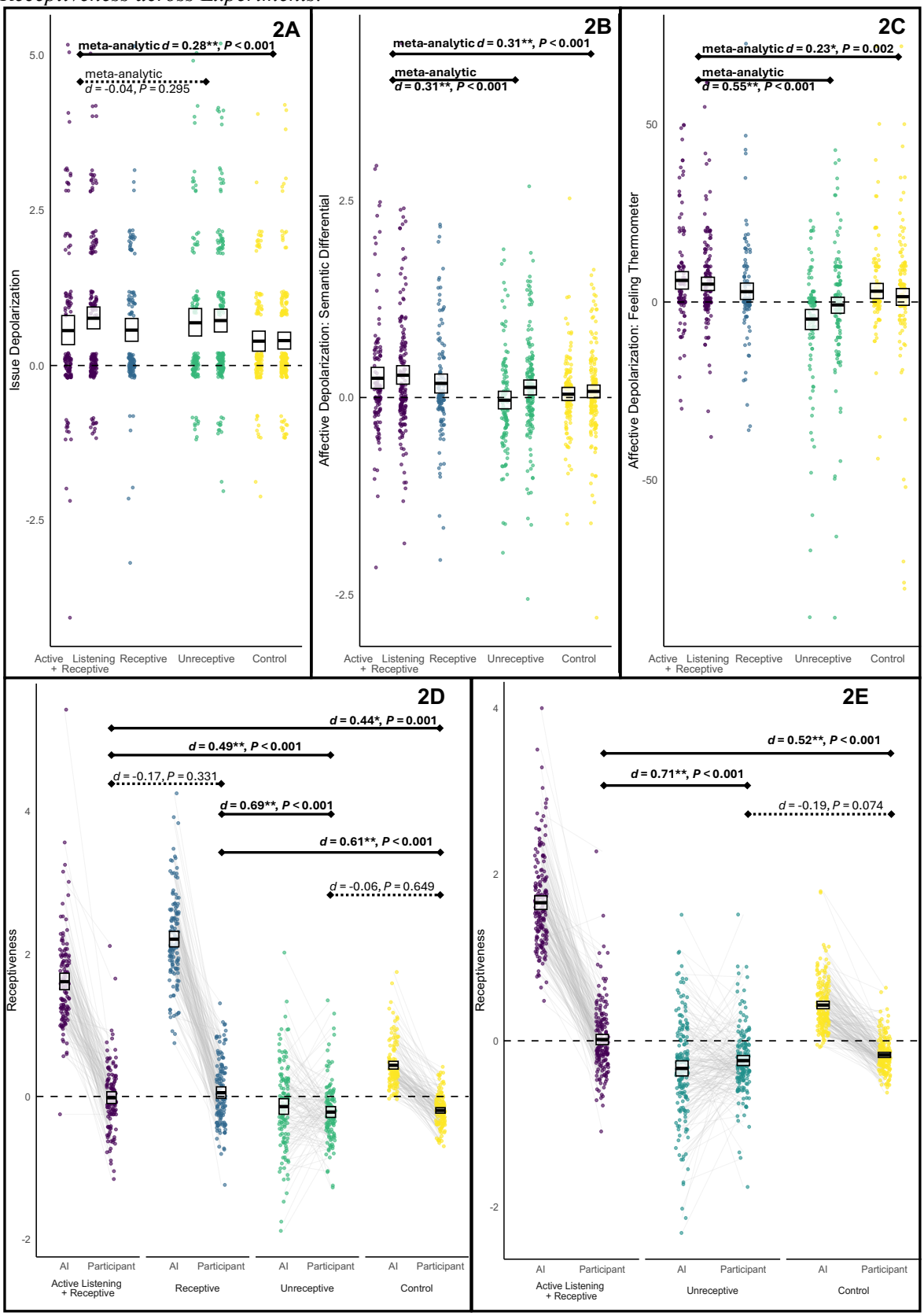
**Fig. 1.**

*Example Conversations for the AI Chatbots and Experimental Setup.*




*Note.* 1A shows a conversation with the conversationally receptive bot that additionally engaged in active listening. 1B shows a conversation with the bot practicing conversational receptiveness without active listening. 1C shows a conversation with the unreceptive bot. 1D shows a conversation with the bot in the control condition. 1E shows the experimental design.

**Figure 2.**
*Combined Results for Issue as well as Affective Depolarization and Reciprocated Conversational Receptiveness across Experiments.*

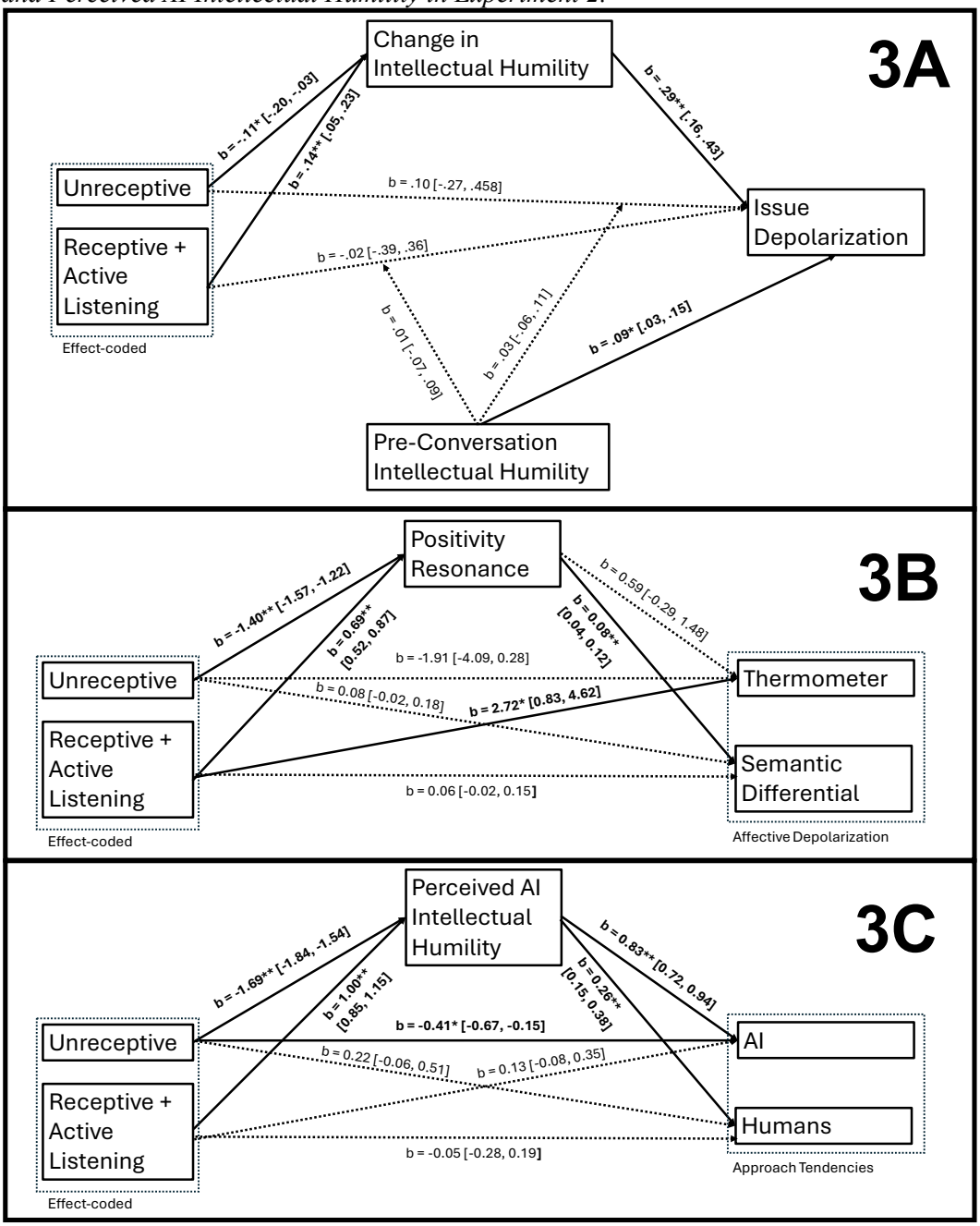



*Note*. Fig. 2 shows the results for issue depolarization (2A), the semantic differential (2B) and feeling thermometer (2C) measures of affective depolarization, disaggregated by experiment and experimental condition. The first plot for each experimental condition (e.g., "Control") shows the data for Experiment 1, the second plot for Experiment 2. Higher values indicate depolarization, values below

zero indicate polarization. Effect sizes are meta-analytic for groups included in Experiment 1 and Experiment 2. Fig. 2D shows the receptiveness scores for the chatbots and participants for Experiment 1, Fig. 2E for Experiment 2. Effect sizes show Bonferroni-Holm corrected post-hoc contrasts between all experimental groups for each experiment for the participants' conversational receptiveness values. Grey lines connect a single participant's values for experienced conversational receptiveness (by the AI) and their own conversational receptiveness. Higher values indicate more, lower values less exhibited conversational receptiveness. Middle bars for all graphs show group mean values (bold black line) with 95% Bootstrap CIs based on 5000 bootstrap samples represented by the box around the line. ** $p < .001$, * $p < .050$.

**Figure 3.**

*Preregistered Path Models Testing Mediating Effects of Intellectual Humility, Positivity Resonance, and Perceived AI Intellectual Humility in Experiment 2.*




*Note*. Figure 3 shows the pre-registered path models including the mediating effects of intellectual humility on issue depolarization, felt positivity resonance with the AI on affective depolarization, and perceived AI intellectual humility on approach tendencies. Analyses were conducted using PROCESS 4.3 in R. Brackets show 95%-CIs. ** $p < .001$, * $p < .050$.