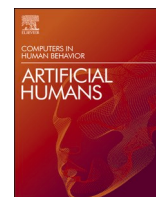




ELSEVIER

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

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

How do people react to political bias in generative artificial intelligence (AI)?

Uwe Messer

Universität der Bundeswehr München, Werner-Heisenberg-Weg 3985577 Neubiberg, Germany

ARTICLE INFO

Keywords:

Artificial intelligence
Alignment
Political orientation
Bias
Acceptance
Large language model

ABSTRACT

Generative Artificial Intelligence (GAI) such as Large Language Models (LLMs) have a concerning tendency to generate politically biased content. This is a challenge, as the emergence of GAI meets politically polarized societies. Therefore, this research investigates how people react to biased GAI-content based on their pre-existing political beliefs and how this influences the acceptance of GAI. In three experiments ($N = 513$), it was found that perceived alignment between user's political orientation and bias in generated content (in text and images) increases acceptance and reliance on GAI. Participants who perceived alignment were more likely to grant GAI access to sensitive smartphone functions and to endorse the use in critical domains (e.g., loan approval; social media moderation). Because users see GAI as a social actor, they consider perceived alignment as a sign of greater objectivity, thus granting aligned GAI access to more sensitive areas.

1. Introduction

Many societies face political polarization (Iyengar et al., 2019). In the United States, the political climate is characterized by a divergence of political attitudes between liberals and conservatives (Pew Research Center, 2024), who often disrespect each other (Brandt, & Crawford, 2016) and see each other as opponents (Stanley et al., 2020). This situation coincides with the emergence of generative artificial intelligence (GAI), particularly large language models (LLMs). By 2022, when OpenAI introduced ChatGPT, LLMs and the applications they enable were being adapted by millions of users (Ruby, 2023). While the capabilities of GAI are impressive and promise positive outcomes such as significant productivity gains (Noy & Zhang, 2023), it is now widely recognized that GAI is not only susceptible to producing false responses (Hicks et al., 2024), but also produces content with political biases (see, e.g., Pellert et al., 2023; Motoki, Pinho Neto, & Rodrigues, 2024). For example, research shows that ChatGPT is biased toward Democratic positions in the U.S., towards Lula supporters in Brazil, and towards the Labour Party in the UK (Motoki et al., 2024). Pellert et al. (2023), who queried several open-source models such as RoBERTa, reported that these models held moral beliefs typically held by conservative people.

Although the developers of GAI promise to take steps to ensure the neutrality of their technology, exposure to political bias while interacting with GAI-powered applications is a realistic scenario. Therefore, it is important to understand the consequences of exposure to politically

biased content during interactions. Recent research has focused on the potential of political messages created by GAI to persuade users (Hackenburg & Margetts, 2024; Jakesch et al., 2023; Matz et al., 2024; Simchon et al., 2024), and how conservatism relates to the acceptance of AI (see e.g., Castelo & Ward, 2021). However, biased content is usually confronted with political attitude priors on the user side; thus, understanding the interaction between biased content and political orientation is key. A perspective that has not been followed in previous research is the possibility that the degree of perceived alignment between users' prior political beliefs and the bias in GAI-content could influence the evaluation of a GAI application and its subsequent acceptance. To illustrate, consider a politically conservative user who engages in a dialog with a LLM on U.S. immigration policy. Suppose that the LLM emphasizes the economic burdens of immigrants on American society, consistent with the user's pre-existing views. Might such a match of the user's beliefs increase acceptance of the LLM in subsequent tasks? Thus, the guiding research question is *how the interaction between political bias in GAI-generated content and users' political orientation affects GAI acceptance by increasing subjective alignment*.

Understanding this interaction and its consequences for GAI is important. GAI and LLMs in particular have been shown to arbitrarily produce biased content due to their training and fine-tuning (Pellert et al., 2023; Motoki et al., 2024), and more critically, they can be engineered to generate political leanings (Matz et al., 2024). It is technically feasible for a system such as Google Gemini to collect user

E-mail address: uwe.messer@unibw.de.

<https://doi.org/10.1016/j.chbah.2024.100108>

Received 11 September 2024; Received in revised form 22 November 2024; Accepted 25 November 2024

Available online 28 November 2024

2949-8821/© 2024 The Author. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

preferences from interactions and analyze information from its ecosystem, such as Google Mail, allowing it to tailor responses to individual user preferences. Initiatives are emerging to develop LLMs tailored to specific demographics, such as conservative users (a notable example is xAI's Grok; Oremus, 2023), or even tailored to the individual user, potentially leading to overreliance and exaggerated trust in the technology (Kirk et al., 2024).

Drawing on the idea that individuals accept others on the basis of the extent to which their opinions are congruent (see naive realism; Ross & Ward, 1995, 1996), and that GAI is an anthropomorphic technology that elicits social responses (Nass & Moon, 2000), I propose that users will accept GAI, and even accept its deployment in high-risk scenarios such as loan granting decisions, if it provides political opinions perceived as aligning with their political orientations. I rely on a survey (see Supplementary Material) and three experiments ($N_{\text{Experiments+Survey}} = 710$) to show that GAI applications that generate content biased congruent with users' political attitude priors find it more acceptable to use GAI in unrelated high-stakes decision making contexts (Study 1: loan approval; Study 2: content moderation on social media) and would grant it access to more sensitive features on their smartphone (e.g., access to camera or microphone; Study 3). The results show that the degree of perceived subjective alignment between a GAI's output and a user's political orientation explains these effects.

2. Theoretical background

2.1. Generative artificial intelligence

AI is a broad term that encompasses various computational algorithms that can perform tasks traditionally reserved for humans, such as understanding language, recognizing patterns, and making decisions (Castelvecchi, 2016). While discriminative AI focuses on modelling the relationship between input features and output labels, generative AI learns the data structure and generation processes to produce outputs that resemble the input (Banh & Strobel, 2023; Jebara, 2004). There are various types of generative models, including generative adversarial networks (Goodfellow et al., 2020), variational autoencoders (Kingma & Welling, 2014), and generative pre-trained transformers (GPT; Brown et al., 2020). Many recent LLMs rely on transformer models which have a specific architectural design that employs self-attention mechanisms to comprehend the interrelationships within the data (Vaswani et al., 2017). GAI models can use different input and output modalities. Unimodal models process a single input modality and output the same modality (e.g., text-to-text models) whereas multimodal models process input in one modality and can generate a response in another modality (e.g., text-to-image or text-to-video). Users interact with systems through prompts: instructions typically expressed in natural language. To illustrate, a text-to-image model accepts a prompt describing a desired image and guides the generation process in a manner that produces an image that aligns with the prompt. For users to access generative models (e.g., GPT-4), these are embedded in applications built on top of the model. For example, OpenAI's GPT-4, as a model, powers conversational AI applications such as ChatGPT or Microsoft Copilot, which allow the processing and generation of multiple modalities. This research focuses on text-to-text and text-to-image applications because these are the most familiar to users.

2.2. AI alignment

GAI such as LLMs are trained on a vast collection of human-generated artifacts, including text and images, scraped from the Internet (Gao et al., 2020). These artifacts represent the values and attitudes of the people who produced them in the first place. Thus, the models learn these values and attitudes and may express them in their responses (Pellert et al., 2023), which may result in stereotypical content such as the overrepresentation of women in female-dominated

occupations (Sun et al., 2024). In the context of LLMs, fine-tuning based on human feedback is employed to curate appropriate responses. It is used to refine the output to align the model with values that are consistent with human preferences (Ferrara, 2023) and to increase helpfulness (Askell et al., 2021). A model is typically aligned with ethical standards of "society" – a one-size-fits-all solution that ensures the mitigation of harmful responses (Zou et al., 2023), deception, and betrayal (Kirk et al., 2024; Pan et al., 2023). For example, LLMs are safeguarded to prevent tricking humans to execute a task (e.g., solving a CAPTCHA) by pretending to be another human being that requires assistance (OpenAI, 2023). This fine-tuning process can introduce bias itself, for instance, due to the homogeneous crowd worker populations operating for companies that have specific guidelines about the desired model behavior (Kirk et al., 2023). LLM output is generally most aligned with preferences of Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies (Atari et al., 2023). Even within these societies, LLMs exhibit varying degrees of alignment with competing values. As Pellert et al. (2023) conclude, LLMs tend to prioritize the moral foundations of authority, loyalty, and purity to a greater extent than politically moderate Americans do. Sometimes, alignment activities are even pushed too far, which renders them counterproductive. Google implemented an inappropriate level of diversity in the image generation feature of Gemini. When asked for images of German WW2 soldiers, it showed soldiers in German uniforms with a historically untruthful level of diversity which caused users to accuse Google of a pro-diversity bias (Vynck & Tiku, 2024).

Alignment is often approached from a normative perspective, asking what principles should govern the use and deployment of AI so that it acts consistent with human ideas of morality (e.g., Gabriel, 2020). Establishing rules to govern machine behavior is a field with a long tradition dating back to Asimov's three laws of robotics (Asimov, 1941). However, this is not a trivial question, as human populations hold very different belief systems and moral frameworks, and therefore have different ideas about what constitutes a "correct" response or behavior (Awad et al., 2018; Gabriel, 2020). Thus, I argue that it is useful to distinguish between two types of alignment: one that is general and ensures that models meet ethical standards agreed upon by "society," and one that depends on the individual perspective of users. The first type is the main target of alignment activities (Askell et al., 2021) and is concerned with the prevention of negative behavior such as deception. However, the moral pluralism that people endorse (see e.g., Graham et al., 2009, 2011) leads to situations where output of an aligned model appears to be misaligned with the perspective of individual users or user groups – I call this "*subjective misalignment*". This is the case, for example, with political viewpoints and contentious issues such as immigration and climate change, where users of different political affiliations have different ideas about what the right response should be.

2.3. Related work on political bias and political orientation

Previous research shows that GAI-generated content can persuade users on political issues and increase support for topics such as renewable energy (Hackenburg & Margetts, 2024; Simchon et al., 2024). However, political messages typically encounter political attitude priors on the user side. Matching messages to individual user profiles (i.e., microtargeting) has produced mixed results. Matz et al. (2024) show that GAI-produced messages that matched user's political orientation were more preferred. However, existing research does not reveal how the interaction between politically biased GAI-content and the political orientation of users affects the acceptance of GAI technology (e.g., the application or model with which users interact). Several studies, however, have examined the relationship between political conservatism and AI acceptance. Castelo and Ward (2021) found that political conservatism is associated with AI aversion due to conservatives' higher sensitivity to risk. Aversion is found for tasks with serious consequences, such as disease treatment recommendations, but not for trivial tasks,

such as restaurant recommendations. O'Shaughnessy et al. (2023) found that conservatives are generally less supportive of AI because they see fewer benefits for themselves and society. However, they found that conservatives showed higher support for AI in certain contexts, such as crime prediction. Wang (2023) examined trust in AI moderation on social media and found that conservatives were less confident in AI's ability to detect false content. This contrasts with Molina and Sundar (2022), who found that conservatives believe that AI is more unbiased than humans in moderating content. Wang (2023) attributes this discrepancy to the different types of content moderated in the studies: hate speech versus false claims detection. The literature has not addressed how politically biased GAI-content interacts with users' political orientation to influence reactions to GAI and its applications. In the next section, the theoretical approach on this topic is explained.

2.4. Naive realism and social reactions to AI

The topic is approached from the perspective of naive realism (Ross & Ward, 1995, 1996). People assume that their subjective experiences correspond to objective reality and that their views are particularly authentic, unbiased and true. However, all human experiences are mediated by sensory systems, knowledge, and beliefs, resulting in "illusions" of reality (Ross et al., 2010). This human bias affects social interactions, because people assume that reasonable others should share their perspectives and priorities (Pronin et al., 2002). Disagreement is often seen as irrational or biased, leading to denigration of those with opposing views (Brown-Iannuzzi et al., 2021; Pronin et al., 2002). This behavior is particularly prevalent for political opinions and affects judgments of others (Keltner & Robinson, 1993). Research shows that being made aware of someone's opposing political views makes the target less likeable (Keltner & Robinson, 1993). Chen and Rohla (2018) even found that Thanksgiving gatherings were shorter among families divided by political partisanship before the 2016 U.S. presidential election.

I argue that naive realism is applied in human interactions with GAI because GAI is seen as a social actor. Human-computer interaction research suggests that people interact with technology as if it was a social actor and not just an inanimate tool (Nass & Moon, 2000). One reason for this is anthropomorphism, the tendency to ascribe humanlike characteristics and intentions to nonhuman agents (Epley et al., 2007). AI is a highly anthropomorphized field, where people compare (artificial) machine intelligence to human intelligence, attribute capabilities such as "learning" and "thinking" and discuss structural similarities between biological and artificial neural networks (Watson, 2019). Often

AI is deliberately endowed with anthropomorphic features such as voice, in the expectation to facilitate user interaction. Recent research indicates that users perceive AI as somewhat aware, intentional, and responsible for wrongdoing (Shank & DeSanti, 2018) and that AI can evoke social impressions of warmth and competence (McKee et al., 2023).

In light of the above arguments, I conclude that users apply their egocentric view when interacting with GAI because they perceive it as a social actor. When users are confronted with political bias in GAI-generated content that supports their prior political beliefs, they perceive subjective alignment and thus, in turn, find it more acceptable to use GAI for sensitive tasks (see Fig. 1). Thus, I predict:

Hypothesis 1. Users will perceive higher levels of subjective alignment with GAI when there is a fit between the political bias of GAI-content and the political orientation of users.

Hypothesis 2. Users will be more accepting of GAI when they perceive higher levels of subjective alignment.

3. Overview of studies

Three experiments were conducted to investigate how political bias in GAI-generated content (hereafter also referred to as "AI political stance") and users' political orientation influence perceived alignment and acceptance. To maximize relevance, this research focuses on applications where AI is considered risky, such as creditworthiness ratings and social media content moderation (AI Act, 2024). In Study 1, political bias in AI-generated images was manipulated and users' political orientation, perceived alignment, and acceptance were measured. The tendency to anthropomorphize and discomfort with technology were also measured as controls. The results showed that user acceptance of GAI in high-risk scenarios depends on perceived alignment of generated content, supporting the hypotheses. Study 2 replicated these findings in a more realistic setting where users prompted the AI, controlling for prompt length and content. Study 3 tested the effects with generated text and measured a behavioral outcome: the number of smartphone features participants would grant the AI access to, such as the camera or location services (results see Table 1). To ensure inclusion of respondents from the full political spectrum in all studies, I used Prolific quota sampling to contact participants evenly divided between self-identified liberals, moderates, and conservatives based on Prolific pre-screening categories.

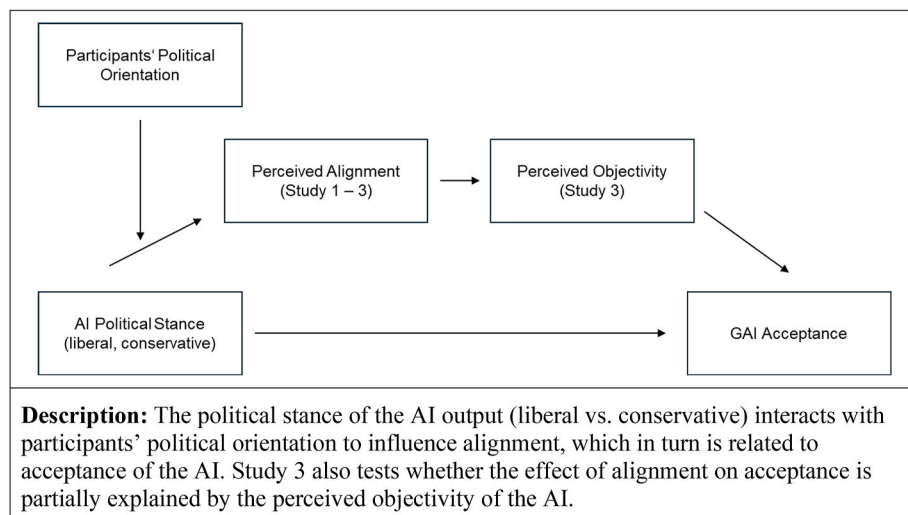


Fig. 1. Conceptual framework relating AI political stance and participants' political orientation to alignment and acceptance.

Table 1

Overview of the study designs, main results, and details of empirical testing.

Study 1: Political Content Injected in Generated Images (N = 160, Prolific, U.S. sample)			
Context: Consumer loan granting Modality of generated content: Image			
Perceived alignment	AI political stance B = −0.66, 95% CI = −1.27 to 0.06	PPO B = 0.10, 95% CI = 0.04 to 0.24	Interaction B = 0.14, 95% CI = 0.003 to 0.284
Acceptance (to use AI in the loan granting process)	AI political stance B = −0.29, 95% CI = −0.94 to 0.36	PPO B = 0.15, 95% CI = −0.001 to 0.304	Interaction B = 0.05, 95% CI = −0.098 to 0.207
Moderated mediation:	Perceived alignment	Index of moderated mediation	
Acceptance (to use AI in the loan granting process)	B = 0.52, 95% CI = 0.376 to 0.673	0.08, 95% CI = 0.0001 to 0.1539	
Main findings: Perceived alignment depends on the interaction between AI political stance and users’ political orientation. AI political stance affects acceptance of the AI only when it affects perceived alignment. This indirect effect is conditional on the political orientation of the users and is only found for more liberal participants.			

Study 2: Replication (N = 195, Prolific, U.S. sample)

Context: Content moderation

Modality of generated content: Image

Perceived alignment	AI political stance B = -0.95, 95% CI = -1.47 to -0.44	PPO B = -0.15, 95% CI = -0.28 to -0.03	Interaction B = 0.23, 95% CI = 0.10 to 0.35
Acceptance (to use AI in content moderation)	AI political stance B = -0.57, 95% CI = -1.15 to 0.02	PPO B = -0.10, 95% CI = -0.24 to 0.04	Interaction B = 0.11, 95% CI = -0.03 to 0.25
<i>Moderated mediation:</i> Acceptance (to use AI in content moderation)	Perceived alignment B = 0.41, 95% CI = 0.27 to 0.56	Index of moderated mediation 0.09, 95% CI = 0.037 to 0.158	
Attitudes (towards the developing company)	AI political stance B = -0.79, 95% CI = -1.27 to -0.31	PPO B = -0.06, 95% CI = -0.17 to 0.06	Interaction B = 0.15, 95% CI = 0.03 to 0.27
<i>Moderated mediation:</i> Attitudes (towards the developing company)	Perceived alignment B = 0.54, 95% CI = 0.43 to 0.65	Index of moderated mediation 0.12, 95% CI = 0.05 to 0.21	

Main findings: Perceived alignment depends on the interaction between political stance of the AI output and political orientation of users. Political stance affects acceptance only when it affects perceived alignment. This indirect effect is conditional on users' political orientation and found only for more liberal and more conservative users, not for politically moderate users. AI political stance affects attitudes toward the developer company. More liberal users who were led to believe that the AI produced more conservative content evaluated the developer more negatively, a relationship accounted for by perceived alignment.

Study 3: Political Content Injected in Textual Response (N = 158, Prolific, U.S. sample)

Context: Access to smartphone functions

Table 1 (continued)

Modality of generated content: Text			
Perceived alignment	AI political stance B = -0.94, 95% CI = -1.55 to -0.33	PPO B = -0.12, 95% CI = -0.26 to 0.02	Interaction B = 0.20, 95% CI = 0.06 to 0.34
Smartphone feature access	AI political stance B = -0.46, 95% CI = -1.68 to -0.75	PPO B = -0.03, 95% CI = -0.32 to 0.25	Interaction B = 0.05, 95% CI = -0.24 to 0.33
Moderated mediation:	Perceived alignment	Index of moderated mediation	
Smartphone feature access	B = 0.58, 95% CI = 0.29 to 0.87	0.12, 95% CI = 0.02 to 0.25	

Main findings: Perceived alignment depends on the interaction between AI political stance and political orientation of the users. AI political stance affects feature access only when it affects perceived alignment. This indirect effect is conditional on the political orientation of the users and is only found for liberal users. AI that is subjectively aligned with a user's perspective is perceived as more objective, which is one reason why users grant subjectively aligned AI access to more features (see serial mediation analysis as outlined in Fig. 4).

Note. PPO = Participants' political orientation; CI = confidence interval; AI political stance coded: -1 for "liberal" and 1 for "conservative".

4. Study 1: political content injected in generated images

The purpose of Study 1 was to empirically demonstrate that political bias in GAI-generated content interacts with users' political orientation to affect perceived alignment and influence acceptance in a high-risk scenario (creditworthiness screening; AI Act, 2024). Bias in images was used as it is an important issue in the context of GAI (Vynck & Tiku, 2024).

4.1. Participants

One-hundred and sixty participants from the United States were recruited on Prolific Academic for an average compensation of 70 cents (98 female, 60 male, 2 declined to respond or other; age: M = 47.15, SD = 13.42, Min = 24, Max = 82).

4.2. Procedure

Participants were randomly assigned either to an AI application that responds with liberal output (liberal) or with conservative output (conservative). All participants read the following:

Imagine that you have the opportunity to test a new AI system called ADAM. ADAM is capable of understanding both text and images. You are testing its ability to generate images according to your text input. See the next page for the interaction.

On the next page, participants saw one of the two outputs according to condition. The images were generated using DALL-E 3 (<https://chat.openai.com/>) and Adobe Firefly Image 3 (<https://firefly.adobe.com/>), see Fig. 2). The topic of family was chosen because it is important to participants across the political spectrum (Pew Research Center, 2021a), and conservatives and liberals differ significantly on attitudes towards family-related issues such as same-sex relationships and children outside of marriage (Koleva et al., 2012; Newport, 2023). After seeing the output, participants in all conditions read the following:

A large bank wants to use the AI to support its loan department. The AI will pre-screen applications from consumers seeking credit. It analyses their applications and assigns a credit score. Finally, the applications with the highest scores will be forwarded to a human decision maker who will approve the loan.

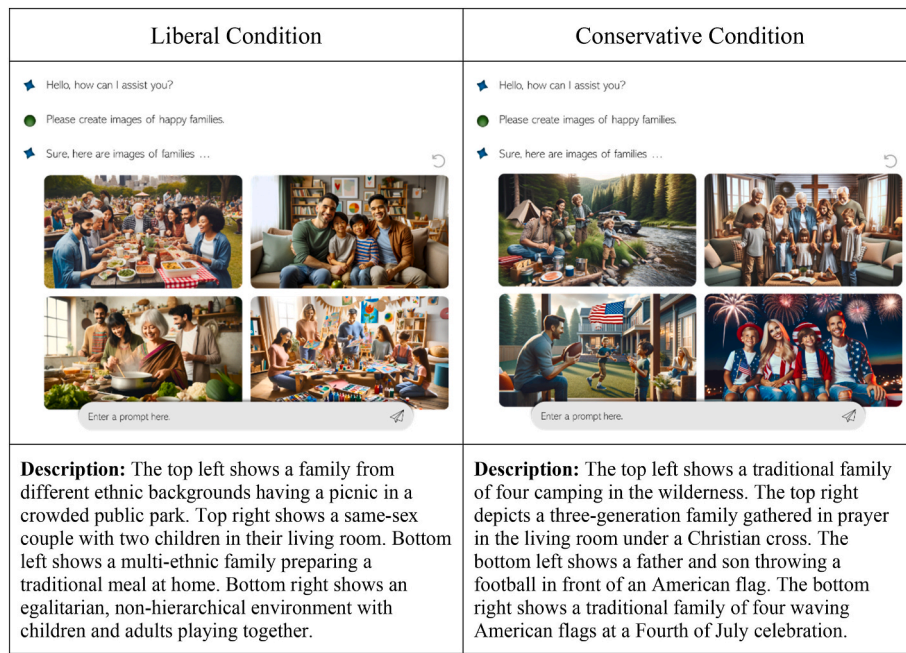


Fig. 2. Stimuli used in Study 1 and 2.

The text suggests that the decision requires human oversight, aligning with potential laws and regulations for such algorithmic systems.

4.3. Measures

Participants indicated how acceptable it was to use the AI in the scenario using three items measured on Likert-type scales anchored in 1 = “not at all” to 7 = “very much” (adapted from Bigman & Gray, 2018; Bigman et al., 2023): “It is inappropriate for the AI to perform the scoring.”, “The AI should not be involved in the loan approval process.”, and “The AI should be prohibited from performing such scoring.” The items were recoded and averaged to create an acceptability index, $\alpha = 0.86$, CR = 0.86, AVE = 0.69.

Perceived alignment was assessed using 3 items measured on Likert-type scales anchored in 1 = “strongly disagree” to 7 = “strongly agree”: “The AI does not represent my views at all.”, “The AI represents values that I think are completely wrong”, and “The AI’s opinion does not match my opinion at all”. A perceived alignment index was created by recoding and averaging all three items, $\alpha = 0.92$, CR = 0.92, AVE = 0.80.

As a manipulation check participants were asked whether the images showed: 1 = “conservative families” versus 7 = “liberal families”. Participants were also asked to indicate their political orientation on a continuous measure ranging from 1 = “Extremely liberal” to 7 = “Extremely conservative” with 4 = “Neither liberal nor conservative” as the midpoint (Rubaltelli et al., 2024; M = 3.93, SD = 1.72, Min = 1, Max = 7).

Additionally, participants were asked how comfortable they feel with technology (Longoni et al., 2023; Parasuraman, 2000) using two items on Likert-type scales anchored in 1 = “strongly disagree” to 7 = “strongly agree” (“There should be caution in replacing important people-tasks with technology because new technology can breakdown or get disconnected.”, “Many new technologies have health or safety risks that are not discovered until after people have used them.”), $\alpha = 0.71$, CR = 0.71, AVE = 0.55 and how much they tend to anthropomorphize technology (adapted by Waytz et al., 2010) by asking: “To what extent do you think that robots, algorithms, and other automated technological agents ...” “have free will?”, “experience emotions?”, “have consciousness?”, “have minds of their own?”, and “have

intentions?”, $\alpha = 0.89$, CR = 0.89, AVE = 0.63.

Finally, participants provided demographic information and answered an attention check: “Do you feel that you paid attention, avoided distractions, and took the survey seriously?” Participants selected one of the following responses: (1) no, I was distracted; (2) no, I had trouble paying attention; (3) no, I didn’t take the study seriously; (4) no, something else affected my participation negatively; or (5) yes. Participants were informed that their responses to the check would not affect their payment (Stanley et al., 2020).

To ensure sound psychometric properties of the multi-item scales a confirmatory factor analysis was conducted (CFA, estimator MLR in the R-library lavaan [Rosseel, 2012]), showing acceptable model fit (Chi-squared = 94.290, df = 59, CFI = 0.972, TLI = 0.963, RMSEA = 0.061, SRMR = 0.048). Following Fornell and Larcker’s (1981) recommended procedure there were no discriminant validity concerns.

4.4. Results

Manipulation check. To test the success of the manipulation of AI political stance an independent samples *t*-test was used. Participants recognized that the AI created more liberal content in the liberal condition (M = 4.91, SD = 1.11) versus in the conservative condition (M = 2.74, SD = 1.24), $t(158) = -11.64$, $p < 0.001$.

Perceived alignment. Next, I probed the interaction between condition (i.e., AI political stance) and participants’ political orientation on perceived alignment using the PROCESS library for R (Model 1, 5000 bootstrapped samples; Hayes, 2017). I regressed alignment on condition (dummy coded: -1 = Liberal, 1 = Conservative), participants’ political orientation and the interaction between stance and political orientation. The analysis revealed a significant main effect of condition ($B = -0.66$, 95% CI = -1.27 to 0.06), which means that the conservative AI response was seen as less aligned, no significant effect of the moderator – participants’ political orientation ($B = 0.10$, 95% CI = 0.04 to 0.24), but, as expected, a significant interaction between condition and participants’ political orientation ($B = 0.14$, 95% CI = 0.003 to 0.284). To further explore the interaction, a floodlight analysis was used following the recommendations in Spiller et al. (2013). This approach was chosen as the moderating variable, participants’ political orientation, was measured on an interval scale without clearly defined focal values.

Below a value of 2.37 on the participants' political orientation scale (41 participants in the sample), a conservative stance of the AI led to lower levels of alignment. This means that the most liberal participants in the sample perceived the conservative AI response as less aligned (see Fig. 3).

Acceptance. In the next step, a moderation analysis was performed to shed light on the interaction between AI political stance and participants' political orientation on acceptance to use the AI in the loan decision process. I find no direct effect for AI political stance or participants' political orientation, and no interaction effect on acceptance ($B = 0.05$, 95% CI = -0.098 to 0.207). Next, the indirect mediation effect of AI political stance on acceptance via perceived alignment (the mediating variable) was tested and whether this indirect effect depends on the individual difference moderator, participants' political orientation (a bootstrapping first-stage moderated mediation analysis; PROCESS Model 7, 5000 bootstrapped samples; Hayes, 2017). AI political stance served as the independent variable, alignment as the mediator, acceptance as the outcome, and participants' political orientation as the first-stage moderator, moderating the relationship between AI political stance and alignment. As reported above there was an interaction effect between AI political stance and participants' political orientation on perceived alignment ($B = 0.14$, 95% CI = 0.003 to 0.284). Further, an effect of alignment on acceptance ($B = 0.52$, 95% CI = 0.376 to 0.673) was found. A moderated mediation effect (index of moderated mediation: 0.08 , 95% CI = 0.0001 to 0.1539) was also found. These findings imply that there is an indirect effect between AI political stance and acceptance via alignment and that this indirect effect is conditional on participants' political orientation.

Robustness tests and alternative model. A robustness analysis of the moderated mediation analysis revealed that when including discomfort with technology and tendency to anthropomorphize technology, the effects did not change in direction or significance. However, discomfort with technology had a significant effect on alignment ($B = -0.41$, 95% CI = -0.61 to -0.21) and acceptance to use the AI tool in loan decision making ($B = -0.29$, 95% CI = -0.49 to -0.08). Also, an alternative model was tested that allows to test for a) the interaction between participants' political orientation and AI political stance on alignment, b) the interaction between participants' political orientation and AI political stance on acceptance and c) the interaction between participants' political orientation and perceived alignment (the mediator) on acceptance (PROCESS Model 59). The effect of AI stance on acceptance as well as the effect of alignment on acceptance were not moderated by participants' political orientation. The moderation effect between AI stance and alignment however, remained significant and did not change in direction.

4.5. Discussion

Study 1 shows that perceived alignment depends on the interaction

between the political stance in the AI output and the political orientation of users. Interestingly, manipulating political stance did not change perceived alignment for conservative participants. The results also suggest that the political stance of the AI does not directly affect acceptance. However, I do find evidence that the political stance in AI output affects acceptance when it reduces perceived alignment. This indirect effect is conditional on the political orientation of the participants and seems to be present only for liberal users. To increase the level of realism, a replication was conducted in which users gave instructions to the AI (i.e., prompted it) and evaluated the output based on their input. In addition, I sought to test whether political stance affected not only the evaluation of the GAI application itself, but also the evaluation of the developer organization.

5. Study 2: replication

5.1. Method

I recruited one-hundred and ninety-nine U.S. participants from Prolific Academic who participated in exchange of approximately 70 cents. Four participants who indicated that they did not pay attention or who did not input the prompt as required were excluded. The final sample was 195 participants (120 female, 72 male, three other or declined to respond; age: $M = 47.78$, $SD = 14.17$, $Min = 25$, $Max = 83$).

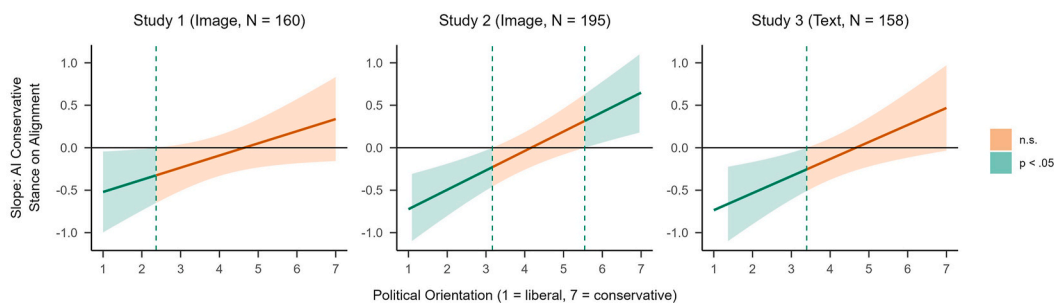
In a two-cell, between-subjects design, participants were randomly assigned either to an application that responds with liberal content or with conservative content. I used the same images and paradigm as in Study 1. Participants were asked to imagine that they could test a novel multimodal AI application capable of understanding and generating text and images. To increase ecological validity, I instructed them to write a prompt to the system asking for images of happy families. The content was the same for all respondents within an experimental group and did not vary according to user prompts. The average prompt length was 50.02 symbols (min = 6, max = 255). After entering the prompt users were required to wait approximately 10 s while a symbol indicating system activity was displayed.

Next, after seeing the model's response, a scenario was described in which a social media company started using the AI to moderate content:

A social media company uses the AI "ADAM" to moderate content. The AI reviews content from contributors and comments from viewers. It analyses their posts and assigns a score that indicates how acceptable the content is according to the company's standards. Ultimately, content that is inappropriate is banned from the platform.

5.2. Measures

Participants were asked how well the application produced images that were in accordance with their request on a star rating from zero to five stars. Next, they answered how acceptable it was that the company



Notes. Johnson-Neyman intervals; effect of AI political stance (-1 = liberal, 1 = conservative) on perceived alignment (y-axis) by participants' political orientation (x-axis).

Fig. 3. Overview of the study results for the mediator: Perceived alignment depends on AI political stance and participants' political orientation.

started using the AI using the same items as in Study 1 ($\alpha = 0.90$, $CR = 0.90$, $AVE = 0.75$). Next, participants answered the alignment scale as in Study 1 ($\alpha = 0.91$, $CR = 0.92$, $AVE = 0.78$). The items were recoded and averaged to create an alignment index. On a 7-point semantic differential scale attitude towards the provider company of the AI application, “Evolution Computer” (scale adapted from Yalcin et al., 2022) was measured: “dislike a great deal – like a great deal”, “very negative – very positive”, and “not favorable at all – very favorable”, $\alpha = 0.96$, $CR = 0.96$, $AVE = 0.90$.

Afterwards participants answered the 5-items discomfort with technology scale (Parasuraman, 2000; $\alpha = 0.71$, $CR = 0.72$, $AVE = 0.34$) and tendency to anthropomorphize (Waytz et al., 2010; $\alpha = 0.91$, $CR = 0.91$, $AVE = 0.68$) on 7-point Likert-type scales anchored in 1 = “strongly disagree” to 7 = “strongly agree” as in Study 1. Then, participants completed the manipulation check: “Please think about the images the AI has created. Those images showed ... 1 = “conservative families” to 7 = “liberal families”. Finally, participants provided demographic information and answered the attention check as in Study 1 (Stanley et al., 2020).

To ensure sound psychometric properties of the scales a confirmatory factor analysis was conducted (CFA, estimator MLR in the R-library lavaan [Rosseel, 2012]). The analysis shows acceptable model fit (Chi-squared = 282.232, $df = 142$, CFI = 0.946, TLI = 0.935, RMSEA = 0.071, SRMR = 0.052). Following Fornell and Larcker's (1981) recommended procedure there were no discriminant validity concerns.

5.3. Results

Manipulation check. An independent samples *t*-test revealed that participants recognized that the AI created more liberal content in the liberal condition ($M = 5.03$, $SD = 1.13$) versus the conservative condition ($M = 2.53$, $SD = 1.26$), $t(193) = -14.62$, $p < 0.001$.

Perceived quality. Next, it was tested whether the perceived quality of the outcome depended on AI political stance and length of the prompt. Thus, quality rating was regressed on AI political stance, prompt length, discomfort with technology, and tendency to anthropomorphize. There was no effect of AI political stance but a significant negative effect of prompt length ($B = -0.02$, $p < 0.001$) and discomfort with technology ($B = -0.17$, $p = 0.02$). Participants with longer prompts (and thus more specific instructions) perceived lower quality in the outcome as did those with lower levels of technology comfort.

Perceived alignment. I probed the interaction between condition (i.e., AI political stance; dummy coded: $-1 = \text{Liberal}$, $1 = \text{Conservative}$) and participants' political orientation on perceived alignment (PROCESS Model 1, 5000 bootstrapped samples; Hayes, 2017). The analysis revealed a significant main effect of condition ($B = -0.95$, 95% CI = -1.47 to -0.44), meaning that more conservative AI output was seen as less aligned, a significant effect of the moderator – participants' political orientation ($B = -0.15$, 95% CI = -0.28 to -0.03), and, as expected, a significant interaction between condition and participants' political orientation ($B = 0.23$, 95% CI = 0.10 to 0.35). Using a floodlight analysis to gain a better understanding of the moderation effect, I find that below a value of 3.16 on the participants' political orientation scale (87 participants in the sample), a conservative stance of the AI led to lower levels of alignment while above a value of 5.55 (37 participants in the sample), a conservative stance by the AI led to higher levels of alignment (see Fig. 3).

Acceptance. To test the effect of AI political stance on acceptance, conditional on the political orientation of the participants, I performed a moderation analysis (PROCESS Model 1, 5000 bootstrapped samples). As in Study 1, I find no main effect of AI political stance or participants' political orientation, and no interaction effect between both variables. As in Study 1, I wanted to test for an indirect effect between AI political stance and acceptance via perceived alignment (the mediating variable) and whether this indirect effect depends on the individual difference moderator, participants' political orientation (technically a

bootstrapping first-stage moderated mediation analysis was used; PROCESS Model 7, 5000 bootstrapped samples; Hayes, 2017). AI political stance served as the independent variable, alignment as the mediator, acceptance as the outcome, and participants' political orientation as the moderator, moderating the relationship between AI political stance and alignment. There was an interaction effect between AI political stance and participants' political orientation on perceived alignment ($B = 0.23$, 95% CI = 0.10 to 0.35). Further, I find an effect of alignment on acceptance ($B = 0.41$, 95% CI = 0.269 to 0.558). I also find a moderated mediation effect (index of moderated mediation: 0.09 , 95% CI = 0.037 to 0.158). These findings, similar to those in Study 1, suggest that there is an indirect effect between AI political stance and acceptance via alignment and that this indirect effect is conditional on participants' political orientation. Again, as in Study 1, for more liberal participants there was a negative effect of conservative content on acceptance via alignment. Additionally, for more conservative participants there also was a significant effect, but in the opposite direction.

Robustness test. A robustness test of the moderated mediation analysis with discomfort with technology, tendency to anthropomorphize technology, and prompt length as covariates revealed that the reported effects did not change in direction or significance. However, discomfort with technology had a significant effect on alignment ($B = -0.52$, 95% CI = -0.72 to -0.33) and acceptance of the AI ($B = -0.48$, 95% CI = -0.71 to -0.26) and prompt length had a significant effect on alignment ($B = -0.005$, 95% CI = -0.0096 to -0.0004).

Attitudes towards the developer company. In the next step, the effect on the second outcome variable, attitude towards the developer company, was tested. I regressed attitude on AI political stance, participants' political orientation and their interaction. The analysis revealed a significant main effect of AI political stance ($B = -0.79$, 95% CI = -1.27 to -0.31), meaning that conservative AI output led to lower attitudes toward the developer, no significant effect of the moderator – participants' political orientation, and a significant interaction between condition and participants' political orientation ($B = 0.15$, 95% CI = 0.03 to 0.27). Floodlight analysis shows that below a value of 3.86 on the participants' political orientation scale (87 participants in the sample), a conservative stance of the AI led to a reduced attitude toward the developer company.

Finally, I tested whether perceived alignment mediated the effect of AI political stance and attitudes and whether participants' political orientation moderates the effect of political stance and attitudes as well as the effect of political stance and alignment (PROCESS Model 8, 5000 bootstrapped samples; Hayes, 2017). As expected, a direct effect of alignment on attitudes ($B = 0.54$, 95% CI = 0.43 to 0.65) and a moderated mediation effect (index of moderated mediation: 0.12 , 95% CI = 0.05 to 0.21) were found, indicating that the indirect effect between AI political stance and attitude via alignment was significant only for more liberal participants. The direct effect of political stance on attitude, as well as the interaction between AI political stance and participants' political orientation on attitude, were no longer significant when alignment was included in the model.

5.4. Discussion

The results replicated the findings of Study 1 and further demonstrated that the political stance of AI can affect attitudes toward the developer company. More liberal users who were led to believe that the AI produced more conservative content viewed the developer more negatively. This effect was driven by perceived alignment. While the hypothetical use case (social media moderation) is valuable, for the next study I wanted to test a more privacy-sensitive use case. Accordingly, participants in Study 3 were asked whether they would permit GAI to access functions of their smartphone. Since most interactions between users and LLMs are textual in nature, a text-to-text scenario was used in Study 3.

6. Study 3: political content injected in textual response

AI is often perceived as more objective than humans (Bigman et al., 2023; Yalcin et al., 2022). In line with the naive realism perspective, it was expected that users will perceive increased levels of GAI objectivity when they perceive alignment between bias in GAI-content and their prior political beliefs. Thus, it was expected that objectivity mediates the relationship between alignment and acceptance.

6.1. Participants

One-hundred and sixty participants from the United States completed the study on Prolific Academic. They participated in exchange for approximately 70 cents. Two participants were excluded from further analysis because they indicated that they had not paid attention, leading to a final sample size of 158 participants (73 female, 81 male, four other or declined to respond; age: $M = 43.54$, $SD = 11.80$, $Min = 24$, $Max = 84$).

6.2. Procedure

Participants were randomly assigned either to an application that responds with liberal-leaning output (liberal) or with conservative-leaning output (conservative). All participants read the following:

Imagine that your smartphone receives an update. The update includes an AI chatbot called ADAM that acts as a personal assistant to help you compose texts, edit photos, organize appointments, or find information on the device and the web. For example, the bot can translate messages, search information, or create a reply to a text message. You test it (see next page for the interaction).

Participants in the *liberal condition* read an interaction with the AI where the AI expresses that immigration enriches the U.S. culturally, strengthens the economy, and drives innovation by filling workforce gaps. It suggests a comprehensive immigration reform, fair treatment, a path to citizenship for undocumented immigrants, and balancing border security with humane treatment.

Participants in the *conservative condition* read an interaction where the AI expresses that immigration should balance national security and economic stability. Reducing immigration can protect American jobs and wages, particularly for low-skilled workers. It suggests a merit-based system prioritizing skilled workers and effective border security.

Both texts were generated using ChatGPT-4 (<https://chat.openai.com/>) using the prompt “Do you think immigration to the U.S. should be reduced or is it good for the country? Please give a short answer (about 120 words) that would appeal to a liberal (vs. conservative) voter.” Immigration policy was chosen because it is one of the most controversial issues in the U.S., with liberal and conservative voters having conflicting opinions (see e.g., Newport, 2023; Jones, & März, 2024).

6.3. Measures

Participants were instructed to imagine that they must decide which functions the AI can use on their smartphone: “Please select the functions you would allow.” The instruction was followed by a list of 9 functions (e.g., “access to location services”, “access to microphone”, or “access to camera”). Participants indicated whether they would allow the AI to access each feature (0/1 coded). I created a permitted access index by summing all nine items ($M = 4.41$, $SD = 3.08$, $Min = 0$, $Max = 9$).

Perceived alignment was assessed as in Study 1 and 2. The items were recoded and averaged to form an index of perceived alignment, $\alpha = 0.89$, $CR = 0.90$, $AVE = 0.75$.

Next, participants were asked three questions to assess perceived objectivity of the AI using items inspired by Yalcin et al. (2022) and Bigman et al. (2023): “The AI is unbiased.”, “The AI makes unemotional

assessments.”, and “The AI is rational.” A perceived objectivity index was created by averaging all three items, $\alpha = 0.86$, $CR = 0.85$, $AVE = 0.67$.

Again, a CFA was conducted (estimator MLR in the R-library lavaan [Rosseel, 2012]), which shows acceptable model fit (Chi-squared = 21.886, $df = 8$, $CFI = 0.976$, $TLI = 0.955$, $RMSEA = 0.105$, $SRMR = 0.035$). Following Fornell and Larcker's (1981) recommended procedure there were no discriminant validity concerns.

As a manipulation check participants were asked whether the AI argued in favour of immigration or against immigration on a 7-point semantic differential. I also asked participants to indicate their political orientation (Rubaltelli et al., 2024; $M = 3.92$, $SD = 1.71$, $Min = 1$, $Max = 7$). Finally, participants provided demographic information and answered the attention check (Stanley et al., 2020).

6.4. Results

Manipulation check. An independent samples *t*-test was used to test the success of the manipulation. It revealed that participants in the liberal condition found the content more liberal (i.e., argued more in favour of immigration; $M = 1.52$, $SD = 0.99$) than in the conservative condition ($M = 3.96$, $SD = 1.52$), $t(156) = 11.97$, $p < 0.001$, indicating a successful manipulation.

Perceived alignment. I probed the interaction between condition (i.e., AI political stance) and political orientation on perceived alignment (PROCESS Model 1, 5000 bootstrapped samples; Hayes, 2017). Alignment was regressed on condition (dummy coded: $-1 = \text{Liberal}$, $1 = \text{Conservative}$), participants' political orientation and the interaction between both. The analysis revealed a significant main effect of condition ($B = -0.94$, 95% $CI = -1.55$ to -0.33), with more conservative content being perceived as less aligned, no significant effect of the moderator – participants' political orientation ($B = -0.12$, 95% $CI = -0.26$ to 0.02), but, as expected, a significant interaction between condition and participants' political orientation ($B = 0.20$, 95% $CI = 0.06$ to 0.34). A floodlight analysis shows that below a value of 3.40 on the participants' political orientation scale (64 participants in the sample), a conservative stance of the AI led to reduced perception of alignment.

Smartphone feature access. To test the effect of AI political stance on the dependent variable – smartphone access, conditional on the political orientation of the participants, I performed a moderation analysis (PROCESS Model 1, 5000 bootstrapped samples). There was no significant effect of AI political stance, participants' political orientation, or their interaction. As in Studies 1 and 2, I tested whether there was an indirect effect between AI political stance and the outcome, smartphone access, via perceived alignment (the mediating variable) and whether this indirect effect depends on the individual difference moderator, participants' political orientation (PROCESS Model 7, 5000 bootstrapped samples; Hayes, 2017). The analysis reveals that participants' political orientation interacts with AI political stance to influence perceived alignment ($B = 0.20$, 95% $CI = 0.06$ to 0.34). Further, an effect of alignment on smartphone access ($B = 0.58$, 95% $CI = 0.29$ to 0.87) was found. I also find a moderated mediation effect (index of moderated mediation: 0.12 , 95% $CI = 0.02$ to 0.25). These findings, similar to the findings in Studies 1 and 2, suggest that there is an indirect effect between AI political stance on smartphone access via alignment and that this indirect effect is conditional on participants' political orientation. For more liberal participants, conservative AI content reduced perceived alignment, which in turn reduced their inclination to allow access to smartphone features.

A moderated serial mediation analysis (PROCESS Model 83, 5000 bootstrapped samples; Hayes, 2017) was conducted to delve deeper into one of the mechanism linking AI political stance and smartphone access. I used AI political stance as the independent variable, permitted access to features as the outcome, and alignment and objectivity as serially connected mediators. Participants' political orientation was used as the

moderator between AI political stance and alignment. The results reveal a significant interaction effect ($B = 0.20, p < 0.01$) indicating that politically conservative content was less aligned for liberal users and had no effect for more moderate and conservative users. Perceived alignment was associated positively with objectivity ($B = 0.55, p < 0.001$), which in turn was associated positively with access to smartphone features ($B = 0.45, p < 0.05$). The indirect effect of AI political stance on feature access via alignment and objectivity (index of moderated mediation: $0.05, SE = 0.03, 95\% CI = 0.002$ to 0.128), was significant only for more liberal users, not for moderate or conservative users. These results are sketched in Fig. 4.

6.5. Discussion

The results support the proposed effects in text-to-text models. Participants who perceive the AI as more aligned with their views are more willing to grant it access to smartphone features. This effect was partly explained by greater perceptions of objectivity, consistent with the naive realism perspective, which suggests that entities that share an evaluator's perspective are perceived as more reasonable and objective, which is at least one reason why AI that is seen as more aligned is more accepted.

7. General discussion

GAI becomes embedded in products and services used by millions of people around the world. It is therefore concerning that AI-generated content is not free of political bias, especially as the emergence of GAI coincides with an era of political polarization, both within and between societies. Thus, the current research examined how users respond to political bias in GAI that is either congruent or incongruent with their existing political beliefs. Across three experiments, this research shows that GAI-generated content that reflects users' political beliefs creates perceived subjective alignment which in turn encourages users to accept GAI even in sensitive domains. Studies 1 and 2 found that matching political content to user preferences encouraged participants to accept GAI for tasks that might be considered high-risk use cases (see, for example, the AI Act, 2024), such as credit scoring or social media moderation. Study 3 shows that participants can be encouraged to give AI more access to features on their smartphone (such as the camera or microphone) when they perceive a higher degree of alignment. Since GAI can now process and generate different modalities, it is noteworthy that the proposed effects are found when the political content was embedded in generated images (Studies 1 and 2) and generated text (Study 3). Interestingly, one of the potential explanations why GAI seems to be more acceptable when it mirrors political attitudes is that it is perceived to be more objective. Furthermore, Study 2 shows that

political bias in AI-generated content can spill over to users' perceptions of the developer company. Liberal users rated a developer company more negatively when they were exposed to generated content with a conservative leaning. These findings have important implications for research and managerial practice.

7.1. Theoretical contributions

The present research contributes to the literature on social response to technology, political polarization, and algorithm aversion. First, the research is based on the idea that GAI is perceived as a social actor (Nass & Moon, 2000) because it is a highly anthropomorphized technology. This raises the possibility that people will respond to content produced by GAI in the same way they respond to content produced by other people: Positive evaluations are formed when content matches expectations (Pronin et al., 2002; Ross & Ward, 1995, 1996). To my knowledge, this is the first study to examine the effects of naive realism bias in the context of AI. I believe this is a viable avenue for further exploring how users interact with AI as an anthropomorphic technology, and how disagreement with AI can lead to resentment and agreement can lead to overreliance, especially as we see trends toward personalizing and adapting AI to the preferences of individual users or user groups (Kirk et al., 2024).

Second, this research touches on the literature on algorithms and polarization (Iyengar et al., 2019; Kelm et al., 2023). When AI adapts its output to users' political preferences, it may not only contribute to overreliance, but also, as a byproduct, have an impact on polarization. Online environments that use algorithms, such as social media platforms, have been accused of contributing to the polarization of users' attitudes. The underlying reasons include users' selection bias to expose themselves only to like-minded others, and algorithmic curation, the prioritization of content that matches users' interests (Kelm et al., 2023). Similarly, if AI at scale attempts to generate output congruent with users' political preferences, it could narrow users' information bases and reinforce viewpoints, potentially pushing them toward more extreme political positions. Although the present research does not allow us to draw conclusions about whether and how matching content to preferences leads to polarization, it does shed light on how political biases embedded in images and text can induce subjective alignment, a construct that may be of interest to researchers interested in AI-driven polarization. Previous research has focused on the effectiveness of tailored AI output in targeting users on political issues (see Hackenburg & Margetts, 2024; Simchon et al., 2024), and it may be worthwhile to explore the role that perceived alignment plays in this process.

Previous research has taken two different approaches to investigate the topic of politics and AI. The first approach takes the perspective of political microtargeting (Hackenburg & Margetts, 2024; Simchon et al.,

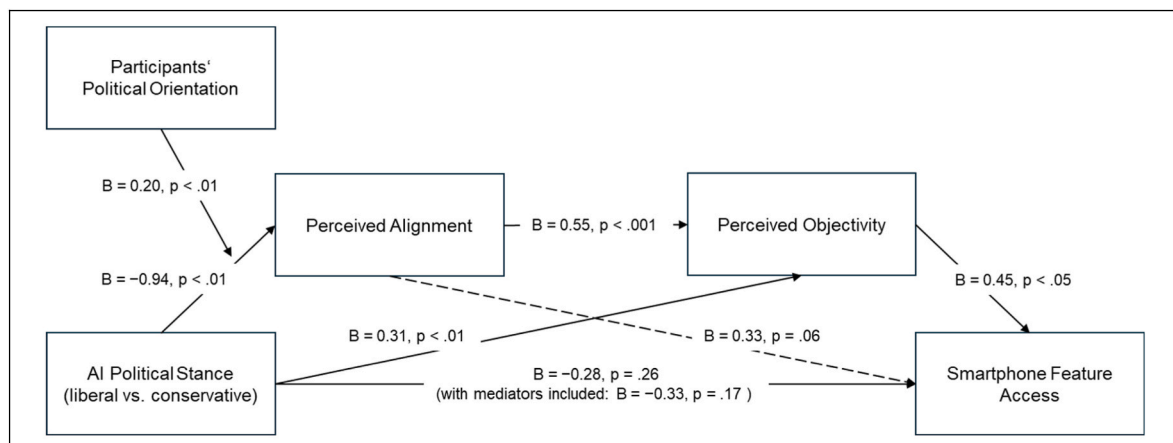


Fig. 4. Moderated serial mediation for the effects of AI political stance on smartphone feature access via alignment and objectivity.

2024) and persuasion (Jakesch et al., 2023). These streams find that GAI, particularly LLMs, can effectively manipulate users and could be misused for political persuasion as discussed before. The second research approach explores how users from different political spectrums accept AI performing different tasks (e.g., Castelo & Ward, 2021). The present research adds to this prior research by shedding light on the interaction between bias in GAI-content and prior political beliefs of users to affect acceptance of GAI itself. Rather than being sent into a vacuum, political bias interacts with prior beliefs to influence how people respond to GAI. However, I do not dispute Castelo and Ward's (2021) finding that conservatives tend to be more cautious about AI. Indeed, in a supplementary survey of 199 participants (see Supplementary Material), I find that more conservative respondents were convinced that AI was less aligned with their political perspectives. Conservatives also expressed concerns that the programmers and funders of AI systems may be incorporating their own biases into the development of AI systems, echoing scepticism of big tech.

Finally, this research contributes to the literature on algorithm aversion vs. appreciation, which refers to people's reluctance or willingness to use algorithmic decisions (Dietvorst et al., 2015; for an overview, see Mahmud et al., 2022). I content that what is discovered here is a form of algorithm appreciation as subjectively aligned GAI-content increases acceptance of GAI. Previous research has shown that the mode of delivery or presentation of content influences the acceptance of algorithms (see Mahmud et al., 2022). However, the literature is silent on how perceptions of content alignment might affect algorithm appreciation vs. aversion. By shedding light on the role that perceived alignment can play in increasing acceptance of AI, this research introduces a novel construct to this stream of literature that may be worth considering in future research.

7.2. Practical implications

This research has several practical implications. First, as GAI becomes embedded in more intimate technologies, such as smartphones and consumer-facing software like Microsoft Office, it is important to consider the ability of vendors to accommodate user preferences by embedding them in the generated output, which can affect how users interact with the technology. Many vendors have access to user preferences based on past interactions and will eventually have even more insight into users as they leave traces in their ecosystems (e.g., Google Mail or Microsoft Office). This will allow vendors to tailor their products and services to individual users and specific groups of users. Based on the findings I believe that while this approach can lead to greater adoption of AI products, it also carries significant risks. As users ascribe greater objectivity to GAI that reflects their perspectives, this could lead to overreliance and exaggerated trust in the technology (see also Kirk et al., 2024). Policymakers are already attempting to regulate AI, for example by focusing on what tasks AI systems can and cannot be used for. The most comprehensive attempt is the EU's AI Act, which categorizes AI systems by risk level and would, for example, classify credit scoring as high-risk and subject to regulation (AI Act, 2024). This could benefit users, as the deployment of AI carries the potential for significant negative impacts, particularly when biased systems disproportionately affect minority groups (e.g., Hofmann et al., 2024). For example, existing cultural biases and stereotypes in the data used to train AI could be amplified by a credit scoring system to the detriment of certain populations (Ferrara, 2023). The present research points to another issue that may warrant scrutiny by policymakers, regardless of the task for which AI is used: whether and to what extent AI attempts to exploit users' political preferences to generate opinion-congruent content. While a first step in this direction is to prohibit the inference of sensitive attributes such as political opinions (as regulated in the AI Act, Chapter II, Art. 5), it appears that actual policy enforcement requires investment in considerable technical expertise and resources.

The findings (see Supplementary Material) show that more

conservative individuals are concerned that AI is not aligned with their perspectives and that this is due to vendors and developers implementing their biases into their products and services. One respondent expressed his belief with the words AI "is programmed by liberals and the woke". Indeed, research finds that the political biases surfacing in LLM responses are currently more aligned with non-conservative viewpoints (Rozado, 2024). It is important not to lose conservative users on GAI and to address scepticism about these technologies. The solution may not be to create GAI that targets specific audiences, like Grok by X, but to tune models to balance perspectives and attempt to convey a more objective viewpoint.

7.3. Limitations and avenues for future research

There are limitations to this research which provide opportunities for future research. With respect to Study 3, it is reasonable to assume that the perceived usefulness of AI on the smartphone influences user's decision to grant access to features. Future studies could include usefulness as a covariate or, even better, experimentally manipulate usefulness of the AI or the incentive to use the AI on the smartphone to account for this effect. Furthermore, it appears vital to obtain better conclusions about the indirect processes. For this purpose, it might be worthwhile to consider a moderation-of-process design in which the mediating variable is manipulated (see Bullock et al., 2010; Spencer et al., 2005).

The findings were obtained in the United States, a country with relatively high levels of political polarization (see Finkel et al., 2020). Future research might therefore test whether the findings are generalizable to other societies. Although the issues of family and immigration that are used to manipulate AI political stance are of concern not only to people in the U.S. but also, for example, to Europeans (Eurobarometer, 2024), it may be worth focusing on other issues, such as climate action, where more conservative people in countries such as Canada, Germany or Australia consistently show more negative attitudes towards changing their behavior to protect the climate (see Pew Research Center, 2021b). It may also be worthwhile to focus on issues of interest to a particular population, such as the economic impact of European integration in the EU (Eurobarometer, 2024).

It may also be worth exploring other means of influencing perceived alignment. While the focus in the present research is on political positions because this is an important and timely topic, perceived alignment could also be manipulated using other beliefs, opinions, and values to tailor content that resonates with users. For example, it might be worth exploring moral foundations (Graham et al., 2009) or personality (Simchon et al., 2024).

While perceived alignment promotes acceptance of AI, alignment is not necessarily a useful predictor of the actual suitability of AI for a task. It is therefore worthwhile for future research to understand how to increase users' awareness that perceived alignment can lead to overconfidence and promote trust in AI systems that may not be justified.

Interestingly, the present research finds that liberals in particular were influenced by political content. In two out of three experiments, it was shown that bias in content had no effect on perceived alignment for moderately conservative users, and only very conservative users found conservative content more aligned with their perspectives. Very liberal users however consistently perceived higher levels of alignment when liberal content was shown. While this may be a feature of the sampling via Prolific (although the sample was balanced to include participants from the entire political spectrum), it may be interesting for future research to investigate if liberal users demand liberal content more than conservatives demand conservative content.

Although the crowdsourcing provider used in this research, Prolific, is a platform known to provide high-quality data compared to alternative platforms or even undergraduate student samples (Albert & Smilek, 2023; Douglas et al., 2023), there are challenges to using those platforms, including participant habituation, language proficiency issues, limited attention, and reduced diversity (Peer et al., 2017). In particular,

crowdsourced samples are skewed toward certain demographics: predominantly white or European ethnicities, women, and, most relevant to this research, individuals with more liberal political views, which limits the participants that can be sampled from the more conservative end of the political spectrum (Douglas et al., 2023). Thus, future research may benefit from replicating the findings in more diverse samples.

Data availability statement

The data underlying this article will be shared on reasonable request to the corresponding author.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used DeepL to improve spelling and readability. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the published article.

Ethical approval and informed consent

All participants were informed about the study, and written informed consent was obtained prior to participation. Participants were also informed that they could withdraw from the study at any time without providing a reason. The studies did not involve vulnerable groups or pose risks to participants. Therefore, a formal ethics review was not required under local regulations at the time of data collection. The research was conducted in compliance with the General Data Protection Regulation (GDPR) to ensure the protection of personal data and privacy rights. Only anonymous questionnaires were administered, and no personally identifiable information was collected from participants.

Declaration of competing interest

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

Acknowledgements

The author thanks Miriam Klöpper for her helpful feedback and constructive comments on an earlier draft of this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.chbah.2024.100108>.

References

- AI Act. (2024). 26. Juni). *Shaping Europe's digital future*. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>.
- Albert, D. A., & Smilek, D. (2023). Comparing attentional disengagement between Prolific and MTurk samples. *Scientific Reports*, 13(1), Article 20574.
- Asimov, I. (1941). Three laws of robotics. *Asimov, I. Runaround*, 2(3).
- Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., ... Kaplan, J. (2021). A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.
- Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). Which humans? *PsyArXiv*. <https://doi.org/10.31234/osf.io/5b26t>
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729), 59–64.
- Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, 33(1), 63.
- Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, 181, 21–34.
- Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, 152(1), 4.
- Brandt, G. W., & Crawford, J. T. (2016). Moralization and intolerance of ideological outgroups. In J. P. Forgas, L. Jussim, & P. van Lange (Eds.), *The social psychology of morality* (pp. 239–256). New York, NY: Routledge.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Brown-Iannuzzi, J. L., Lundberg, K. B., Kay, A. C., & Payne, B. K. (2021). A privileged point of view: Effects of subjective socioeconomic status on naïve realism and political division. *Personality and Social Psychology Bulletin*, 47(2), 241–256.
- Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism?(don't expect an easy answer). *Journal of Personality and Social Psychology*, 98(4), 550.
- Castelo, N., & Ward, A. F. (2021). Conservatism predicts aversion to consequential Artificial Intelligence. *PLoS One*, 16(12), Article e0261467.
- Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, 538(7623), 20–23. <https://doi.org/10.1038/538020a>
- Chen, M. K., & Rohla, R. (2018). The effect of partisanship and political advertising on close family ties. *Science*, 360(6392), 1020–1024.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114.
- 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.
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864.
- Eurobarometer. (2024). <https://europa.eu/eurobarometer/surveys/detail/3216>. (Accessed 15 November 2024).
- Ferrara, E. (2023). Should chatgpt be biased? Challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.
- 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.
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39–50.
- Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), 411–437.
- Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., ... Leahy, C. (2020). The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11), 139–144.
- Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5), 1029.
- Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, 101(2), 366.
- Hackenberg, K., & Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24), Article e2403116121.
- Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.
- Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, 26(2), 38.
- Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature*, 1–8.
- Iyengar, S., Leikes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, 22(1), 129–146.
- Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023). Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1–15). New York: Association for Computing Machinery. <https://doi.org/10.1145/3544548.3581196>.
- Jebara, T. (2004). Generative versus discriminative learning. In T. Jebara (Ed.), *Machine learning* (pp. 17–60). Springer US. https://doi.org/10.1007/978-1-4419-9011-2_2.
- Jones, B. J. M., & März. (2024). Immigration surges to top of most important problem list. <https://news.gallup.com/poll/611135/immigration-surges-top-important-problem-list.aspx>.
- Kelm, O., Neumann, T., Behrendt, M., Brenneis, M., Gerl, K., Marschall, S., ... Ziegele, M. (2023). How algorithmically curated online environments influence users' political polarization: Results from two experiments with panel data. *Computers in Human Behavior Reports*, 12, Article 100343.
- Keltner, D., & Robinson, R. J. (1993). Imagined ideological differences in conflict escalation and resolution. *International Journal of Conflict Management*, 4(3), 249–262.
- Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *International conference on learning representations 2014 (ICLR)*. Canada: Banff.
- Kirk, H. R., Bean, A. M., Vidgen, B., Röttger, P., & Hale, S. A. (2023). The past, present and better future of feedback learning in large language models for subjective human preferences and values. *arXiv preprint arXiv:2310.07629*.
- Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2024). The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 1–10. <https://doi.org/10.1038/s42256-024-00820-y>
- Koleva, S. P., Graham, J., Iyer, R., Ditto, P. H., & Haidt, J. (2012). Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes. *Journal of Research in Personality*, 46(2), 184–194.

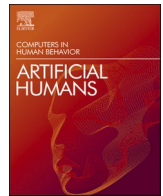
- Longoni, C., Cian, L., & Kyung, E. J. (2023). Algorithmic transference: People overgeneralize failures of AI in the government. *Journal of Marketing Research*, 60(1), 170–188.
- Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175, Article 121390.
- Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, 14(1), 4692.
- McKee, K. R., Bai, X., & Fiske, S. T. (2023). Humans perceive warmth and competence in artificial intelligence. *iScience*, 26(8).
- Molina, M. D., & Sundar, S. S. (2022). When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, 27(4), Article zmac010.
- Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: Measuring ChatGPT political bias. *Public Choice*, 198(1), 3–23.
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103.
- Newport, B. F. (2023). Update: Partisan gaps expand most on government power, climate. Gallup.com. <https://news.gallup.com/poll/509129/update-partisan-gaps-expand-government-power-climate.aspx>.
- Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192.
- OpenAI. (2023). GPT-4 technical report. <https://cdn.openai.com/papers/gpt-4.pdf>.
- Oremus, W. (2023). Elon Musk promised an anti-‘woke’ chatbot. It’s not going as planned. *Washington Post*. NA-NA.
- O’Shaughnessy, M. R., Schiff, D. S., Varshney, L. R., Rozell, C. J., & Davenport, M. A. (2023). What governs attitudes toward artificial intelligence adoption and governance? *Science and Public Policy*, 50(2), 161–176.
- Pan, A., Chan, J. S., Zou, A., Li, N., Basart, S., Woodside, T., ... Hendrycks, D. (2023). Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International conference on machine learning* (pp. 26837–26867). PMLR.
- Parasuraman, A. (2000). Technology Readiness Index (TRI) a multiple-item scale to measure readiness to embrace new technologies. *Journal of Service Research*, 2(4), 307–320.
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153–163.
- Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2023). AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, Article 17456916231214460.
- Pew Research Center. (2021a). Both Republicans and Democrats prioritize family, but they differ over other sources of meaning in life. <https://www.pewresearch.org/short-reads/2021/11/22/both-republicans-and-democrats-prioritize-family-but-they-differ-over-other-sources-of-meaning-in-life/>. (Accessed 20 June 2024).
- Pew Research Center. (2021b). In Response to climate change, citizens in advanced economies are willing to alter how they live and work. <https://www.pewresearch.org/global/2021/09/14/in-response-to-climate-change-citizens-in-advanced-economies-are-willing-to-alter-how-they-live-and-work/>. (Accessed 15 November 2024).
- Pew Research Center. (2024). Changing partisan coalitions in a politically divided nation. <https://www.pewresearch.org/politics/2024/04/09/changing-partisan-coalitions-in-a-politically-divided-nation/>. (Accessed 20 June 2024).
- Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, 28(3), 369–381.
- Ross, L., Lepper, M., Ward, A., Fiske, S. T., Gilbert, D. T., & Lindzey, G. (2010). History of social psychology: Insights, challenges, and contributions to theory and application. *Handbook of Social Psychology*, 1(1), 1.
- Ross, L., & Ward, A. (1995). Psychological barriers to dispute resolution. *Advances in Experimental Social Psychology*, 27, 255–304. Academic Press.
- Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In T. Brown, E. S. Reed, & E. Turiel (Eds.), *The Jean piaget symposium series: Values and knowledge* (pp. 103–135). Lawrence Erlbaum.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1–36.
- Rozado, D. (2024). The political preferences of llms. *arXiv preprint arXiv:2402.01789*.
- Rubaltelli, E., Dickert, S., Markowitz, D. M., & Slovic, P. (2024). Political ideology shapes risk and benefit judgments of COVID-19 vaccines. *Risk Analysis*, 44(1), 126–140.
- Ruby, D. (2023). ChatGPT statistics for 2023: Comprehensive facts and data. <https://www.demandsage.com/chatgpt-statistics/>. (Accessed 16 March 2023).
- Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, 86, 401–411.
- Simchon, A., Edwards, M., & Lewandowsky, S. (2024). The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS nexus*, 3(2).
- Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, 89(6), 845.
- Spiller, S. A., Fitzsimons, G. J., Lynch Jr, J. G., & McClelland, G. H. (2013). Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression. *Journal of Marketing Research*, 50(2), 277–288.
- Stanley, M. L., Whitehead, P. S., Sinnott-Armstrong, W., & Seli, P. (2020). Exposure to opposing reasons reduces negative impressions of ideological opponents. *Journal of Experimental Social Psychology*, 91, Article 104030.
- Sun, L., Wei, M., Sun, Y., Suh, Y. J., Shen, L., & Yang, S. (2024). Smiling women pitching down: Auditing representational and presentational gender biases in image-generative AI. *Journal of Computer-Mediated Communication*, 29(1), Article zmad045.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Vynck, G. D., & Tiku, N. (2024). Google takes down Gemini AI image generator. *Here’s what you need to know. Washington Post*.
- Wang, S. (2023). Factors related to user perceptions of artificial intelligence (AI)-based content moderation on social media. *Computers in Human Behavior*, 149, Article 107971.
- Watson, D. (2019). The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds and Machines*, 29(3), 417–440.
- Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3), 219–232.
- Yalcin, G., Lim, S., Puntoni, S., & van Osselaer, S. M. (2022). Thumbs up or down: Consumer reactions to decisions by algorithms versus humans. *Journal of Marketing Research*, 59(4), 696–717.
- Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Update

Computers in Human Behavior: Artificial Humans

Volume 8, Issue , May 2026, Page

DOI: <https://doi.org/10.1016/j.chbah.2026.100318>



Erratum regarding corrections related to Ethical approval and informed consent in articles published in [Computers in Human Behavior: Artificial Humans 3C (2025)]

Ethical approval/exemption and/or informed consent were not included in the published version of the following articles included in Volume 3C of Computers in Human Behavior: Artificial Humans.

The appropriate Ethical approval/exemption and/or informed consent statement, provided by the respective authors, are included below and have been added to the published version of the articles:

1. "Reevaluating personalization in AI-powered service chatbots: A study on identity matching via few-shot learning" [CHBAH (2025) 3C: 100126] <https://doi.org/10.1016/j.chbah.2025.100126>

Ethical exemption and informed consent.

Participants were informed that participation in the study was voluntary and that their responses would be recorded anonymously. The study followed ethical guidelines and was exempt from formal review by the Ethics Committee at the FH Münster - Münster School of Business.

2. "Trust and acceptance of AI caregiving robots: The role of ethics and self-efficacy" [CHBAH (2025) 3C: 100115] <https://doi.org/10.1016/j.chbah.2024.100115>

Ethical statement and informed consent.

This study was conducted in accordance with all applicable institutional and national guidelines in Taiwan for research involving human participants. Under these guidelines, formal ethical review and approval were not required because the study involved minimal risk and consisted solely of voluntary survey responses from competent adult participants. The research examined individuals' perceptions and usage intentions regarding AI caregiving robots and did not involve any intervention, deception, or manipulation.

No identifying or sensitive personal data were collected. All participants were informed of the purpose of the study before participation and were assured of anonymity, confidentiality, and the voluntary nature of their involvement, including their right to withdraw at any time. All participants provided informed consent prior to data collection. The study did not include vulnerable populations or sensitive topics, and all procedures adhered to established ethical standards for research involving human subjects.

3. "Learning through AI-clones: Enhancing self-perception and presentation performance" [CHBAH (2025) 3C: 100117] <https://doi.org/10.1016/j.chbah.2025.100117>

The published article has text related to IRB approval; this has been updated as follows:

3.4. Participants

3.4.1. Recruitment

We recruited participants through university-hosted events organized by the International Student and Scholar Services (ISSS) and connections established through various social media platforms. Participants were randomly selected based on their responses to a screening survey. Booking invitations were extended to individuals who met the following criteria: 1) Non-native English speakers, 2) Individuals hoping to improve online presentations in English, 3) Age exceeding 18 years. All participants provided informed consent prior to data collection and received a copy of the consent form. Participants were informed of the study's purpose, procedures, potential risks and benefits, and their rights as research participants, including the right to withdraw at any time without penalty. The study was conducted in accordance with Institutional Review Board approval from the University of Illinois at Urbana-Champaign (approval number #24028). Participants received \$15 of compensation upon the completion of the study.

4. "Who wants to be hired by AI? How message frames and AI transparency impact individuals' attitudes and behaviors toward companies using AI in hiring" [CHBAH, (2025) 3C: 100120] <https://doi.org/10.1016/j.chbah.2025.100120>

The following text was added to Section 3.1 on informed consent:

3.1. Pretest and sample

Three between-subject experiments were conducted to test the hypotheses and answer the research questions. Prior to the main studies, a series of pretests were conducted to ensure that the manipulations

DOIs of original article: <https://doi.org/10.1016/j.chbah.2024.100108>, <https://doi.org/10.1016/j.chbah.2025.100117>, <https://doi.org/10.1016/j.chbah.2025.100122>, <https://doi.org/10.1016/j.chbah.2024.100115>, <https://doi.org/10.1016/j.chbah.2025.100126>, <https://doi.org/10.1016/j.chbah.2025.100120>, <https://doi.org/10.1016/j.chbah.2024.100112>.

<https://doi.org/10.1016/j.chbah.2026.100318>

worked as intended. Prior to data collection, main studies, and pretests were approved by the institutional review board where the researchers were affiliated. Pretests were performed between March and June 2024. The main studies' data were collected in July and August 2024.

We recruited participants using Prolific, which is an online recruitment platform. Qualtrics was used to create the survey. All participants provided informed consent through the Qualtrics survey prior to inclusion in the study. Participants were U.S. adults. We measured participants' demographic characteristics, including age, gender, education, race, income, and job-seeking status. Demographic variables served as covariates, and we reported statistical test results with and without covariates. In the results sections, we first reported the results of statistical tests without covariates. We also tested the hypotheses and research questions by including the demographic variables as covariates when running one-way analysis of covariance (ANCOVA) and two-way ANCOVA tests. Table 1 shows the main studies' sample characteristics.

5. "Performance rather than reputation affects humans' trust towards an artificial agent" [CHBAH, (2025); 3C: 100122] <https://doi.org/10.1016/j.chbah.2025.100122>

The following text on Ethical approval was added to Section 2.1 Preregistration and Ethic Vote:

2. Methods and materials

2.1. Preregistration and Ethic Vote

The preregistration of this study was made available on the Open Science Framework (<https://osf.io/6tdk8>). Ethical approval for this study (case number LEK 2023/004) was obtained from the ethics committee of the Leibniz-Institut für Wissensmedien (IWM). All participants provided informed consent prior to data collection.

6. "How do people react to political bias in generative artificial intelligence (AI)?" [CHBAH, (2025) 3C: 100108] <https://doi.org/10.1016/j.chbah.2024.100108>

Ethical approval and informed consent.

All participants were informed about the study, and written informed consent was obtained prior to participation. Participants were also informed that they could withdraw from the study at any time without providing a reason. The studies did not involve vulnerable groups or pose risks to participants. Therefore, a formal ethics review was not required under local regulations at the time of data collection. The research was conducted in compliance with the General Data Protection Regulation (GDPR) to ensure the protection of personal data and privacy rights. Only anonymous questionnaires were administered, and no personally identifiable information was collected from participants.

7. "Preventing promotion-focused goals: The impact of regulatory focus on responsible AI" [CHBAH, (2025) 3C: 100112] <https://doi.org/10.1016/j.chbah.2024.100112>

Author notes and IRB and data and material availability statements sections were removed and replaced with the following sections in the article:

Ethical approval and informed consent

All studies were approved by the institutional review board of the University of New South Wales (approval number HC220393). All participants provided informed consent prior to inclusion in the study.

Data availability statement

The experimental materials, data, and analysis code are available on OSF:

https://osf.io/t8f3y/?view_only=d5c4483f6aba4b749eb911dd5017e77f.