Article

SM+S
social media + society

Social Media + Society
October-December 2025: 1–14

DOI: 10.1177/20563051251405069
journals.sagepub.com/home/sms
Sage

Public Opinion on the Politics of AI Alignment: Cross-National Evidence on Expectations for AI Moderation From Germany and the United States

Andreas Jungherr[1] **and Adrian Rauchfleisch**[2]

Abstract

Recent advances in generative AI have raised public awareness, shaping expectations and concerns about their societal implications. Central to these debates is the question of AI alignment—how well AI systems meet public expectations regarding safety, fairness, and social values. However, little is known about what people expect from AI-enabled systems and how these expectations differ across national contexts. We present evidence from two surveys of public preferences for key functional features of AI-enabled systems in Germany ($n = 1800$) and the United States ($n = 1756$). We examine support for four types of alignment in AI moderation: accuracy and reliability, safety, bias mitigation, and the promotion of aspirational imaginaries. U.S. respondents report significantly higher AI use and consistently greater support for all alignment features, reflecting broader technological openness and higher societal involvement with AI. In both countries, accuracy and safety enjoy the strongest support, while more normatively charged goals—like fairness and aspirational imaginaries—receive more cautious backing, particularly in Germany. We also explore how individual experience with AI, attitudes toward free speech, political ideology, partisan affiliation, and gender shape these preferences. AI use and free speech support explain more variation in Germany. In contrast, U.S. responses show greater attitudinal uniformity, suggesting that higher exposure to AI may consolidate public expectations. These findings contribute to debates on AI governance and cross-national variation in public preferences.

Keywords

Artificial Intelligence, alignment, moderation, survey, international comparison, safety, bias, imaginaries

Public Opinion on AI Alignment in Germany and the United States

Recent successes of AI-enabled systems[1] such as *ChatGPT*, *DALL·E*, and *Midjourney* have significantly raised public awareness of generative AI. These systems, and others like them, have made generative models more accessible and user-friendly, leading to widespread adoption and increased visibility. As a consequence, both the capabilities and limitations of these technologies have entered public discourse, fueling growing expectations as well as mounting concerns about their societal impact.

Among the key concerns emerging in this debate is the issue of AI alignment—that is, the extent to which AI-enabled systems act in accordance with the intentions and expectations of their developers (Amodei et al., 2016; Bommasani et al., 2022; Gabriel, 2020; Hendrycks et al., 2022). Whereas AI governance refers to the institutional, legal, and political frameworks that shape how AI is developed and deployed, alignment covers technical and ethical efforts to ensure that AI systems work in accordance with human intentions or values. In practice, alignment measures can be understood as one element or object of governance: public authorities and private actors seek to ensure that model outputs accord with normative and societal expectations. Our study examines public attitudes toward such alignment goals, recognizing

[1]University of Bamberg, Germany
[2]National Taiwan University

Corresponding Author:
Andreas Jungherr, Institute for Political Science, University of Bamberg, Feldkirchenstraße 21, Bamberg 96052, Germany.
Email: andreas.jungherr@uni-bamberg.de

that these attitudes may indirectly inform, but not determine, broader governance choices.

Understanding these attitudes is essential because the politics of alignment is inseparable from the politics of governance: the way people think AI systems should behave also shapes what forms of regulation they will find legitimate. Yet, despite the growing relevance of AI alignment, public attitudes toward features of AI-enabled systems remain underexplored, though such attitudes matter. Public opinion on digital governance often reflects—though not always straightforwardly—partisan affiliations and deeper ideological commitments (Jang et al., 2024; Rauchfleisch & Jungherr, 2024). This is likely to hold true for AI governance as well, although empirical insights remain scarce, especially across countries (Rauchfleisch et al., 2025).

This article addresses this research gap by presenting a comparative study of public preferences for specific goals in AI alignment: accuracy and reliability, safety, bias mitigation, and the promotion of aspirational imaginaries. We selected these features for their relevance in the larger AI alignment debate. We compare survey responses from Germany ($n = 1800$) and the U.S. ($n = 1756$) and examine the role of key factors that may shape governance preferences on the individual, group, and systems levels. In doing so, our study contributes to the expanding literature on public attitudes in digital governance more broadly and AI governance specifically (Neyazi et al., 2024; Rauchfleisch et al., 2025; Rauchfleisch & Jungherr, 2024; Riedl et al., 2021, 2022; Vogler et al., 2025).

We find that respondents in Germany and the U.S. differ significantly in their use of AI, with U.S. participants reporting higher usage levels. These patterns reflect broader differences in technological adoption between the two countries. When evaluating preferences for AI alignment, both German and U.S. participants strongly support features aiming for accuracy and reliability, as well as safety. However, support declines for interventions aimed at mitigating bias or promoting aspirational imaginaries, especially in Germany. U.S. respondents consistently show stronger support across all categories, consistent with their higher societal involvement with AI. Experience with AI, political ideology, gender, and free speech attitudes all shape individual preferences, though their associations vary by country. In Germany, both AI experience and free speech attitudes are stronger predictors of support than in the U.S., where attitudes appear more uniformly developed. Political ideology is more influential in the U.S. when it comes to aspirational portrayals. These findings suggest that in contexts with lower AI exposure, individual-level factors play a greater role in shaping attitudes, while in high-exposure contexts like the U.S., public views on AI moderation are more consolidated.

Explaining Public Preferences for Features of AI-Enabled Systems

AI Alignment

As AI-enabled systems move from research contexts into widespread professional and public use, people's preferences for how these systems are governed become increasingly important for the success of AI models, governance frameworks, and broader public trust or skepticism toward the growing integration of AI into society. This introduces an attitudinal and psychological dimension to the broader debate on AI alignment (Amodei et al., 2016; Bommasani et al., 2022; Hendrycks et al., 2022).

AI alignment broadly refers to the challenge of ensuring that AI models act in accordance with the intentions and values of their developers or deployers, rather than deviating from them in harmful or unintended ways (Gabriel, 2020; Ji et al., 2023; Leike et al., 2018; Ngo et al., 2022). This is challenging enough in settings where goals can be clearly specified, but alignment becomes even more challenging and contested when extended to the expectations of the broader public—especially given the significant variation in public attitudes, levels of expertise, and degrees of engagement (Achintalwar et al., 2024; Padhi et al., 2024). A key source of divergence is the debate over how much control developers should exert over model outputs and for what purposes that control should be applied.

Preferences in the Use of AI-Enabled Systems: Accuracy and Reliability, Safety, Bias Mitigation, and Promotion of Aspirational Imaginaries

Accuracy and Reliability. People vary in their expectations toward AI and the preferences they hold in using AI-enabled systems. On a foundational level, an important characteristic of such systems is *accuracy*. In other words, AI models should be bound by facts. However, it is essential to recognize that this can only mean facts as represented in data, not facts as such (Fourcade & Healy, 2024; Smith, 2019).

Much of public debate around generative AI centers on the risks associated with factually incorrect or misleading outputs—whether these stem from flawed training data, faulty inference, or so-called "hallucinations," in which models generate information that appears authoritative but is entirely fabricated or inaccurate (Maleki et al., 2024; Maynez et al., 2020). This is particularly problematic in domains such as search, information retrieval, public discourse, education, democratic participation, and professional usage contexts where users rely on AI for accurate and trustworthy information (Jungherr, 2023; Jungherr et al., 2024; Jungherr & Schroeder, 2023; Spitale et al., 2023; Weidinger et al., 2022).

Generative models can, for example, fabricate historical events, misattribute facts, disseminate false claims about individuals, or simply return wrong responses to factual or analytical queries. While these outputs often appear plausible, their misleading or fictional nature can result in real-world harms.

Expecting accuracy and reliability from AI-enabled systems seems a natural precondition for their widespread and systematic use. At the same time, there are good reasons to favor alignment principles that permit the adjustment of purely data-driven outputs, allowing systems to depart from the facts as represented in data.

Safety. We expect people to voice a preference for AI-enabled systems that take harm prevention and safety seriously and moderate their model outputs accordingly. Concerns about preventing *harm* and ensuring *safety* often justify intentional deviations from strictly data-driven model outputs (Anwar et al., 2024; Askell et al., 2021; Hao et al., 2023; Hendrycks et al., 2022; Phuong et al., 2024). This includes filtering outputs that could pose direct risks—such as generating instructions for building weapons, carrying out terrorist attacks, or engineering pathogens. However, safety concerns also extend to outputs that violate individual rights, such as privacy breaches or unauthorized use of intellectual property. In these instances, moderating or restricting factual content based on model-learned patterns becomes a necessary intervention.

Safety-oriented moderation is not a new concept in digital governance. Public support for such moderation is well documented in debates over digital speech. While there is variation between countries in what is perceived as severely harmful content (Jiang et al., 2021), there is general support to moderate harmful content on online platforms (Kozyreva et al., 2023; Pradel et al., 2024). Thus, the safety-focused moderation of AI systems might appear to be a comparable case. However, the impersonal and machine-generated nature of AI content may introduce new dynamics in how people perceive and evaluate such moderation. These differences could lead to different explanatory factors shaping public attitudes and preferences in the context of AI.

Bias Mitigation. Another feature people might be looking for in their preferences for AI-enabled systems is the *mitigation of bias* to promote *fairness* (Askell et al., 2021; Barocas et al., 2023; Hao et al., 2023). This intervention necessarily follows the workings of AI and therefore has no direct equivalent in discussions of speech moderation in other digital contexts. AI models do not access objective facts about the world directly, but only representations of those facts as encoded in data (Fourcade & Healy, 2024; Hand, 2004; Smith, 2019). As a result, they are prone to reproducing the biases embedded in their training data (Bianchi et al., 2023; Friedrich, Brack, et al., 2024; Friedrich, Hämmerl, et al., 2024; Hofmann et al., 2024; Tao et al., 2024; Weidinger et al., 2021). When AI systems are bound to such biased representations, their outputs may reinforce harmful distortions or inequalities.

Bias, in this context, refers to a divergence between the distribution of a variable in AI training data or outputs and its true distribution in the world. Fairness-oriented moderation seeks to adjust such outputs so that they better reflect a more accurate or equitable distribution rather than reproduce the skewed patterns found in data. This may involve correcting underrepresentation, countering stereotypes, or highlighting marginalized perspectives.

While such interventions are often normatively desirable, they are politically and ethically contested (Binns, 2018; Gabriel, 2020). Efforts to mitigate bias inevitably raise questions about what constitutes a fair or accurate representation of the world—and who gets to decide which absences or distortions in data should be corrected. In this sense, fairness-oriented alignment confronts not only technical challenges but also deeper disagreements about knowledge, representation, and justice. This, in turn, is likely to shape who supports or actively demands such interventions in AI-enabled systems.

Promoting Aspirational Imaginaries. Another reason to adjust the outputs of AI models builds on the idea that generative systems might not only show the world as it is, but also contribute to envisioning the world as it could or should be. In this view, moderation may serve to embed *aspirational imaginaries*—collectively held visions of desirable social futures (Taylor, 2004)—and thereby enact them in technical practice.

Imaginary-based alignment can serve two distinct functions. From a science and technology studies perspective, sociotechnical imaginaries are institutionally stabilized and publicly performed visions of social order that become materialized in policies, standards, and infrastructures (Jasanoff & Kim, 2015). Some approaches to the moderation of model outputs can thus be seen as an expression of dominant social orders, while other approaches used in different models might adopt alternative imaginaries that contest established orders.

One driver of such contestation draws on pragmatist and cultural theory traditions that treat imagination as a driver of moral reflection, inclusion, and social change (Dewey, 1916, 1934; Rorty, 1989). Following this view, moderation can be understood as an effort to shape AI outputs toward inclusive or solidaristic ideals rather than mirror existing distributions or inequalities—even when those reflect empirical patterns in data.

Both forms of imaginary-oriented alignment—whether expressing prevailing orders or envisioning alternative moral horizons—raise questions of legitimacy and authority: who defines these ideals, and should AI systems promote them? As a result, aspirations-based moderation represents a contested and value-laden approach within the broader debate on AI alignment, and public support for such interventions is unlikely to be universal.

Preferences for the Alignment of AI-Enabled Systems. These observations point to likely differences in the preferences for principles guiding the adjustment of outputs of AI-enabled systems. Specifically, attitudes are likely to vary depending on the underlying rationale. Moderation aimed at ensuring accuracy and reliability, or safety—such as preventing harm, illegal activity, or violations of individual rights—is likely to enjoy broad public support, as these goals align with widely accepted norms and relatively uncontroversial forms of risk prevention. In contrast, interventions designed to mitigate bias or promote aspirational imaginaries introduce greater ambiguity, normative complexity, and potential for political disagreement. Bias mitigation involves contested judgments about what constitutes fairness or representational accuracy (Binns, 2018; Gabriel, 2020), while aspirational goals go further by seeking to reshape cultural narratives or advance particular visions of a better future. Such aims can trigger concerns about legitimacy, overreach, and ideological bias. Accordingly, the motivation behind a given moderation decision is likely to influence how it is received by the public, with support declining as the rationale shifts from widely shared safety concerns to more contested and value-laden objectives.

> Research Question 1 (RQ1): To what extent do public preferences vary between different principles of AI governance, namely accuracy and reliability, safety, the mitigation of bias, or the promotion of aspirational societal values?

Explaining Preferences: The Role of Involvement

We propose a model in which preferences regarding the adjustment of AI-generated outputs are shaped by varying degrees of personal and collective involvement. This involvement can manifest on multiple levels:

> At the individual level, it may reflect personal experiences with AI and relevant related attitudes.
>
> At the group level, it may be influenced by membership in a group that is particularly affected by or sensitive to AI-based interventions.
>
> At the systemic level, it may depend on whether an individual resides in a country with a strong or weak technological infrastructure and relationship to digital innovation.

In the following, we elaborate on the rationale behind each of these dimensions.

Individual-Level Involvement: Experience and Values. Individual support for AI content moderation is shaped not only by how frequently people use AI technologies but also by the values and beliefs they bring to evaluating their use. We distinguish two dimensions of individual-level involvement with AI: experiential and ideational.

Experiential Involvement. Personal involvement with AI can take different forms. One case involves individuals who actively use AI for personal or professional reasons. These users experience the technology firsthand and can develop more elaborate and specific preferences regarding its regulation than those who do not (Horowitz et al., 2024; Horowitz & Kahn, 2021). We expect this familiarity to translate into differentiated preferences toward AI moderation.

People with limited exposure to or interaction with AI-enabled systems may approach them with greater skepticism or uncertainty. This skepticism may lead to a heightened demand for external safeguards, particularly those framed around safety concerns. Moderation aimed at reducing bias or promoting aspirational portrayals, however, may appear unnecessary or overly intrusive to individuals with low AI involvement since these interventions presuppose familiarity with how AI systems operate.

In contrast, individuals with greater hands-on experience using AI-enabled systems may have more specific views on the systems' capabilities and limitations. This familiarity may foster a greater appreciation for moderation tasks that are specific to AI-generated content—particularly interventions aimed at bias mitigation or the promotion of aspirational social values.

Ideational Involvement. In addition to usage-based experience, individual attitudes toward AI moderation are shaped by broader normative commitments and political values. AI moderation can be seen as a special case of speech governance more broadly (Dabhoiwala, 2025; Kosseff, 2023; Mchangama, 2022). As such, individual support for AI interventions is likely to reflect how people understand and prioritize free expression.

People who view free speech as a foundational democratic right may oppose moderation efforts—especially those perceived as ideological or normative in nature. In contrast, those who understand speech as something that can and should be regulated in the interest of societal fairness or safety may be more supportive of AI content moderation (Rauchfleisch & Jungherr, 2024; Riedl et al., 2021, 2022).

We therefore expect individuals who strongly support free speech to be more critical of AI moderation interventions aimed at shaping content in terms of fairness or aspirational values, while potentially supporting moderation grounded in factual accuracy or safety.

Beyond attitudes toward free speech, broader political ideology—particularly along the liberal–conservative spectrum—also informs support for different types of AI moderation. Conservatives, who tend to prioritize order, safety, and personal responsibility, may support moderation that prevents harmful or illegal content. Liberals—here referring to

the progressive or left-leaning orientation in U.S. political discourse, rather than classical or European liberalism—especially in recent years, may place greater emphasis on equity, inclusivity, and social justice, and thus may be more supportive of interventions that promote fairness (bias mitigation) or progressive values (aspirational imaginaries) (Chong et al., 2024; Chong & Levy, 2018).

> Research Question 2 (RQ2): How do individual-level factors—including personal experience with AI, support for free speech, and political ideology—shape public preferences for different goals of AI governance, such as accuracy and reliability, safety, bias mitigation, and the promotion of aspirational societal values?

Group-Level Involvement: Partisanship & Gender. People can also experience AI through the lens of group-level involvement. Such involvement occurs when group membership shapes exposure to AI technologies or to the societal debates surrounding their regulation. This study examines two forms of group-level involvement: political partisanship and gender.

Partisanship. Partisanship increasingly shapes attitudes toward digital governance. In the U.S., the issue of speech moderation—particularly on digital platforms—has become highly politicized. Republican political elites have framed moderation efforts as ideologically biased and as threats to free speech (McCabe & Kang, 2020). This discourse often targets progressive actors as overreaching in their regulation of online content. In the context of AI, this has culminated in the politicized framing of so-called "woke AI," shorthand for alleged left-leaning or socially progressive bias in automated systems (Roose, 2025). As a result, we expect Republican supporters to oppose forms of AI content moderation that are framed as ideologically motivated—such as bias mitigation and the promotion of aspirational imaginaries. However, moderation in the name of safety may find greater acceptance among Republicans, as it aligns with conservative discourses of security and protection.

In contrast, Democratic partisans are more likely to view moderation as a tool for fostering equity, inclusivity, and representation. We therefore expect them to show greater support for AI moderation focused on bias reduction and aspirational imaginaries. In Germany, digital content governance is less politicized than in the U.S. However, the Green Party has been a leading advocate of strong regulatory measures to counter misinformation, hate speech, and inequality online (e.g., Künast & Geese, 2020). Although these initiatives do not explicitly target AI alignment, they reflect a broader orientation toward normative regulation in the digital sphere. We therefore expect this tendency to carry over, with Green Party supporters showing comparatively high support for AI content moderation across all justifications. For other parties in Germany, where digital policy debates are less divided, we do not expect systematic differences.

Gender. Group-level involvement may also arise from shared experiences of harm or vulnerability. One such case is gender. Women are disproportionately exposed to digital risks (De Ruiter, 2021; Wang & Kim, 2022). These experiences may sensitize them to the potential harms of AI-generated content and increase their support for interventions designed to moderate it (Vogler et al., 2025). We therefore expect women to show greater support for AI content moderation, particularly for justifications grounded in safety and bias reduction.

> Research Question 3 (RQ3): How do group-level characteristics such as partisanship and gender shape public preferences for different goals of AI governance, such as accuracy and reliability, safety, bias mitigation, and the promotion of aspirational societal values?

System-Level Involvement: Country. Countries differ in their openness toward new technologies (Comin & Hobijn, 2010; Ding, 2024) and perceptions of technological risk (Douglas & Wildavsky, 1982). This is also evident when we examine the uses of generative AI. In the U.S.—a country with a world-leading digital technology sector and comparatively strong openness toward new technologies—33% of respondents in a survey representative of American adults said in August 2024 they had used AI-enabled chatbots like ChatGPT or Google Gemini (McClain et al., 2025). In contrast, in September 2024 in Germany—a country without a strong digital technology sector and more hesitant in its approach to new technology—25% of respondents to a representative survey of Germans age 16 and older said they had used AI-enabled services like ChatGPT or Google Gemini (IfD-Allensbach, 2024). These figures illustrate that countries differ in how citizens engage with emerging technologies.

Country-specific differences in AI use can also extend to attitudes toward new technology and associated phenomena. For example, people vary across countries in their views on the benefits and risks of AI (Kelley et al., 2021). Similar differences can be expected for regulatory preferences for AI and digital technology more broadly (Riedl et al., 2021; Theocharis et al., 2025). We argue that these national differences in engagement and preferences translate into varying levels of societal involvement with AI, which we define as the extent to which AI technologies are integrated into daily life, institutional practices, and public debate.

We further assume that societal involvement conditions the role of individual-level involvement. In highly involved societies, we expect public opinion to be relatively uniform, such that highly and weakly involved individuals hold similar views. In contrast, in less involved societies, attitudes

Table 1. Descriptive Statistics for All Variables.

	U.S.		Germany	
	M (*SD*)	*n*	*M* (*SD*)	*n*
Accuracy	6.02 (1.28)	1756	5.14 (1.66)	1800
Safety	5.65 (1.68)	1756	5.29 (1.75)	1800
Mitigating bias	5.54 (1.61)	1756	4.75 (1.71)	1800
Imaginaries	4.55 (1.78)	1756	4.43 (1.69)	1800
AI use (α U.S. = 0.70; DE = 0.80, Spearman-Brown U.S. = 0.70; DE = 0.80)	3.18 (1.59)	1756	2.70 (1.61)	1800
Free speech (α U.S. = 0.90; DE = 0.85, Spearman-Brown U.S. = 0.90; DE = 0.86)	5.56 (1.34)	1756	5.87 (1.16)	1800
Political orientation	3.66 (1.86)	1756	3.86 (1.15)	1800
Democratic Party/Green Party ID	47.4%	1756	14.6%	1800
Gender (female)	50.3%	1756	50.0%	1800
Education (high)	17.4%	1756	19.6%	1800
Age	45.92 (15.94)	1756	45.58 (15.48)	1800

toward AI moderation should differ more strongly depending on personal involvement.

To examine these expectations, we compare public attitudes in the U.S. (representing a high-involvement context) and Germany (representing a low-involvement context).

Research Question 4 (RQ4): How does the system-level variable country shape public preferences for different goals of AI governance, such as accuracy and reliability, safety, bias mitigation, and the promotion of aspirational societal values?

Research Question 5 (RQ5): How does the relationship between individual- and group-level involvement with AI and preferences for AI governance vary across countries with differing levels of societal involvement with AI?

Methods

We collected data in the U.S. and Germany through online panels. The study was approved by the IRB of the University of Bamberg. In the U.S., 1800 participants were recruited from the survey research company Prolific (collected between 1 and 6 March 2024). We used a representative quota sample for the U.S. for sex, age, and political affiliation (see Supplementary Information A). Participants had to be U.S.-based and aged 18 or older to participate in the study. Participants were paid £0.75 (an hourly rate of £9; we ran the survey through Prolific's European platform) for their study participation, which took around 5 minutes to complete. Forty-four participants who failed a simple attention check at the beginning of the study were excluded, resulting in a sample of 1756. On the starting page, we informed participants about their rights (for example, that they could withdraw from the study at any point by simply closing the browser) and asked them to provide their consent. None of the questions asked for personally identifiable information. In Germany, we also recruited 1800 participants from the survey research company Bilendi (collected between 14 and 18 March 2024). We used a quota for age, gender, and regions in Germany (16 states). The only difference from the U.S. survey was that we could directly filter out participants who failed the attention check and thus ran the survey until we achieved 1800 successful complete responses.[2]

The descriptive statistics for all measured variables are reported in Table 1 (for a complete table on the item level with the question wording, see Supplementary Information B.1). We measured AI moderation preferences by asking respondents: "Please indicate how important the following criteria are for you when choosing AI-enabled services". We measured responses on a 7-point scale (1 = not important at all; 7 = very important) for the four concepts related to AI moderation. Accuracy and reliability ("The AI service consistently provides accurate and reliable information or results based on its analysis and data-driven insights."), safety ("Measures are in place to prevent the AI from generating or promoting illegal, dangerous, or harmful content."), bias mitigation ("Efforts are made to identify and reduce biases in AI outputs, ensuring fairness and equity in treatment and decision-making across different groups of people."), and aspirational imaginaries ("The AI aims to highlight and encourage positive societal values, portraying an aspirational view of society.").

AI use was measured with two items: one asking about AI use in the professional or work context and one assessing AI use in personal life and spare time (1 = never; 7 = very often). The two items were combined into a mean index. Support for free speech was measured using two items from Riedl et al. (2021), which were adapted from Rojas et al. (1996). Political orientation was measured with a single scale (U.S. = 1-liberal; 7-conservative; Germany = 1-left; 7-right). To identify supporters of the Democratic Party in the U.S. and the Green Party in Germany, we recoded the answers to a question about the general leaning toward a party in the country.[3] For education, we recoded responses into two categories:

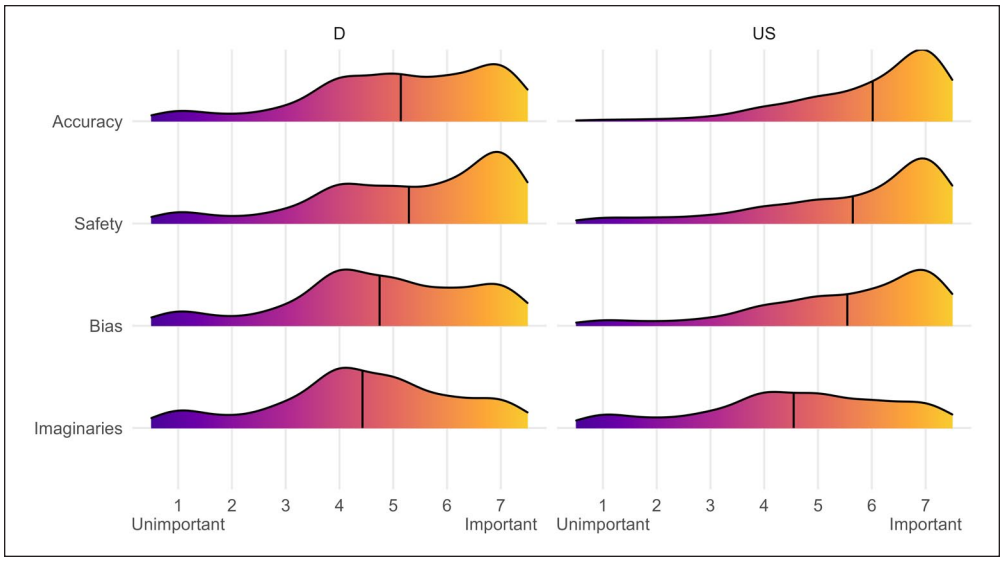

Figure 1. Distribution for the four outcome variables for Germany (left) and the U.S. (right). Vertical lines indicate the mean.

"postgraduate degree or higher" and "other" (with "other" serving as the reference category).

As an analytical strategy, we use both datasets together for the regression analysis. We first estimate, for each outcome, a model with all predictors—including a country variable (Germany = 0; U.S. = 1)—to check whether single predictors have an overall association with the outcome variable. We then estimate a second model in which we enter all variables (mean-centered) as interaction terms with the country variable. This also allows us to test, through the interactions, whether there is a country difference in the explanatory strength of the predictors. A positive estimate for the interaction term would indicate that the predictor is stronger in the U.S., whereas a negative estimate would indicate a stronger predictor in Germany. As these interaction terms are difficult to interpret, we will visualize them as marginal effect plots in the results section (we also report single-country regression models in Supplementary Information C.2 and a specification curve analysis in Supplementary Information D).

Results

RQ1 Preferences

We start with a descriptive analysis of respondents' alignment preferences when choosing AI-enabled services. Figure 1 displays the distribution of responses in Germany and the U.S. when participants were asked how important the respective alignment goal is in selecting an AI-enabled service.

As expected, support varies systematically across approaches. Public support is strongest for alignment goals oriented toward accuracy, reliability, and safety. Both U.S. and German respondents assign high importance to accuracy. Similarly, preventing harm—defined as preventing the generation of illegal, dangerous, or harmful content—is widely supported. These safety-oriented adjustments to model outputs appear to be largely noncontroversial, likely due to their alignment with conventional risk regulation in digital communication environments.

Support declines, however, as motives behind adjustments to model output shift from safety to more normative goals. Bias mitigation, which aims to promote fairness and equity, is still positively received but exhibits more variation, especially among German respondents.

Importantly, country-level differences in support patterns (see Figure 2) align with broader national trends in technology adoption and risk perception. A Welch two-sample t-test indicates a significant difference, $t(3553.4) = -8.98$, $p < .001$, in AI use between the U.S. and Germany. Participants in the U.S.—home to a world-leading digital technology sector and generally higher openness to new technologies (Comin & Hobijn, 2010; Ding, 2024)—reported higher AI use scores ($M = 3.18$, $SD = 1.59$) than those in Germany ($M = 2.70$, $SD = 1.61$) where adoption of generative AI tools remains more cautious and public discourse often emphasizes potential risks. Furthermore, 26.8% of respondents in Germany reported that they never use AI-supported applications for either personal or professional purposes, compared to only 10.8% in the U.S.

These differences in usage correspond with differences in preference. The U.S. consistently shows greater support across all categories. In contrast, Germany shows more reserved or varied support, particularly for normatively driven alignment goals. These differences reflect varying levels of societal involvement with AI.

Explaining Preferences for AI Alignment

We now examine how different explanatory factors shape preferences for alignment principles of AI-enabled systems among respondents in Germany and the U.S. Figure 2 displays

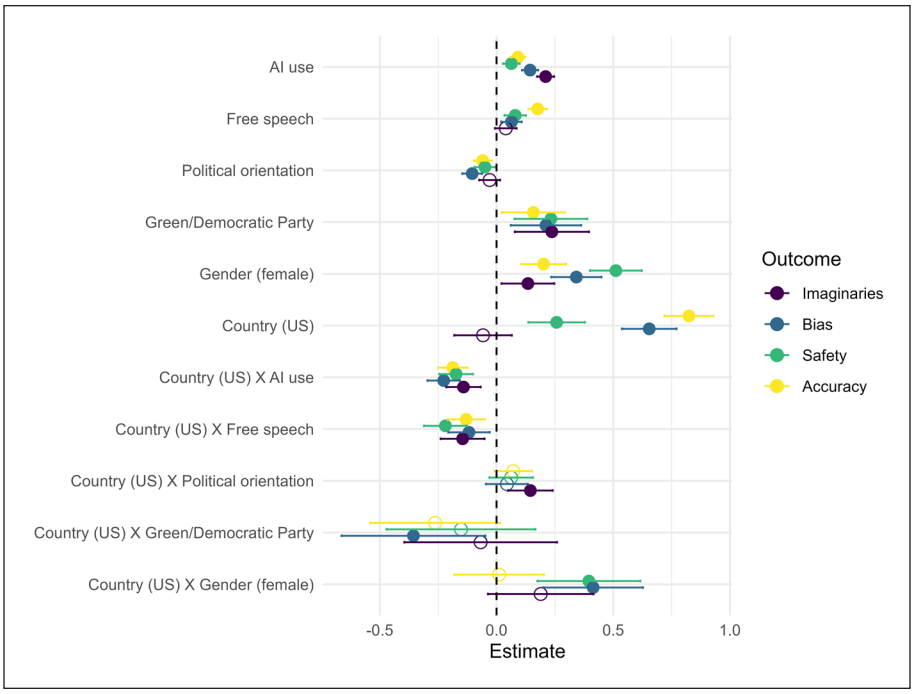

Figure 2. Estimates with 95% CIs for all four outcome variables. Non-significant estimates are indicated as empty dots. Significant single estimates indicate an overall association. Significant negative interaction estimates indicate that the association is relatively stronger in Germany, while significant positive interaction estimates indicate that it is relatively stronger in the U.S.

the estimated coefficients from our models (for the complete tables for the models, see Supplementary Information C.1).

RQ2 Individual-Level Factors. At the individual level, personal experience with AI consistently predicts support for all tested features across both countries. The association is particularly strong for alignment goals that are specific to AI systems, such as bias mitigation (b = 0.14, $p < .001$, 95% CI [0.11, 0.18]) and aspirational imaginaries (b = 0.21, $p < .001$, 95% CI [0.17, 0.25]). This indicates that direct experience with AI fosters a more nuanced understanding of its societal implications, making individuals more receptive to content interventions targeting AI-specific harms and potentials.

We also find a significant role for free speech attitudes, though the pattern is somewhat counterintuitive. Overall, stronger support for free speech is associated with greater support for accuracy (b = 0.18, $p < .001$, 95% CI [0.14, 0.22]), safety-related moderation (b = .08, $p < .001$, 95% CI [0.03, 0.13]), and bias reduction (b = .06, $p = .005$, 95% CI [0.02, 0.11]), but not for the promotion of aspirational imaginaries (b = .04, $p = .095$, 95% CI [−0.01, 0.09]). Political ideology also aligns with our expectations: in both countries, individuals who identify as liberal or left are more supportive of AI moderation than those who identify as conservative. Only for aspirational imaginaries, the estimate is not significant (b = −0.03, $p = .201$, 95% CI [−0.07, 0.02]).

RQ3 Group-Level Involvement. Turning to group-level factors, we observe that partisan allegiance plays a role consistent with party cues. Overall, self-identified supporters of the Green Party (Germany) and the Democratic Party (U.S.) are more likely to support all four forms of AI moderation than supporters of other parties (see Figure 2). These patterns mirror the elite discourse within these parties, suggesting that elite signaling helps structure public attitudes toward AI governance.

We also find that gender shapes moderation preferences, as women see all four forms of AI moderation as important. Primarily for safety (b = 0.51, $p < .001$, 95% CI [0.40, 0.62]) and bias reduction (b = .34, $p < .001$, 95% CI [0.23, 0.45]), women, on average, are more likely than men to support interventions. This supports the argument that group-level exposure to digital risks—such as online harassment and misrepresentation—can translate into greater support for protective content interventions.

RQ4 Cross-National Differences. The most pronounced differences between countries emerge for preferences related to accuracy and reliability (b = 0.82, $p < .001$, 95% CI [0.72, 0.93]), safety (b = 0.26, $p < .001$, 95% CI [0.14, 0.38]), and bias mitigation (b = 0.65, $p < .001$, 95% CI [0.54, 0.77]). As expected, respondents in the U.S.—a context characterized by higher societal involvement with AI—express stronger support for these alignment goals compared to respondents in Germany. This finding supports the idea that greater societal exposure to AI corresponds with increased public demand for moderation practices tailored to the specific risks and opportunities associated with these systems.

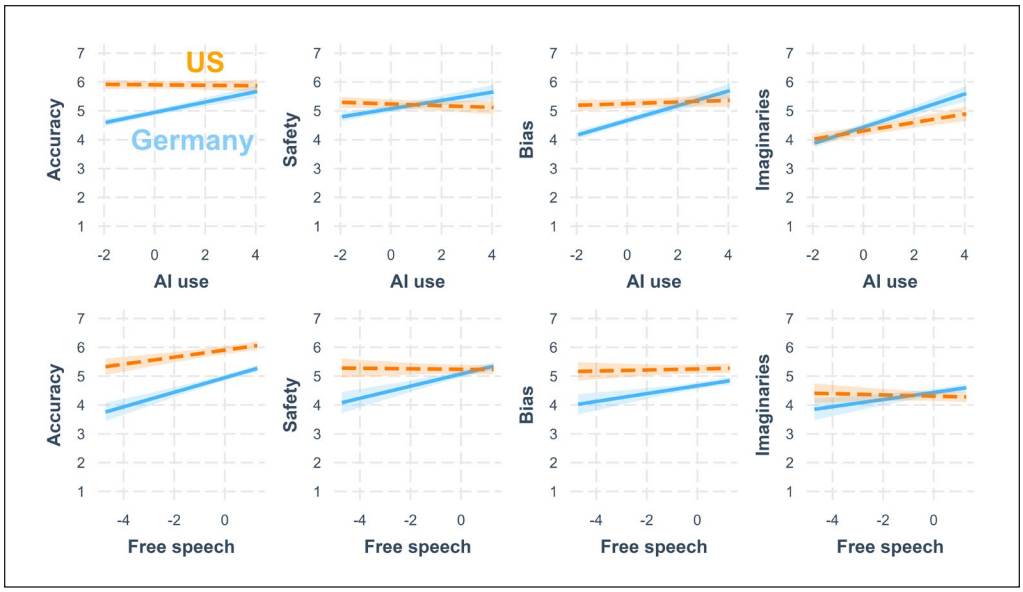

Figure 3. Interactions for AI use and free speech with country for all four outcome variables. All interaction terms are significant.

Interestingly, this pattern does not extend to support for aspirational imaginaries (b = −0.06, p = .358, 95% CI [−0.18, 0.07])—that is, AI-generated content promoting particular visions of society. For this type of intervention, we observe no significant difference between the two countries. This suggests that support for aspirational content moderation may be driven more by ideological values than by levels of societal AI involvement.

RQ5 Involvement Differences by Country. Figure 2 also illustrates the interaction terms from our second model, indicating whether the influence of a given variable differs significantly across countries. Negative interaction estimates suggest a stronger association in Germany; positive estimates suggest a stronger association in the U.S.

We find that individual-level differences in AI use have a significant impact in Germany but a considerably smaller effect in the U.S. The same is true for support for free speech: these attitudes are more predictive of preferences in Germany than in the U.S. In contrast, political ideology only has a differential association in the U.S., where it significantly predicts support for aspirational imaginaries (b = 0.15, p = .003, 95% CI [0.05, 0.24]). This suggests that in the U.S., the political discourse around aspirational imaginaries of society—particularly those embedded in AI systems—is more developed and divided.

Regarding group-level variables, partisan differences do not exhibit consistent cross-national interaction terms, with only a difference between countries for mitigating bias (b = −0.36, p = .024, 95% CI [−0.66, −0.05]) as the association is stronger in Germany. For gender, however, the associations are more pronounced in the U.S. for preferences related to safety (b = 0.40, p < .001, 95% CI [0.18, 0.62]) and bias mitigation (b = 0.41, p < .001, 95% CI [0.20, 0.63]).

Figure 3 further illustrates these dynamics by visualizing the interactions of the country variable with AI use and free speech attitudes. The figure shows that in the U.S., there is little difference between individuals with high and low AI experience in terms of their alignment preferences. In Germany, however, these differences are much more pronounced—particularly among respondents with low AI experience, who are significantly less supportive of moderation. As AI experience increases, the gap between German and U.S. respondents narrows.

A similar pattern emerges for free speech attitudes. Again, this supports our broader argument: in high-involvement societies like the U.S., attitudes toward AI alignment are more uniformly developed, reducing the explanatory power of individual-level variation. In contrast, in low-involvement societies, such as Germany, individual experiences and values play a larger role in shaping attitudes.[4]

Discussion

This study brings a comparative and attitudinal perspective to the debate on AI alignment by examining how users evaluate key features of AI-enabled systems. We show that individuals hold distinct preferences for moderation mechanisms that influence model outputs and that these preferences vary systematically between countries. Respondents in the U.S. report significantly higher levels of AI use than those in Germany, reflecting broader national differences in technology adoption and societal engagement with AI. Across both contexts, accuracy, reliability, and safety receive the strongest public support—indicating a shared baseline of expectations for trustworthy and safe systems. In contrast, support is more conditional for interventions to mitigate bias or promote aspirational imaginaries, especially among German

respondents. This aligns with the view that bias correction involves normative judgments that can be politically charged and subject to disagreement. Similarly, the lower support of interventions promoting aspirational imaginaries indicates hesitancy about the active role of AI in shaping cultural narratives—and potentially associated concerns about legitimacy, ideological overreach, and value alignment. U.S. participants consistently express higher support across all dimensions, which corresponds with their greater exposure to and involvement with AI technologies. The difference in support between Germany and the U.S. could be an expression of the maturity of the public discourse and awareness about the functioning of AI-enabled systems, indicating a lower awareness among German respondents about related issues.

Our analysis further shows that personal experience with AI, political ideology, gender, and support for free speech shape attitudes toward AI alignment—but with varying strength across countries. In Germany, both AI experience and free speech attitudes are stronger predictors of support, suggesting that in contexts with lower exposure, individual-level factors play a more decisive role. In the U.S., where AI technologies are more deeply embedded in public and institutional life, views on AI moderation appear more consolidated, with political ideology particularly influencing support for aspirational interventions.

In the U.S., the stronger ideological structuring of responses—especially regarding aspirational imaginaries—likely reflects not only higher levels of engagement with AI but also the more pronounced and divided media and elite discourse surrounding digital technologies. Public attitudes toward AI moderation may therefore mirror existing partisan divides in how technological innovation, speech regulation, and social values are debated in the public arena. In this sense, the observed divisions are less a product of direct experience with AI than of the broader cultural framing through which AI enters politicized discussion.

Our findings on the role of free speech support and preferences for adjustments of model outputs are especially interesting, since they contrast with previous findings on digital content moderation, where free speech concerns often predict greater resistance to moderation interventions by companies or states (Jang et al., 2024; Rauchfleisch & Jungherr, 2024). One possible interpretation is that respondents do not view generative AI output as equivalent to human speech. That is, the normative privilege of free speech may not extend, in the public's view, to AI-generated content. These findings suggest that assumptions from earlier debates about digital content moderation cannot be automatically transferred to the case of AI. Future policy and public debate on AI moderation should take these differences into account.

Our study is subject to several important limitations. First, there is a temporal dimension to consider. As AI-enabled systems become more prevalent in daily life, both individual experience with these technologies and public discourse around them are likely to evolve. The cross-national differences we identify may, therefore, be time-bound and could diminish over time as country-level involvement with AI converges internationally.

Second, our analysis is limited to just two countries. Future research should broaden the comparative scope to include a more diverse set of countries, particularly those with varying levels of technological integration and public attitudes toward AI. In this context, we see particular value in examining countries in Asia, where both the pace and form of AI adoption differ substantially from Western contexts. Moreover, our operationalization of "technology involvement" is relatively coarse. Future studies should develop and test more nuanced and systematic measures—such as indicators of public discourse, regulatory activity, or the economic significance of AI in a given country.

Our research design is cross-sectional and based on self-reported data. This limits the causal inferences that can be drawn and may be subject to bias in participants' self-assessments of AI use and preferences. Future work should incorporate more objective measures of AI experience and leverage experimental or longitudinal designs to capture how individuals respond to concrete AI interventions rather than relying solely on abstract descriptions or stated preferences (Rauchfleisch & Jungherr, 2025).

Finally, our findings show how publics in two democracies perceive different principles of AI alignment, but they do not imply that public opinion should directly determine technical or regulatory choices. Rather, these attitudes provide insight into the social legitimacy of competing alignment goals and can help policymakers anticipate which forms of AI governance are likely to encounter support or resistance. Future work should explore how democratic institutions can balance expert-driven alignment decisions with public expectations in the broader governance of AI.

In this sense, our findings should not be read as a call for either more or less moderation of AI outputs, but as an empirical mapping of where publics draw the boundaries of legitimate intervention. Understanding these boundaries is essential for designing governance arrangements that are both democratically responsive and epistemically sound. If one were to translate these findings into practical terms, they would suggest prioritizing accuracy and safety as default expectations for trustworthy systems, offering optional or transparent controls for fairness-oriented adjustments, and approaching aspirational or value-promoting settings with particular caution, given their divided public support.

Our findings point to a set of important and more general considerations that should be taken into account and pursued further. This includes consideration of geopolitical competition and conflict, the role of companies, and the deep opaqueness and non-assessability of the model provision pipeline. For example, our study highlights substantial cross-national differences in public attitudes toward AI. These differences are particularly significant in today's AI landscape, where

U.S. or Chinese companies develop the most widely used systems. As a result, public expectations for AI moderation may not only reflect concerns about functionality and fairness, but also the perceived degree of foreign versus domestic control over digital environments. Like dynamics observed in international trade (Jungherr et al., 2018), attitudes toward the countries of origin of AI technologies can "contaminate" perceptions of the technologies themselves. This dimension warrants close attention in a geopolitical climate marked by intense competition and strategic rivalry.

Moreover, access to AI models is shaped by the strategic and commercial interests of the companies that develop them. Whether driven by profit or geopolitical considerations, these motivations may influence both the design and availability of AI systems in ways that affect public trust. Importantly, AI moderation is just one step in a longer, largely opaque chain of decisions made during model development, training, deployment, and evaluation. At each stage, political values—intentionally or not—may become embedded in technical systems. To ensure global legitimacy and public trust in AI, these decision-making chains must become more transparent, assessable, and where appropriate, open to public negotiation and contestation.

Currently, model training and moderation practices remain largely hidden from public scrutiny. This lack of visibility risks eroding public confidence and enabling politicized narratives about AI bias or hidden agendas. In a context of growing diversity in AI development—spanning open-source and commercial models, varying origins, and geopolitical alignments (Buyl et al., 2024)—there is an urgent need for a more mature, structured debate about legitimate approaches to model adjustment, both during training and in real-time operation.

Without such a debate, we risk stumbling from one controversy to the next, fostering a general climate of suspicion toward AI. Transparency alone is not enough; societies must also articulate clear expectations of what they want from AI systems and how those systems should be governed. Therefore, companies, policymakers, and researchers must take active responsibility for documenting and debating the principles, procedures, and techniques underpinning justified AI moderation. As O'Neill (2021) argues, digital systems must be made assessable to users. If moderation practices in the adjustment of AI outputs remain opaque, public trust will deteriorate—especially when high-profile errors are framed as evidence of hidden political or cultural agendas.

Ultimately, realizing the societal benefits of AI will depend on building a public governance framework that allows for visibility, legitimacy, and accountability in model development and output adjustments. Failing to do so risks deepening skepticism and undermining AI's long-term viability in democratic societies.

Acknowledgments

The authors used ChatGPT 5 for language editing.

ORCID iDs

Andreas Jungherr https://orcid.org/0000-0003-2598-2453
Adrian Rauchfleisch https://orcid.org/0000-0003-1232-083X

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Adrian Rauchfleisch's work was supported by the National Science and Technology Council, Taiwan (R.O.C.) (grant no. 113-2628-H-002-018 and 114-2628-H-002-007) and by the Taiwan Social Resilience Research Center (grant no. 114L9003) from the Higher Education Sprout Project by the Ministry of Education in Taiwan.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability

Replication code and data are publicly available at https://osf.io/rpfbn.

Supplemental material

Supplemental material for this article is available online.

Notes

1. We define AI-enabled systems as technologies or services that apply methods from Artificial Intelligence—such as machine learning and deep learning, including recent advances in generative modeling—to perform tasks intelligently. Intelligence is understood here "as that quality that enables an entity to function appropriately and with foresight in its environment" (Nilsson, 2009, p. xiii). These systems are capable of learning from data, recognizing patterns, making context-aware decisions, and generating content—often autonomously or with minimal human input—across a range of applications, including natural language processing, computer vision, and robotics.
2. Replication code and data are publicly available at https://osf.io/rpfbn.
3. The percentage of Democratic Party supporters is higher for this question than that reported for the party ID used for quota sampling, due to the wording of the question: "In the US, many people lean towards a particular party for a long time, although they may occasionally vote for a different party. How about you, do you in general lean towards a particular party? If so, which one?" This higher percentage is attributable to independents who lean toward the Democratic Party.
4. We also report single models for each country and outcome variable in Supplementary Information C.1. They support the overall interpretations of the analysis with the interaction terms.

References

Achintalwar, S., Baldini, I., Bouneffouf, D., Byamugisha, J., Chang, M., Dognin, P., Farchi, E., Makondo, N., Mojsilović, A., Nagireddy, M., Natesan Ramamurthy, K., Padhi, I., Raz,

O., Rios, J., Sattigeri, P., Singh, M., Thwala, S. A., Uceda-Sosa, R. A., & Varshney, K. R. (2024). Alignment studio: Aligning large language models to particular contextual regulations. *IEEE Internet Computing*, *28*(5), 28–36. https://doi.org/10.1109/MIC.2024.3453671

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in ai safety*. OpenAI. https://doi.org/10.48550/arXiv.1606.06565

Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., Lubana, E. S., Jenner, E., Casper, S., Sourbut, O., Edelman, B. L., Zhang, Z., Günther, M., Korinek, A., Hernandez-Orallo, J., Hammond, L., Bigelow, E., Pan, A., Langosco, L., & Krueger, D. (2024). Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, *9*. https://doi.org/10.48550/arXiv.2404.09932

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., & Kaplan, J. (2021). *A general language assistant as a laboratory for alignment*. Anthropic. https://doi.org/10.48550/arXiv.2112.00861

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. The MIT Press.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1493–1504). ACM. https://doi.org/10.1145/3593013.3594095

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp. 149–159). PMLR.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S., von Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., & Liang, P. (2022). *On the Opportunities and Risks of Foundation Models*. Center for Research on Foundation Models (CRFM), Stanford Institute for Human-Centered Artificial Intelligence (HAI). https://doi.org/10.48550/arXiv.2108.07258

Buyl, M., Rogiers, A., Noels, S., Dominguez-Catena, I., Heiter, E., Romero, R., Johary, I., Mara, A.-C., Lijffijt, J., & Bie, T. D. (2024). *Large language models reflect the ideology of their creators* (No. arXiv:2410.18417). arXiv. https://doi.org/10.48550/arXiv.2410.18417

Chong, D., Citrin, J., & Levy, M. (2024). The realignment of political tolerance in the United States. *Perspectives on Politics*, *22*(1), 131–152. https://doi.org/10.1017/S1537592722002079

Chong, D., & Levy, M. (2018). Competing norms of free expression and political tolerance. *Social Research: An International Quarterly*, *85*(1), 197–227. https://doi.org/10.1353/sor.2018.0010

Comin, D., & Hobijn, B. (2010). An exploration of technology diffusion. *American Economic Review*, *100*(5), 2031–2059. https://doi.org/10.1257/aer.100.5.2031

Dabhoiwala, F. (2025). *What is free speech? The history of a dangerous idea*. Belknap Press.

De Ruiter, A. (2021). The distinct wrong of deepfakes. *Philosophy & Technology*, *34*(4), 1311–1332. https://doi.org/10.1007/s13347-021-00459-2

Dewey, J. (1916). *Democracy and education. An introduction to the philosophy of education*. Macmillan.

Dewey, J. (1934). *Art as experience*. Minton, Balch & Co.

Ding, J. (2024). *Technology and the rise of great powers: How diffusion shapes economic competition*. Princeton University Press.

Douglas, M., & Wildavsky, A. B. (1982). *Risk and culture: An essay on the selection of technological and environmental dangers*. University of California Press.

Fourcade, M., & Healy, K. (2024). *The ordinal society*. Harvard University Press.

Friedrich, F., Brack, M., Struppek, L., Hintersdorf, D., Schramowski, P., Luccioni, S., & Kersting, K. (2024). Auditing and instructing text-to-image generation models on fairness. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00531-5

Friedrich, F., Hämmerl, K., Schramowski, P., Brack, M., Libovicky, J., Kersting, K., & Fraser, A. (2024). *Multilingual text-to-image generation magnifies gender stereotypes and prompt engineering may not help you*. arXiv. https://doi.org/10.48550/arXiv.2401.16092

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Hand, D. J. (2004). *Measurement theory and practice: The world through quantification*. Wiley.

Hao, S., Kumar, P., Laszlo, S., Poddar, S., Radharapu, B., & Shelby, R. (2023, June 9). *Safety and fairness for content moderation in generative models* [Conference session]. 1st Workshop on Multimodal Content Moderation. CVPR'23: The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023. https://doi.org/10.48550/arXiv.2306.06135

Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2022). *Unsolved problems in ML safety*. arXiv. https://doi.org/10.48550/arXiv.2109.13916

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature*, *633*(8028), 147–154. https://doi.org/10.1038/s41586-024-07856-5

Horowitz, M. C., & Kahn, L. (2021). What influences attitudes about artificial intelligence adoption: Evidence from U.S. local officials. *PLOS ONE*, *16*(10), Article e0257732. https://doi.org/10.1371/journal.pone.0257732

Horowitz, M. C., Kahn, L., Macdonald, J., & Schneider, J. (2024). Adopting AI: How familiarity breeds both trust and contempt. *AI & SOCIETY*, *39*(4), 1721–1735. https://doi.org/10.1007/s00146-023-01666-5

IfD-Allensbach. (2024). *Fast Food Wissen und virtuelle Liebe. KI-Assistenten und wir* [Fast food knowledge and virtual love. AI assistants and us]. Institut für Demoskopie Allensbach im Auftrag der Telekom. https://www.telekom.com/de/medien/medieninformationen/detail/mit-ki-auf-du-und-du-was-macht-das-mit-uns-1082132

Jang, H., Barrett, B., & McGregor, S. C. (2024). Social media policy in two dimensions: Understanding the role of anti-establishment beliefs and political ideology in Americans' attribution of responsibility regarding online content. *Information, Communication & Society*, *27*(6), 1047–1072. https://doi.org/10.1080/1369118X.2023.2234970

Jasanoff, S., & Kim, S.-H. (Eds.). (2015). *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. The University of Chicago Press.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K. Y., Dai, J., Pan, X., O'Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., & Gao, W. (2023). *AI alignment: A Comprehensive Survey*. arXiv. https://doi.org/10.48550/arXiv.2310.19852

Jiang, J. A., Scheuerman, M. K., Fiesler, C., & Brubaker, J. R. (2021). Understanding international perceptions of the severity of harmful content online. *PLOS ONE*, *16*(8), Article e0256762. https://doi.org/10.1371/journal.pone.0256762

Jungherr, A. (2023). Artificial intelligence and democracy: A conceptual framework. *Social Media + Society*, *9*(3), 1–14. https://doi.org/10.1177/20563051231186353

Jungherr, A., Mader, M., Schoen, H., & Wuttke, A. (2018). Context-driven attitude formation: The difference between supporting free trade in the abstract and supporting specific trade agreements. *Review of International Political Economy*, *25*(2), 215–242. https://doi.org/10.1080/09692290.2018.1431956

Jungherr, A., Rauchfleisch, A., & Wuttke, A. (2024). *Artificial Intelligence in Election Campaigns: Perceptions, Penalties, and Implications*. arXiv. https://doi.org/10.48550/arXiv.2408.12613

Jungherr, A., & Schroeder, R. (2023). Artificial intelligence and the public arena. *Communication Theory*, *33*(2–3), 164–173. https://doi.org/10.1093/ct/qtad006

Kelley, P. G., Yang, Y., Heldreth, C., Moessner, C., Sedley, A., Kramm, A., Newman, D. T., & Woodruff, A. (2021). Exciting, useful, worrying, futuristic: Public perception of Artificial Intelligence in 8 countries. In *AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 627–637). ACM. https://doi.org/10.1145/3461702.3462605

Kosseff, J. (2023). *Liar in a crowded theater: Freedom of speech in a world of misinformation*. Johns Hopkins University Press.

Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, *120*(7), Article e2210666120. https://doi.org/10.1073/pnas.2210666120

Künast, R., & Geese, A. (2020). *Digital Services Act – Die EU muss die Demokratie im Netz schützen* [Digital Servies Act – The EU must protect democracy on the internet]. Handelsblatt. https://www.handelsblatt.com/meinung/gastbeitraege/gastkommentar-digital-services-act-die-eu-muss-die-demokratie-im-netz-schuetzen/26287432.html

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). *Scalable agent alignment via reward modeling: A research direction*. DeepMind. https://doi.org/10.48550/arXiv.1811.07871

Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI hallucinations: A misnomer worth clarifying. In *CAI'24: 2024 IEEE Conference on Artificial Intelligence* (pp. 133–138). https://doi.org/10.1109/CAI59869.2024.00033

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *ACL'20: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). https://doi.org/10.18653/v1/2020.acl-main.173

McCabe, D., & Kang, C. (2020, October 28). Republicans blast social media C.E.O.s while democrats deride hearing. *The New York Times*. https://www.nytimes.com/2020/10/28/technology/senate-tech-hearing-section-230.html

McClain, C., Kennedy, B., Gottfried, J., Anderson, M., & Pasquini, G. (2025). *How the U.S. public and AI experts view artificial intelligence*. Pew Research Center. https://www.pewresearch.org/wp-content/uploads/sites/20/2025/04/pi_2025.04.03_us-public-and-ai-experts_report.pdf

Mchangama, J. (2022). *Free speech: A history from Socrates to social media*. Basic Books.

Neyazi, T. A., Nadaf, A. H., Tan, K. E., & Schroeder, R. (2024). Does trust in government moderate the perception towards deepfakes? Comparative perspectives from Asia on the risks of AI and misinformation for democracy. *Government Information Quarterly*, *41*(4), 1–11. https://doi.org/10.1016/j.giq.2024.101980

Ngo, R., Chan, L., & Mindermann, S. (2022, August 30). The alignment problem from a deep learning perspective. In *ICLR'24: Proceedings of the 12th International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2209.00626

Nilsson, N. J. (2009). *The quest for Artificial Intelligence: A history of ideas and achievements*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819346

O'Neill, O. (2021). *A philosopher looks at digital communication*. Cambridge University Press.

Padhi, I., Dognin, P., Rios, J., Luss, R., Achintalwar, S., Riemer, M., Liu, M., Sattigeri, P., Nagireddy, M., Varshney, K. R., & Bouneffouf, D. (2024). ComVas: Contextual moral values alignment system. In *IJCAI '24: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (pp. 8759–8762). https://doi.org/10.24963/ijcai.2024/1026

Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., Lindner, D., Rahtz, M., Assael, Y., Hodkinson, S., Howard, H., Lieberum, T., Kumar, R., Raad, M. A., Webson, A., Ho, L., Lin, S., Farquhar, S., Hutter, M., & Shevlane, T. (2024). *Evaluating frontier models for dangerous capabilities*. Google DeepMind. https://doi.org/10.48550/arXiv.2403.13793

Pradel, F., Zilinsky, J., Kosmidis, S., & Theocharis, Y. (2024). Toxic speech and limited demand for content moderation on social media. *American Political Science Review*, *118*(4), 1895–1912. https://doi.org/10.1017/S000305542300134X

Rauchfleisch, A., & Jungherr, A. (2024). Blame and obligation: The importance of libertarianism and political orientation in the public assessment of disinformation in the United States. *Policy & Internet*, *16*(4), 801–817. https://doi.org/10.1002/poi3.407

Rauchfleisch, A., & Jungherr, A. (2025). *The politics of Artificial Intelligence alignment: Public reactions to AI moderation in the case of Google's Gemini* [Working paper].

Rauchfleisch, A., Jungherr, A., & Wuttke, A. (2025). Explaining public preferences for regulating Artificial Intelligence in election campaigns: Evidence from the U.S. and Taiwan. *Telecommunications Policy*. https://doi.org/10.1016/j.telpol.2025.103072

Riedl, M. J., Naab, T. K., Masullo, G. M., Jost, P., & Ziegele, M. (2021). Who is responsible for interventions against problematic comments? Comparing user attitudes in Germany and the United States. *Policy & Internet*, *13*(3), 433–451. https://doi.org/10.1002/poi3.257

Riedl, M. J., Whipple, K. N., & Wallace, R. (2022). Antecedents of support for social media content moderation and platform regulation: The role of presumed effects on self and others. *Information, Communication & Society*, *25*(11), 1632–1649. https://doi.org/10.1080/1369118X.2021.1874040

Rojas, H., Shah, D. V., & Faber, R. J. (1996). For the good of others: Censorship and the third-person effect. *International Journal of Public Opinion Research*, *8*(2), 163–186. https://doi.org/10.1093/ijpor/8.2.163

Roose, K. (2025, July 23). The Chatbot culture wars are here. *The New York Times*. https://www.nytimes.com/2025/07/23/technology/trump-ai-chatbots-bias.html

Rorty, R. (1989). *Contingency, irony, and solidarity*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804397

Smith, B. C. (2019). *The promise of Artificial Intelligence: Reckoning and judgment*. The MIT Press.

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances*, *9*(26), eadh1850. https://doi.org/10.1126/sciadv.adh1850

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, *3*(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

Taylor, C. (2004). *Modern social imaginaries*. Duke University Press.

Theocharis, Y., Kosmidis, S., Zilinsky, J., Quint, F., & Pradel, F. (2025). *Content warning: Public attitudes on content moderation and freedom of expression*. Content Moderation Lab at TUM Think Tank.

Vogler, D., Rauchfleisch, A., & de Seta, G. (2025). Support for deepfake regulation: The role of third-person perception, trust, and risk. *Studies in Communication and Media*, *14*(4).

Wang, S., & Kim, S. (2022). Users' emotional and behavioral responses to deepfake videos of K-pop idols. *Computers in Human Behavior*, *134*(C). https://doi.org/10.1016/j.chb.2022.107305

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., & Gabriel, I. (2021). *Ethical and social risks of harm from Language Models*. DeepMind. https://doi.org/10.48550/arXiv.2112.04359

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., & Gabriel, I. (2022). Taxonomy of risks posed by language models. *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). https://doi.org/10.1145/3531146.3533088

Author biographies

Andreas Jungherr (Dr. rer. pol. University of Bamberg) holds the Chair for Political Science, especially Digital Transformation at the University of Bamberg and is a Director at the Bavarian Research Institute for Digital Transformation (bidt). His research interests include the impact of digital media on politics and society, with a special focus on Artificial Intelligence, political communication, and governance.

Adrian Rauchfleisch (PhD University of Zurich) is a Professor at the Graduate Institute of Journalism, National Taiwan University. His research focuses on the interplay of politics, technology, and journalism in Asia, Europe, and the United States. His new project explores Artificial Intelligence's influence on society across different cultural contexts.