Forthcoming in *Cognition*

Disagreement Judgments Capture More than Differences in Beliefs

Kerem Oktar*, John Branson Byers*, Tania Lombrozo

Author Note

We have no known conflicts of interest to disclose.

*These authors contributed equally and share first-authorship.

All pre-registrations, materials, analysis scripts, and data available at https://osf.io/f3ve6/

Abstract

Decades of research have examined the consequences of disagreement, both negative (harm to relationships) and positive (fostering learning opportunities). Yet the psychological mechanisms underlying disagreement judgments themselves are poorly understood. Much research assumes that disagreement tracks divergence: the difference between two individuals' beliefs with respect to a proposition, where this can be understood in binary terms (individuals either agree or disagree) or as a continuous difference (for instance, in subjective probability). We test divergence as an account of interpersonal disagreement judgments using predictive modeling ($N = 238$). Our results indicate that while judgments of disagreement track divergence, other properties of beliefs—such as their extremity—play a significant role. Moreover, these additional factors are necessary for predicting key social consequences of disagreement, including inferences of bias, coldness, and incompetence. These results suggest that the assumption that disagreement judgments merely track differences in belief is empirically unjustified.

Keywords: disagreement, divergence, judgment; belief; social cognition

Disagreement Judgments Capture More than Differences in Beliefs

Disagreement is ubiquitous: from petty arguments over where to order dinner to heated debates about fiscal policy, we frequently find ourselves at odds with one another. Though much research has studied the *consequences* of disagreement, we know surprisingly little about what underlies *judgments* of disagreement in the first place, including judgments of "interpersonal" disagreement, or disagreement with another person. This is because most relevant research across decades and disciplines has operationalized interpersonal disagreement judgments as a function of divergence in belief, without examining the validity of this operationalization. Divergence itself has been operationalized either in binary terms (e.g., if Alex believes that climate change is real, and Sam thinks it is not, they disagree) or as a continuous measure (e.g., if Alex assigns a probability of .7 to the claim that climate change is real, and Sam assigns .3, their disagreement is .4). In this paper, we first provide theoretical reasons for suspecting that interpersonal disagreement judgments might not be reducible to a difference in beliefs. We then present data and modeling demonstrating that judgments of disagreement are in fact sensitive to a rich array of reasonable inputs (such as the extremity of participants' views) that drive disagreement judgments in conjunction with divergence, and that characterizing disagreement in this way is important, as it supports better predictions of key social judgments (such as warmth, competence, and bias).

The Study of Disagreement

Psychologists have studied the impact of disagreement on behavior for decades. Social psychologists have primarily focused on the negative interpersonal consequences of disagreement, from failures of communication (Ziembowicz et al., 2023) to escalating conflict (Kennedy & Pronin, 2008), as well as negative intrapersonal consequences, such as discomfort (Matz & Wood, 2005) and lower self-esteem (Pool et al., 1998). Cognitive and developmental psychologists have

instead often focused on the positive consequences of encountering diverse, conflicting opinions, examining the benefits of diverse inputs across domains of cognition, such as perception (Bahrami et al., 2010), problem solving (Smaldino et al., 2023), judgment (Soll & Larrick, 2009), and learning (Blakey & Ronfard, 2026; Harris, 2012). The study of disagreement is also of central interest to philosophers (Frances & Matheson, 2019), political scientists (Iyengar & Westwood, 2015), and beyond (see Table 1; see also Oktar & Lombrozo, 2026).

Note that while some of this work focuses on disagreement with particular propositions (e.g., consider a public opinion researcher measuring disagreement with claims about the economy), other research focuses on explaining interpersonal disagreement (e.g., consider a psychologist measuring how much people disagree with members of opposing parties). Though distinct, these two kinds of disagreement are intertwined, with propositional disagreement often used as a foothold for operationalizing or grounding interpersonal disagreement.

Disagreement as Divergence

A plausible, initial hypothesis about judgments of interpersonal disagreement is that they correspond to the *divergence* of belief regarding some proposition. In the simplest case, we can think of divergence in binary terms: following the previous example, if Alex and Sam express the same view (e.g., both say "I believe in climate change"), they agree. If they express different views (e.g., "I believe in climate change" vs. "I do not believe in climate change"), they disagree.

Yet such *binary divergence* seems to poorly characterize disagreement in cases where two people have the same general attitude towards a proposition (e.g., they both say they believe in climate change), but have differing levels of certainty with which they hold this belief. For instance, Alex might be entirely certain that climate change is happening, while Sam might have a weak sense that climate change is probably happening. In this case, Alex and Sam seem to

disagree more than, say, Alex and Alexa, who are both entirely certain that climate change is happening. This suggests that disagreement judgments come in degrees, with some people slightly disagreeing with our views, while others are in total disagreement. Thus, a more complete characterization of disagreement judgments could introduce the notion of *continuous divergence*, whereby disagreement is captured by differences in degrees of belief.

These intuitive notions of binary and continuous divergence are often leveraged to operationalize disagreement in empirical studies. For example, in experimental manipulations, disagreements over art have been operationalized through binary divergence in two people's views concerning the value of particular pieces (Cheek et al., 2021), whereas disagreements with others over statistics (such as differences in the forecasts of advisors) have been operationalized through continuous divergence in estimates across participants (Budescu et al., 2003).

Notions of binary and continuous divergence are also used to ground theoretical debates (Palmira, 2018). For instance, formal research in social science frequently formalizes disagreement as a difference in belief. As Bullock (2009) writes: "Let μ_{Dt} be the mean of voter D's belief about μ at time t. Let μ_{Rt} be the mean of voter R's belief about μ at time t. If $\mu_{Dt} = \mu_{Rt}$, D and R agree at time t. If $\mu_{Dt} \neq \mu_{Rt}$, they disagree at time t." In this formalization, disagreement corresponds to any mismatch in binary or continuous belief. Similarly, Macfarlane (2007) identifies binary divergence as the intuitive notion of disagreement that philosophers might appeal to: "The obvious thing to say is that [two people] disagree just in case there is a proposition that one party accepts and the other rejects. (…) Perhaps it is because [this] is such an obvious answer that philosophers have not wasted much ink on the question of what it is to disagree."

Table 1 presents a broader collection of references that either rely on divergence to operationalize or conceptualize disagreement, or that identify divergence as an intuitive way to do

so (whether or not they ultimately endorse that intuition). These references are heterogenous, as divergence has been used as a measure of disagreement for many reasons, from theoretical commitments (e.g., to a particular definition of disagreement in a formal model) to practical necessity (e.g., using binary operationalizations in developmental research to communicate dissent to children). The table is not meant to be exhaustive, but merely to make the point that many scholars across disciplines appeal to or use divergence as a default notion of disagreement. To our knowledge, prior work rarely goes beyond binary or continuous divergence and therefore leaves unanswered central questions about the factors that contribute to judgments of interpersonal disagreement.

Table 1

Across Fields, Disagreement is Often Conceptualized and Operationalized as Divergence

Field	Binary Divergence	Continuous Divergence
Psychology	(Amemiya et al., 2024; Langenhoff et al., 2023)	(Budescu et al., 2003; Ren & Schaumberg, 2024)
Political Science	(Nir, 2011; Price et al., 2002)	(Bullock, 2009; Hopmann et al., 2020)
Computer Science	(Haghtalab et al., 2021; Kozitsin, 2022)	(Aaronson, 2005; Aumann, 1976)
Philosophy	(Egan, 2010; MacFarlane, 2007)	(Christensen, 2007; Kelly, 2010)
Linguistics	(Kakava, 2002; Rees-Miller, 2000)	(Pham & Buchsbaum, 2020; Stromer-Galley et al., 2015)

Note. The field-based classification is not meant to be definitive as interdisciplinary studies were placed in convenient cells of the table. These references are illustrative rather than exhaustive; we include them here to make the point that across fields, binary and continuous divergence are often taken to be intuitive or default operationalizations of disagreement. Note that each reference

appeals to binary and/or continuous divergence as part of an argument or operationalization, but not all authors ultimately (or explicitly) endorse that notion of disagreement.

What Lies Beyond Divergence

While continuous divergence can capture the graded nature of disagreement, it is nevertheless a highly constrained operationalization: it constrains the inputs of disagreement judgments to the disputed belief (and fails to consider, for instance, the emotions expressed by the individuals), it constrains the function relating these inputs to subtraction (and fails to consider richer mappings between two people's beliefs and disagreement), and it constrains the output to a point estimate—for instance, that Sam and Alex disagree by 4 points on a probability scale (and fails to consider how judgments of disagreement might manifest in different verbal expressions of dissent).

Here, we focus on the functional constraint. Note that subtraction is a *linear* and *symmetric* operation. Linearity implies that a one-point difference in belief, independent of the initial extremity of one's own view, should correspond to the same amount of disagreement. Similarly, symmetry implies that a one-point difference in belief, in any direction, should correspond to the same amount of disagreement. Yet judgments of disagreement seem to be non-linear and asymmetric. To illustrate non-linearity, imagine that Alex, Sam, and Maya assign the following probabilities to the proposition that climate change is happening: Alex = .52, Sam = .76, and Maya = 1. Intuitively, Alex and Sam seem to disagree more than Sam and Maya, perhaps because Alex holds a middling view—yet the continuous divergence across them is identical, suggesting that disagreement judgments might be non-linear. Now consider Casey, who assigns a .38 probability to climate change. Alex and Casey have the same continuous divergence (.52 - .28 = .24) as Sam

and Alex (.76 - .52 = .24), yet they might seem to disagree more, perhaps because Casey is on the 'opposing side,' suggesting that disagreement judgments might be asymmetric.

Note that these nuances are rarely captured in existing research, and that this omission is sometimes due to measurement limitations, and sometimes due to analysis limitations. For instance, if disagreement is operationalized through continuous divergence in some real quantity (e.g., perceived number of homicides), there is no natural midpoint, so capturing these nuances is difficult due to measurement limitations. On the other hand, researchers might use a continuous agreement scale, but not explicitly include effects for extremity in their analysis, and hence nuances in disagreement would be lost due to analytical limitations.

In line with the possibility that disagreement judgments are non-linear, research on attitude strength has shown that in many settings middling views lead to different judgments than extreme views (Howe & Krosnick, 2017). In this literature, attitude or belief extremity is operationalized as the distance to the midpoint of a given judgment on a scale, with the endpoints of the scale indicating maximal extremity, and the midpoint indicating minimal extremity (Abelson, 1995). Note that belief extremity is thus different from the extremity of the claim itself. Consider an extreme claim (e.g., "most people are taller than 7 ft"); someone who holds a middling view about whether this claim is true has more credence in this extreme claim than most people do, but their degree of belief is not itself extreme. Toner et al. (2016) found that participants with more extreme political beliefs tend to believe that their views are superior to disagreeing others' views at much higher rates than participants with middling political beliefs. Such lack of humility is inversely associated with the tendency to try to understand others' perspectives in greater depth prior to judging their views (Koetke et al., 2024). Thus, extremity might encourage the construal of

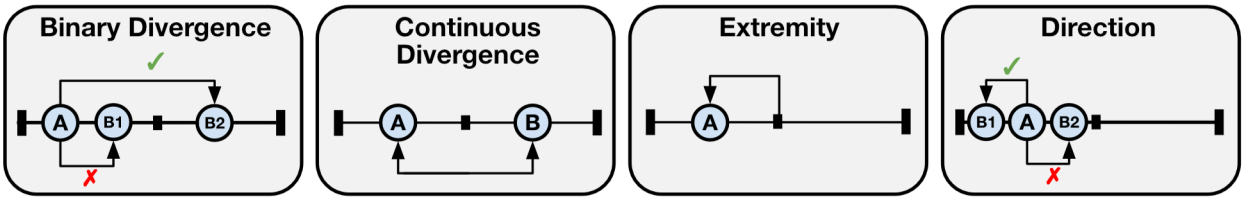
differences in views as *disagreement*, rather than, say, judging such differences to reflect miscommunication or nuances in interpretations (Cronin & Weingart, 2019).

Research on social judgment suggests that judgments of disagreement might also be sensitive to the direction of deviation, as people prefer to associate with others who have more extreme political views than their own (rather than more moderate views; Goldenberg et al., 2023). Related research on group deviance suggests that people evaluate deviation from group norms towards the extreme more favorably than deviations towards more moderate views (Abrams et al., 2002; Morrison & Miller, 2008). If people generally favor deviance towards aligned yet extreme views, they might be less likely to perceive such deviance as disagreement.

These points raise the possibility that even continuous divergence might not, on its own, capture important nuances in judgments of disagreement, including (but not limited to) extremity and direction. To investigate this possibility we present an empirical study of disagreement judgments, and we compare models of those judgments that incorporate different combinations of four predictor: *binary divergence* (which reflects whether A and B are on the same side of the midpoint), *continuous divergence* (which captures the absolute value of the difference between A and B), *extremity* (which we formalize as the distance of a belief from the midpoint), and *direction* (which indicates whether B is at least as extreme as A *and* on the same side of the midpoint; see Figure 1).

Figure 1

Visualizing Divergence, Extremity, and Direction



Note. In this schematic representation, A and B correspond to different beliefs, where their locations on the line correspond to their subjective probability, between 0 and 100 (with 50 as the midpoint).

Experiment

Participants were first assigned to one of four domains (religion, morality, politics, or science) and indicated their beliefs about three key issues within that domain. They then encountered 12 characters with differing beliefs about these issues, rated how much they disagreed with the characters, and made a variety of social judgments about these characters (explained further below), including their competence, warmth, similarity, perspective taking, and bias (see also Samuelson & Dahl, 2026). The study was pre-registered (see our OSF repository for all pre-registrations, materials, data, and analysis scripts, available at https://osf.io/f3ve6/); any departures from pre-registered analyses are noted).

Methods

Participants

Participants were 238 adults (91 men, 143 women, 4 other, mean age = 39.7 years) recruited on Prolific in exchange for monetary compensation ($1.58 for an 8-minute study). Participation across all studies was restricted to users currently residing in the United States with

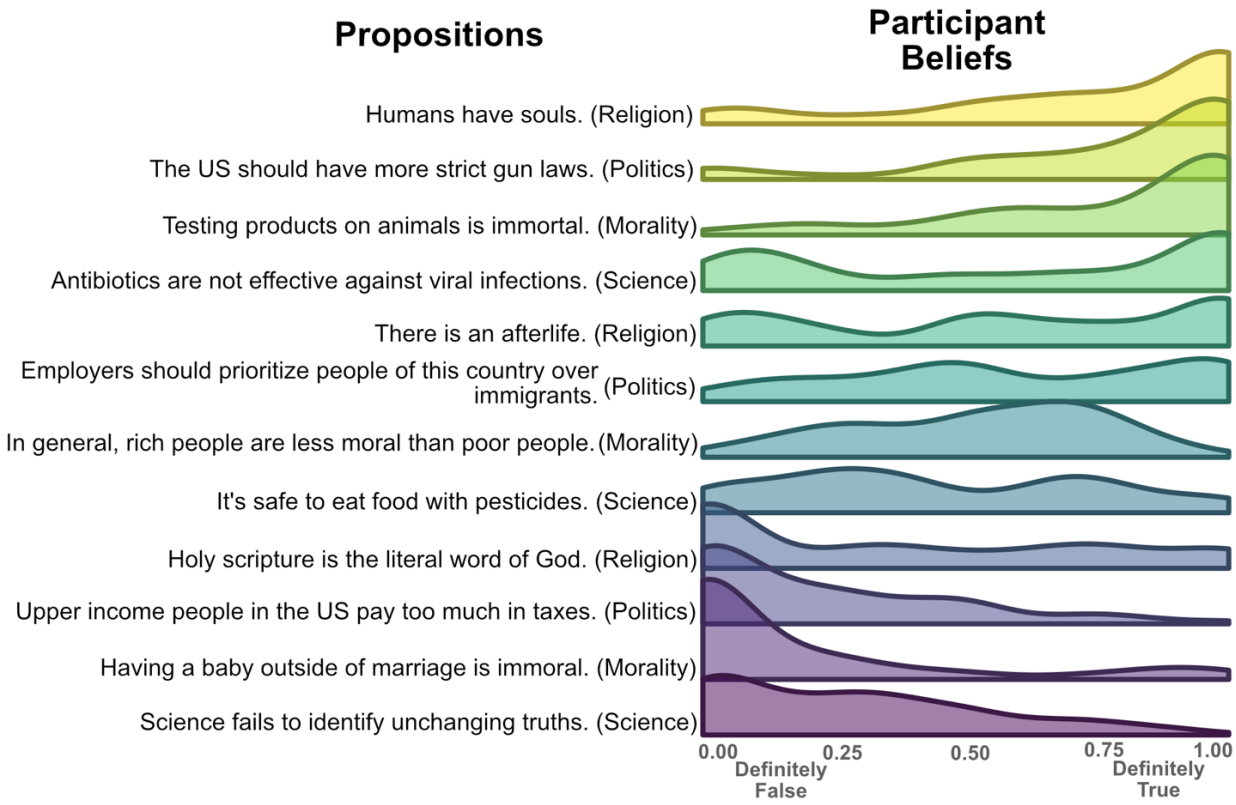
an approval rating ≥ 98% on at least 100 tasks. The number of participants was determined through a power analysis based on a previous pilot, such that we would have 95% power to observe our key effect of interest, whether the Rich Belief model outperforms the Extremity model with 95% certainty (these models are explained below; see Supplementary Materials Appendix A for the power analysis). Repeat participation within and across related studies was restricted using the Prolific platform. Two additional participants were excluded for failing an attention check. The sample size and exclusion criteria were pre-registered. The study was approved by the Institutional Review Board (IRB) at Princeton University.

Materials and Procedure

Participants were first randomly assigned to one of four domains (science, religion, politics, morality), and told that they would be presented with the beliefs of randomly selected Americans "collected from 'Attitudes in America,' a project investigating American beliefs across a wide range of issues." Participants indicated their beliefs about three issues within their assigned domain on slider scales from 'Definitely False' [0] to 'Definitely True' [100]. The issues were selected from Oktar and Lombrozo (2024) to span a variety of domains and degrees of societal agreement (see Figure 2).

Figure 2

Statements Used in the Study



Note. Statements were piloted and chosen such that participants were likely to provide a wide range of truth ratings across them (from *Definitely False* to *Definitely True*). Figure shows density plots for truth judgments across all items and the domains they are drawn from.

Participants then saw other characters' beliefs indicated on the same scale they used to indicate their own beliefs. These beliefs were generated to span levels of divergence: For each statement, participants saw four characters, each with a belief randomly sampled from one quartile [0-25; 26-50; 51-75; 76-100]. Participants indicated "whether and how much [they] agree or

disagree with [character]," using a slider from 'Totally Agree' [0] to 'Totally Disagree' [100], with 'Neither Agree nor Disagree' as a neutral midpoint [50].

In addition to providing disagreement judgments, participants also made a variety of social judgments: They rated disagreeing others' competence (Cuddy et al., 2008; "[character] would be viewed as highly competent by others who share my perspective on the issue"), warmth (Cuddy et al., 2008; "[character] would be viewed as highly warm by those who share my perspective on the issue"), similarity (Barnidge, 2018; "Aside from their belief on the issue, I would expect [character] to be very similar to me"), perspective taking (Davis, 1980; "I find it difficult to see things from [character]'s point of view"), and bias (Kennedy & Pronin, 2008; "[character] was influenced by careful consideration of the facts when arriving at their perspective on the issue"). These items were selected from a broader set tested in pilot studies because they captured the most unique variance (see Oktar et al., 2024). Note that judgments of warmth and competence maintain a third-person framing in line with the seminal literature that uses warmth and competence to study individual beliefs about stereotypes (Cuddy et al., 2008).

Finally, participants answered demographic questions (age, sex, educational background, level of religiosity, political affiliation) and received debriefing information.

Results

Analytical Strategy: Nested Model Comparison and Predictive Accuracy

The aim of our analysis was to use predictive modeling to test whether interpersonal disagreement judgments are better predicted by models that incorporate aspects of belief beyond binary and continuous divergence. Specifically, we tested whether taking into account "extremity" and "direction" significantly increased accuracy in predicting disagreement judgments. Since additional degrees of freedom allow more complex models to better predict observed data (i.e.,

they might overfit data), all models were evaluated using out-of-sample predictive power. That is, models were iteratively fit and evaluated on "unseen" data, which implies that an increase in model performance results from truly generalizing to predicting new data, rather than overfitting to data provided during model fitting.

We employed a nested model comparison, which assessed whether each regressor (beyond binary and continuous divergence) significantly improved predictions of human disagreement judgements. In a nested model comparison, a simple model is made more complex with the addition of another regressor (and any appropriate interaction effects). The simple and more complex models are then compared to test if the additional regressor significantly improved the predictions of the complex model over the simpler one. This process is then repeated for each regressor. (See the next section on Quantifying Predictive Power for more detail on evaluating improvements of model performance.)

In total, we evaluated the predictive power of six different regressors (Figure 3) in describing disagreement judgments: "binary divergence," "continuous divergence," "extremity," "direction," and the interactions of direction and extremity with continuous divergence. Note that for each additional predictor, our key question is whether that predictor introduces predictive signal *above and beyond the effects of divergence*. We therefore refer to the models that include these additional predictors in terms of the additional predictor that they include (e.g., the "Extremity Model" includes extremity in addition to binary and continuous divergence).

"Binary divergence" was coded as a "1" when participants' beliefs were on opposite sides of the midpoint and as "0" otherwise. "Continuous divergence" treated differences in belief as the absolute value of the difference in the beliefs of the participant and the character. While "extremity" was coded as the distance of the *participant's* belief from the midpoint, the interaction

between extremity and continuous divergence is a more meaningful predictor. Specifically, the interaction of extremity with continuous divergence describes how differences in beliefs might be amplified when the participant's beliefs are more extreme. (Extremity on its own simply adds the distance from the midpoint as a regressor.) Finally, "direction" describes whether the character's belief was more or less extreme than the participant's belief on the participant's side of the scale (e.g., if the participant reported a belief of 0.7 and the character's belief was 0.8, this would be coded as a 1 for direction).

Quantifying Predictive Power with Information Criteria

We quantified the predictive power of each model by estimating the (out-of-sample) likelihood of each model producing the observed data. We employed LOO (Pareto-Smoothed Importance-Sampling of Leave-One-Out Cross-Validation) as our likelihood-based information criterion of choice. We note that LOO is comparable to methods like AIC and BIC, which also evaluate out-of-sample prediction using Bayesian likelihood. However, LOO relaxes two assumptions: (1) the number of observations $n \rightarrow \infty$ and (2) each regressor contributes equally to overfitting. Since nested model comparison relies on accurately penalizing a model for overfitting using finite data, LOO provides a more accurate estimate of model performance. We also note that while AIC and BIC provide point estimates of model performance, LOO additionally provides a measure of uncertainty (through importance-sampling of cross-validation).

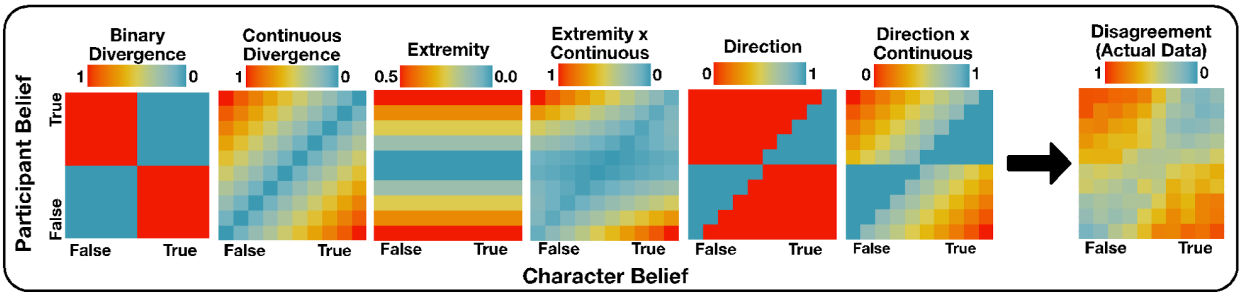
All reported models converged as predicted by our pre-registered power analysis (Pareto $k < 1$). As pre-registered, we used a pilot study to power our analysis to be 95% confident (using a 95% credibility interval as our criterion for significantly better model performance) that the Rich Belief model (which included all of our predictors) performed better than the Extremity model (which included binary divergence, continuous divergence, and extremity, but not direction).

Model Specification: Zero-One-Inflated Beta-Regression

We employed Zero-One Inflated Beta Regression (ZOIB) as our model family for capturing human disagreement judgements (see Supplementary Materials for model formulation). As depicted in Figure 4C, our data are bounded (i.e., can be normalized between 0 and 1) with many values hitting exactly the bounds. Since linear regression assumes an unbounded domain, it cannot specifically describe data at the bounds, making ZOIB a more appropriate model. ZOIB can be compared to linear regression; it finds the best-fit beta distribution instead of finding the best-fit line. Furthermore, logistic regression describes data points that are at exactly zero or one. A dedicated parameter describes the probability that a data point is generated using the beta distribution versus the logistic curve – making ZOIB a hybrid model of beta-regression (for middling points) and logistic regression (for points at the bounds). More details on model parameterization are available in Appendix B. We note that for regressors such as Extremity, modeling extreme judgements at the bounds is particularly relevant.

Figure 3

Visualizing the Predictors Tested in Model Comparisons Across Belief Space



Note. We visualize each regressor across participant and character belief. Actual disagreement judgments (far right) can be compared to each factor. If disagreement were fully predicted by a

single regressor (e.g., *Binary Divergenc*e), we would expect actual disagreement judgments to match the predictions of that regressor. Instead, disagreement appears as a summed combination of multiple regressors, not mapping precisely to a single factor.

Our nested model comparison proceeded as follows. First, *Binary Divergence* [B], was only sensitive to whether the two beliefs were on the same side of the midpoint. The second model, *Continuous Divergence* [B+C], further included the effect of continuous divergence. The third model, *Extremity* [B+C+E], additionally captured how disagreement judgments might be amplified as participant belief moves farther away from the midpoint. The fourth model, *Direction* [B+C+D], captured directional asymmetries in disagreement as well as divergence. Finally, we considered *Rich Belief* [B+C+E+D], which captured whether extremity and direction jointly provide additional predictive value. Note that the last three models also include interactions with continuous divergence. Nested model comparisons indicated that adding such interaction effects improved model performance (see Figure 7S in Supplementary Materials). The nested structure of the models allows us to examine the additional contribution of each factor through model comparisons (see Figure 4A).

Is Divergence Sufficient to Capture Disagreement?

The Rich Belief model, which contained all regressors, was the best performing model. The Extremity model also performed well (reported as the mean ± 95% credibility interval; ELPD-13.8 ± 10.8 less than the Rich Belief model; less negative is better), whereas Continuous Divergence and Direction failed to capture important variance (ELPD difference; -160.8 ± 37.6 and -165.2 ± 37.5, respectively), and Binary Divergence performed very poorly (ELPD difference; -755.6 ± 85.5). These analyses show that predicting disagreement judgments requires going

beyond divergence, and that extremity in particular plays a large role in characterizing judgments of disagreement.

Beyond examining the adequacy of divergence across our entire dataset, we can ask whether *individual* responses are also best characterized by the Rich Belief model. This is important, as it is possible for models to better capture population-level trends without describing any particular participant's behavior well (c.f. Eberhardt & Danks, 2011). To investigate the alignment between individual responses and model predictions, we performed the following analysis. We simulated model predictions for each participant's judgments using the parameters from the best-fitting population-level models. We then computed, for each participant, the correlation between their 12 disagreement judgments and model predictions across those judgments, as well as the difference between their judgments and predictions (the mean squared error). We finally computed the proportion of participants that each model best characterized along both metrics (see Supplementary Materials C for further details). This analysis revealed that the Rich Belief model better characterized individual participants' estimates than other models, with Extremity trailing closely behind, though each model best characterized some subset of participants (see Figure 4B).

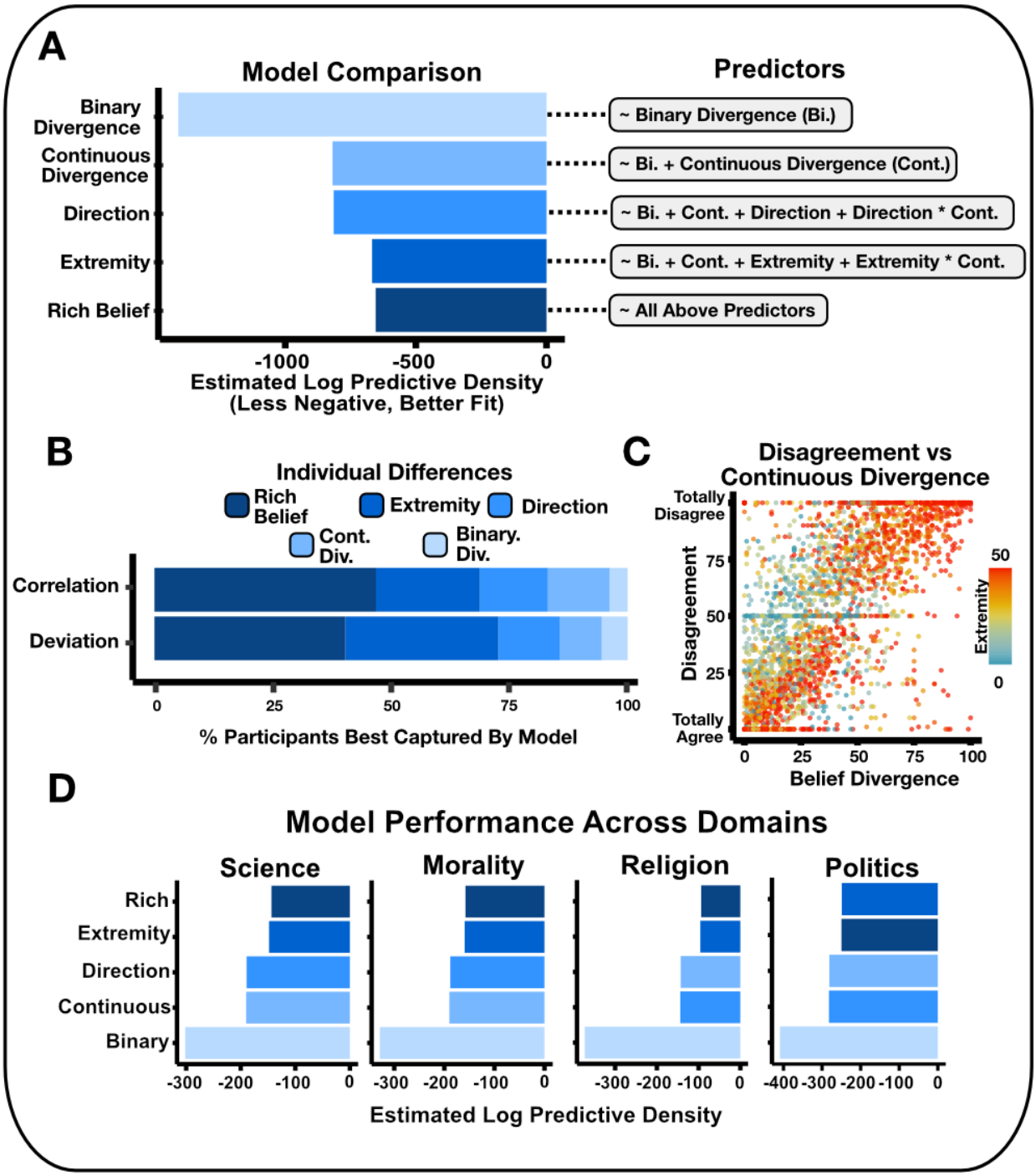
We also note that a logistic regression predicting a binarized agreement vs. disagreement measure (computed based on whether participants' disagreement ratings fell above or below the midpoint) replicates the analyses reported above, reinforcing our conclusion that factors beyond divergence do not merely play a role in moderating inferences about the strength of disagreement, but also play a role in people's judgments of what constitutes disagreement (vs. agreement) itself. In other words, it could have been the case that the additional predictors we examined were useful for predicting the extent of disagreement, but would not have been useful for predicting whether

another's belief would be judged as disagreement—but our data do not support this possibility, and instead suggest that richer models better predict both the presence and extent of disagreement than binary or continuous divergence.

Lastly, there remains the possibility that participants interpret the probability scale as a logit scale, which could offer an alternative explanation for the non-linear effects captured by extremity (e.g., a jump from .50 to .75 would be smaller than a jump from .75 to 1.00). A logit interpretation of continuous divergence performs 260.9 ± 47.2 ELPD (95% Credibility Interval) *worse* than the Extremity model – implying that extremity and a logit scale are not interchangeable, and that our formulation of extremity better predicts disagreement judgements.

Figure 4

Predicting Disagreement Judgments



Note. (A) Larger (less negative) ELPD values indicate better predictive performance. The Rich Belief model performs the best. (B) The Rich Belief model best describes a plurality of

participants. (C) Disagreement generally tracks continuous divergence, but with an additional effect of extremity (red) pushing judgments to the bounds. This illustrates why a richer model of belief (i.e., one that takes extremity into account) can better predict disagreement. (D) The Rich Belief model consistently performs well in predicting disagreement judgments across domains, but does not significantly differ from Extremity.

Does Relying on Divergence Yield Poor Predictions of Social Consequences?

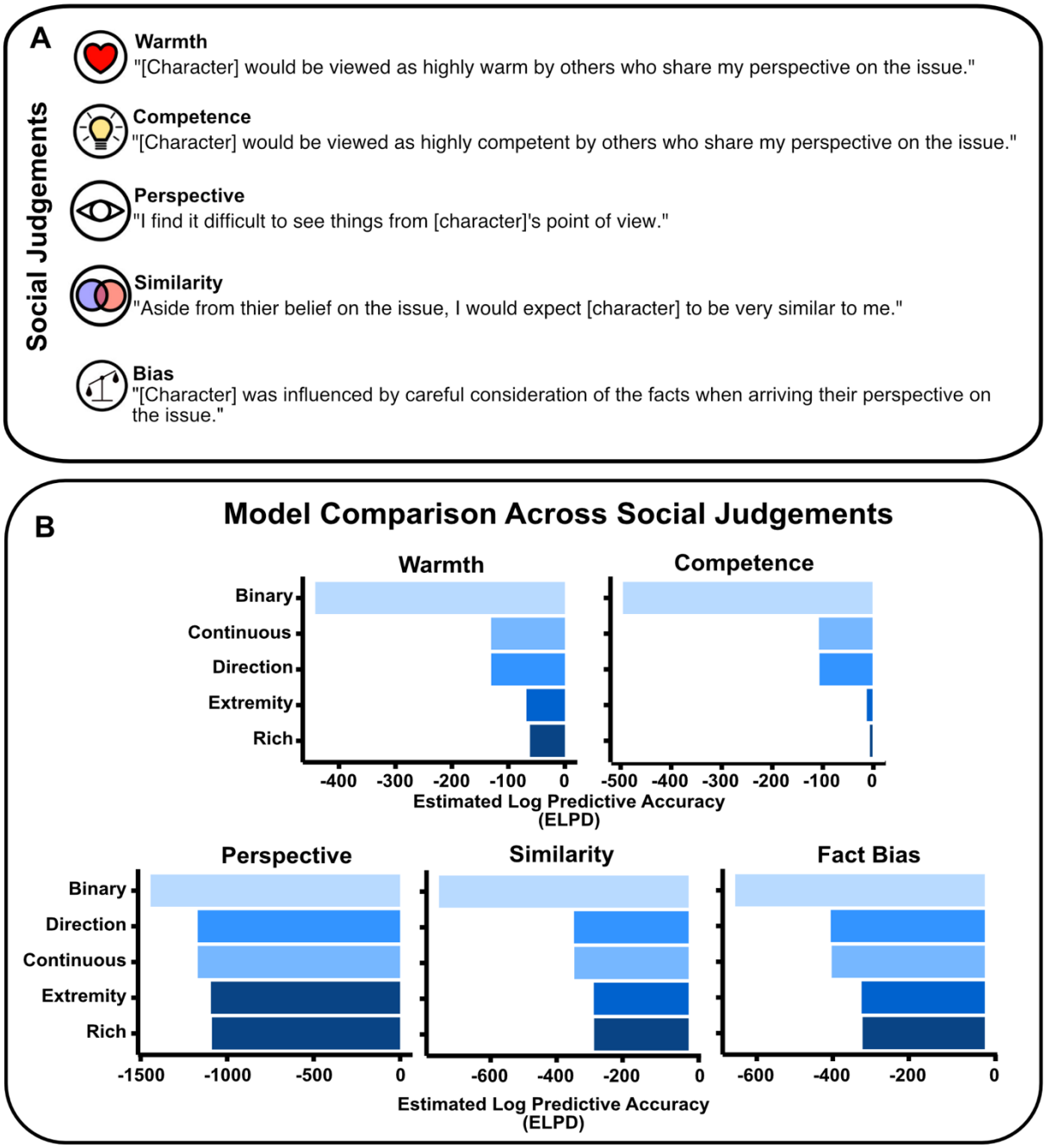
Our first analyses indicate that divergence alone is not sufficient for predicting people's judgements of interpersonal disagreement. However, improved predictions of disagreement do not necessarily imply improved prediction of, for example, social judgements that might ensue as a result of differences in beliefs. Additional analyses allow us to ask whether a rich representation of belief also yields improved predictions of social judgements (not just disagreement). To answer this question, we used a similar nested model comparison approach to test whether the same aspects of belief – extremity and direction – are important to predicting social consequences of disagreement (as opposed to disagreement alone). The social consequences that we consider are judgments of warmth, competence, bias, similarity, and perspective taking.

While the Rich Belief model performed best numerically, it did not significantly outperform the Extremity model (see Figure 5B, or Table 1S for exact values). This suggests that, for the social judgements included here, extremity likely plays an important role in predicting social outcomes. But unlike our first analysis, direction did not play a statistically significant role. Overall, our nested model comparison still suggests that a more nuanced representation of belief better predicts social judgements than a model based on divergence alone. This finding – that the Rich Belief and Extremity models outperform simpler alternatives on social judgement prediction

– persists across domains, suggesting that our findings hold broadly (Figure 5B; See supplementary table 2S for exact values).

Figure 5

Predicting Social Judgments



Note. (A) Alongside making disagreement judgments, participants made additional social judgments about another person after seeing a reported belief from them. (B) We conducted the same nested model comparisons from Figure 4A for social outcomes. The Rich Belief model numerically wins for each comparison, though Rich Belief and Extremity were statistically indistinguishable in model performance. Overall, these results support the idea that factors beyond divergence predict social outcomes, with a significant role for extremity.

General Discussion

From motivating scientists to conduct experiments that advance debates (Lamers et al., 2021) to motivating politicians to jail dissidents or protestors (Watts, 2019), judgements of disagreement can have major consequences. While past research has shed light on such consequences, our research investigates how people judge whether and how strongly others disagree with them to begin with. Our findings show that the intuitive notion of disagreement judgments as capturing differences in beliefs is a necessary yet insufficient characterization. Instead, disagreement judgments capture rich properties of beliefs, from how extreme one's own views are to the direction in which others diverge. Moreover, the differences between divergence and perceived disagreement are not merely statistically significant, but also practically consequential: models of belief that incorporate extremity in addition to divergence better predict key social outcomes, from whom we judge as cold or incompetent to whom we are willing to extend our empathy to.

These results raise substantive theoretical questions about the nature of disagreement (that speak to large literatures on disagreement and social learning across disciplines, reviewed in Oktar & Lombrozo, 2025). Most pressingly, *why* do these additional factors of extremity and direction

affect disagreement judgments? Borrowing from an existing framework for why disagreements rarely change views (Oktar & Lombrozo, 2026), we can broadly think of three kinds of explanations: there might be informational, functional, and ontological reasons for why disagreement judgments benefit from rich representations of beliefs.

From an informational (i.e., epistemic) standpoint, we might consider why tracking belief extremity (beyond divergence) could provide useful information about the nature of an interpersonal disagreement. One possibility is that extremity functions as a proxy for disagreement about a broader range of relevant issues. For example, if one person holds an extreme view about vaccination, they might be more likely to cast those with moderate views as disagreeing because they expect them to disagree about a broader range of issues beyond vaccination itself, for instance concerning autonomy and public health. Most expansively, disagreement can potentially be cast as an evaluation of the misalignment between two individuals' worldviews, not just individual beliefs (Oktar, Sucholutsky, et al., 2024). To the extent that an extreme view enables stronger inferences about other views that someone might hold (Dellsén, 2024), people could have a rational basis for considering extremity in their disagreement judgments.

From a functional or social standpoint, we can ask what role disagreement judgments play in guiding interpersonal behavior and outcomes, and how extremity or direction might matter for this role. For instance, consider the role beliefs play as signals of group membership (Golman, 2016; see also Vesga et al., 2025; Westra, 2023). Judging another individual as disagreeing with oneself could mark them as an out-group member and guide social behaviors (e.g., that one should avoid affiliating with them). For this role, the direction of a difference in views might be highly informative, as individuals with more extreme views on one's own side of a debate might be seen

as more prototypical in-group members that one might wish to affiliate with, and hence should avoid disagreeing with (Goldenberg, 2023).

From an ontological standpoint, learning that someone has an extreme view on an issue could suggest that they view themselves as having the objectively correct belief on that issue (as opposed to merely having a subjective opinion), much as learning about high levels of consensus on an issue can lead to the inference that there is an objective fact of the matter with respect to it (Heiphetz & Young, 2017; Ayars & Nichols, 2020). Given that subjectivity can shield people from engaging deeply with differences in views (Kivy, 2015), extremity might influence disagreement judgments by licensing different inferences about the way in which others conceptualize and hold their differing beliefs, with middling views suggesting subjectivity and hence weakening judgments of disagreement.

It is likely that all three considerations (and possibly their interactions) are necessary to explain why judgments of disagreement capture more than mere differences in beliefs, much as all three are crucial for understanding how people respond to dissent. A key direction for future research is thus to uncover *when* and *how* each consideration matters—for instance, by examining whether disagreement judgments in different domains (e.g., science vs. religion) are influenced by different informational, functional, and ontological drivers.

Importantly, while we focused our attention on the constraints that divergence imposes on the *functional* relationship between disagreement and belief—namely, that divergence presupposes linearity and symmetry—we did not examine the many constraints that this characterization imposes on the *inputs* and *outputs* of disagreement. For instance, it is plausible that disagreement judgments also take affective states as inputs (e.g., whether a dispute is angry or peaceful). In terms of outputs, a plausible interpretation of our social judgment results is that when

we perceive disagreements, we also draw a swath of social inferences about disagreeing others, including their individual properties and relationships with others. Our findings suggest that further studies of these constraints on the inputs, function, and outputs of disagreement judgments will yield fruitful insights into what disagreement is, as well as its causes and consequences. In particular, moving beyond single linear scales (such as those used in our studies) to examining sets of issues, or even asking participants to provide rich, qualitative descriptions of their stances or perceptions of others, will be fruitful in providing a richer picture of the mechanisms and consequences of disagreement.

In keeping with the literature on disagreement, our experiment considered beliefs as view*points*, rather than exploring the broader space afforded by representing beliefs as *distributions* of probabilities (for instance, KL-divergence over probability distributions can offer a richer operationalization of disagreement). Future research should explore a broader set of paradigms for eliciting beliefs, and broader sets of issues, to examine the robustness of our findings. Relatedly, there are many ways of formulating non-linearities in the functional relationship between divergence and disagreement, including those beyond what we examined. For instance, the midpoint non-linearity might only be relevant when the midpoint-crossing happens for people who are initially some sufficient distance away from one another. These and related questions are worth examining in future research.

While many questions remain, our findings suggest that across these contexts, judgments of disagreement are likely to be woven into the rich tapestry of social inferences that guide people's lives, and unlikely to merely track differences in belief.

References

Aaronson, S. (2005). The complexity of agreement. *Proceedings of the Thirty-Seventh Annual ACM Symposium on Theory of Computing*, 634–643. https://doi.org/10.1145/1060590.1060686

Abelson, R. P. (1995). Attitude extremity. In *Attitude strength* (pp. 25-41). Psychology Press.

Amemiya, J., Heyman, G. D., & Gerstenberg, T. (2024). Children use disagreement to infer what happened. *Cognition*, *250*, 105836. https://doi.org/10.1016/j.cognition.2024.105836

Aumann, R. J. (1976). Agreeing to Disagree. *The Annals of Statistics*, *4*(6), 1236–1239.

Ayars, A., & Nichols, S. (2020). Rational learners and metaethics: Universalism, relativism, and evidence from consensus. Mind & Language, 35(1), 67-89.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally Interacting Minds. *Science*, *329*(5995), 1081–1085. https://doi.org/10.1126/science.1185718

Barnidge, M. (2018). Social Affect and Political Disagreement on Social Media. *Social Media + Society*, *4*(3), 2056305118797721. https://doi.org/10.1177/2056305118797721

Blakey, K. H., & Ronfard, S. (2026). How to Increase Children's and Adults' Interest in Learning From Disagreement. *Developmental Science*, *29*(4), e70237.

Budescu, D. V., Rantilla, A. K., Yu, H.-T., & Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, *90*(1), 178–194. https://doi.org/10.1016/S0749-5978(02)00516-2

Bullock, J. G. (2009). Partisan Bias and the Bayesian Ideal in the Study of Public Opinion. *The Journal of Politics*, *71*(3), 1109–1124. https://doi.org/10.1017/S0022381609090914

Cheek, N. N., Blackman, S. F., & Pronin, E. (2021). Seeing the subjective as objective: People perceive the taste of those they disagree with as biased and wrong. *Journal of Behavioral Decision Making*, *34*(2), 167–182. https://doi.org/10.1002/bdm.2201

Christensen, D. (2007). Epistemology of Disagreement: The Good News. *The Philosophical Review*, *116*(2), 187–217.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. In *Advances in Experimental Social Psychology* (Vol. 40, pp. 61–149). Academic Press. https://doi.org/10.1016/S0065-2601(07)00002-0

Davis, M. H. (1980). *Interpersonal Reactivity Index*. https://doi.org/10.1037/t01093-000

Dellsén, F. (2024). Interthematic polarization. *American Philosophical Quarterly*, *61*(1), 45-58.

Eberhardt, F., & Danks, D. (2011). Confirmation in the Cognitive Sciences: The Problematic Case of Bayesian Models. *Minds and Machines*, *21*(3), 389–410. https://doi.org/10.1007/s11023-011-9241-3

Egan, A. (2010). Disputing about Taste. In R. Feldman & T. A. Warfield (Eds.), *Disagreement* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199226078.003.0011

Frances, B., & Matheson, J. (2019). Disagreement. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/disagreement/

Goldenberg, A., Abruzzo, J. M., Huang, Z., Schöne, J., Bailey, D., Willer, R., Halperin, E., & Gross, J. J. (2023). Homophily and acrophily as drivers of political segregation. *Nature Human Behaviour*, *7*(2) https://doi.org/10.1038/s41562-022-01474-9

Golman, R., Loewenstein, G., Moene, K. O., & Zarri, L. (2016). The preference for belief consonance. *Journal of Economic Perspectives*, *30*(3), 165-188.

Haghtalab, N., Jackson, M. O., & Procaccia, A. D. (2021). Belief polarization in a complex world: A learning theory perspective. *Proceedings of the National Academy of Sciences*, *118*(19), e2010144118. https://doi.org/10.1073/pnas.2010144118

Harris, P. L. (2012). *Trusting What You're Told: How Children Learn from Others*. Harvard University Press.

Heiphetz, L., & Young, L. L. (2017). Can only one person be right? The development of objectivism and social preferences regarding widely shared and controversial moral beliefs. Cognition, 167, 78–90. https://doi.org/10.1016/j.cognition.2016.05.014Hopmann, D. N., Bjarnøe, C., & Wonneberger, A. (2020). Responding to Interpersonal Political Disagreement. *International Journal of Public Opinion Research*, *32*(1), 66–88. https://doi.org/10.1093/ijpor/edz011

Howe, L. C., & Krosnick, J. A. (2017). Attitude Strength. *Annual Review of Psychology*, *68*(1), 327–351. https://doi.org/10.1146/annurev-psych-122414-033600

Iyengar, S., & Westwood, S. J. (2015). Fear and Loathing across Party Lines: New Evidence on Group Polarization. *American Journal of Political Science*, *59*(3), 690–707. https://doi.org/10.1111/ajps.12152

Kakava, C. (2002). Opposition in Modern Greek discourse: Cultural and contextual constraints. *Journal of Pragmatics*, *34*(10), 1537–1568. https://doi.org/10.1016/S0378-2166(02)00075-9

Kelly, T. (2010). Peer Disagreement and Higher Order Evidence. In R. Feldman & T. A. Warfield (Eds.), *Disagreement*. Oxford University Press.

Kennedy, K. A., & Pronin, E. (2008). When Disagreement Gets Ugly: Perceptions of Bias and the Escalation of Conflict. *Personality and Social Psychology Bulletin*, *34*(6), 833–848. https://doi.org/10.1177/0146167208315158

Kivy, P. (2015). De gustibus: Arguing about taste and why we do it. Oxford University Press. https://doi.org/10.1093/acprof:Oso/9780198746782.001.0001Kozitsin, I. V. (2022). A general framework to link theory and empirics in opinion formation models. *Scientific Reports*, *12*(1), 5543. https://doi.org/10.1038/s41598-022-09468-3

Lamers, W. S., Boyack, K., Larivière, V., Sugimoto, C. R., van Eck, N. J., Waltman, L., & Murray, D. (2021). Meta-Research: Investigating disagreement in the scientific literature. *Elife*, *10*, e72737.

Langenhoff, A. F., Engelmann, J. M., & Srinivasan, M. (2023). Children's developing ability to adjust their beliefs reasonably in light of disagreement. *Child Development*, *94*(1), 44–59. https://doi.org/10.1111/cdev.13838

MacFarlane, J. (2007). Relativism and disagreement. *Philosophical Studies*, *132*(1), 17–31. https://doi.org/10.1007/s11098-006-9049-9

Matz, D. C., & Wood, W. (2005). Cognitive dissonance in groups: The consequences of disagreement. *Journal of Personality and Social Psychology*, *88*(1), 22–37. https://doi.org/10.1037/0022-3514.88.1.22

Nir, L. (2011). Disagreement and Opposition in Social Networks: Does Disagreement Discourage Turnout? *Political Studies*, *59*(3), 674–692. https://doi.org/10.1111/j.1467-9248.2010.00873.x

Oktar, K., Byers, J. B., & Lombrozo, T. (2024). Are Disagreements Just Differences in Beliefs? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0). https://escholarship.org/uc/item/2k81k4qs

Oktar, K., & Lombrozo, T. (2026). How beliefs persist amid controversy: The paths to persistence model. *Psychological Review, 133*(3), 636–665. https://doi.org/10.1037/rev0000583

Oktar, K., Sucholutsky, I., Lombrozo, T., & Griffiths, T. L. (2024). Dimensions of disagreement: Divergence and misalignment in cognitive science and artificial intelligence. *Decision*, 01–12. https://doi.org/10.1037/dec0000244

Palmira, M. (2018). Disagreement, Credences, and Outright Belief. *Ratio*, *31*(2), 179–196. https://doi.org/10.1111/rati.12163

Pham, T., & Buchsbaum, D. (2020). Children's use of majority information is influenced by pragmatic inferences and task domain. *Developmental Psychology*, *56*(2), 312–323. https://doi.org/10.1037/dev0000857

Pool, G. J., Wood, W., & Leck, K. (1998). The self-esteem motive in social influence: Agreement with valued majorities and disagreement with derogated minorities. *Journal of Personality and Social Psychology*, *75*(4), 967–975. https://doi.org/10.1037/0022-3514.75.4.967

Price, V., Cappella, J. N., & Nir, L. (2002). Does Disagreement Contribute to More Deliberative Opinion? *Political Communication*, *19*(1), 95–112. https://doi.org/10.1080/105846002317246506

Rees-Miller, J. (2000). Power, severity, and context in disagreement. *Journal of Pragmatics*, *32*(8), 1087–1111. https://doi.org/10.1016/S0378-2166(99)00088-0

Ren, Z. (Bella), & Schaumberg, R. (2024). Disagreement Gets Mistaken for Bad Listening. *Psychological Science*, *35*(5), 455–470. https://doi.org/10.1177/09567976241239935

Samuelson, A., & Dahl, A. (2026). Can reasonable people disagree about right and wrong? The determinants and implications of perceiving moral reasonableness. *Cognition*, *274*, 106597.

Smaldino, P. E., Moser, C., Pérez Velilla, A., & Werling, M. (2023). Maintaining Transient Diversity Is a General Principle for Improving Collective Problem Solving. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 17456916231180100. https://doi.org/10.1177/17456916231180100

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780–805. https://doi.org/10.1037/a0015145

Stromer-Galley, J., Bryant, L., & Bimber, B. (2015). Context and Medium Matter: Expressing Disagreements Online and Face-to-Face in Political Deliberations. *Journal of Deliberative Democracy*, *11*(1), Article 1. https://doi.org/10.16997/jdd.218

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2024). Pareto Smoothed Importance Sampling. *Journal of Machine Learning Research*, *25*(72), 1–58.

Vesga, A., Van Leeuwen, N., & Lombrozo, T. (2025). Evidence for multiple kinds of belief in theory of mind. *Journal of Experimental Psychology: General*, *154*(8), 2241.

Watts, R. (2019). Criminalizing Dissent: The Liberal State and the Problem of Legitimacy (1st ed.). Routledge. https://doi.org/10.4324/9781351039581

Westra, E. (2023). Symbolic belief in social cognition. *Philosophical perspectives*, *37*(1), 388-408.

Ziembowicz, K., Rychwalska, A., & Nowak, A. (2023). Arguments at Odds—Dyadic Turn-Taking and Conflict Development in Consensus-Making Groups. *Small Group Research*, *54*(4), 551–589. https://doi.org/10.1177/10464964221118674