PNAS Nexus, 2024, 3, pgae393
https://doi.org/10.1093/pnasnexus/pgae393
Advance access publication 15 October 2024
Research Report

How rational inference about authority debunking can curtail, sustain, or spread belief polarization

Setayesh Radkani[a], Marika Landau-Wells[b] and Rebecca Saxe[a,c,*]

[a]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[b]Travers Department of Political Science, University of California, Berkeley, CA 94705, USA
[c]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*To whom correspondence should be addressed: Email: saxe@mit.edu
Edited By Eugen Dimant

Abstract

In polarized societies, divided subgroups of people have different perspectives on a range of topics. Aiming to reduce polarization, authorities may use debunking to lend support to one perspective over another. Debunking by authorities gives all observers shared information, which could reduce disagreement. In practice, however, debunking may have no effect or could even contribute to further polarization of beliefs. We developed a cognitively inspired model of observers' rational inferences from an authority's debunking. After observing each debunking attempt, simulated observers simultaneously update their beliefs about the perspective underlying the debunked claims and about the authority's motives, using an intuitive causal model of the authority's decision-making process. We varied the observers' prior beliefs and uncertainty systematically. Simulations generated a range of outcomes, from belief convergence (less common) to persistent divergence (more common). In many simulations, observers who initially held shared beliefs about the authority later acquired polarized beliefs about the authority's biases and commitment to truth. These polarized beliefs constrained the authority's influence on new topics, making it possible for belief polarization to spread. We discuss the implications of the model with respect to beliefs about elections.

Keywords: polarization, Bayesian inference, credibility, misinformation, inverse planning

Significance Statement

Polarized beliefs about election fairness put democracy at risk. Debunking is one of only a few tools for combating the polarization of beliefs in a divided society. We develop a cognitively inspired model of how observers interpret an authority's debunking actions. The act of debunking can lead rational observers' beliefs to converge or remain polarized, and also affects the reputation of the authority. An authority's acquired reputation in turn can spread polarization to new topics. Applied to the case of beliefs about election fraud, our model provides insights into when and how debunking fails to reduce polarization, and delineates conditions under which authorities perceived as independent can offer an antidote to divided societies.

Introduction

A functioning democracy requires that citizens accept election results when the elections are administered fairly (1–3). One of the most notable indicators of partisan polarization in the United States has been the divergence in beliefs between Democrats and Republicans about the legitimacy of the 2020 presidential election (4–6). Belief in "the big lie" among Republicans has persisted despite a wave of debunking efforts by public officials of both parties and by the media (7, 8). Why have debunking efforts in this domain failed? And under what conditions are debunking efforts likely to be successful in general?

Prior research has shown that confidence in election outcomes can be enhanced by the presence of independent observers who certify that electoral procedures were respected and followed (9–11). These observers also debunk false claims regarding election manipulation (e.g. (12)). Authority figures in other domains such as public health and climate science also act to debunk harmful forms of misinformation (13–15).

However, an authority's debunking efforts are not always effective, and the changes in beliefs induced by authorities' efforts are both variable and hard to predict (16–21). Perceptions of partisan bias (e.g. (22–24)) or less-than-scrupulous motives (e.g. (25)) can reduce the impact of factually accurate debunking. Of course, not all authority debunking is consistent with the truth, particularly when the authority derives a benefit from a piece of misinformation, e.g. (26–28). Moreover, people may be initially uncertain

Competing Interest: The authors declare no competing interests.
Received: April 8, 2024. **Accepted:** August 19, 2024

about their prior beliefs (29–32). Many beliefs pertain to domains where knowledge accumulates slowly and lacks epistemic certainty at the time debunking occurs (e.g. public health, climate change, election fairness monitoring in real-time and claims about election results prior to final vote tabulation, for example the 2024 Venezuelan election (33)). Fundamental uncertainty (as opposed to measurement error) can influence the extent to which debunking results in converging or diverging beliefs (34–36).

In this article, we construct a formal Bayesian model of the effects of debunking on observers' beliefs to synthesize these diverse observations. The unwillingness to update beliefs given disconfirming information has sometimes been characterized as evidence of irrationality (e.g. (37)). However, following recent work in Bayesian network models (e.g. (38)), our model reveals many conditions under which debunking rationally fails to reduce confidence in a debunked perspective and instead engenders divergent beliefs about the authority (i.e. new polarization). Our model also shows how polarization may rationally spread into additional topics.

Bayesian models posit that people interpret and explain evidence by updating their mental causal model of how the evidence was generated (38–42). A Bayesian approach highlights that after observing evidence, people jointly update graded beliefs probabilistically across the whole network. Thus, even when two groups of people see the same evidence, and share the same mental model of potential causal pathways leading to that evidence, they may reach different conclusions about the most likely cause of the evidence if they hold different prior beliefs about the causal pathways. In some cases, the same piece of evidence can even strengthen directly opposing beliefs (38). In this sense, we use the term "rational" to refer to the process by which beliefs are updated, rather than the truth of the resulting beliefs, consistent with common practice in cognitive science, e.g. (43, 44).

In contrast to prior Bayesian models designed specifically to capture political polarization (e.g. (45)), we use a generic cognitive framework to model observers' inferences from others' actions. In this framework, the underlying generative model posits that people plan actions to achieve their desires given their beliefs, or put differently, that people choose the action that will maximize their own expected utility (the principle of rational action (46, 47)). By inverting this generative model, known as inverse planning (46), observers estimate the unobservable internal states (i.e. the actor's beliefs and desires) that most likely generated the observed actions. Inverse planning models provide detailed quantitative fits to human judgements in many settings (46–49). The model applied here was initially developed and validated to capture inferences from observing an authority's choice to punish, or not punish, a potentially harmful action (Radkani et al. (50)).

We propose that observers interpret authority debunking, like other actions, by inverse planning. To simulate the effects of authority debunking, we start with minimal assumptions: two subgroups of observers within a society differ only in their beliefs about one topic (e.g. election security). Then, an authority faces the choice of whether to debunk one claim, drawn from one perspective on the disputed topic (e.g. perspective: the election was stolen; claim: voter rolls in a specific district were manipulated). Observers assume that the authority chooses between debunking and not debunking the claim based on the authority's own expected utility. All observers share an intuitive theory of the authority's decision-making process: authorities may be motivated to debunk a claim because they believe it is likely to be false (if the authority gains utility from the accuracy of accepted claims) and/or because the claim harms their own interests (if the authority gains utility from undermining the perspective or source of the claim). Given this intuitive theory, all observers then rationally update their beliefs about the authority based on the observed choices of that authority over time, here a sequence of debunking five claims, all of which originate from the same perspective (e.g. state that dead people did not vote; state that illegal immigrants did not vote).

Using simulations, we illustrate the effects of debunking on the evolution of beliefs about both the topic and the authority. In each simulation, the two groups initially share beliefs on all dimensions except about their perspective on the focal topic. This allows us to (i) identify the conditions leading to convergence or continued polarization of beliefs about the topic; (ii) demonstrate that debunking can lead to divergent and polarized beliefs about the authority's epistemic motivation and bias, despite identical initial beliefs; and (iii) investigate how inferences about the authority can spill over and cause a rational divergence of beliefs in a new topic area where observers are initially uncertain and the authority attempts to engage in debunking.

This approach yields several contributions. First, we establish a set of minimal conditions for belief convergence or continued polarization among observers. Our model adopts a neutral stance regarding the underlying truth of the claim evaluated by observers and does not assume group differences in either rationality or ability. Instead, the two groups in our model differ only in their perspectives on one initial topic, which may be correlated with any other descriptive difference (e.g. gender, religion, political partisanship). Though differences between groups in their epistemic values or abilities could contribute to belief convergence or polarization (e.g. (51)), our simulations show that they are not necessary for either outcome. These minimal conditions allow our model to speak to a wide variety of real-world divisions, not limited to those where belief accuracy and group membership are correlated, e.g. (52).

Second, our model provides insight into the cognitive processes associated with belief change. Experimental and observational studies have demonstrated the significance of a source's credibility for persuasion (53–55) and debunking (52, 56–59). An inverse planning approach allows us to unbundle two components of credibility from an observer's perspective (epistemic motivation and bias) and then demonstrate how these components interact with belief uncertainty to produce nuanced effects of the authority's reputation on belief updating. This approach also extends the formal literature on the processing of political information (e.g. (45, 60, 61)) and contributes to the integration of cognitive science into the study of political behavior, e.g. (62–64).

Computational framework

To adapt the inverse planning framework to the context of debunking, we must characterize how authorities are *perceived* to make debunking decisions. That is, we must define the observers' mental model of the authority's expected consequences of debunking or not debunking (i.e. each option's utilities), and how much the authority values and pursues those consequences (i.e. their motives), in choosing whether or not to debunk a claim. The inverse planning model (of observers' inferences) thus begins with a generative planning model (of authority's choices) (Fig. 1A). Note that for this approach, the planning model need not be a correct or complete model of the authority's actual planning process; it only has to capture the planning process as imagined by the observers (48).

The generative model treats debunking as causing overall harmful consequences for the perspective from which the claims

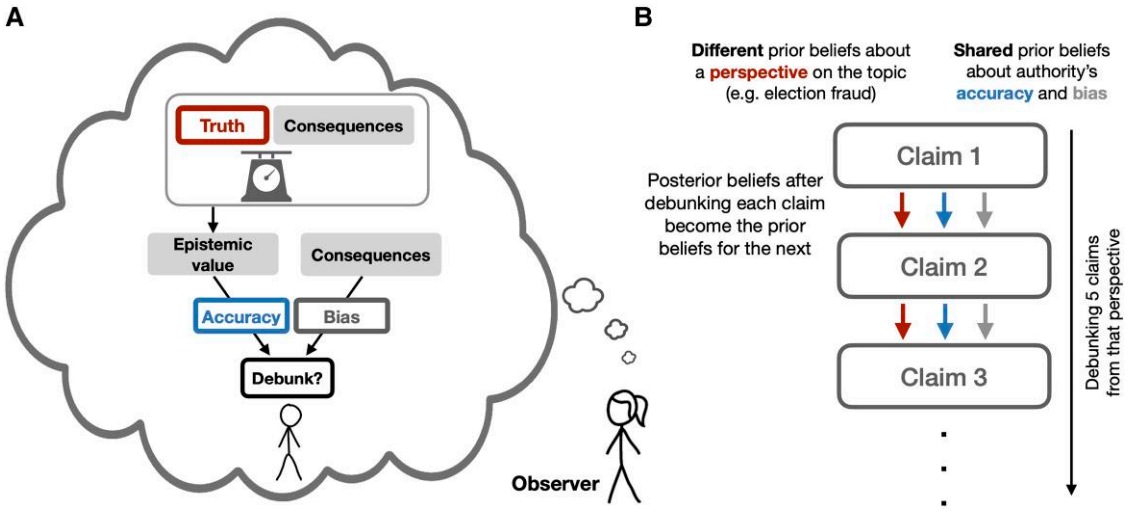

Fig. 1. A) Conceptual depiction of the inverse planning model and B) schematic of model simulations. In each pair of simulations, two subgroups initially hold different beliefs about a topic, but hold shared beliefs about the authority's accuracy motive and bias. After observing each of the authority's debunking decisions (i.e. debunking the claims from one perspective), observers use their mental model of the authority's decision-making to simultaneously update their beliefs about the perspective (i.e. its truth) and the authority's accuracy motive and bias. Observers' posterior beliefs after each debunking decision then serve as the prior beliefs for judging the next decision.

are drawn. This harm (negative utility) could be understood as making the claims less likely to be transmitted, or the perspective less likely to be accepted, or as imposing a reputational cost on the source of the claims.

Authorities consider two kinds of utility when planning whether to debunk a claim and thus cause harm to the perspective, or not. First, different authorities place different weights on this negative utility. These weights, here called bias, range from bias against the perspective (a negative weight, i.e. debunking harms the perspective and thus benefits the authority) to bias in favor of the perspective (a positive weight, i.e. debunking harms the perspective and thus also harms the authority). The bias could also be understood as towards the source of the claim, for example gaining utility from harming political competitors or from protecting political allies. Bias creates a motivation to debunk claims if the authority is biased against the claims or their source, and not to debunk claims if the authority is biased in favor of the claims or their source.

Second, the consequences of debunking can be balanced against whether the perspective is likely true or false. This can be understood as an epistemic value, creating higher utility if the debunked claims are likely to be false, and lower utility if the debunked claims are likely to be true. The epistemic value of not debunking is the converse: a higher utility if the claims are likely true and a lower utility if the claims are likely false. Different authorities weigh the epistemic value of debunking or not debunking to different degrees, depending on how much they care about transmitted claims being accurate and truthful, here called their accuracy motive.

To formalize these ideas, we write the authority's expected utility over each action "a" (i.e. debunking or not) as:

$$U_{\text{total}}(a) = \alpha_{\text{target}} U_{\text{target}} + \alpha_{\text{accuracy}} U_{\text{accuracy}}, \tag{1}$$

where U_{target} is the utility associated with the consequences of each possible action "a" for the target. Debunking is harmful, so U_{target} is negative for debunking and zero for doing nothing.

U_{accuracy} denotes the epistemic value of "a," and is defined as the balance between the consequences of "a" and the claim's truth likelihood; i.e. U_{accuracy} is highest when the response is proportional to the likelihood of the claim's truth (varying continuously between 0 and 1), and decreases the more the response is either too lenient (doing nothing when the claim is likely false) or too harsh (debunking when the claim is likely true):

$$\begin{aligned} \text{proportionality}(a) &= (L_{\text{claim truth}} - 1) - U_{\text{target}}(a) \\ U_{\text{accuracy}}(a) &= \|\text{proportionality}(a)\| \end{aligned} \tag{2}$$

Authorities can differ in the weight placed on each of these utilities. An authority's accuracy motive (α_{accuracy}) is sampled continuously between 0 and 1, and bias towards the perspective (α_{target}), is sampled between −0.5 and 0.5 to denote bias against and in favor of the perspective, respectively.

An authority plans to debunk a claim, or not, by comparing the overall utilities of these two options using the softmax decision rule, whereby the β parameter controls the noisiness, or the degree of rationality of the authority in choosing the decision with highest utility.

$$P(a \mid L_{\text{claim truth}}, \alpha_{\text{accuracy}}, \alpha_{\text{target}}, U_{\text{target}}) \propto e^{\beta U_{\text{total}}(a)}. \tag{3}$$

By inverting this generative model of an authority's debunking decision, observers simultaneously update their beliefs about the truth of the claims' perspective (or source), and the authority's motives. That is, an observer with prior beliefs about the truthfulness of the perspective and authority's motives, $P(L_{\text{claim truth}}, \boldsymbol{\alpha})$, as well as their appraisal of the cost of debunking for the target, U_{target}, uses Bayesian inference to update their beliefs about the perspective and authority's motives, simultaneously.

$$\begin{aligned} &P(L_{\text{claim truth}}, \boldsymbol{\alpha} \mid a, U_{\text{target}}) \\ &\quad \propto P(a \mid L_{\text{claim truth}}, \boldsymbol{\alpha}, U_{\text{target}}) \times P(L_{\text{claim truth}}, \boldsymbol{\alpha}), \end{aligned} \tag{4}$$

where $P(a|L_{\text{claim truth}}, \boldsymbol{\alpha}, U_{\text{target}})$ is the authority's policy derived in Eq. 3.

These equations determine observers' belief updates after observing one debunking decision by one authority. To simulate the evolution of beliefs given evidence of a series of debunking decisions within one domain, we iterate Bayesian inferences; that is, the posterior belief after observing one decision becomes the prior belief for the next observation, and this sequence is repeated five times.

Results

We simulated societies in which two subgroups of observers have different initial perspectives on a topic: a proponent subgroup who believe claims drawn from one perspective to be on average likely true, and an opponent subgroup who believe claims drawn from that perspective to be on average likely false. Critically, observers in both subgroups initially have shared beliefs about an authority. Everyone then observes the authority sequentially debunk five claims from the same perspective. After each time the authority chooses to debunk a claim, observers in both subgroups rationally update their beliefs about the causes of the authority's choice (Fig. 1B). We simulate and study the resulting evolution of observers' beliefs about the topic (i.e. the distribution of likelihoods that any given claim from that perspective is true) and about the authority's accuracy motive and bias.

Note that our model is agnostic as to the ground-truth of the claim. Indeed, our model applies equally to observations of debunking a claim (1) when the claim is false (e.g. debunking the claim that a fair election was unfair); (2) despite the claim being true (e.g. debunking the claim that a fair election was fair); and (3) when there is genuine uncertainty about the claim (e.g. debunking a claim that an election was fair when fairness cannot be established).

We ran 243 pairs of simulations (i.e. one pair contains a proponent and an opponent subgroup, with the same initial beliefs about the authority, making the same observations), testing 3^5 variations of prior belief parameters. Across all simulation pairs, the two subgroups differed in their initial perspectives on the topic by the same amount; the proponent subgroup initially believed that claims from one perspective on the topic are on average 80% likely to be true, while the opponent subgroup believed that claims are on average 20% likely to be true. We systematically varied the subgroups' certainty in their perspective (symmetrically; both subgroups had the same level of uncertainty in a given pair of simulations), as well as the value and uncertainty of their shared beliefs about the authority's motive to be accurate, and bias for or against the perspective (see "Simulations" in Materials and methods).

These simulations support a set of general principles about polarization in these settings. Debunking led to convergence of perspectives on the topic in a minority of simulations. The modal outcome, however, was that perspectives on the topic remained divergent between the two groups, with beliefs about the authority's accuracy and bias additionally becoming polarized in some of those simulations.

Belief dynamics

We demonstrate the evolution of beliefs in two contrasting example settings (Fig. 2). In the first example, both subgroups share priors that the authority's motives include a high motive for accuracy, no bias in favor or against the claims' perspective, and both subgroups are uncertain about their initial beliefs about the topic. In this situation, debunking is effective: after five debunking acts, the proponent subgroup strongly updates their beliefs about the topic so both subgroups eventually consider the claims more likely false than true, and still view the authority as highly motivated by accuracy and quite impartial. This happens because both subgroups believe that the debunking choices are highly motivated by accuracy and not motivated by bias against the claims' perspective or source; therefore, debunking carries a lot of information about the truthfulness of the claims (i.e. that the claims are likely false). Therefore, especially when initial beliefs about the topic are uncertain, debunking causes most of those who believed in the debunked perspective to change their mind. Note that for the proponent subgroup, debunking is initially evidence against the authority's accuracy motive and impartiality. However, beliefs about the authority's motives are hard to change because of their high certainty. Maintaining this high certainty in authority's accuracy and impartiality contributes to the convergence of perspectives on the topic.

By contrast, in the second example, shared observations of the authority's debunking lead to divergence of the subgroups' beliefs. When the subgroups are very confident in their differing beliefs about the topic, and uncertain in their shared beliefs about the authority, the debunking is not effective. The subgroups do not change their perspectives on the topic, and at the same time they acquire very different beliefs about the authority. The subgroup that initially believed the claims still does so, but additionally believes that the authority is not motivated by accuracy and is highly biased against the source. By contrast, the opponent subgroup believes the authority cares a lot about accuracy and is at most slightly biased against the claims' perspective or source.

These two examples demonstrate how the value and uncertainty of initial beliefs about the topic and the authority affect the evolution of these beliefs after observing debunking. Using this formal model, we can go beyond these two examples and study the evolution of subgroup beliefs in a continuous space of shared and differing priors that can vary in both value and confidence.

Drivers of belief polarization

Results from the full set of 243 simulations are shown in (Fig. 3). Focusing first on the convergence of beliefs about the topic, three prominent patterns were evident (Figs. S3 and S4). First, debunking induced convergence of beliefs about the topic when those beliefs were initially uncertain. The more the two groups were certain about their initial beliefs about the topic, the more these beliefs remained different despite debunking. Second, debunking was more effective when both groups initially believed the authority was highly motivated to give accurate information. The more the groups initially suspected the authority of low accuracy motives, the more their beliefs remained polarized despite debunking. Third, debunking was more effective when the authority was initially believed to be impartial or biased in favor of the debunked perspective. The more the groups initially suspected the authority of bias against the debunked perspective, the more their beliefs remained polarized despite debunking. In a small subset of simulations, authority's debunking led to an increase in disagreement about the topic. In these simulations, both subgroups updated their beliefs rationally by decreasing their beliefs in the truth of the claims' perspective, but the opponent subgroup updated their beliefs more strongly than the proponent subgroup.

However, all of these prominent patterns were modulated by systematic interactions, based on the uncertainty as well as the value of the other beliefs. For example, consider the interactions of beliefs about the topic and the authority's accuracy (Fig. 4). When the authority was initially believed to be highly motivated by accuracy with some uncertainty, then the effect of debunking was very sensitive to the groups' confidence in their differing beliefs about the topics. If those beliefs were certain, debunking failed and the groups quickly acquired very polarized beliefs about the authority's accuracy motives. If beliefs about the topic were uncertain, debunking was successful, the perspectives converged and confidence in the authority's accuracy motive

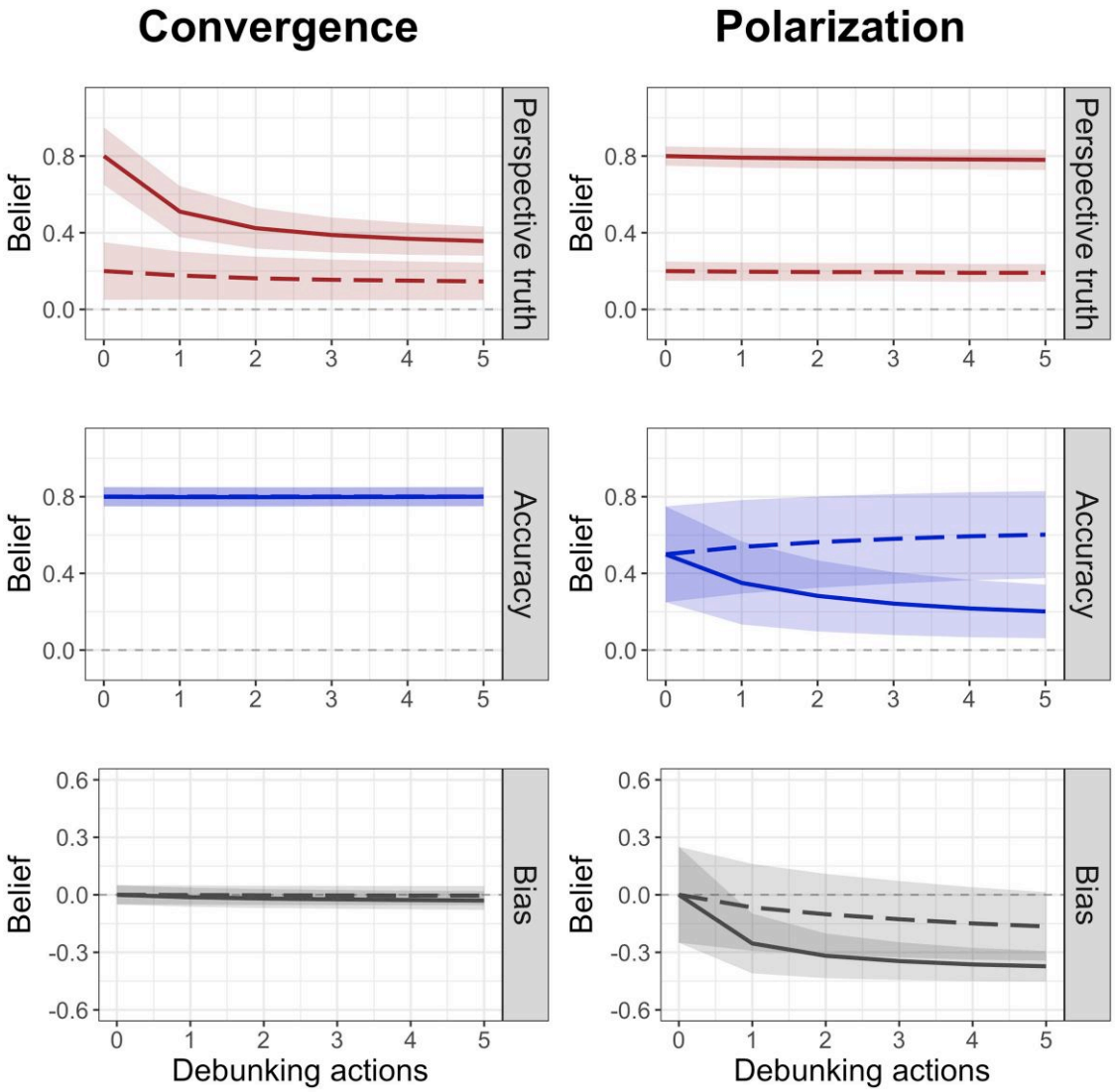

Fig. 2. Model simulations for evolution of beliefs in two example settings. The y-axis represents the mean of the belief distribution, and the x-axis shows the number of observed debunking actions; debunking actions 0 represents the prior belief on each variable. Truth of the perspective varies between 0 and 1, with larger values representing a higher likelihood of truth. Accuracy varies between 0 and 1, with larger values representing a higher motivation to respond in proportion to the truthfulness of the claims—debunk likely false content and do nothing in response to likely true content. Bias varies between −0.5 and 0.5, with positive values representing bias in favor and negative values representing bias against the perspective—e.g. based on the political orientation of the perspective. The shaded ribbon shows the standard deviation of belief distribution. The solid line represents the subgroup who a priori believes the perspective of the claims is likely true (i.e. the proponent subgroup), and the dashed line represents the subgroup who a priori believes the perspective of the claims is likely false (i.e. the opponent subgroup). Left column: A setting in which the subgroups' prior differing beliefs about the topic are uncertain, and their shared prior beliefs about authority's legitimacy are confident—high accuracy and impartiality, leads to effective debunking of false content. Right column: A setting in which the subgroups' prior differing beliefs about the topic are confident and their shared prior beliefs about the authority's motivations are uncertain, leads to polarization of beliefs about the authority in addition to beliefs about the topic remaining polarized.

remained high. The influence of uncertainty was itself modulated by the value of beliefs about authority's accuracy. The influence of uncertainty about the topic disappeared if, holding everything else constant, the authority was initially believed to have low accuracy motive. In that case, beliefs about the topic remained polarized, and beliefs about the authority's accuracy slowly became partially polarized, regardless of the initial certainty of the beliefs about the topic.

Another systematic interaction arose between beliefs about the topic and the authority's bias (Fig. 5). For example, when the authority was initially believed with high certainty to be biased in favor of the debunked perspective (and moderately motivated to be accurate), then the effect of debunking was very sensitive to the groups' certainty in their initial beliefs about the topic. If those beliefs were uncertain, debunking was successful, the beliefs about the topic converged and both groups' confidence in the authority's accuracy motive increased. If the beliefs about the topic were certain, debunking failed and the groups acquired polarized beliefs about the authority's accuracy motive. However, if holding everything else constant, initial beliefs about the authority's bias in favor of the content were uncertain, then beliefs about the topic remained polarized regardless of their initial certainty, and the proponent group instead quickly came to believe that the authority was actually biased against the debunked perspective. If the groups initially suspected that the authority was biased against the debunked perspective, then debunking also failed regardless of certainty about the differing beliefs about the topic or the authority's bias (Fig. S2).

Overall these simulations suggest that the effects of debunking on observers' beliefs could be enormously variable, depending sensitively on the value and uncertainty of prior beliefs about both the topic and the authority's motives (see also Figs. S3 and S4 for a larger set of simulations).

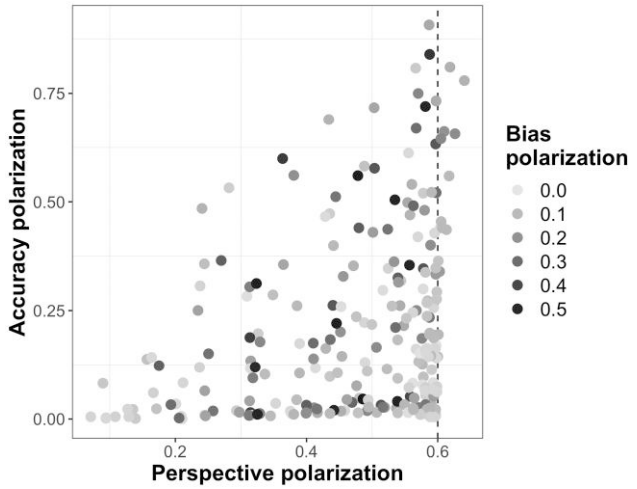

Fig. 3. Relationship between the final belief polarization about the topic (perspective), the authority's accuracy and bias, after five debunking actions. Each data point represents one pair of simulations; therefore, there are 243 points shown in the plot (see Fig. S5 for labels). Belief polarization is defined as the absolute value of the difference between the two subgroups' beliefs (i.e. mean of belief distributions). The x- and y-axis represent polarization in beliefs about the perspective on the topic and the authority's accuracy, respectively; the points are color-coded by polarization in beliefs about the authority's bias. The vertical dashed line represents the initial level of perspective belief polarization.

Polarization in new topic areas

The acquired polarization in beliefs about the authority's motives has implications beyond a single topic and can propagate belief divergence into new topic areas. To explore this effect, we used the final beliefs about authority's motives at the end of the simulations described above (243 pairs of simulations) as prior beliefs for a new topic area where the two subgroups start with shared uncertain beliefs (both groups believe the claims from one perspective are 50% likely to be true but are very uncertain about the probability that any given claim from this perspective is true). We ran the simulations for a perspective on a new topic that is related to the previous one, such that beliefs about the authority's bias would be generalizable across the two topics. We studied the resulting polarization in beliefs, as the authority debunks claims about a perspective in this new topic five times (Fig. 6A).

Divergence in beliefs about the new topic after 5 debunking actions ranged from about 0 to 0.3 (first Qu. = 0.0389, median = 0.0771, third Qu. = 0.134), creating at most about half the magnitude of divergence as in the original domain, or 60% of the maximum possible divergence in the new topic. The polarization in beliefs about the authority's motives (accuracy and bias) was a major determinant of how much beliefs about the new topic would polarize. If the two groups had acquired a larger difference in their beliefs about the authority's accuracy, and/or their beliefs about the authority's bias, they ended up more polarized in their beliefs about the new topic (Fig. 6B). Differences in beliefs about accuracy or bias alone can lead to divergence in beliefs about the new topic (e.g. simulations 62 and 76).

Beliefs about the new topic polarize particularly when the proponent subgroup (in the original domain) have acquired either certain beliefs that the authority does not care about accuracy at all (e.g. simulation 106), or relatively certain beliefs that the authority is highly biased against such perspectives (e.g. simulation

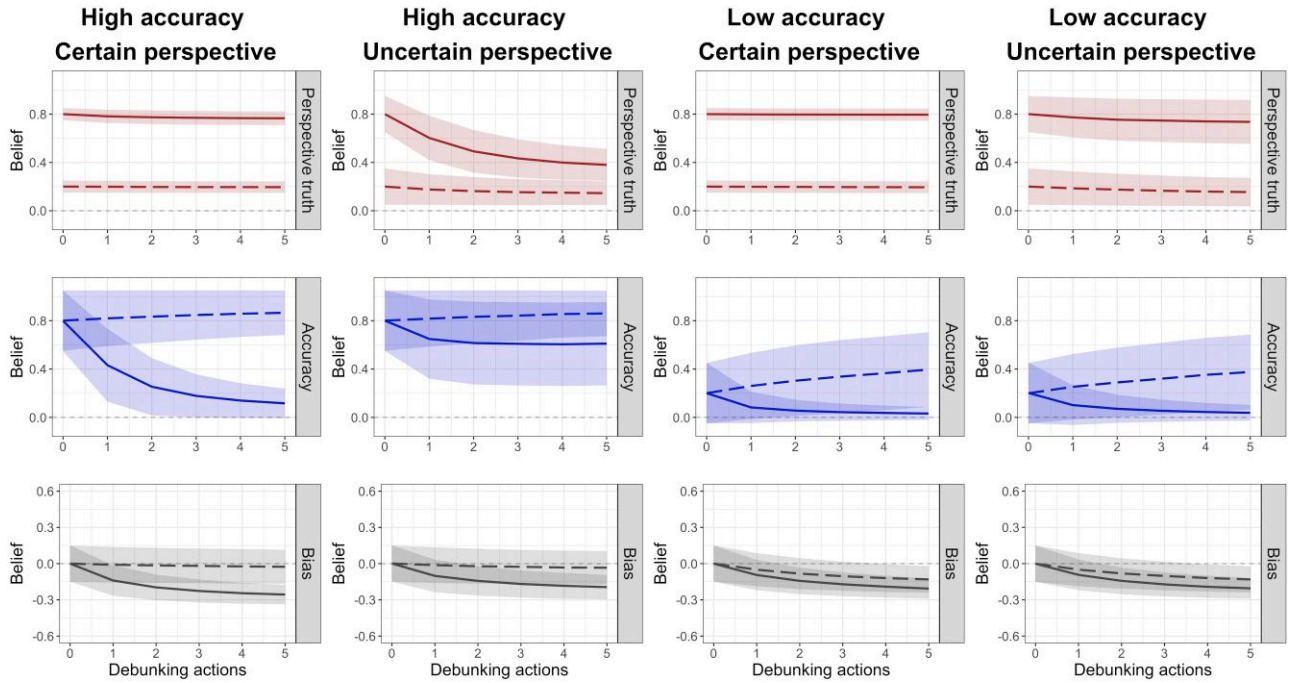

Fig. 4. Interaction between accuracy belief value and perspective belief uncertainty in determining the polarization of beliefs. In all four simulations, both subgroups are initially somewhat certain that the authority is impartial (mean = 0, std = 0.15), and both subgroups are quite uncertain about their accuracy beliefs as well. Low and high accuracy correspond to belief values (i.e. distribution mean) of 0.2 and 0.8, respectively. Certain and uncertain perspective beliefs correspond to belief distributions with standard deviation of 0.05 and 0.15, respectively.

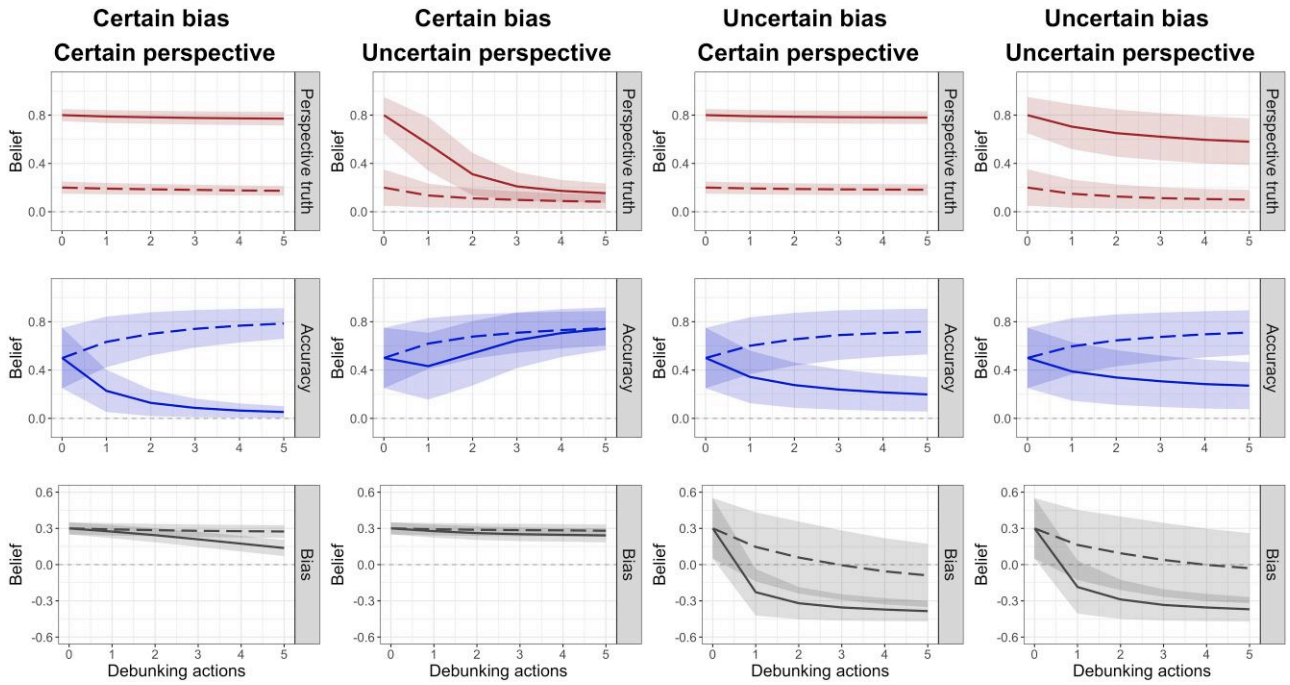

Fig. 5. Interaction between bias belief uncertainty and perspective belief uncertainty in determining the polarization of beliefs. In all four simulations, both subgroups are initially uncertain that the authority is somewhat accurate (mean = 0.5, std = 0.25), and both subgroups believe the authority is biased in favor of the target. Certain and uncertain perspective beliefs correspond to standard deviations of 0.05 and 0.15, respectively. Certain and uncertain bias beliefs correspond to standard deviations of 0.05 and 0.25, respectively.

80). However, if the proponent group's acquired beliefs about the authority's accuracy motive are slightly higher and more uncertain, especially when the authority is believed to be biased in favor of the perspective (e.g. simulations 52, 70, and 133), the authority could be successful in debunking in the new domain; and the authority may even be able to gain back the proponent subgroup's trust, through recovering their beliefs about authority's accuracy (e.g. simulation 70 and 133).

Discussion

Using an inverse planning approach to model observers' inferences from an authority's debunking, we showed how observers' perspectives and their beliefs about the authority interact to shape the observers' beliefs. The results of our simulations offer several insights. First, we show that it is possible for debunking to work as intended by shifting beliefs. The relative rarity of this outcome across our simulations reveals important constraints, however. The certainty of initial beliefs about the topic constrains belief change, even in the presence of an authority who is seen as truth-motivated and unbiased. When there is some initial uncertainty about the topic, several factors can independently influence whether two groups ultimately converge on the same beliefs. When beliefs do not converge, polarization can spread to beliefs about the authority's commitment to the truth and to perceptions of their biases. These acquired polarized perceptions can spread to new topics as well, coloring the effects of future debunking by that same authority.

Our simulations identified three configurations of initial beliefs that facilitated successful debunking. Initial belief polarization was reduced by debunking when: (i) initially differing beliefs about the topic were held with greater uncertainty; (ii) the authority was believed to be highly motivated by accuracy (especially with greater certainty); and (iii) the authority was initially believed to be biased in favor of the debunked claims or their source (especially with greater certainty). The current model thus contributes to the formal literature by synthesizing these findings into a single cognitive framework, and by emphasizing the importance of estimating the certainty of beliefs (both beliefs about the topic and beliefs about the authority; (35)), as well as distinguishing between the authority's perceived accuracy and bias (i.e. the dimensions of credibility), instead of measuring a single variable of trust or credibility.

These three configurations are consistent with the existing experimental and observational literature on belief change. First, uncertainty about a claim's truthfulness can be correlated with successful persuasion (35, 65). For example, in domains where scientific knowledge accumulates slowly and is contested (e.g. public health), debunking can have limited impact, e.g. (66, 67). In the case of US elections, uncertainty (rather than certainty) about fairness characterizes views about future election security (68), which suggests that our simulations capturing this uncertainty are relevant.

Second, perceptions of a commitment to accuracy can enhance the persuasiveness of a source (69, 70). In the case of elections outside the US context, some monitors commit significant resources to accurately characterizing election processes (27, 71), which suggests they are aware that perceived accuracy is distinct from perceived bias.

Third, our finding that a presumed bias in favor of debunked claims (or their source) heightens the persuasiveness of the authority's debunking actions also has precedent in the literature on political persuasion. Just as "only Nixon could go to China," so have studies found that debunking efforts and other persuasive claims carry more weight when the identity of the person making the correction or claim is a surprise (e.g. (16, 52, 55)), although this effect is not always found among the strongest partisans (e.g. (54)). In the context of elections, information provided by

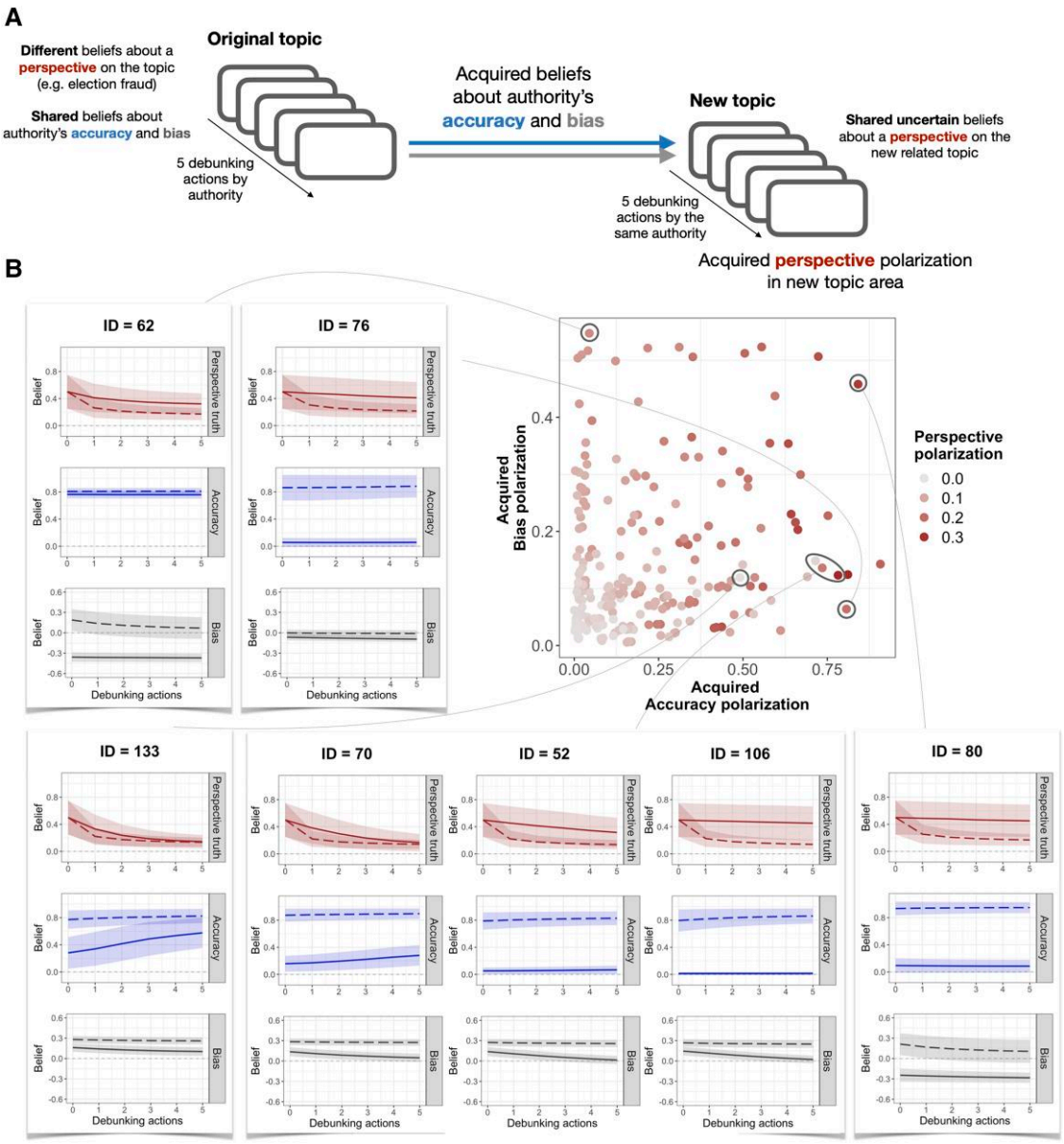

Fig. 6. Acquired polarized beliefs about the authority in one domain can propagate to and polarize initially shared beliefs in new domains. A) Simulation procedures in the original and new topic area for each pair of subgroups. In the original domain, two subgroups start with differing beliefs about a perspective on a topic area (i.e. proponent and opponent), but shared beliefs about the authority's accuracy motive and bias. After observing five debunking actions by an authority, the two subgroups update their perspective beliefs as well as their beliefs about the authority, referred to as acquired beliefs about the authority's accuracy and bias. The same authority (i.e. the same person or organization) then chooses to enter and debunk claims about a related perspective in a new topic area, therefore the acquired beliefs about the authority transfer to the new domain. We assume the two subgroups initially have shared beliefs about the new topic, but they are quite uncertain in their perspective beliefs (mean = 0.5, std = 0.25). We simulate the evolution of beliefs and quantify the potential acquired polarization in perspective beliefs in the new topic area, as a result of their acquired polarized beliefs about the authority from the original domain. B) The scatter plot shows the results of 243 pairs of simulations corresponding to the 243 pairs of simulations with various prior belief settings in the original domain. Each data point represents one pair of simulations; therefore, there are 243 points shown in the plot (see Fig. S6 for labels). Belief polarization is defined as the absolute value of the difference between the two subgroups' beliefs (i.e. mean of belief distributions). The x- and y-axis represent the acquired polarization in beliefs about authority's accuracy and bias, respectively, before observing the authority's decisions in the new topic domain; darker points indicate larger polarization in beliefs about the new topic after five debunking actions by the same authority. The side panels show the evolution of beliefs in seven example pairs of simulations.

authorities perceived as biased (e.g. Fox News calling the state of Arizona for Joe Biden in the 2020 election) (26) or debunking efforts by authorities perceived as biased (e.g. Arizona's Republican governor later denying voter fraud claims in that state) (72) also appear to be more persuasive (57).

Yet, in many settings, debunking did not reduce the initial disagreement, and instead induced polarization in beliefs about the authorities' motivations. At the start of our simulations, all observers had shared beliefs about the authority. This situation is not far-fetched and could arise for several reasons. First, knowledge about well-established authorities might simply be low, as it is across many aspects of politics (e.g. (73, 74)). Second, nonpartisan or newly formed authorities may not have established reputations (e.g. (75, 76)). Third, well-known authorities may enter

a new domain where priors about them are not yet established, such as when celebrities aim to translate their popularity to political influence (e.g. (77, 78)). Finally, authorities may have established reputations within the domain based on characteristics that do not correlate with the source of group differences in belief about the claim. For example, outside the US context, it is not uncommon for multiple organizations to be invited to monitor elections (10) and there is evidence that the monitor's sponsor (EU, United States, regional states) influences prior beliefs about the authority's capability and biases, rather than the observers' preferred election outcomes (e.g. Tunisia as discussed in Ref. (79)). Despite these initially shared beliefs, however, the final state of our simulations frequently included polarized beliefs about the authorities, following the same societal fault lines as the initial disagreement.

Given the frequent reputational damage to debunking authorities in our simulations, authorities could rationally anticipate these consequences, and thus be reluctant to engage in debunking (80). Yet our model reflects only the perspective of observers; the expected consequences for the authority's reputation *from the authority's own perspective* might be different. Observably, authorities often do engage in debunking, and do experience reputational damage as a consequence. These authorities may have underestimated the likelihood of reputational damage (e.g. because they overestimated observers' prior beliefs in their epistemic motives), or may have deliberately accepted this risk because of very strong epistemic motivations (e.g. the public health officials surveyed in Ref. (13), or the Republicans who denied voter fraud claims after the 2020 elections). To capture the authority's perspective, future work could embed the current model of observers, recursively, in a model of *authority* choice to better understand this dynamic and study how authorities value the trade-off between correcting beliefs and managing their reputation. This type of recursive model is standard in cognitive science for studying language and communication (81, 82), and has been recently used to model an authority's punitive decision-making (e.g. see Ref. (83)).

A primary contribution of this work is to illustrate how a set of minimal conditions can give rise to both belief convergence and divergence. The initial conditions in our model were minimal: two groups with differing beliefs in one domain. The source of this initial divergence could itself be either rational (e.g. exposure to different evidence) (60, 84) or irrational (e.g. motivated beliefs that preserve self-image) (85, 86). In either case, the process of belief updating using an intuitive causal model can entrench these initial differences into a robust structure of polarized supporting beliefs.

We see several additional extensions for our minimal conditions framework. In addition to modeling authorities' decision-making processes, this model could be extended to capture additional cognitive processes among observers. For example, our model does not predict one form of polarization often taken as the best evidence for irrationality, known as backfiring. In our model, debunking could reduce observers' belief in the debunked claims, or not reduce their belief, but debunking never backfired by directly increasing beliefs in the debunked claims (17, 38, 59). To predict backfiring, the current model could include observers who have a generative mental model that an authority may gain utility specifically by spreading false claims (40, 59). Another extension could incorporate alternative updating mechanisms through which belief polarization can occur, including differences in goals and utilities (e.g. motivated reasoning) and differences in belief updating mechanisms (e.g. ignoring/lapses in considering the evidence, cognitive ability).

So far, we have explored the effects of a fixed series of five actions, always debunking claims from the same perspective (e.g. the election was stolen). Five debunked claims provides enough information in our model to generate convergence or polarization using a wide variety of priors. However, our simulations hint that a major effect of beliefs about the authority's accuracy is on the speed of convergence between the two groups, and the uncertainty of their final beliefs. Considering these other features of belief distributions in defining and measuring polarization could prove useful in certain contexts, such as when the authority has limited opportunities for debunking, so that the speed of convergence would be more important.

Formal simulations and empirical studies could also investigate the effect of observing a mixture of decisions to debunk and decisions to say nothing about the truth of the claim, i.e. abstain. The same model developed here can be used to simulate the evolution of beliefs in these mixed situations. For example, Fig. S7 illustrates the effect of a single debunking (in round 3) and four abstentions using the same initial priors as Fig. 2. When observers have high confidence in either their views about the authority (left) or about the perspective's truth (right), choices to abstain from debunking do not affect these certain beliefs. However, when observers have higher levels of uncertainty, the confusion introduced by the mixed debunking causes beliefs to evolve non-monotonically. Indeed, in the case where convergence occurs, the combination of four abstentions and one debunking yields convergence towards the debunked perspective, contrary to the pattern observed in Fig. 2. While fully evaluating the complexity introduced by mixed cases is outside the scope of this manuscript, the extreme case of one debunking action and four abstentions illustrates that the impact of failing to debunk all claims from a particular perspective can both undermine the opportunity to converge on true beliefs, and influence the impressions of the authority's epistemic motives and biases.

The most significant limitation of the current work is the lack of validation by direct comparison to human inferences. The intuitive causal model of authority's decision-making process used here is adapted from a generic framework for modeling human action understanding, known as inverse planning, that has been empirically validated in many contexts (46–49). In particular, the current model is derived directly from a model of human observers' inferences from an authority's punitive decisions (50). Similar experimental approaches could be used to calibrate the instantiation of our modeling framework, to test alternative specifications of the authorities' utilities, and to test the predictions of this model in hypothetical or actual scenarios, for political as well as non-political beliefs.

Belief polarization arose in our simulations through entirely rational belief updating. Applied to the case of election-related debunking in a polarized political environment, our results provide useful if sobering insights. First, while independent observers have persuasively debunked claims of election fraud in non-US contexts (9, 11, 75), the chance that such efforts will succeed among those who have extreme confidence in their beliefs is small. However, there is hope for the type of nonpartisan election monitoring that has been trialed on a small scale in the United States. Our model shows that organizations able to establish a clear reputation across partisan groups for commitment to the truth and unbiasedness can engage in successful debunking. This debunking can achieve its purpose: enhancing confidence in election outcomes. Nevertheless, as our simulations show, the act of debunking itself will affect the reputation of those organizations. When organizations with reputations derived from

other domains move into election monitoring and debunking (e.g. former Democratic president Jimmy Carter's foundation the Carter Center) their reputations come with them, and the perceptions of bias in particular can reduce the impact of their debunking efforts in the elections domain.

Materials and methods

Belief distributions

We modeled observers' beliefs about the topic area and the authority's motives as random variables with beta distributions. The beta distribution is a family of continuous probability distributions defined on the interval [0, 1]. We modeled observers as assigning a truth likelihood (i.e. probability) to a set of claims drawn from a perspective on a topic (continuous value between 0 and 1). Choosing to model beliefs about claims as the truth likelihood of the claim (rather than a binary belief about whether the claim is true or false) is inspired by empirical studies which find people's beliefs in a claims' truth vary continuously (e.g. between 0 and 100, (59)). We then modeled their beliefs about the topic as a distribution of truth likelihood values for various claims drawn from one perspective on that topic. For modeling beliefs that follow other distributions (e.g. binary beliefs) one could simply change the class of belief distribution, without changing any other computational machinery in the model. Although we primarily conceptualized the belief distribution as implementing the hierarchical relationship between claims and the topic (i.e. claims are sampled from one perspective on the topic, and the belief distribution is the distribution of claim truth likelihoods from that perspective), the topic belief distribution could alternatively be conceptualized as the reliability of beliefs about the perspective, especially in topics where the perspective could be represented as a continuous variable (e.g. estimates of economic growth).

We modeled beliefs about the authority's motives as beta distributions, over [0, 1] for accuracy and [−0.5, 0.5] for bias (by first sampling from a beta distribution and then shifting the samples by −0.5). We primarily conceptualized the distribution over authority's motives as the certainty of beliefs about authority. However, the belief distribution could also be conceptualized as implementing a hierarchical relationship between a single authority (observed in each action) drawn from a set of authorities, for example if the debunking is done by a different individual journalist from the same newspaper, scientist in the same discipline, or observer working with the same NGO.

We model beliefs using the beta distribution because it is flexible and can model a wide variety of distribution shapes over a bounded interval. Moreover, the shape of the distribution can be controlled and uniquely specified (i.e. "a" and "b" shape parameters) by the mean (to represent belief value) and standard deviation (to represent belief uncertainty) of the distribution.

Utilities

We modeled the authority as considering a binary decision between doing nothing in response to a claim, or debunking it. The utility of each response, associated with the consequences of the response for the target U_{target}, is modeled as a deterministic value: U_{target} is set to −1 for debunking and zero for doing nothing. By setting U_{target} of debunking to −1, we assumed that debunking is a proportional response to a claim that is surely false (truth likelihood = 0; see Eq. 2); that is, debunking a surely false claim would maximize $U_{accuracy}$. Similarly, by setting U_{target} of doing nothing to 0, we assumed that doing nothing is a proportional response to a claim that is surely true (truth likelihood = 1); that is, doing nothing in response a surely true claim would maximize $U_{accuracy}$. For a claim with truth likelihood of 0.5, the authority would be agnostic between doing nothing and debunking, in terms of the epistemic value of their response. In that case, the authority's bias towards the target would be the main driver of their response.

Implementation

The probabilistic programming language `WebPPL` (the local installation from https://webppl.org) was used to code the inverse planning model. `Python` (3.7.6) and `R` (4.1.1) were used for processing the simulation results and generating the plots.

We used Markov chain Monte Carlo (MCMC) sampling, as implemented within `WebPPL`, to estimate the observer's posterior belief distributions after observing each response by the authority. The program is structured such that the observer has a generative model of the authority, who decides whether to do nothing or debunk the claim, given their accuracy motive, bias, claim truth likelihood and U_{target} of each possible response. The observer has prior belief distributions over authority's accuracy, bias, and the claims' truth likelihood. In each iteration of the MCMC algorithm, the observer samples accuracy motive, bias, and claim truth from their belief distribution and then simulates how this specific hypothetical authority (with the sampled accuracy motive and bias) would behave in response to this specific claim (with the sampled truth likelihood). The observer's posterior belief distribution is estimated by conditioning the program on the observed authority's response (debunking in our simulations). We used 400,000 MCMC iterations for estimating observers' posterior beliefs.

The estimated posterior belief distributions are saved after each time the authority responds to a claim by debunking in the form of samples from posterior distributions and their corresponding probabilities. Then, three beta distributions are fit to the posterior data (accuracy, bias, and perspective truth), and are input to the same `WebPPL` program to serve as the prior belief distributions for the next authority's decision.

Simulations

We ran two series of simulations, with each series containing 243 pairs of simulations. Each simulation pair consists of two simulations with IDs "i" and "i"+243 (in the code base); the simulation pair is referenced by ID = "i" in the manuscript.

In series 1 (Within-Topic), each pair of simulations featured two subgroups who initially had different belief values about one perspective on the topic (mean of belief distribution about the perspective in the opponent subgroup = 0.2, proponent subgroup = 0.8), but every other aspect of their beliefs was shared (uncertainty of belief distribution, mean, and uncertainty of beliefs about authority). The 243 simulations systematically sampled different settings of the shared beliefs (five dimensions, each taking three possible values, hence all the combinations would amount to $3^5 = 243$ pairs). Note the value and uncertainty of beliefs can be continuously varied, however for the purpose of illustration and tractability we used three values to denote low, medium, and high levels for each variable. The standard deviation (i.e. uncertainty) of topic beliefs varied in [0.05, 0.1, 0.15]. The value and uncertainty of accuracy motive prior beliefs varied in [0.2, 0.5, 0.8] and [0.05, 0.15, 0.25], respectively; similarly, the value and uncertainty of bias prior beliefs varied in [−0.3, 0, 0.3] and [0.05, 0.15, 0.25], respectively. Figure S1 shows the belief distributions with these parameters. These initial belief distributions served as the

prior beliefs for the first debunking action (in the figures, they are depicted as beliefs when debunking actions = 0). Each pair of simulations consist of five iterated epochs, that is observations of five debunking actions by the authority. The posterior beliefs from each epoch serve as the prior beliefs for the next. There is no independent effect of trial within the model, so all trials are computationally identical.

In series 2 (Cross-Topic), each pair of simulations featured two subgroups who initially had shared beliefs about a new topic area. Both subgroups believed that the claims from the new topic area are somewhat likely to be true (mean of perspective belief distribution in both subgroups = 0.5), but they were quite uncertain (standard deviation of perspective belief distribution in both subgroups = 0.25). We assumed the same authority is debunking claims in the new topic area, and the new topic is related to the original topic enough that the subgroups' acquired beliefs about authority's accuracy and bias would generalize to this new domain. So, for each pair of simulations, the two subgroups' beliefs about authority were set to be the posterior beliefs after epoch 5 of the corresponding pair from the Within-Topic simulation.

Across all simulations, the rationality parameter (beta) was set to 10.

Acknowledgments

We thank David Rand and Emily Falk for helpful suggestions in the development of the project, and Andrew Little for feedback on the earlier versions of the manuscript.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

This work is supported in part by the Patrick J. McGovern Foundation and by the Guggenheim Foundation.

Author Contributions

S.R. (Conceptualization, Data curation, Methodology, Formal analysis, Visualization, Writing—original draft, Writing—review & editing), M.L.W. (Conceptualization, Writing—original draft, Writing—review & editing), R.S. (Conceptualization, Methodology, Project administration, Supervision, Writing—original draft, Writing—review & editing).

Preprints

A preprint of this article is published at https://doi.org/10.31234/osf.io/8dq9x.

Data Availability

The code to simulate the model and generate the figures, as well as the simulation data are available on https://github.com/sradkani/inverse-planning-polarization and https://osf.io/jg87r/.

References

1 Fearon JD. 2011. Self-enforcing democracy. *Q J Econ*. 126(4):1661–1708.
2 Przeworski A. 2005. Democracy as an equilibrium. *Public Choice*. 123(3):253–273.
3 Weingast BR. 1997. The political foundations of democracy and the rule of the law. *Am Polit Sci Rev*. 91(2):245–263.
4 Holliday D, Grimmer J, Lelkes Y, Westwood S. 2023. Who are the election skeptics? Evidence from the 2022 midterm elections. *Election Law J*. Forthcoming.
5 Fahey JJ. 2023. The big lie: expressive responding and misperceptions in the united states. *J Exp Polit Sci*. 10(2):267–278.
6 McCarthy J. 2022. Confidence in election integrity hides deep partisan divide. *Gallup News*. https://news.gallup.com/poll/404675/confidence-election-integrity-hides-deep-partisan-divide.aspx
7 Arceneaux K, Truex R. 2023. Donald trump and the lie. *Perspect Polit*. 21(3):863–879.
8 Canon DT, Sherman O. 2021. Debunking the "big lie": election administration in the 2020 presidential election. *Pres Stud Q*. 51(3):546–581.
9 Bush SS, Prather L. 2017. The promise and limits of election observers in building election credibility. *J Polit*. 79(3):921–935.
10 Hyde SD, Marinov N. 2014. Information and self-enforcing democracy: the role of international election observation. *Int Organ*. 68(2):329–359.
11 Nevitte N, Canton SA. 1997. The rise of election monitoring: the role of domestic observers. *J Democracy*. 8(3):47–61.
12 Organization for Security and Cooperation in Europe. 2020. International Election Observation Mission, United States of America General Elections, 3 November 2020: Statement of Preliminary Findings and Conclusions. Technical Report.
13 Ittefaq M. 2023. "It frustrates me beyond words that I can't fix that": health misinformation correction on Facebook during COVID-19. *Health Commun*. https://doi.org/10.1080/10410236.2023.2282279
14 Turney C, van der Linden S. 2024. Can we be inoculated against climate misinformation? Yes – if we prebunk rather than debunk. The Conversation.
15 World Health Organization. 2022. COVID-19 Mythbusters – World Health Organization.
16 Berinsky AJ. 2017. Rumors and health care reform: experiments in political misinformation. *Br J Polit Sci*. 47(2):241–262.
17 Nyhan B, Reifler J. 2010. When corrections fail: the persistence of political misperceptions. *Polit Behav*. 32(2):303–330.
18 Swire-Thompson B, Miklaucic N, Wihbey JP, Lazer D, DeGutis J. 2022. The backfire effect after correcting misinformation is strongly associated with reliability. *J Exp Psychol: Gen*. 151(7):1655–1665.
19 Gelfand M, *et al.* 2022. Persuading republicans and democrats to comply with mask wearing: an intervention tournament. *J Exp Soc Psychol*. 101:104299.
20 Kozyreva A, *et al.* 2024. Toolbox of individual-level interventions against online misinformation. *Nat Hum Behav*. 8(6):1044–1052.
21 Zhang FJ. 2023. Political endorsement by nature and trust in scientific expertise during covid-19. *Nat Hum Behav*. 7(5):696–706.
22 Cohen MJ, Sheagley G. 2021. Partisan poll watchers and Americans' perceptions of electoral fairness. *Public Opin Q*. 88(SI):536–560.
23 Swire B, Berinsky AJ, Lewandowsky S, Ecker UKH. 2017. Processing political misinformation: comprehending the trump phenomenon. *R Soc Open Sci*. 4(3):160802.
24 Painter DL, Fernandes J. 2024. "The big lie:" how fact checking influences support for insurrection. *Am Behav Sci*. 68(7):892–912.
25 Lewandowsky S, Oberauer K, Gignac GE. 2013. NASA faked the moon landing—therefore, (climate) science is a hoax: an anatomy of the motivated rejection of science. *Psychol Sci*. 24(5):622–633.
26 Baker P. 2023 Mar. Inside the panic at Fox News after the 2020 election. The New York Times.

27 Bush SS, Cottiero C, Prather L. 2024. Zombies ahead: explaining the rise of low-quality election monitoring. *Rev Int Organ*. https://doi.org/10.1007/s11558-024-09554-3
28 López A. 2023. Gaslighting: fake climate news and big Carbon's network of Denial. In: Fowler-Watt K, McDougall J, editors. The Palgrave handbook of media misinformation. Cham: Springer International Publishing. p. 159–177.
29 Bullock JG. 2011. Elite influence on public opinion in an informed electorate. *Am Polit Sci Rev*. 105(3):496–515.
30 Freeder S, Lenz GS, Turney S. 2019. The importance of knowing "what goes with what": reinterpreting the evidence on policy attitude stability. *J Polit*. 81(1):274–290.
31 Goldstein DAN, Wiedemann J. 2022. Who do you trust? The consequences of partisanship and trust for public responsiveness to COVID-19 orders. *Perspect Polit*. 20(2):412–438.
32 Zaller JR. 1992. *The nature and origins of mass opinion*. Cambridge (UK): Cambridge University Press.
33 Gunson P. 2024. *Venezuela: what next after its election uproar?* International Crisis Group Q&A.
34 Graham MH. 2023. Measuring misperceptions? *Am Polit Sci Rev*. 117(1):80–102.
35 Li J, Wagner MW. 2020. The value of not knowing: partisan cue-taking and belief updating of the uninformed, the ambiguous, and the misinformed. *J Commun*. 70(5):646–669.
36 Tormala ZL. 2016. The role of certainty (and uncertainty) in attitudes and persuasion. *Curr Opin Psychol*. 10:6–11.
37 Lord CG, Ross L, Lepper MR. 1979. Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. *J Pers Soc Psychol*. 37(11):2098.
38 Jern A, Chang K-MK, Kemp C. 2014. Belief polarization is not always irrational. *Psychol Rev*. 121(2):206.
39 Bhui R, Gershman SJ. 2020. Paradoxical effects of persuasive messages. *Decision*. 7(4):239.
40 Cook J, Lewandowsky S. 2016. Rational irrationality: modeling climate change belief polarization using Bayesian networks. *Top Cogn Sci*. 8(1):160–179.
41 Botvinik-Nezer R, Jones M, Wager TD. 2023. A belief systems analysis of fraud beliefs following the 2020 US election. *Nat Hum Behav*. 7(7):1106–1119.
42 Powell D, Weisman K, Markman EM. 2023. Modeling and leveraging intuitive theories to improve vaccine attitudes. *J Exp Psychol: Gen*. 152(5):1379.
43 Griffiths TL, Tenenbaum JB. 2006. Optimal predictions in everyday cognition. *Psychol Sci*. 17(9):767–773.
44 Griffiths TL, Chater N, Kemp C, Perfors A, Tenenbaum JB. 2010. Probabilistic models of cognition: exploring representations and inductive biases. *Trends Cogn Sci (Regul Ed)*. 14(8):357–364.
45 Bullock JG. 2009. Partisan bias and the Bayesian ideal in the study of public opinion. *J Polit*. 71(3):1109–1124.
46 Baker CL, Saxe R, Tenenbaum JB. 2009. Action understanding as inverse planning. *Cognition*. 113(3):329–349.
47 Jara-Ettinger J, Gweon H, Schulz LE, Tenenbaum JB. 2016. The naïve utility calculus: computational principles underlying commonsense psychology. *Trends Cogn Sci (Regul Ed)*. 20(8):589–604.
48 Houlihan SD, Kleiman-Weiner M, Hewitt LB, Tenenbaum JB, Saxe R. 2023. Emotion prediction as computation over a generative theory of mind. *Philos Trans R Soc A*. 381(2251):20220047.
49 Strouse DJ, McKee K, Botvinick M, Hughes E, Everett R. 2021. Collaborating with humans without human data. *Adv Neural Inf Process Syst*. 34:14502–14515.
50 Radkani S, Saxe R. 2023. What people learn from punishment: joint inference of wrongness and punisher's motivations from observation of punitive choices. In: Goldwater M, Anggoro FK, Hayes BK, Ong DC, editors. *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Vol. 45. p. 1027–1034.
51 Jost JT, Baldassarri DS, Druckman JN. 2022. Cognitive–motivational mechanisms of political polarization in social-communicative contexts. *Nat Rev Psychol*. 1(10):560–576.
52 Benegal SD, Scruggs LA. 2018. Correcting misinformation about climate change: the impact of partisanship in an experimental setting. *Clim Change*. 148(1):61–80.
53 Flanagin A, Metzger MJ. 2017. Digital media and perceptions of source credibility in political communication. In: Kenski K, Jamieson KH, editors. *The Oxford handbook of political communication*. Oxford: Oxford University Press. p. 417–436.
54 Grossman G, Kim S, Rexer JM, Thirumurthy H. 2020. Political partisanship influences behavioral responses to governors' recommendations for COVID-19 prevention in the united states. *Proc Natl Acad Sci*. 117(39):24144–24153.
55 Mattes M, Weeks JLP. 2019. Hawks, doves, and peace: an experimental approach. *Am J Pol Sci*. 63(1):53–66.
56 Berinsky AJ. 2017. Rumors and health care reform: experiments in political misinformation. *Br J Polit Sci*. 47(2):241–262.
57 Carey J, Fogergty B, Gehrke M, Nyhan B, Reifler J. 2024. Prebunking and credible source corrections increase election credibility: evidence from the U.S. and Brazil. Working Paper.
58 Martel C, Rand DG. 2024. Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nat Hum Behav*. https://doi.org/10.1038/s41562-024-01973-x
59 Reinero DA, Harris EA, Rathje S, Duke A, Van Bavel JJ. 2023. Partisans are more likely to entrench their beliefs in misinformation when political outgroup members fact-check claims. Working Paper.
60 Gerber A, Green D. 1999. Misperceptions about perceptual biases. *Ann Rev Polit Sci*. 2:189–210.
61 Little AT. 2023. Bayesian explanations for persuasion. *J Theor Polit*. 35(3):147–181.
62 Bendor J. 2020. Bounded rationality in political science and politics. In: Mintz A, Terris L, editors. *Oxford handbook of behavioral political science*. Oxford: Oxford University Press. p. 37–68.
63 Landau-Wells M, Saxe R. 2020. Political preferences and threat perception: opportunities for neuroimaging and developmental research. *Curr Opin Behav Sci*. 34:58–63.
64 McGraw KM. 2000. Contributions of the cognitive approach to political psychology. *Polit Psychol*. 21(4):805–832.
65 Druckman JN. 2022. A framework for the study of persuasion. *Ann Rev Polit Sci*. 25: 65–88.
66 Cappella JN, Maloney E, Ophir Y, Brennan E. 2015. Interventions to correct misinformation about tobacco products. *Tob Regul Sci*. 1(2):186.
67 Smith P, *et al*. 2011. Correcting over 50 years of tobacco industry misinformation. *Am J Prev Med*. 40(6):690–698.
68 Allen J, Orey R, Sanchez T. 2024 Feb. Who voters trust for election information in 2024. Technical Report, Bipartisan Policy Center, Washington, D.C.
69 Buczel KA, Szyszka PD, Siwiak A, Szpitalak M, Polczyk R. 2022. Vaccination against misinformation: the inoculation technique reduces the continued influence effect. *PLoS One*. 17(4):e0267463.
70 Walter N, Brooks JJ, Saucier CJ, Suresh S. 2021. Evaluating the impact of attempts to correct health misinformation on social media: a meta-analysis. *Health Commun*. 36(13):1776–1784.
71 Hyde SD. 2012. Why believe international election monitors? In: Lake DA, Stein JG, Gourevitch PA, editors. The credibility of

transnational NGOs: when virtue is not enough. Cambridge: Cambridge University Press. p. 37–61.
72 Caldwell LA, Dawsey J, Sanchez YW. 2023 Jul. Trump pressured Arizona Gov. Doug Ducey to overturn 2020 election. Washington Post.
73 Angelucci C, Prat A. 2024. Is journalistic truth dead? Measuring how informed voters are about political news. *Am Econ Rev*. 114(4):887–925.
74 Borelli G, Gracia S. 2023. *What Americans know about their government*. Pew Research Center.
75 Barker J, Samet O, Hyde SD. 2024. Citizen election observation and public confidence in U.S. elections. Working Paper.
76 Darnall N, Ji H, Vázquez-Brust DA. 2018. Third-party certification, sponsorship, and consumers' ecolabel use. *J Bus Ethics*. 150(4):953–969.
77 Grynbaum MM. 2024 Jan. Fox News to Taylor swift: 'don't get involved in politics!'. The New York Times.
78 Towler CC, Crawford NN, Bennett RA. 2020. Shut up and play: black athletes, protest politics, and black political action. *Perspect Polit*. 18(1):111–127.
79 Bush SS, Prather L. 2018. Who's there? Election observer identity and the local credibility of elections. *Int Organ*. 72(3):659–692.
80 Choshen-Hillel S, Shaw A, Caruso EM. 2020. Lying to appear honest. *J Exp Psychol: Gen*. 149(9):1719.
81 Frank MC, Goodman ND. 2012. Predicting pragmatic reasoning in language games. *Science*. 336(6084):998–998.
82 Goodman ND, Frank MC. 2016. Pragmatic language interpretation as probabilistic inference. *Trends Cogn Sci (Regul Ed)*. 20(11):818–829.
83 Radkani S, Tenenbaum J, Saxe R. 2022. Modeling punishment as a rational communicative social action. In: Culbertson J, Perfors A, Rabagliati H, Ramenzoni V, editors. *Proceedings of the 44th Annual Conference of the Cognitive Science Society*. Vol. 44. p. 1040–1047.
84 Achen CH. 2002. Parental socialization and rational party identification. *Polit Behav*. 24(2):151–170.
85 Kim S-y, Taber CS, Lodge M. 2010. A computational model of the citizen as motivated reasoner: modeling the dynamics of the 2000 presidential election. *Polit Behav*. 32(1):1–28.
86 Little AT. 2019. The distortion of related beliefs. *Am J Pol Sci*. 63(3):675–689.