

A belief systems analysis of fraud beliefs following the 2020 US election

Received: 19 August 2022

Accepted: 24 February 2023

Published online: 10 April 2023



Rotem Botvinik-Nezer¹✉, Matt Jones^{2,3}✉ & Tor D. Wager¹✉

Beliefs that the US 2020 Presidential election was fraudulent are prevalent despite substantial contradictory evidence. Why are such beliefs often resistant to counter-evidence? Is this resistance rational, and thus subject to evidence-based arguments, or fundamentally irrational? Here we surveyed 1,642 Americans during the 2020 vote count, testing fraud belief updates given hypothetical election outcomes. Participants' fraud beliefs increased when their preferred candidate lost and decreased when he won, and both effects scaled with partisan preferences, demonstrating partisan asymmetry (desirability effects). A Bayesian model of rational updating of a system of beliefs—beliefs in the true vote winner, fraud prevalence and beneficiary of fraud—accurately accounted for this partisan asymmetry, outperforming alternative models of irrational, motivated updating and models lacking the full belief system. Partisan asymmetries may not reflect motivated reasoning, but rather rational attributions over multiple potential causes of evidence. Changing such beliefs may require targeting multiple key beliefs simultaneously rather than direct debunking attempts.

On 6 January 2021, the US Congress assembled in the Capitol for the electoral vote count that would formalize Joe Biden's election as the new president. Outside the Capitol building, thousands of Americans participated in a riot motivated by claims that the election was fraudulent. Many of them broke into the Capitol building, endangering members of Congress and forcing an evacuation. These events were preceded, and followed, by widespread attempts to debunk beliefs in election fraud with substantial contradictory evidence. These attempts largely failed.

How do people update beliefs in light of new information, and why do false beliefs often persist despite countervailing evidence? An epistemically rational agent makes evidence-based inferences guided by the pursuit of accuracy¹. However, belief updating can exhibit desirability effects, such that, given evidence, beliefs are more often updated towards a desired state. For example, people update their beliefs about an expected election winner less following polls that show a projected loss (versus win) for their preferred candidate, discounting undesired outcomes². In politics, desirability effects can take the form of partisan asymmetries, with members of opposing

groups interpreting evidence in a way that favours their party. Such effects can lead to belief polarization, with partisan groups increasingly diverging in their beliefs^{3–5}, even if they are exposed to similar evidence^{6–8}. Beyond politics, this dynamic extends to many of the most important issues of our time, from vaccine uptake to climate change^{9–13}.

According to current theories, these desirability effects are the result of irrational, biased reasoning ('directional motivated reasoning'^{14,15}); beliefs are selectively updated to support desired conclusions. Often the conclusions are those that help maintain positive emotions, self-esteem and social identities^{4,16}. Put simply, people 'believe what they want to believe'. If this is the case, the role of evidence in combatting false beliefs is unclear.

When studying beliefs in isolation, as in most previous work on belief updating, observations of updating in opposing directions given the same evidence¹⁷ or preferential updating of desired beliefs¹⁵ seem irrational. We consider an alternative possibility based on a systems perspective, in which multiple beliefs combine or compete to explain observed evidence^{18–20}. Desirability effects might emerge

¹Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA. ²Department of Psychology and Neuroscience and Institute of Cognitive Science, University of Colorado Boulder, Boulder, CO, USA. ³Research performed in part while at Google Research, Brain Team, Mountain View, CA, USA. ✉ e-mail: rotemb9@gmail.com; mcj@colorado.edu; Tor.D.Wager@dartmouth.edu

from the dynamics of rational updating over this system. Belief systems are typically modelled in a Bayesian framework (for example, ‘belief nets’²¹). As Bayesian belief updating strictly follows normative rules of probabilistic inference, Bayesian models describe the behaviour of rational agents free of bias towards a preferred belief (that is, unbiased belief updating).

Surprisingly, under some conditions, Bayesian belief updating can produce desirability effects as an emergent property. Consider a simple belief system in which an official election outcome is attributed to two possible causes: (1) winning the true (fraudless) vote or (2) winning by fraud. Upon observing the outcome, a rational Bayesian agent will update both causal beliefs in proportion to how diagnostic the observed outcome is about each causal variable (for mathematical details, see Methods). This process entails two predictions that are perhaps surprising when considering beliefs in isolation: (1) beliefs (for example, in fraud) can be updated in the absence of direct evidence for or against them and (2) different people can update a belief in opposing directions given the same evidence, because updates depend not only on the prior belief and outcome, but also on other beliefs²⁰. For example, the more strongly a losing candidate was expected to win beforehand, the more likely an alternative explanation such as fraud becomes. Therefore, if prior beliefs are polarized in a partisan fashion (for example, both sides expect their candidate to win the true vote), desirability effects in fraud beliefs could emerge as a result of rational, Bayesian updating.

Distinguishing between these two possible mechanisms—biased updating arising from directional motivated reasoning and rational updating of polarized prior beliefs—has important implications for real-world policies and campaigns for combatting false beliefs across domains. In particular, they point to different strategies, both of which go beyond straightforward attempts to debunk false beliefs with counter-evidence. If people are biased to believe what they want to believe, the value of new evidence is limited, and interventions might best focus on incentives or on inducing rational thinking. Alternatively, if updating is rational but depends on a system of interrelated beliefs, then one needs to identify the key beliefs and target them simultaneously to elicit change.

In this Article, we examine the dynamics of belief polarization in a high-stakes, real-world setting: fraud beliefs in the context of the 2020 US presidential election. We address three key questions. First, are fraud beliefs affected simply by presentation of electoral outcomes, in the absence of any evidence for or against fraud per se? Second, are beliefs about fraud in the 2020 election subject to desirability effects—that is, partisan asymmetry in how the election outcome is interpreted? Third, if these two effects are observed, are they better explained by directional motivated reasoning (believing in fraud to protect a desired belief about the true winner) or by rational updating of polarized priors (believing in fraud to explain an otherwise unexpected result)?

To address these questions, we surveyed a large online sample of 1,642 Americans from Amazon’s Mechanical Turk (MTurk) via Cloud Research²² (‘Use of MTurk for data collection’ in Supplementary Information) during the 2020 US presidential election on 4–5 November, while the winner was not yet determined and votes were being counted in key states. The sample included participants from each state, whose distribution strongly correlated with the population distribution (on the basis of the 2020 US census: Pearson’s $r = 0.938$, 95% confidence interval (CI) 0.893 to 0.964, $P < 0.001$) and age ranging 18–84 years (mean 41.6 (s.d. 13.1) years; Extended Data Fig. 1). Participants reported their preference for president (Biden or Trump) and strength of preference, probability of a win by their preferred candidate (prior win belief), and probability that election fraud would play a significant role in the outcome (prior fraud belief; Supplementary Table 1). We then showed each participant a hypothetical map showing the winner in each state, randomized to indicate either a Republican (Trump) or Democratic (Biden) victory (Fig. 1a,b). Participants reported the probability

that fraud played a significant role in the outcome (posterior fraud belief). Comparing posterior and prior fraud beliefs allowed us to test for desirability effects on belief updating. Approximately 11 weeks later, a subsample of the same participants ($N = 828$) completed a follow-up survey reporting their beliefs about the true vote winner and fraud beneficiary (Supplementary Tables 2 and 3). This allowed us to formulate a Bayesian model under which participants postulate two possible causes of the election outcome (a true win or fraud) and test whether it accurately predicted individual differences in fraud belief updates.

We found an asymmetric updating of fraud beliefs (Fig. 1) depending on the partisan group (direction desirability effect), which was stronger in those with stronger preferences (magnitude desirability effect). The Bayesian model accurately predicted these patterns (Fig. 2) and outperformed models that either lacked the full system of beliefs or included an extra Bayesian motivated reasoning mechanism (Fig. 3). Together, these results suggest that strong partisan asymmetries in how the election outcome is interpreted are consistent with a rational updating process operating on politically polarized prior beliefs.

Results

Prior fraud and win beliefs

Participants’ preferences for one candidate were strong across the sample, with 52.7% indicating the strongest preference and only 4.8% in the lower half of the scale. Preferences were stronger for Biden’s compared with Trump’s supporters (henceforth labelled Democrats and Republicans, respectively; Democrats: mean 89.6 (s.d. 18.4) on a 100-point scale; Republicans: mean 85.8 (s.d. 19.6); two-sample t -test: 95% CI 1.915 to 5.689, $t_{(1640)} = 3.95$, $P < 0.001$; Extended Data Fig. 2a). The prior win beliefs indicated that most participants believed their preferred candidate would win the election, with stronger beliefs for Biden than Trump supporters, in line with pre-election polls as well as the partial results from some states available at the time of the survey (estimated probability of preferred candidate’s win: Democrats, mean 77.5% (s.d. 15.5%); Republicans, mean 50.3% (s.d. 26.4%); Welch two-sample t -test, 95% CI 24.956 to 29.551, $t_{(861.49)} = 23.28$, $P < 0.001$; Extended Data Fig. 2b). Before the outcome presentation, fraud beliefs were higher for Republicans than for Democrats, as was expected on the basis of allegations made by Republican politicians before election day (estimated probability of fraud playing a significant role in the election outcome: Democrats, mean 19.9% (s.d. 24.4%); Republicans, mean 30.9% (s.d. 30.7%); Welch two-sample t -test 95% CI 41.068 to 46.794, $t_{(1,061.6)} = -30.11$, $P < 0.001$; Extended Data Fig. 2c).

Desirability effects on fraud belief updates

The hypotheses and analysis plan were pre-registered after data collection but before inspection of the belief data (<https://osf.io/kucsw>; for more details, see Methods). As expected, we observed strong partisan asymmetry in fraud belief updating. Participants’ fraud beliefs significantly increased after a preferred candidate loss (one-sample t -test: mean 12.97 (s.d. 31.11), 95% CI 10.844 to 15.098, $t_{(823)} = 11.97$, $P < 0.001$) and significantly decreased after a preferred candidate win (one-sample t -test: mean -17.35 (s.d. 29.41), 95% CI -19.365 to -15.327, $t_{(817)} = -16.87$, $P < 0.001$; Fig. 1c). That is, for either hypothetical outcome (Biden wins or Trump wins), Democrats and Republicans updated their fraud beliefs in opposite directions. The directional update was weaker in the ‘Biden wins’ scenario (Biden wins, mean 11.37 (s.d. 16.91); Trump wins, mean 30.22 (s.d. 31.28); Welch two-sample t -test, 95% CI -21.282 to -16.408, $t_{(1255)} = -15.17$, $P < 0.001$), potentially because many Republicans had maximal prior fraud beliefs that could not increase given a Republican loss, creating a ceiling effect, and many Democrats had minimal prior fraud beliefs that could not decrease given a Democratic win, creating a floor effect.

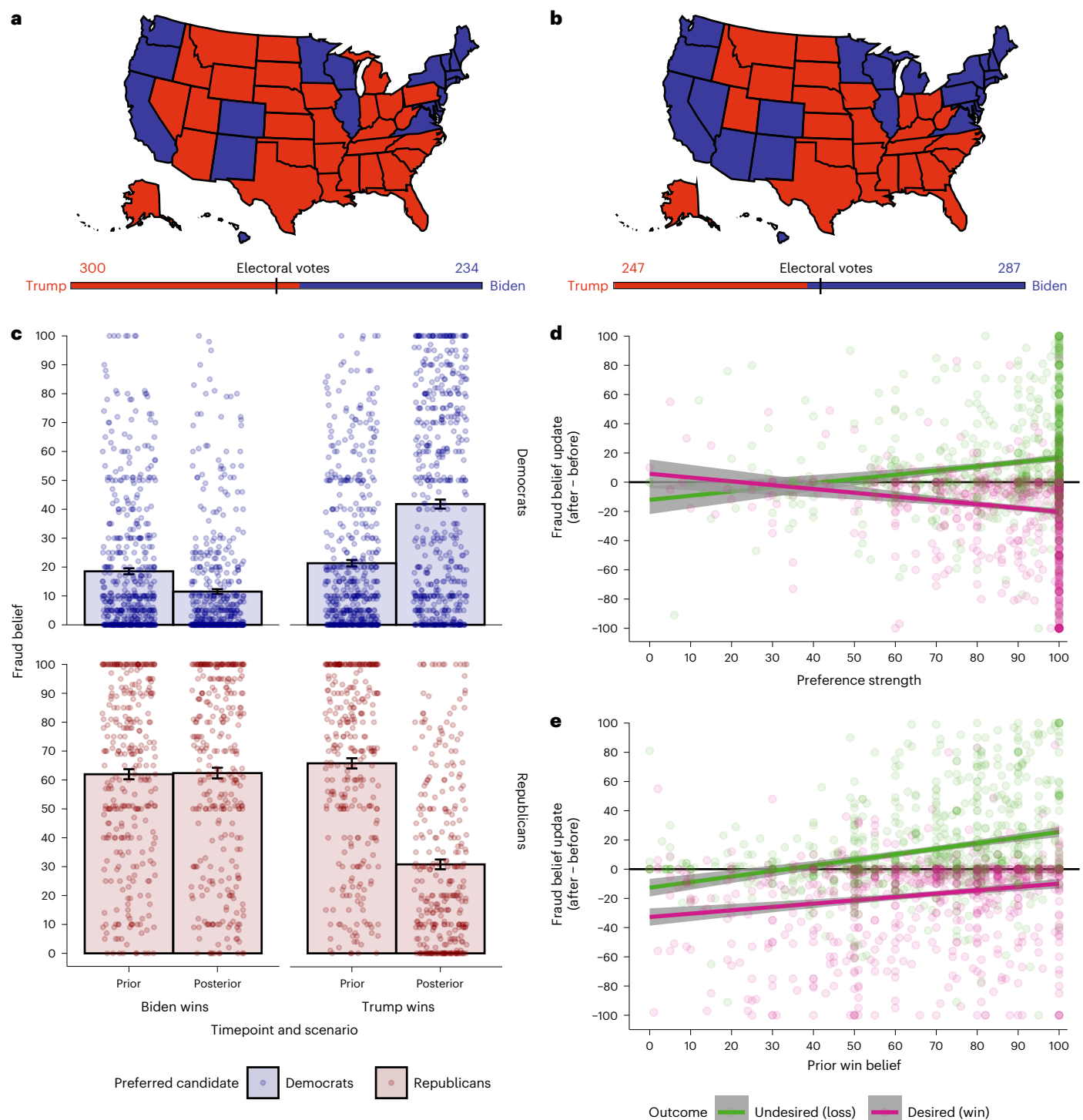


Fig. 1 | Empirical fraud belief update given hypothetical election outcome maps. a, b. Participants were randomly presented with either a map showing a Republican win (Trump, red) (**a**) or a map showing a Democratic win (Biden, blue) (**b**). The bottom scale in each map shows the distribution of electoral votes (the party receiving more than 270 electoral votes wins the presidential election). Maps were created by flipping four out of the six states that were considered critical at the time of the survey with votes still being counted, using the election tool provided by the fivethirtyeight.com website: <https://projects.fivethirtyeight.com/trump-biden-election-map/>. The maps presented here are an illustration of the maps used in the experiment. **c.** Prior and posterior (following presentation of the hypothetical maps) fraud beliefs are shown separately for Democrats (top, blue, $N=1,032$ participants) and Republicans (bottom, red, $N=610$ participants), for each scenario. Points represent single participants, and error bars represent s.e.m. across participants. Fraud beliefs increased following a hypothetical loss and decreased

following a win, except for Republicans following a loss, potentially because many Republican participants' beliefs were already maximal before the hypothetical outcome (that is, a ceiling effect). **d.** Fraud belief updates, aggregated across partisan groups, as a function of strength of preference towards the preferred candidate (x axis), separately for preferred candidate loss (green) and win (purple). Increases in fraud beliefs following a loss and decreases following a win (as shown in **c**) both scale with preference strength: updates are close to 0 with weak preferences and increase in magnitude with stronger preferences (neutral, 0; the crossover at the left is an artefact of linearity constraints). **e.** Fraud belief updates as a function of the prior win belief (probability of the preferred candidate's win; equal probability, 50). Updates are larger when the outcome is less expected (that is, when the prior win belief is higher but the preferred candidate loses, or when the prior win belief is lower but the preferred candidate wins). Points represent single participants. The lines represent the linear fit, and the grey shading represents 95% CIs of the linear fit.

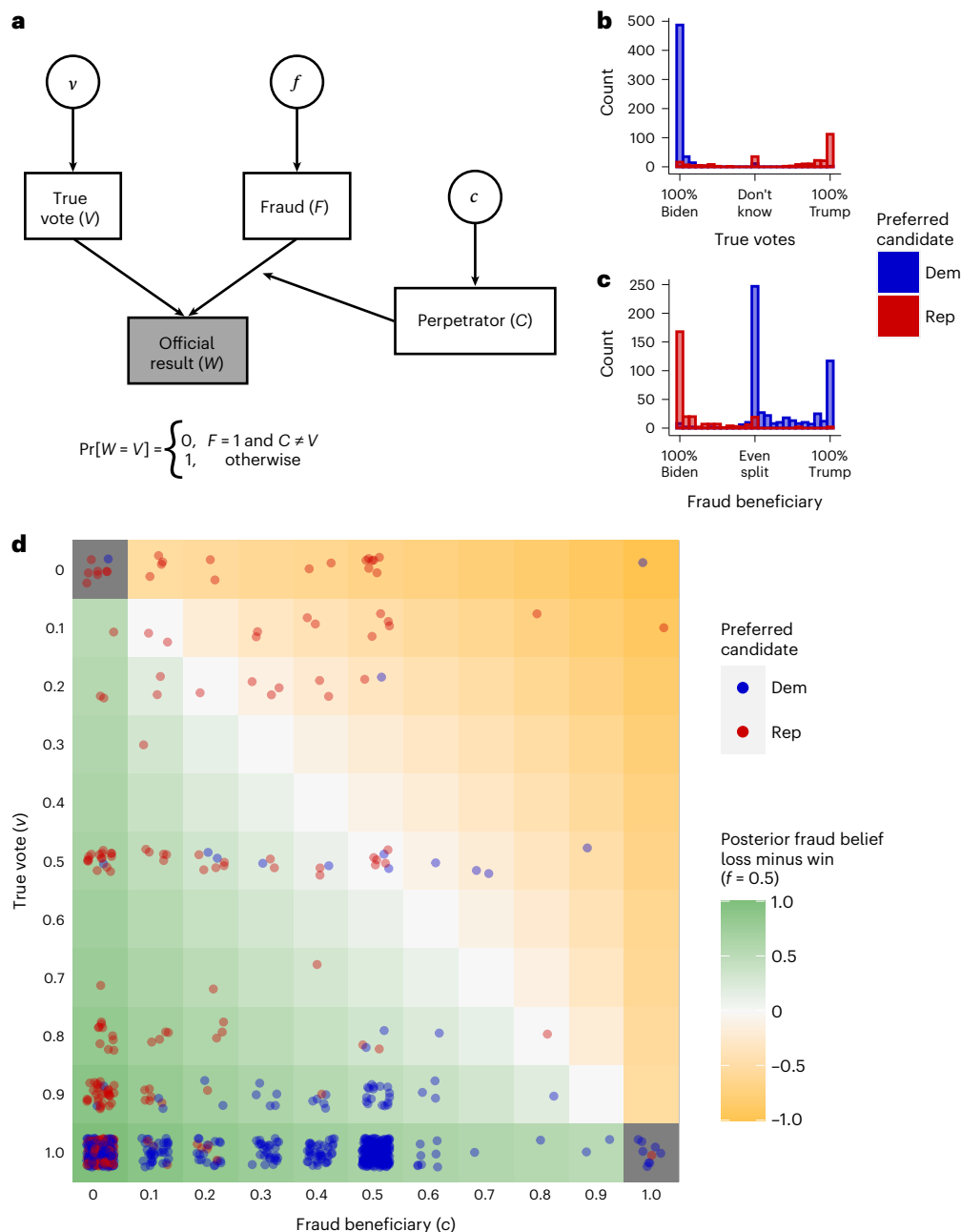


Fig. 2 | Bayesian model, priors and predictions. a, The generative model.

The observer begins with a prior belief state comprising the probability v that candidate A (the participant's preferred candidate) will win the true vote, the probability f of substantial fraud and the probability c that fraud (if present) favours candidate A. After observing the official election outcome, these beliefs are updated according to Bayes' rule (Methods). **b, c**, Histograms of participants' empirically measured prior beliefs (from the follow-up sample) regarding the true (fraudless) winner (v) (**b**) and the candidate benefiting more from fraud (c) (**c**). Both beliefs strongly diverge on the basis of partisan preferences. Dem, Democratic; Rep, Republican. **d**, Heat map of the model's predicted difference in posterior fraud belief between loss and win scenarios, as a function of v and c , when prior fraud belief $f = 0.5$. Larger values of v indicate a stronger belief that

candidate A will win the true vote. Larger values of c indicate a stronger belief that fraud favours candidate A. Values of 0.5 reflect no fraud, or fraud from both sides with neither candidate favoured. We define A as the preferred candidate of each partisan group for convenience, a choice of notation that does not affect the model's structure or predictions. Thus, high v and low c (shaded in green) indicate prior beliefs that the preferred candidate will win the true vote but the opposing side will cheat. In this part of the prior belief space, fraud beliefs increase for a loss and decrease for a win (a direction desirability effect). Most participants are located in this space, and this was especially true for participants with stronger preferences (a magnitude desirability effect; Extended Data Fig. 7). Coloured points represent single participants based on empirically measured priors (binned and then jittered within each bin for visibility).

These directional updates in fraud beliefs were significantly associated with participants' strength of preference for one candidate and prior win beliefs (each controlling for the other in a multiple regression model; Fig. 1d,e). Following a preferred candidate loss, increases in fraud beliefs were greater for those with stronger partisan preferences ($\beta = 0.10$, 95% CI 0.04 to 0.15, $t_{(800)} = 3.20$, $P = 0.001$), and

when the outcome was less expected ($\beta = 0.16$, 95% CI 0.09 to 0.23, $t_{(800)} = 4.31$, $P < 0.001$; Supplementary Table 4). Following a preferred candidate win, fraud beliefs decreased more strongly with stronger preferences ($\beta = -0.11$, 95% CI -0.16 to -0.07 , $t_{(799)} = -5.05$, $P < 0.001$), but updates were not significantly related to prior win belief ($\beta = -0.03$, 95% CI -0.08 to 0.02 , $t_{(799)} = -1.11$, $P = 0.265$; Supplementary Table 5).

Results were robust with respect to model variants and partisan groups ('Partisan asymmetry in belief updating' and 'Update of fraud belief: robustness across model variants and covariates' in Supplementary Information, and Extended Data Fig. 2e). Furthermore, the empirical desirability effects remained significant when limiting the analysis to the follow-up subsample, except for the scaling of the fraud belief update with preferences following a loss ('Update of fraud belief, restricted to the follow-up sub-sample' in Supplementary Information).

In sum, we observed three empirical effects: (1) participants believed in fraud less following a win, and more following a loss (a 'direction desirability effect'). These updates were larger (2) when the preference was stronger (a 'magnitude desirability effect') and (3) when the outcome was less expected, at least for the loss scenario (an 'expectancy effect'). These effects generalized to members of both political parties.

A Bayesian model predicts the observed desirability effects

At first glance, these results suggest directionally motivated reasoning, that is, biased updating in favour of preferred beliefs. However, as explained in the introduction, desirability effects might emerge from rational belief updating as a result of how fraud beliefs interact with other beliefs^{1,18,19,23,24}. As Jern et al.²⁰ previously demonstrated, using a model structurally identical to the one we report below, Bayesian updating can yield different conclusions from the same evidence if people hold divergent prior beliefs. In a belief systems framework, fraud beliefs result from abductive inferences ('inference to best explanation') about which of two potential causal explanations (true votes or fraud) is the most likely explanation for the outcome. If both sides expect their candidate to win, then normative updating could explain the observed desirability effects.

To test this proposition, we formulated a Bayesian attribution model (not pre-registered), which formalizes the abductive inference concept by describing a Bayesian observer of an election outcome with two candidates (A and B) jointly determined by three binary causal variables (Fig. 2a): winner of the true votes (V : election outcome given no fraud), presence of substantial fraud that could affect the outcome (F) and the side perpetrating any such fraud (C). Thus, V represents a causal explanation for the observed outcome, and the combination of F and C together represent an alternative causal explanation.

The observer's prior belief constitutes a probability distribution over these variables, taken to be independently distributed (within each individual) with subjective probabilities ν , the probability that candidate A will win the most true votes; f , the probability of substantial fraud; and c , the probability that fraud (if present) favours candidate A. The candidate winning the true vote will win the election unless substantial fraud exists and favours the other candidate, in which case the latter wins. The model provides quantitative predictions by simultaneously updating beliefs in the three binary variables (true vote winner, fraud and fraud beneficiary), as dictated by the mathematics of Bayesian inference. The posterior fraud belief given a win by candidate A ($W = A$) or by candidate B ($W = B$) as predicted by the model is as follows (for full derivation and explanations, see Methods, and for visualizations, see Extended Data Fig. 3):

$$\Pr[F = 1|W = A] = \frac{\Pr[F = 1] \Pr[W = A|F = 1]}{\Pr[W = A]} = \frac{fc}{\nu + f(c - \nu)}$$

$$\Pr[F = 1|W = B] = \frac{\Pr[F = 1] \Pr[W = B|F = 1]}{\Pr[W = B]} = \frac{f(1 - c)}{1 - \nu + f(\nu - c)}$$

The posterior true vote belief is given by (Extended Data Fig. 4):

$$\Pr[V = A|W = A] = \frac{\Pr[V = A] \Pr[W = A|V = A]}{\Pr[W = A]} = \frac{\nu(1 - f) + \nu fc}{\nu + f(c - \nu)}$$

$$\Pr[V = A|W = B] = \frac{\Pr[V = A] \Pr[W = B|V = A]}{\Pr[W = B]} = \frac{\nu f(1 - c)}{1 - \nu + f(\nu - c)}$$

For convenience, we define A as the preferred candidate, a purely notational choice that aligns interpretation of ν and c across partisan groups. We used the empirically obtained ν and c from the follow-up survey (subsample, $N = 828$, not pre-registered) to obtain model predictions. The model accurately predicted individual participants' fraud belief updates in the original survey, with no free parameters. The correlation between predicted and empirical fraud belief updates was $r = 0.63$ (95% CI 0.586 to 0.669, $t_{(826)} = 23.27$, $P < 0.001$, $N = 828$, root mean squared error (RMSE) 0.336; Extended Data Fig. 5). Democratic and older participants from the original sample were more likely to complete the follow-up survey (logistic regression, preferred candidate: log odds ratio -0.462 , 95% CI -0.75 to -0.18 , $z = -3.21$, $P < 0.001$; age: log odds ratio 0.03, 95% CI 0.02 to 0.03, $z = 6.15$, $P < 0.001$). However, sensitivity analyses revealed that the model fits similarly well for both partisan groups and that the quality of the fit was insensitive to age and preference strength ('Sensitivity analyses of the Bayesian model' in Supplementary Information), indicating robustness to the sample characteristics available in our data.

Importantly, although the model did not have access to participants' preferences, it successfully captured the desirability effects we observed (Fig. 3 and Extended Data Fig. 6): model predictions on the basis of individual participants' beliefs show increased fraud beliefs after a loss ($N = 398$, mean 38.59, 95% CI 35.89 to 41.28, $t_{(397)} = 28.16$, $P < 0.001$), decreased fraud after a win ($N = 430$, mean -18.54 , 95% CI -21.10 to -15.99), $t_{(429)} = -14.27$, $P < 0.001$), and scaling of both effects with the strength of partisan preferences (loss, $\beta = 0.10$, 95% CI 0.06 to 0.15, $t_{(384)} = 4.19$, $P < 0.001$; win, $\beta = -0.12$, 95% CI -0.17 to -0.07 , $t_{(417)} = -4.96$, $P < 0.001$). The model captures these effects and predicts individual differences in belief updating because participants' desires and preferences are correlated with their beliefs in the true winner (ν) and fraud beneficiary (c).

More specifically, the model predicts that a rational agent observing a win by their preferred candidate will decrease fraud beliefs if $\nu > c$ and increase fraud beliefs if $c > \nu$, and vice versa for a win by the dispreferred candidate (Fig. 2d and Extended Data Fig. 3a). Thus, a direction desirability effect arises in a rational observer when the observer believes their preferred candidate is more likely to win the true votes (ν) than to benefit from fraud (c). Indeed, almost all participants from both parties (Fig. 2b–d and Extended Data Fig. 7a,b) believed that (1) their candidate would have been the true winner in the absence of fraud ($\nu \approx 1$) and (2) if there was fraud, it favoured the dispreferred candidate ($c \leq 0.5$). Thus, nearly all participants occupied a portion of the (ν, c) parameter space that, under rational updating, would lead to a direction desirability effect (that is, partisan asymmetry), as we observed empirically ('Follow-up survey: divergent beliefs' in Supplementary Information). Moreover, the desirability effect becomes stronger as ν increases and c decreases (Fig. 2d), and indeed participants with stronger preferences reported significantly higher ν and lower c values, accounting for the magnitude desirability effect ('Prediction of magnitude desirability effects' in Supplementary Information).

In sum, the Bayesian model provides a potential mechanism for the desirability effects observed in the empirical data (Fig. 1c,d), which comprises a rational belief-updating process operating on polarized prior beliefs that co-vary with preference strength, rather than motivated updating. While Democrats expect Biden to win the true vote and fraud (if any) to favour Trump or favour both sides equally, Republicans expect Trump to win the true vote and fraud to favour Biden (Fig. 2b,c). These priors are clearly biased, since both groups cannot be correct. As for the expectancy effect (Fig. 1e), the relationship between prior win belief and fraud update is indirect according to the Bayesian model, and depends on how prior win belief co-varies with ν, f and c . Nevertheless,

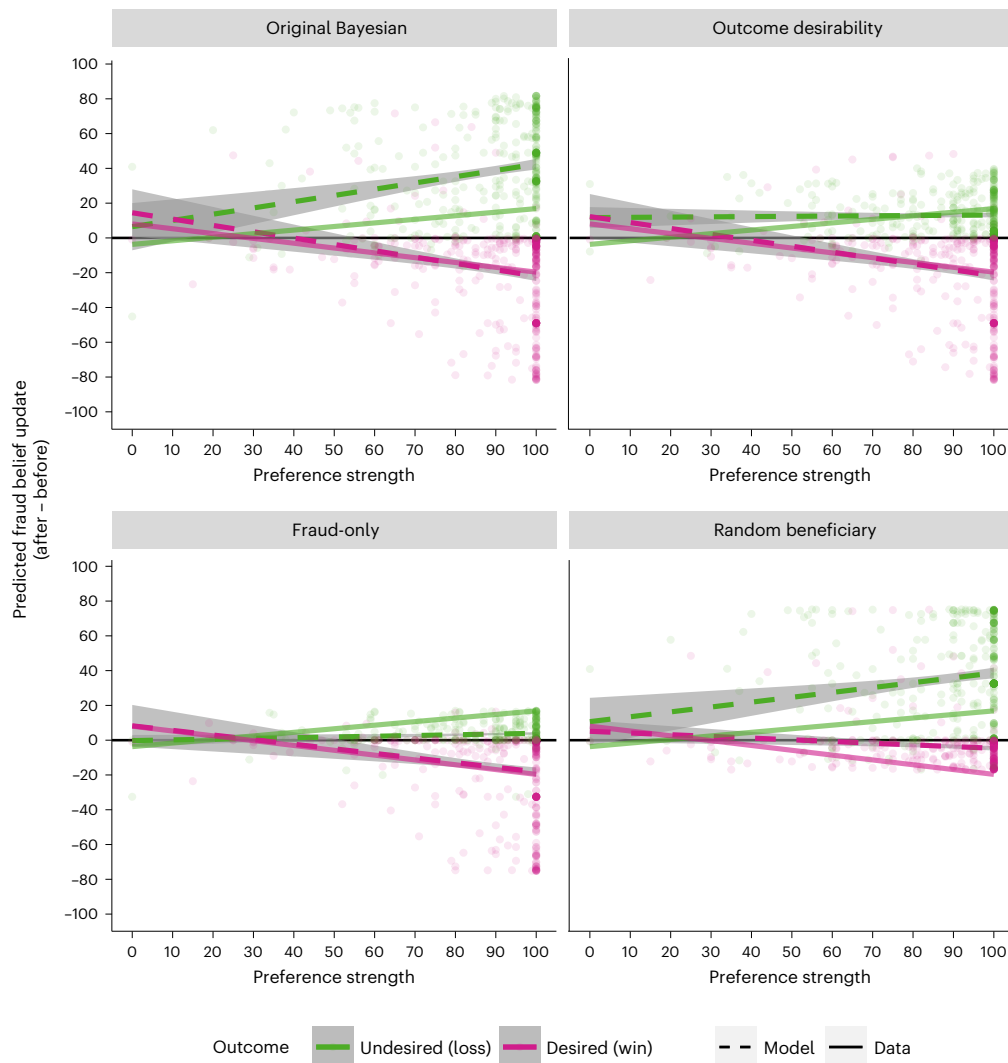


Fig. 3 | Predicted fraud belief update as a function of preference strength across models. Paralleling the empirical results from Fig. 1d, the model-based predictions of the fraud belief update for each participant are presented as a function of the empirical strength of preference towards the preferred candidate. Dashed lines show linear fits to models' predictions, with 95% confidence regions. For comparison, solid lines show the linear fits of the observed empirical patterns (from Fig. 1d). The original Bayesian model accounts for all desirability effects: increase of fraud belief following a loss, decrease following a win and

scaling of both with preference. The outcome desirability model improved the fit by tempering the over-prediction of fraud belief increase following a loss. The fraud-only model does not account for the increase of fraud belief following a loss, and the random beneficiary model does not account for the decrease of fraud belief following a win. Note that the plot for the Bayesian model is also the plot for the hypothesis desirability model, because fitting the latter to the data yields $\alpha = 0$.

the Bayesian model also accounts for the observed expectancy patterns (Extended Data Fig. 8). The one notable discrepancy, addressed further below, is that the model over-predicts the direction desirability effect by producing greater increases in fraud belief in the loss scenario than was empirically observed. Thus, if participants were optimal Bayesian observers, they would have been expected to increase fraud beliefs following a loss more than they actually did.

Comparisons with competing models

To test whether rational belief updating operating on polarized priors provides a better explanation than motivated updating, we considered two kinds of irrational updating biases. We formulated extended models that included each of these two non-Bayesian biases in addition to the attribution mechanism in the core Bayesian model (Methods). One is a hypothesis desirability bias, under which people update more in favour of beliefs they prefer (that is, people believe what they want to be true), in line with the notion that beliefs have value beyond their epistemic content¹⁶. In the 'hypothesis desirability' model, probabilities

of hypotheses (joint values of V , F and C) are weighted by their desirability, such that posterior beliefs are biased towards desired hypotheses. The second is outcome desirability bias, under which people update more on the basis of preferred outcomes, and information from undesired outcomes is discounted. In the 'outcome desirability' model, likelihoods for undesired outcomes are under-weighted and thus elicit reduced belief updates compared with more desired outcomes^{15,25,26}.

The difference between these two models is that, in the hypothesis desirability model, desired hypotheses receive a greater (more positive or less negative) update than do undesired hypotheses, regardless of what evidence is observed. By contrast, in the outcome desirability model, all hypotheses receive a stronger update if the evidence (that is, election outcome) is desired and a weaker update (closer to zero) if the outcome is undesired. In both of these models, the strength of the bias is determined by a free parameter, α , and the models reduce to the original Bayesian model when $\alpha = 0$.

The hypothesis desirability model did not improve the fit compared with the original Bayesian model, resulting in a best-fitting

α of zero (equivalent to the original Bayesian model). Positive α values, corresponding to biased inference towards desired hypotheses, further increased the over-prediction of observed desirability effects in the loss scenario. This finding supports the conclusion that the empirical desirability effects arise from rational updating. That is, we find not only that the data are predicted reasonably well by the Bayesian model but also that incorporating motivated reasoning worsens the fit. Therefore, we did not consider the hypothesis desirability model further.

In contrast, the outcome desirability model fits better than the original Bayesian model, with an estimated α of 0.3 (RMSE 0.293; bootstrapped 95% CI $0.188 \leq \alpha \leq 0.470$; F -test, $F_{(1,827)} = 364.37$, $P < 0.001$; for model comparison statistics, see Supplementary Table 13). Importantly, however, the biased updating in this model did not drive the observed desirability effects. Instead, it mitigated the over-prediction of fraud updates present in the original model by reducing the strength of the desirability effects in the loss scenario (that is, it tempers rather than explains the desirability effects produced by the Bayesian attribution process; Fig. 3). Rational attribution over-polarized prior beliefs remained the best explanation of the observed desirability effects. We discuss alternative explanations for the over-prediction of desirability effects in the original Bayesian model below (Discussion).

To test whether all three prior beliefs (f , v and c) in our modelling framework are necessary to account for the empirically observed desirability effects, we formulated two reduced Bayesian models (Methods, Fig. 3 and Supplementary Table 13). The fraud-only model, which includes beliefs only about fraud (F and C) and not about the true vote (V), fits the data relatively well but does not account for the increase of fraud beliefs following an undesired loss. The random beneficiary model, which includes V and F but omits C , fits the data worse than the other models and does not account for the decrease in fraud belief following a desired win. Thus, all three elements of the belief system appear essential to the models' explanation of the observed desirability effects.

Update of beliefs in the legitimate winner

The goal of an election is to convince the population to accept the winner as their legitimate leader. Our models also predict how people update their belief in the legitimate winner after observing the outcome (Extended Data Fig. 4). Bayesian observers update their beliefs about V and F (and C) simultaneously upon observing an outcome, following the rules of Bayesian inference.

In our sample, both the outcome desirability model and the original Bayesian model predict that 36% of the participants will be 'election proof', in the sense that more than 90% of the overall update following an official loss by their preferred candidate will be to increase fraud beliefs rather than to change belief in the true winner. The models both classify 29% of Democrats and 51% of Republicans in this way. Importantly, our sample is not strictly representative of the entire US population, which is necessary for quantitatively estimating the share of the population that holds specific beliefs or is 'election proof'. Therefore, these specific estimates, which also rely on certain assumptions about reports of extreme probabilities (Methods), should be treated only as qualitative approximations.

Nevertheless, they suggest that a considerable proportion of the population will give essentially no credence to official election results, even if their reasoning is epistemically rational, based on their prior beliefs. Moreover, our model-based estimate is corroborated by both our follow-up data (39% of Republicans in our sample were still certain that Trump was the true winner weeks after Biden was declared as the official winner) and national polls that consistently find that about one-third of Americans continue to believe that Biden won only due to fraud even months after the election (for example, https://www.monmouth.edu/polling-institute/reports/monmouthpoll_us_062121/ and <https://www.cnn.com/2021/04/30/politics/cnn-poll-voting-rights/index.html>).

Discussion

Our findings illustrate how specific combinations of beliefs can prevent rational people from accepting the results of democratic elections. Beliefs in election fraud have played a substantial role in undermining democratic governments worldwide²⁷, and have grown and remained strikingly prevalent in the United States^{28,29}. Belief in fraud undermines both motivation to vote and acceptance of election results, which bear directly on the viability of an elected government.

Our results suggest that fraud beliefs and other beliefs do not exist in isolation, but are part of systems of beliefs that interact to guide how new evidence is interpreted. These beliefs can withstand countervailing evidence, even if the belief-updating process is rational, depending on other prior beliefs that determine how credit for observed evidence (for example, an election winner) is distributed across potential causes. Real-world evidence that directly bears on these causal beliefs is difficult at best for individuals to obtain. For example, direct evidence about V requires a fraudless election, and the 'gold standard' evidence about F (and C) is the result of an unbiased investigation. Rhetoric that increases belief in fraud or challenges the credibility of the investigation can thus create a 'short circuit': if one side is believed to be cheating disproportionately (C), then an election outcome (W) might not be taken as evidence regarding the true winner (V), and might instead serve almost solely to update fraud beliefs (F and C). Our model-based analysis suggests this occurred in a considerable proportion of our sample.

The belief systems framework has important implications for efforts to counteract fraud beliefs in the United States and beyond. Such efforts often focus on directly 'debunking' them (providing direct evidence against fraud). However, these debunking efforts largely fail³⁰. Our model affords alternative ways of thinking about how to change beliefs, in accordance with other perspectives emphasizing the worldview of the person holding the beliefs and the importance of alternative causal explanations^{31–33}. Specifically, to reduce belief in election fraud, in addition to (1) directly targeting them with evidence that fraud is very rare (decreasing f in our model), it is advantageous to simultaneously provide evidence that (2) Americans have diverse preferences and one's dispreferred candidate truly is supported by many (decreasing v), and (3) to the extent fraud does occur, its benefits are distributed across both candidates (increasing c). Furthermore, our model could be useful for distributing resources across interventions targeting these three prior beliefs, by predicting how each would impact people's willingness to update fraud beliefs (for a formal proof of concept, see 'Comparing among interventions targeting prior beliefs' in Supplementary Information).

Real-world political campaigns already consider multiple beliefs in this fashion to varying degrees, though perhaps without the explicit formulation embodied in our model. For example, the Trump campaign simultaneously targeted these same three beliefs before and during the 2020 election³⁴: (1) 'It's rigged': election fraud is rampant (increase f); (b) 'We're winning': Trump is popular and enjoys the true support of most Americans (increase v for Trump) and (c) 'The Democrats are crooked': fraud overwhelmingly favours the opposition (decrease c). Although these messages may initially seem unrelated, our model shows that they combine to form a coordinated strategy to create evidence-resistant beliefs about the true winner. Moreover, the campaign also cast suspicion on the unprecedented use of mail-in ballots in the 2020 election (due to the coronavirus disease 2019 pandemic), which probably made fraud more plausible as an alternative explanation. While some political actors already seem able to intuitively apply this approach in effective campaigns, a formal understanding such as we offer here could be beneficial to studying and preventing misinformation campaigns, and promoting belief change for the sake of society.

A few caveats and limitations deserve additional mention. First, our sample was not strictly representative of the US population. Thus, it is not suited for quantifying population-average beliefs (or belief updates) across the population as election polls intend to do, and does

not precisely estimate prevalence of fraud beliefs across the entire US population. Moreover, Democratic and older participants were more likely to complete the follow-up survey, which could potentially bias numerical estimates. However, the presence of desirability effects and the model's goodness of fit did not vary across partisan groups and age, suggesting that our qualitative conclusions are robust to these sample characteristics.

Second, there are at least two reasons that participants' reported prior beliefs might have been more extreme than their actual priors during the main survey: (1) reports of v and c were collected from the follow-up survey and thus may have been impacted by intervening events (for example, further claims by the Trump campaign that he was the true winner but lost because of fraud, and consensus among Democrats that Biden won in the absence of fraud). Indeed, fraud beliefs after the 2020 election have been shown to become more polarized with time³⁵. (2) Participants might have reported more extreme prior beliefs than they actually held because they were motivated to communicate partisanship (expressive responding³⁶). If reported prior beliefs were artificially polarized for either of these reasons, this could undermine the conclusion that participants updated their beliefs rationally. That is, our Bayesian model holds that the desirability effects in participants' posterior beliefs are rooted in the asymmetries in their priors, and that argument could fail if the asymmetries were not genuine. However, it is unlikely that the model's predictive success was artificially inflated in this way. First, recent studies, including a comprehensive analysis of fraud beliefs in the context of the 2020 US election, found little or no support for expressive responding in reported political beliefs^{37,38}. Second, we conducted an additional analysis that assumed participants' prior beliefs were less extreme than reported. This analysis adjusted the reported priors by shrinking them towards the undesired state before passing them to the models to calculate predictions ('Simulations of extreme priors' in Supplementary Information). This improved, rather than worsened, the Bayesian model's fit, by reducing over-predictions of fraud belief increases in the loss scenario. While these results are consistent with some degree of expressive responding or polarization over time, they show that the model's ability to capture the desirability effects in our primary analysis was despite such effects, not because of them.

Third, the present models are relatively simple and do not explicitly incorporate many beliefs, such as confidence in election polls or in news sources. Interestingly, beliefs about many such additional latent variables are not expected to change the models' predictions regarding fraud belief, because they are not directly causally related to the official winner (Additional beliefs not included in the models' in Supplementary Information).

Finally, the unique settings of our data collection may elicit pre-treatment effects, whereby participants may appear to be insensitive to experimentally presented information because they have already been exposed to similar information. However, we compare between participants who were given contrasting observations (rather than certain information versus no information), and the results clearly show that participants did update their beliefs. Furthermore, a strength of our Bayesian approach is that updates in Bayesian models depend on both the prior beliefs and the new evidence, and thus they account for pre-treatment effects that would arise when those two are aligned (that is, when priors are already strongly in the direction of the new evidence). Nevertheless, future studies should examine whether our findings generalize to other situations, and particularly to situations with milder prior beliefs.

In sum, our findings indicate that the desirability effects in fraud beliefs we observed during the 2020 US presidential election are best attributed to rational updating from polarized prior beliefs rather than to biased updating. This suggests that partisan groups tend to stereotype opposing groups as irrational because they do not know (or disregard) their prior beliefs. Accurately taking them into account might help bring people closer together by showing they are in fact

often rational. Such polarized prior beliefs could be driven by rational and/or irrational processes, including many social, technological and psychological processes beyond the scope of our model and study. For example, they could result from differential exposure to evidence driven by social media, which has been shown to reinforce belief polarization⁵ and create 'filter bubbles' and 'echo chambers'^{39,40}. In this case, the partisan differences we found in people's priors may be rational, in that individuals are simply drawing on the information available to them. However, the process of information selection may itself be irrationally biased, and indeed people have been found to select information that reinforces desired beliefs^{41–43}. Prior beliefs are also shaped powerfully by social norms and group identity^{4,44–46}. This could be interpreted as an instrumental bias (agreeing with in-group members has value) or potentially as rational, if people place more trust in their in-group to deliver accurate information. Our model does not attempt to explain these varied and powerful processes, but it helps to understand the dynamics of their influence.

More broadly, our findings highlight the importance of a systems perspective in understanding human beliefs across domains, and in particular the critical role of attribution. The dynamics we observed here are probably at play in multiple areas crucial for human wellbeing, including beliefs about the self and others that shape individual mental health⁴⁷, beliefs in the efficacy and safety of vaccines⁴⁸, and beliefs in the need for action to address climate change^{49,50}. In each of these domains, evidence must be attributed across multiple latent causes, shaping which beliefs are reinforced by observed evidence. Our approach provides a blueprint that can be applied to beliefs in these diverse domains, identifying key beliefs and making quantitative predictions about how they will interact.

Methods

Participants

A large online sample of 1,760 American citizens was collected from MTurk (Amazon) via Cloud Research²². The sample size was chosen on the basis of available funds aiming for a large sample across the United States. Participants were 18 years old or older, provided their consent for participation online before accessing the survey questions and were paid \$1 for their participation. The study was approved by Dartmouth College's Committee for the Protection of Human Subjects.

Thirty-two participants were excluded due to failure to correctly answer a simple attention check included in the survey and one due to not meeting the participation requirement of American citizenship or permanent residency. Eighty-five additional participants were excluded from analyses because they chose 'Other' for their preferred Presidential candidate (and not 'Donald Trump' or 'Joe Biden'). Analyses included 1,642 participants. Using G*Power version 3.1 (ref. 51), we found that the minimal detectable effect size for either predictor (preference strength or prior win belief) in our main regression analysis of fraud belief update, based on $\alpha = 0.05$, power of 0.9 and $N = 818$ (the number of included participants who were randomly assigned to the loss scenario, which is lower than the 824 participants in the win scenario), is a small effect size of Cohen's $f^2 = 0.016$ (ref. 52).

Out of the 1,642 included participants, 1,032 preferred Joe Biden and 610 preferred Donald Trump. Partisan affiliation was generally, but not perfectly, aligned with the preferred candidate. We use Democrat and Republican to refer to participants according to their preferred candidates (Biden and Trump, respectively). The sample included 1–138 participants from each of the 50 US states and the District of Columbia. The age of participants ranged from 18 to 84 years old (mean 41.6, s.d. 3.1 years) and Republicans were older than Democrats on average (Democrats, mean 39.9, s.d. 12.8 years; Republicans, mean 44.5, s.d. 13.1 years).

Materials and procedures

Each participant completed a survey distributed on the Qualtrics platform (Qualtrics). The US election took place on 3 November

2020. The survey was available online from 4 November at 20:00 until 5 November at 10:30 Eastern Time. During that time, votes were still being counted and there was no conclusive winner. Six states were considered critical ‘swing states’ with no clear winner and votes still being counted: Nevada, Arizona, Michigan, Pennsylvania, North Carolina and Georgia. We created two hypothetical maps that differed by flipping four of these six states: Nevada, Arizona, Michigan and Pennsylvania, using the election tool provided by [fivethirtyeight.com](https://projects.fivethirtyeight.com/trump-biden-election-map/): <https://projects.fivethirtyeight.com/trump-biden-election-map/>. Each map provided a graphical representation of the states’ final results (red for Republican victory and blue for Democrat victory), sum of electoral votes for Republicans and Democrats and the winner: one map with a Republican (Donald Trump) win and one with a Democrat (Joe Biden) win (for an illustration, see Fig. 1ab).

We collected participants’ demographic information (age and state of residency), political affiliation (Republican, Democrat, Independent or Other), candidate preference (Donald Trump, Joe Biden or Other), preference strength (continuous scale from 0 (don’t care either way) to 100 (extremely strong preference)), prior subjective probability of the preferred candidate’s win (prior win belief; continuous scale from 0 (no chance) to 100 (definite win)) and prior subjective probability of fraud playing a significant role in the election outcome (prior fraud belief; continuous scale from 0 (not at all) to 100 (extremely)); for full description of the survey questions, see Supplementary Table 1). Then, each participant was randomly presented with one of the two maps described above and was asked to indicate their belief about fraud affecting the election outcome if it were to turn out as in the map shown to them.

Analysis

All statistical tests reported in the manuscript and Supplementary Information were two sided.

Pre-registration. We pre-registered our analysis plan and predictions on the Open Science Framework (<https://osf.io/kucsw>) after data collection but before inspection of beliefs-related data. We did examine some portions of the data, such as the number of participants supporting each candidate, as well as demographic information. These procedures did not inform us about the effects of interest. Rather, they were meant to monitor data quality and sample characteristics.

Deviations from pre-registration. We report a different multiple linear regression model than the one we pre-registered for the update of fraud belief, to simplify the model and interpretations. The differences between the two models, along with the results of the pre-registered model, which are in line with those of the revised one, are described in ‘Updates in fraud belief: pre-registered analysis’ in Supplementary Information.

The pre-registration was based mainly on the biased belief updating hypothesis and focused on the empirical data from the original survey. After observing the data of the original survey, we developed the Bayesian model and associated variants. A follow-up survey was then collected to test the model and expand the findings from the original survey. The follow-up survey and models were not pre-registered.

Updating of fraud belief. At the time of the pre-registration, we predicted, on the basis of the desirability bias perspective and learning theories, that the update of fraud belief (posterior minus prior subjective probability of fraud playing a significant role in the election outcome) would be affected by two factors: (1) participants’ preferences, such that fraud belief would increase when the hypothetical results were undesired, with a larger update for stronger preferences and (2) participants’ prior win belief, such that fraud belief will increase when the hypothetical results are less expected.

We tested our predictions with multiple linear regression predicting update of fraud belief as a function of the preference

strength and the prior win belief, separately for the loss scenario (that is, following a map in which the preferred candidate lost) and for the win scenario (that is, following a map in which the preferred candidate won). The prior fraud belief, preferred candidate (Biden/Trump), participant’s age and order of survey submission during data collection were included in both models as covariates. The participant’s state of residence was not included as it did not significantly reduce residual variance while substantially increasing model complexity. All numeric variables included in the model (dependent and independent) were z scored. The same models were also tested separately for each partisan subgroup, where the preferred candidate covariate was omitted (‘Partisan asymmetry in belief updating’ in Supplementary Information). Additionally, a one-sample two-sided *t*-test was used to test whether the mean change in fraud belief was significantly non-zero in each scenario.

Bayesian model. To test whether the general patterns of belief update we found can reflect rational reasoning from possibly polarized prior beliefs, we formulated a Bayesian model that assumes participants attribute the election outcome to the combination of the true votes (that is, election outcome given no fraud) and fraud (Fig. 2a). The candidates are labelled A and B, with A representing the participant’s preferred candidate. Let $W, V \in \{A, B\}$ be the official and true winners, respectively. Let $F \in \{0, 1\}$ indicate whether there was substantial fraud that could overturn the outcome, and let $C \in \{A, B\}$ indicate which side committed fraud (only meaningful when $F = 1$). Therefore, the official winner will match the true winner unless there was substantial fraud by the opposition ($F = 1$ and $C \neq V$):

$$\Pr[W = V] = \begin{cases} 0 & F = 1 \text{ and } C \neq V \\ 1 & \text{otherwise} \end{cases}$$

We assume the participant has an independent prior on V , F and C , and denote $\Pr[V = A] = v$, $\Pr[F = 1] = f$, $\Pr[C = A] = c$. For convenience, we define $R \in \{\emptyset, A, B\}$ by $R = \emptyset$ when $F = 0$ and $R = C$ when $F = 1$; that is, R indicates who committed substantial fraud, with \emptyset denoting no one. Then the joint prior on V, R is given by

$\Pr[V, R]$	$V = A$	$V = B$
$R = \emptyset$	$v(1 - f)$	$(1 - v)(1 - f)$
$R = A$	$vf c$	$(1 - v) f c$
$R = B$	$vf(1 - c)$	$(1 - v) f(1 - c)$

The likelihoods, as described above, are given by

$\Pr[W = A V, R]$	$V = A$	$V = B$
$R = \emptyset$	1	0
$R = A$	1	1
$R = B$	0	0

The joint posteriors are then given by

$\Pr[V, R W = A]$	$V = A$	$V = B$
$R = \emptyset$	$v(1 - f) / Z_A$	0
$R = A$	$vf c / Z_A$	$(1 - v) f c / Z_A$
$R = B$	0	0

and

$\Pr[V, R W = B]$	$V = A$	$V = B$
$R = \emptyset$	0	$(1 - v)(1 - f) / Z_B$
$R = A$	0	0
$R = B$	$vf(1 - c) / Z_B$	$(1 - v)f(1 - c) / Z_B$

with partition function

$$Z_A = \Pr[W = A] = v + f(c - v)$$

$$Z_B = \Pr[W = B] = 1 - v + f(v - c)$$

Marginalizing the joint posteriors gives the posteriors on F :

$$\Pr[F = 1 | W = A] = \frac{fc}{v + f(c - v)} = \frac{c}{c + (1 - f)(v - c)} \Pr[F = 1]$$

$$\Pr[F = 1 | W = B] = \frac{f(1 - c)}{1 - v + f(v - c)} = \frac{1 - c}{1 - c + (1 - f)(c - v)} \Pr[F = 1]$$

Therefore, fraud belief increases if the official winner is more likely to be the cheater than to be the true winner (that is, if $c > v$ for $W = A$, and if $c < v$ for $W = B$).

We also derived the posteriors for V , which are given by

$$\Pr[V = A | W = A] = \frac{v(1 - f) + vfc}{v + f(c - v)}$$

and

$$\Pr[V = A | W = B] = \frac{vf(1 - c)}{1 - v + f(v - c)}$$

Importantly, as can be seen above, the model simultaneously updates beliefs about V , F and C , as dictated by Bayes rule. For example, the posterior probability that candidate A is the true vote winner given that A is the official winner ($\Pr[V = A | W = A]$) depends on the prior probability of V ($\Pr[V = A]$, which is v), the likelihood of A being the official winner given that A is the true vote winner ($\Pr[W = A | V = A]$, which depends on f and c) and the marginal probability of A being the official winner ($\Pr[W = A]$, which depends on v , f and c). The belief systems perspective comes into play primarily in the term $\Pr[W = A | V = A]$. This likelihood term masks hidden complexity, because it depends on other causal explanations. In our model, its value depends on f and c : for example, $\Pr[W = B | V = A] = f(1 - c)$. Thus, the updates regarding V , F and C are simultaneous and inter-dependent. Updates about V depend on f and c (in addition to v), and updates about F depend on v and c (in addition to f). In some cases, depending on the prior beliefs, it is rational not to take the election outcome as diagnostic of the true winner. For example, if an agent's prior beliefs are $f = 1$ and $c = 0$, the agent is certain that substantial fraud is committed in favour of candidate B, and therefore that B will be the official winner regardless of who wins the true vote. In such a case, a win by candidate B is not diagnostic of V , and therefore a Bayesian (rational) agent will not update beliefs about V .

Follow-up survey. To assess the participants' beliefs about the true vote winner and who commits fraud, we collected a follow-up survey, about 11 weeks after the original survey. The follow-up survey was again distributed on Qualtrics via Cloud Research and was available to all participants who completed the original survey. Although participants did not originally sign up for a multi-session study, and the follow-up survey was available for less than 24 h, more than half of the original

sample completed it ($N = 937$; nine additional ones failed the attention check). Participants were notified about the follow-up survey via email (sent via Cloud Research) and were paid an additional \$1 upon completion of the follow-up survey. For the purpose of testing the models and assessing the priors, we further excluded 44 participants who indicated 'Other' as their preferred candidate during the original survey, 22 participants who indicated 'Other' as their preferred candidate during the follow-up survey and 43 additional participants who changed their preferences between the surveys (all switched from preferring 'Trump' to preferring 'Biden'). We obtained the priors and applied the models on the remaining $N = 828$ participants.

Participants included in the final follow-up sample were 18–77 years old with at least one participant from each state. The follow-up survey was accessible between the evening of 19 January 2021 and the inauguration of Joe Biden (around 12:00 Eastern Time on 20 January 2021), and included questions about the participants' preferences (both at the time of the follow-up survey completion and during the election), their beliefs about the winner of the election, their beliefs about fraud and the activities they considered when answering the questions about fraud (for all survey questions, see Supplementary Table 2). Participants' values for v were computed from their answer to the question 'Who would have won in a fraudless election?' along with their reported confidence in that answer (continuous scale 0–100). The confidence level was multiplied by -1 for participants who indicated the opponent candidate as the fraudless winner, and v was set as the confidence level after rescaling to a range of $[0, 1]$. For participants who answered 'Don't know' with regard to the fraudless winner, we set their v to 0.5, indicating uncertainty. Participants' values for c were computed from their answer to the question 'Who benefited most from fraudulent voting activity?' (continuous scale 0 (Trump/Biden) to 100 (Biden/Trump); the labels' order was counterbalanced across participants). The value of c was set as the reported probability of fraud benefiting their preferred candidate, divided by 100 for a range of $[0, 1]$. The correlations between the preference strength and v , and between the preference strength and c , were computed with Pearson correlation. Prior fraud belief values (f) from the original survey were also scaled to a range of $[0, 1]$ to be used in the models.

We derived the models' predicted posteriors and magnitude of updating of fraud beliefs on the basis of the empirically measured priors for the 828 participants who successfully completed the follow-up survey, preferred either Trump or Biden (that is, not 'Other') and did not change their preference between the two surveys. Many participants reported extreme prior beliefs (f , v and/or c of 0 or 1), which are unlikely to indicate true mathematical certainty and instead may result from lack of granularity in reports, biases towards anchors or a tool to communicate partisanship or support^{36,53,54}. Furthermore, for some participants, these extreme prior beliefs led to undefined posterior beliefs in the model, because they implied zero probability for the hypothetical outcome the participant observed (for example, viewing a win by the preferred candidate after having a 100% belief that the opposition would reverse the election through fraud). Therefore, we regularized extreme priors to the range $[0.01, 0.99]$ by converting values that are lower than 0.01 to 0.01, and values that are higher than 0.99 to 0.99 when computing the models' predictions. Following this step, we could derive predictions for all 828 follow-up participants. Importantly, we verified that the choice of regularization value (no regularization, regularization to $[0.01, 0.99]$ or regularization to $[0.05, 0.95]$ did not affect the study's conclusions ('Regularization of extreme prior beliefs' in Supplementary Information). To assess the model fit, we calculated the RMSE and Pearson correlation between the empirical and predicted values of fraud belief updates.

Non-Bayesian models. To examine potential deviations from the original Bayesian model, we formulated and tested two extended models, each adding a particular form of non-Bayesian desirability

bias to the model's belief updating rule. The Bayesian model uses the standard Bayes' rule

$$\Pr[h|W] \propto \Pr[h] \Pr[W|h]$$

where h is the hypothesis (combinations of V , F and C in our models), W is the observed outcome and \propto stands for proportionality. As an analogue of correct probabilities in true Bayesian models, $\Pr[\cdot]$, we introduce the notation $\mathbb{B}[\cdot]$ to stand for belief in non-Bayesian models.

Hypothesis desirability model. One possible form of bias in belief updating comes from desirability of hypotheses, whereby hypotheses are weighted according to their subjective utility when determining posterior beliefs. We implemented this by assuming each hypothesis h is weighted by a linear function $1 + \alpha U(h)$, where U denotes subjective utility and α is a free parameter determining the strength of the desirability bias:

$$\mathbb{B}[h|W] \propto (1 + \alpha U(h)) \Pr[h] \Pr[W|h]$$

When $\alpha = 0$, the hypothesis desirability model reduces to normative Bayesian inference. As $\alpha \rightarrow \infty$, the desirability bias becomes maximal in that every hypothesis is weighted exactly according to its utility (that is, the $\alpha U(h)$ term dominates the 1 term, and then α cancels out due to normalization).

Under the assumption that participants care about the true vote winner and not about fraud per se, we defined the utility of each hypothesis by the utility of the corresponding true vote winner, $U(h) = U(V)$. These utilities were then taken as $U(A) = u$ and $U(B) = 1 - u$, where u denotes the participant's reported preference strength towards their preferred candidate (A), scaled to lie in $[1/2, 1]$. In other words, $U(h) = u$ for all hypotheses for which $V = A$, and $U(h) = 1 - u$ for all hypotheses for which $V = B$. Negative values of α lead to down-weighting of desired hypotheses relative to undesired ones, and therefore only values of 0 or greater were considered.

These assumptions lead to the following posterior beliefs:

$\mathbb{B}[V, R W = A]$	$V = A$	$V = B$
$R = \emptyset$	$(1 + \alpha u) v (1 - f) / Z_A$	0
$R = A$	$(1 + \alpha u) v f c / Z_A$	$(1 + \alpha - \alpha u) (1 - v) f c / Z_A$
$R = B$	0	0

and

$\mathbb{B}[V, R W = B]$	$V = A$	$V = B$
$R = \emptyset$	0	$(1 + \alpha - \alpha u) (1 - v) (1 - f) / Z_B$
$R = A$	0	0
$R = B$	$(1 + \alpha u) v f (1 - c) / Z_B$	$(1 + \alpha - \alpha u) (1 - v) f (1 - c) / Z_B$

with partition function

$$Z_A = v - v f + f c + \alpha (u v + f c - u v f - u f c - v f c + 2 u v f c)$$

$$Z_B = 1 - v + v f - f c + \alpha (1 - u - v + u v - f c + u v f + u f c + v f c - 2 u v f c)$$

Marginalizing the joint posteriors gives the posteriors on F :

$$\mathbb{B}[F = 1|W = A] = \frac{f c + \alpha (f c - u f c - v f c + 2 u v f c)}{v - v f + f c + \alpha (u v + f c - u v f - u f c - v f c + 2 u v f c)}$$

$$\mathbb{B}[F = 1|W = B]$$

$$= \frac{f (1 - c) + \alpha (f - u f - v f - f c + 2 u v f + u f c + v f c - 2 u v f c)}{1 - v + v f - f c + \alpha (1 - u - v + u v - f c + u v f + u f c + v f c - 2 u v f c)}$$

Outcome desirability model. Another possible form of bias in belief updating comes from desirability of outcomes. Specifically, people have been shown to discount evidence from undesired outcomes¹. We implemented discounting of evidence by shrinking the effective likelihood of the outcome toward a uniform distribution ($\Pr[W = A|h] = \Pr[W = B|h] = 1/2$ for all h), to a degree determined by a variable $\lambda \in [0, 1]$

$$\mathbb{B}[h|W] \propto \Pr[h] ((1 - \lambda) \Pr[W|h] + \lambda/2)$$

When $\lambda = 0$, this update rule reduces to standard Bayesian inference, $\mathbb{B}[h|W] \propto \Pr[h] \Pr[W|h]$. When $\lambda = 1$, the outcome is completely disregarded, and the posterior belief is the same as the prior (the $1/2$ factor is absorbed into normalization).

Desirability bias is implemented by assuming the degree of discounting depends negatively on the desirability of the observed outcome:

$$\lambda = \alpha(1 - U(W))$$

We define the utility of the official winner (W) in the same way as we define that of the true vote winner in the hypothesis desirability model: $U(A) = u$ and $U(B) = 1 - u$, where u is the participant's preference strength towards their preferred candidate (A), scaled to lie in $[1/2, 1]$ (note that the uses of u in the two models do not together imply that participants' utility functions for the true vote winner and the official winner are the same. These are two separate candidate models, embodying different proposals for how participants' prior beliefs and preferences are related to their posterior beliefs). As before, α determines the strength of the bias. When $\alpha = 0$, any outcome leads to $\lambda = 0$, and the outcome desirability model reduces to the original Bayesian model. When $\alpha = 1$, maximally undesired outcomes (for example, a loss observed by a participant for whom $u = 1$) are completely disregarded ($\lambda = 0$). Negative values of α lead to hyper-weighting of evidence ($\lambda < 0$) that is stronger for undesired outcomes, and can also produce nonsensical negative beliefs ($\mathbb{B}[h|W] < 0$). Values of α greater than 1 lead to reverse weighting of undesired outcomes ($\lambda > 1$), meaning the posterior belief favours hypotheses that are inconsistent with the evidence. Therefore, only values of α in $[0, 1]$ were considered.

The preceding assumptions lead to the following posterior beliefs:

$\mathbb{B}[V, R W = A]$	$V = A$	$V = B$
$R = \emptyset$	$(2 - \alpha + \alpha u) v (1 - f) / Z_A$	$\alpha (1 - u) (1 - v) (1 - f) / Z_A$
$R = A$	$(2 - \alpha + \alpha u) v f c / Z_A$	$(2 - \alpha + \alpha u) (1 - v) f c / Z_A$
$R = B$	$\alpha (1 - u) v f (1 - c) / Z_A$	$\alpha (1 - u) (1 - v) f (1 - c) / Z_A$

and

$\mathbb{B}[V, R W = B]$	$V = A$	$V = B$
$R = \emptyset$	$\alpha u v (1 - f) / Z_B$	$(2 - \alpha u) (1 - v) (1 - f) / Z_B$
$R = A$	$\alpha u v f c / Z_B$	$\alpha u (1 - v) f c / Z_B$
$R = B$	$(2 - \alpha u) v f (1 - c) / Z_B$	$(2 - \alpha u) (1 - v) f (1 - c) / Z_B$

with partition function:

$$Z_A = 2(v - v f + f c) + \alpha(1 - u)(1 - 2v + 2v f - 2f c)$$

$$Z_B = 2(1 - v + v f - f c) - \alpha u(1 - 2v + 2v f - 2f c)$$

Marginalizing the joint posteriors gives the posteriors on F :

$$\mathbb{B}[F = 1|W = A] = \frac{2f c + \alpha(1 - u)f(1 - 2c)}{2(v - v f + f c) + \alpha(1 - u)(1 - 2v + 2v f - 2f c)}$$

$$\mathbb{P}[F = 1|W = B] = \frac{2f(1-c) - \alpha u f(1-2c)}{2(1-v + vf - fc) - \alpha u(1-2v + 2vf - 2fc)}$$

Similarly, the posteriors on V are:

$$\mathbb{P}[V = A|W = A] = \frac{2v(1-f+fc) - \alpha(1-u)v(1-2f+2fc)}{2(v-vf+fc) + \alpha(1-u)(1-2v+2vf-2fc)}$$

$$\mathbb{P}[V = A|W = B] = \frac{2vf(1-c) + \alpha uv(1-2f+2fc)}{2(1-v + vf - fc) - \alpha u(1-2v + 2vf - 2fc)}$$

Reduced Bayesian models. To test whether all component beliefs of the Bayesian model are needed to account for the four qualitative aspects of the direction and magnitude desirability effects (increase in fraud beliefs following a loss, decrease in fraud beliefs following a win, and a correlation with preference strength for each), we formulated two versions of reduced Bayesian models.

Fraud only. The first reduced model includes only the beliefs about fraud (F) and the side committing fraud (C), without considering the true winner (V). The side committing fraud wins. If no fraud is committed, then the winner is determined randomly. This model is equivalent to the original Bayesian model with a fixed value of $v = 0.5$ (that is, maximal uncertainty regarding the true winner). The posterior fraud belief is thus:

$$\Pr[F = 1|W = A] = \frac{2fc}{1-f+2fc}$$

$$\Pr[F = 1|W = B] = \frac{2f(1-c)}{1+f-2fc}$$

Random beneficiary. The second reduced model includes only the beliefs about fraud (F) and the true winner (V), without considering the fraud beneficiary (C). Thus, the candidate benefiting more from fraud is random. This model is equivalent to the original Bayesian model with a fixed value of $c = 0.5$. The posterior fraud belief is thus:

$$\Pr[F = 1|W = A] = \frac{f}{2v-2vf+f}$$

$$\Pr[F = 1|W = B] = \frac{f}{2-2v+2vf-f}$$

Model comparison

The free parameter of the hypothesis desirability and outcome desirability models, α , was optimized for each model based on the sum of squared errors (SSE) between the empirical and predicted fraud belief update for all participants in the follow-up sample. All models were compared on their ability to reproduce the qualitative observed desirability effects (increase in fraud beliefs following a loss, decrease following a win and scaling of both with preference strength) and based on the sum of squared errors (transformed to RMSE in Supplementary Table 13). Since its optimized α was different from 0, we further tested whether the outcome desirability model significantly improves the fit compared with the Bayesian model, in two ways: first, we performed bootstrapping with 10,000 resamples of the follow-up sample and optimized α as explained above for each resample. We then computed the 95% CI of the bootstrapped α with the percentile method⁵⁵ to test whether the value 0, which represents the Bayesian model, is included in the CI. Second, we used a nested-model F -test to compare the explained variance of the two models.

Update of beliefs in the legitimate winner

To estimate the proportion of participants who are considered 'election-proof', we calculated for each participant the predicted updates of beliefs about V and F in the loss scenario. We then computed how much of the overall update (sum of absolute updates regarding V and F) was allocated to each of these two beliefs. We report the percentage of participants for whom more than 90% of the overall update was attributed to fraud rather than to the true vote. The estimates from the outcome desirability model and the original Bayesian model are identical. Note that the regularized prior beliefs (to 0.01 and 0.99) were used, as for all other analyses with the models.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data are publicly shared at <https://doi.org/10.5281/zenodo.5730630>, release version 5.0.0.

Code availability

Code is publicly shared (along with the data) at <https://doi.org/10.5281/zenodo.5730630>, release version 5.0.0. All analyses were performed with R version 3.6.3 (<https://www.r-project.org>). For reproducibility, we used the *checkpoint* package, which installs all needed R packages as they were on a specific date. We set the date to June 30, 2021.

References

- Hahn, U. & Harris, A. J. L. in *Psychology of Learning and Motivation* 41–102 (Elsevier, 2014).
- Tappin, B. M., van der Leer, L. & McKay, R. T. The heart trumps the head: desirability bias in political belief revision. *J. Exp. Psychol. Gen.* **146**, 1143–1149 (2017).
- Rollwage, M., Zmigrod, L., de-Wit, L., Dolan, R. J. & Fleming, S. M. What underlies political polarization? A manifesto for computational political psychology. *Trends Cogn. Sci.* **23**, 820–822 (2019).
- Van Bavel, J. J. & Pereira, A. The partisan brain: an identity-based model of political belief. *Trends Cogn. Sci.* **22**, 213–224 (2018).
- Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C. & Sternisko, A. How social media shapes polarization. *Trends Cogn. Sci.* **25**, 913–916 (2021).
- Lord, C. G., Ross, L. & Lepper, M. R. Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. *J. Pers. Soc. Psychol.* **37**, 2098–2109 (1979).
- Batson, C. D. Rational processing or rationalization? The effect of disconfirming information on a stated religious belief. *J. Pers. Soc. Psychol.* **32**, 176–184 (1975).
- Munro, G. D. & Ditto, P. H. Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Pers. Soc. Psychol. Bull.* **23**, 636–653 (1997).
- Finkel, E. J. et al. Political sectarianism in America. *Science* **370**, 533–536 (2020).
- Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. & Westwood, S. J. The origins and consequences of affective polarization in the United States. *Annu. Rev. Polit. Sci.* **22**, 129–146 (2019).
- Rutjens, B. T., Sutton, R. M. & van der Lee, R. Not all skepticism is equal: exploring the ideological antecedents of science acceptance and rejection. *Pers. Soc. Psychol. Bull.* **44**, 384–405 (2018).
- Gollwitzer, A. et al. Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nat. Hum. Behav.* **4**, 1186–1197 (2020).
- Doell, K. C., Pärnamets, P., Harris, E. A., Hackel, L. M. & Van Bavel, J. J. Understanding the effects of partisan identity on climate change. *Curr. Opin. Behav. Sci.* **42**, 54–59 (2021).

14. Kunda, Z. The case for motivated reasoning. *Psychol. Bull.* **108**, 480–498 (1990).
15. Sharot, T. & Garrett, N. Forming beliefs: why valence matters. *Trends Cogn. Sci.* **20**, 25–33 (2016).
16. Bromberg-Martin, E. S. & Sharot, T. The value of beliefs. *Neuron* **106**, 561–565 (2020).
17. Taber, C. S., Cann, D. & Kucsova, S. The motivated processing of political arguments. *Polit. Behav.* **31**, 137–155 (2009).
18. Dorfman, H. M., Bhui, R., Hughes, B. L. & Gershman, S. J. Causal inference about good and bad outcomes. *Psychol. Sci.* **30**, 516–525 (2019).
19. Gershman, S. J. How to never be wrong. *Psychon. Bull. Rev.* **26**, 13–28 (2019).
20. Jern, A., Chang, K.-M. K. & Kemp, C. Belief polarization is not always irrational. *Psychol. Rev.* **121**, 206–224 (2014).
21. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, 2014).
22. Litman, L., Robinson, J. & Abberbock, T. TurkPrime.com: a versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behav. Res. Methods* **49**, 433–442 (2017).
23. Kim, M., Park, B. & Young, L. The psychology of motivated versus rational impression updating. *Trends Cogn. Sci.* **24**, 101–111 (2020).
24. Bhui, R. & Gershman, S. J. Paradoxical effects of persuasive messages. *Decision* **7**, 239–258 (2020).
25. Eil, D. & Rao, J. M. The good news–bad news effect: asymmetric processing of objective information about yourself. *Am. Econ. J. Microecon.* **3**, 114–138 (2011).
26. Mobius, M. M., Niederle, M., Niehaus, P. & Rosenblat, T. S. Managing self-confidence: theory and experimental evidence. *Management Science* **68**, 7793–7817 (2022).
27. Levitsky, S. & Ziblatt, D. *How Democracies Die* (Crown Publishing Group, 2018).
28. Berlinski, N. et al. The effects of unsubstantiated claims of voter fraud on confidence in elections. *J. Exp. Polit. Sci.* 1–16 (2021); <https://doi.org/10.1017/XPS.2021.18>
29. Clayton, K. et al. Elite rhetoric can undermine democratic norms. *Proc. Natl Acad. Sci. USA* **118**, e2024125118 (2021).
30. Enders, A. M. et al. The 2020 presidential election and beliefs about fraud: continuity or change? *Elect. Stud.* **72**, 102366 (2021).
31. Lewandowsky, S. et al. *The Debunking Handbook 2020* (2020); <https://doi.org/10.17910/b7.1182>
32. Hyman, I. E. Jr & Jalbert, M. C. Misinformation and worldviews in the post-truth information age: commentary on Lewandowsky, Ecker, and Cook. *J. Appl. Res. Mem. Cogn.* **6**, 377–381 (2017).
33. Nyhan, B. & Reifler, J. Displacing misinformation about events: an experimental test of causal corrections. *J. Exp. Polit. Sci.* **2**, 81–93 (2015).
34. Benkler, Y. et al. *Mail-in Voter Fraud: Anatomy of a Disinformation Campaign*. Berkman Center Research Publication (2020); <https://doi.org/10.2139/ssrn.3703701>
35. Grant, M. D., Flores, A., Pedersen, E. J., Sherman, D. K. & Van Boven, L. When election expectations fail: polarized perceptions of election legitimacy increase with accumulating evidence of election outcomes and with polarized media. *PLoS ONE* **16**, e0259473 (2021).
36. Schaffner, B. F. & Luks, S. Misinformation or expressive responding? *Public Opin. Q.* **82**, 135–147 (2018).
37. Graham, M. H. & Yair, O. *Expressive Responding and Trump's Big Lie* [Paper presentation]. Midwest Political Science Association Annual Meeting (2022); https://m-graham.com/papers/GrahamYair_BigLie.pdf
38. Berinsky, A. J. Telling the truth about believing the lies? Evidence for the limited prevalence of expressive survey responding. *J. Polit.* **80**, 211–224 (2018).
39. Bakshy, E., Messing, S. & Adamic, L. A. Political science. Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).
40. Flaxman, S., Goel, S. & Rao, J. M. Filter bubbles, echo chambers, and online news consumption. *Public Opin. Q.* **80**, 298–320 (2016).
41. Charpentier, C. J., Bromberg-Martin, E. S. & Sharot, T. Valuation of knowledge and ignorance in mesolimbic reward circuitry. *Proc. Natl Acad. Sci. USA* **115**, E7255–E7264 (2018).
42. Iyengar, S. & Hahn, K. S. Red media, blue media: evidence of ideological selectivity in media use. *J. Commun.* **59**, 19–39 (2009).
43. Sharot, T. & Sunstein, C. R. How people decide what they want to know. *Nat. Hum. Behav.* **4**, 14–19 (2020).
44. Pereira, A., Harris, E. & Van Bavel, J. J. Identity concerns drive belief: the impact of partisan identity on the belief and dissemination of true and false news. *Group Process. Intergr. Relat.* **26**, 24–47 (2021).
45. Vlasceanu, M. & Coman, A. The impact of social norms on health-related belief update. *Appl. Psychol. Health Well Being* **14**, 453–464 (2022).
46. Jerit, J. & Barabas, J. Partisan perceptual bias and the information environment. *J. Polit.* **74**, 672–684 (2012).
47. Beck, J. S. & Beck, A. T. *Cognitive Behavior Therapy* (Guilford Press, 2020).
48. Johnson, N. F. et al. The online competition between pro- and anti-vaccination views. *Nature* **582**, 230–233 (2020).
49. Druckman, J. N. & McGrath, M. C. The evidence for motivated reasoning in climate change preference formation. *Nat. Clim. Change* **9**, 111–119 (2019).
50. Bertoldo, R. et al. Scientific truth or debate: on the link between perceived scientific consensus and belief in anthropogenic climate change. *Public Underst. Sci.* **28**, 778–796 (2019).
51. Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* **41**, 1149–1160 (2009).
52. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Academic Press, 2013).
53. Bullock, J. G. & Lenz, G. Partisan bias in surveys. *Annu. Rev. Polit. Sci.* **22**, 325–342 (2019).
54. Prior, M., Sood, G. & Khanna, K. You cannot be serious: the impact of accuracy incentives on partisan bias in reports of economic perceptions. *Quart. J. Polit. Sci.* **10**, 489–518 (2015).
55. Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans* (SIAM, 1982).

Acknowledgements

We thank J. M. Carey, R. Muirhead and B. J. Nyhan for providing comments on a previous version of this manuscript. R.B.-N. is an Awardee of the Weizmann Institute of Science—Israel National Postdoctoral Award Program for Advancing Women in Science. M.J. was supported by a National Science Foundation (NSF) grant (2020906). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

Study design: R.B.-N. and T.D.W. Data collection: R.B.-N. Data analysis: R.B.-N. Supervision, expertise and feedback: M.J. and T.D.W. Bayesian model formulation: M.J. Conceptual framework development: R.B.-N., M.J. and T.D.W. Writing: R.B.-N., M.J. and T.D.W.

Competing interests

This research was conducted in part while M.J. was a Visiting Faculty Researcher at Google Research, Brain Team (Mountain View, CA, USA). This work was not part of a commercial project. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41562-023-01570-4>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41562-023-01570-4>.

Correspondence and requests for materials should be addressed to Rotem Botvinik-Nezer, Matt Jones or Tor D. Wager.

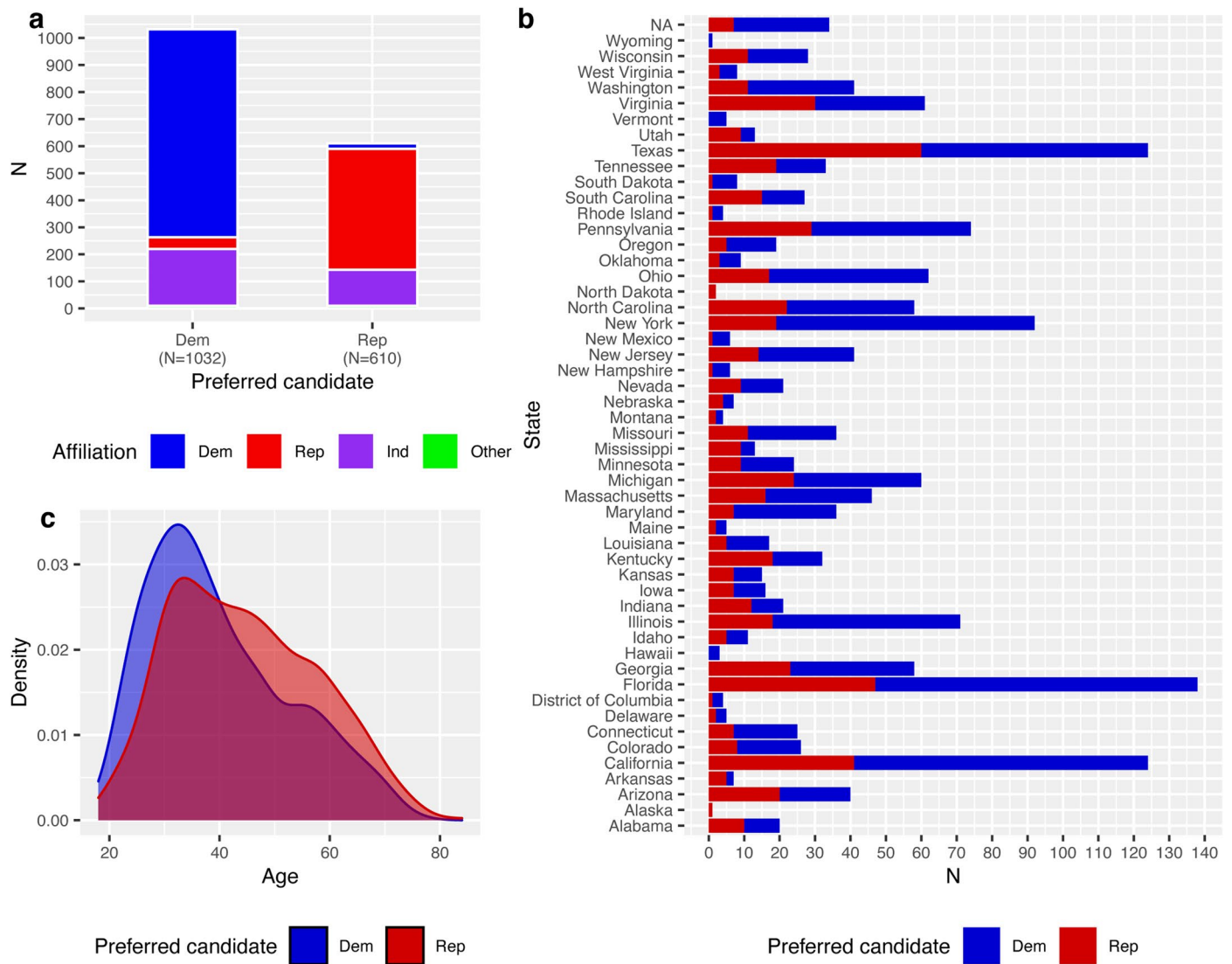
Peer review information *Nature Human Behaviour* thanks Eric Groenendyk and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

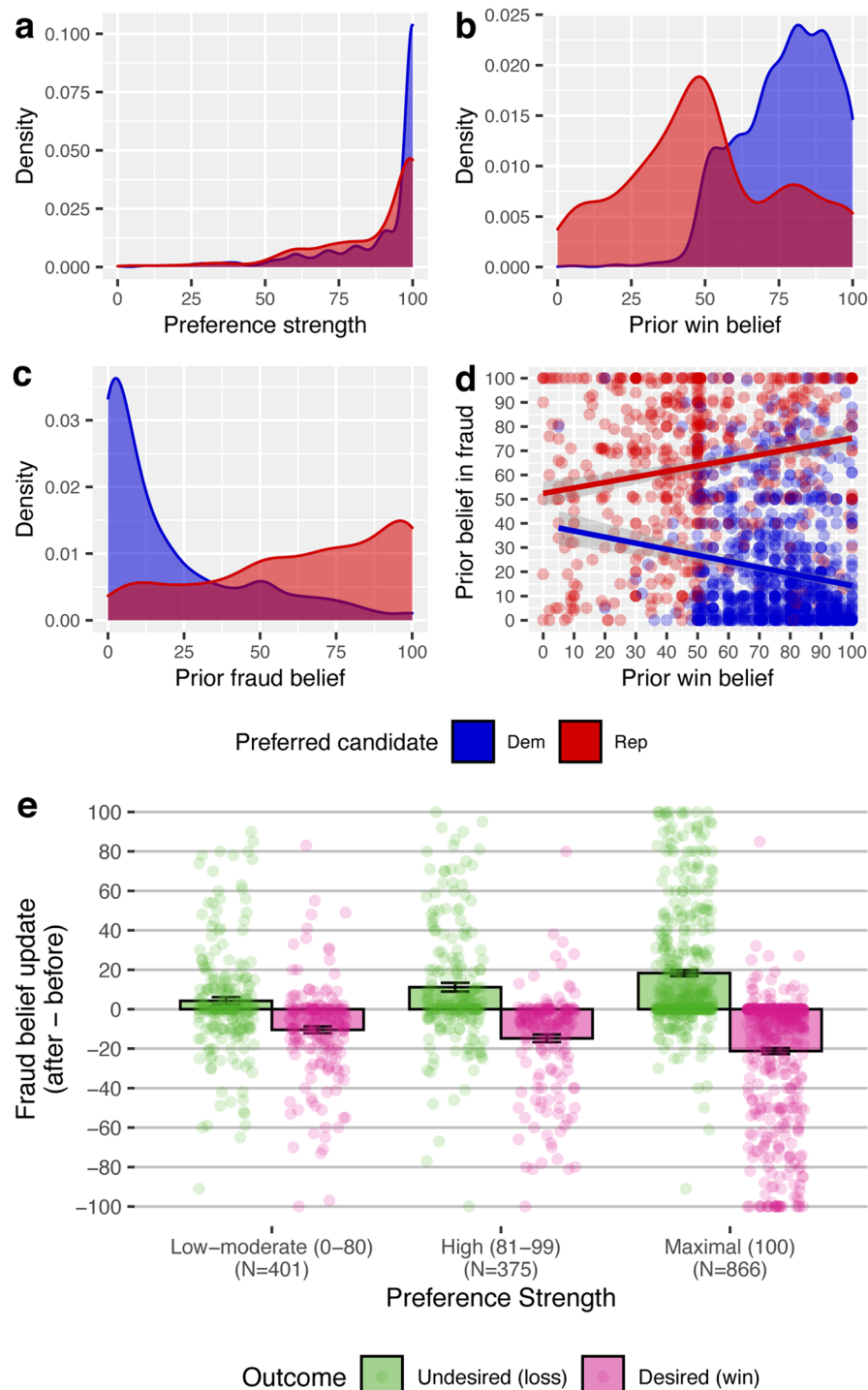
Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023



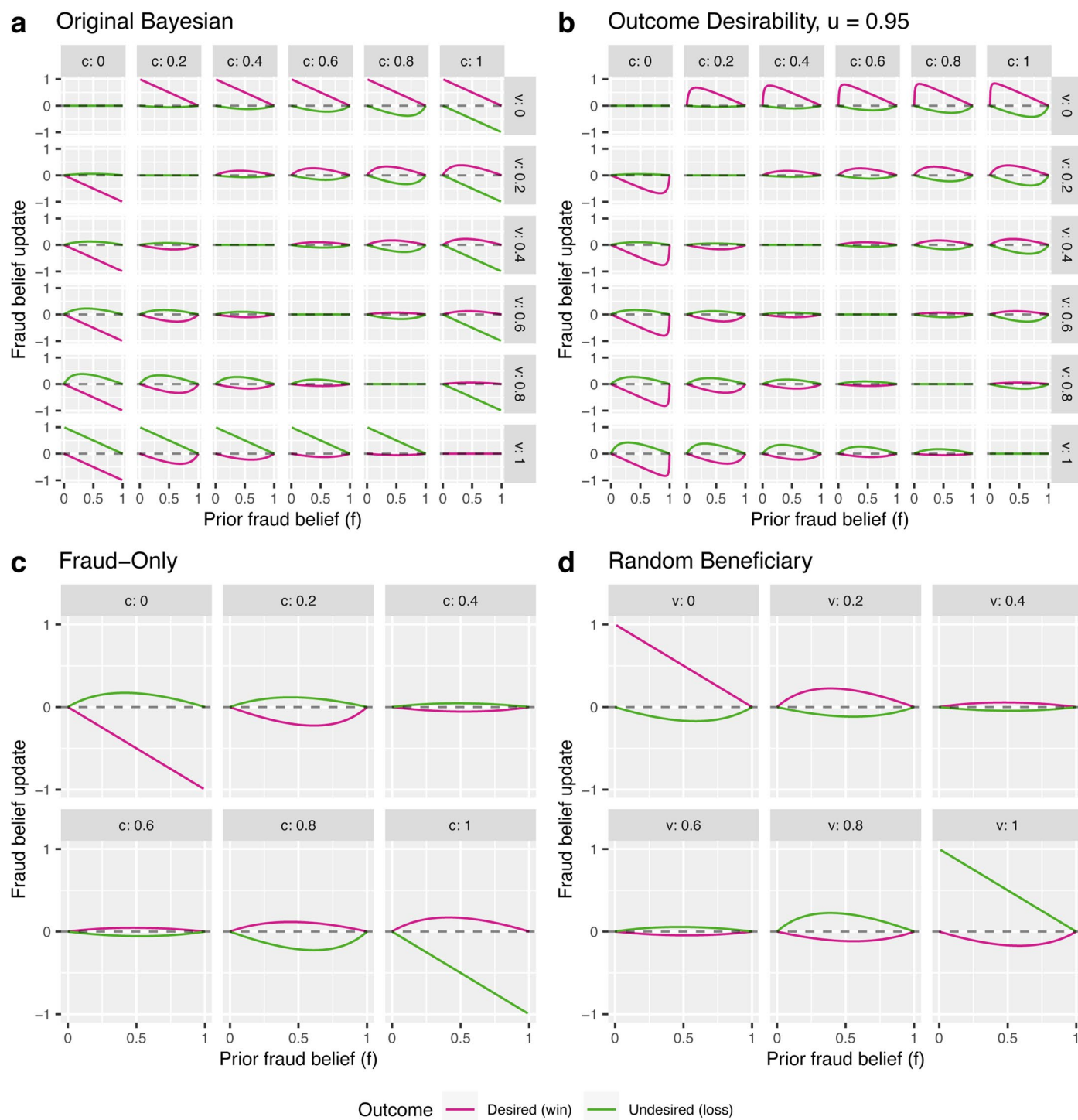
Extended Data Fig. 1 | Demographic and partisan affiliation of participants. (a) Number of participants for each combination of preferred candidate (x axis) and political affiliation (color). (b) Number of participants from each state who preferred each candidate (color). (c) Kernel density plot of participants' age as a

function of the preferred candidate (color). Preferred candidate: Dem = Biden; Rep = Trump; Affiliation: Dem = Democrat; Rep = Republican; Ind = Independent; Other.



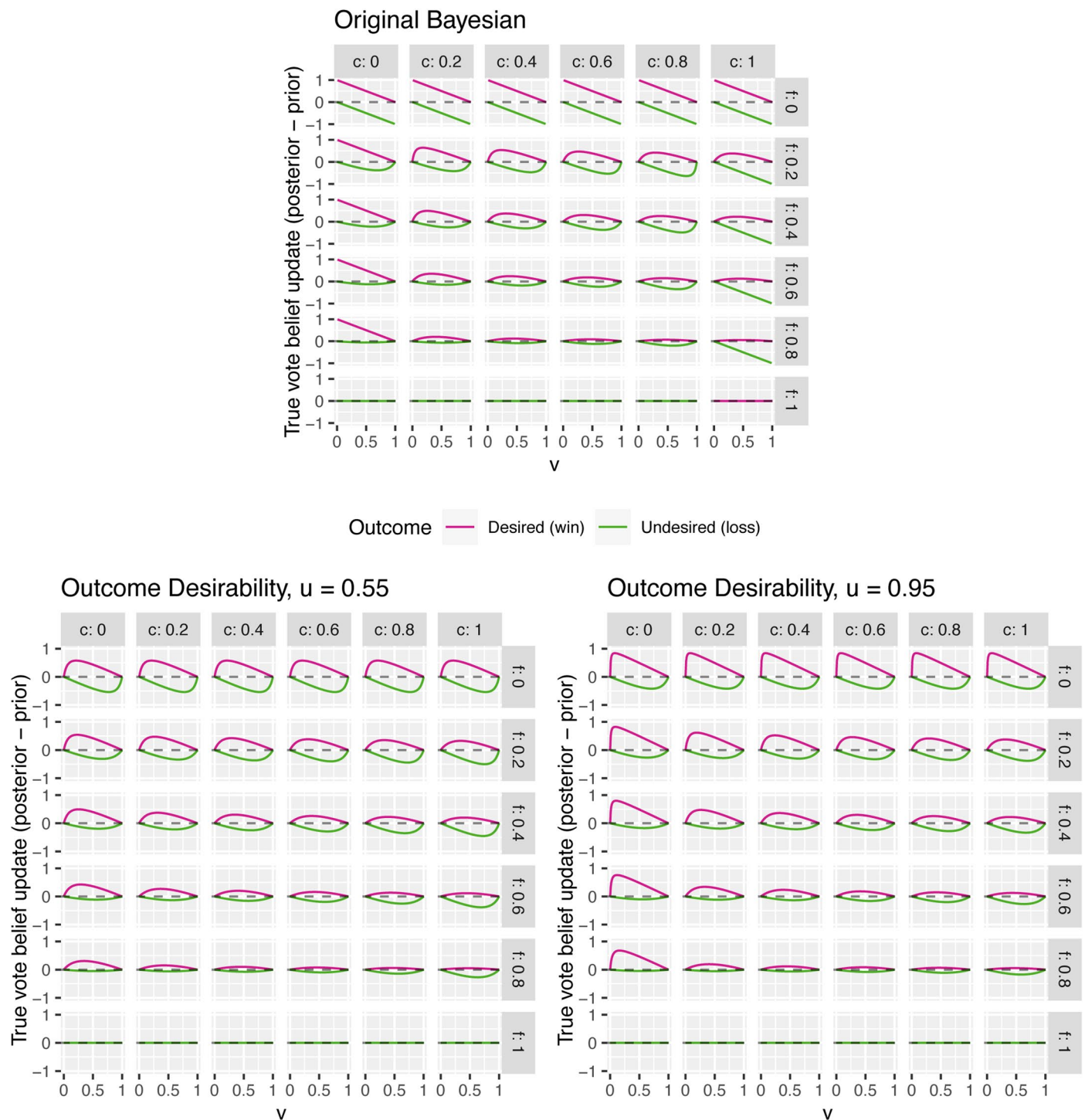
Extended Data Fig. 2 | Preference and prior belief data. Kernel density plots of: (a) preference strength; (b) prior subjective probability of win by the preferred candidate; and (c) prior fraud belief. (d) Scatterplot for prior fraud belief as a function of prior subjective probability of win by the preferred candidate. Gray area represents 95% confidence interval; each point represents a single participant (with 30% opacity). Color represents the preferred candidate:

Rep = Trump (red); Dem = Biden (blue). (e) Fraud belief update as a function of the scenario (loss or win according to the hypothetical map) and the preference strength (categorized into three ordered categories). Points represent single participants, and error bars represent standard error of the mean across participants.



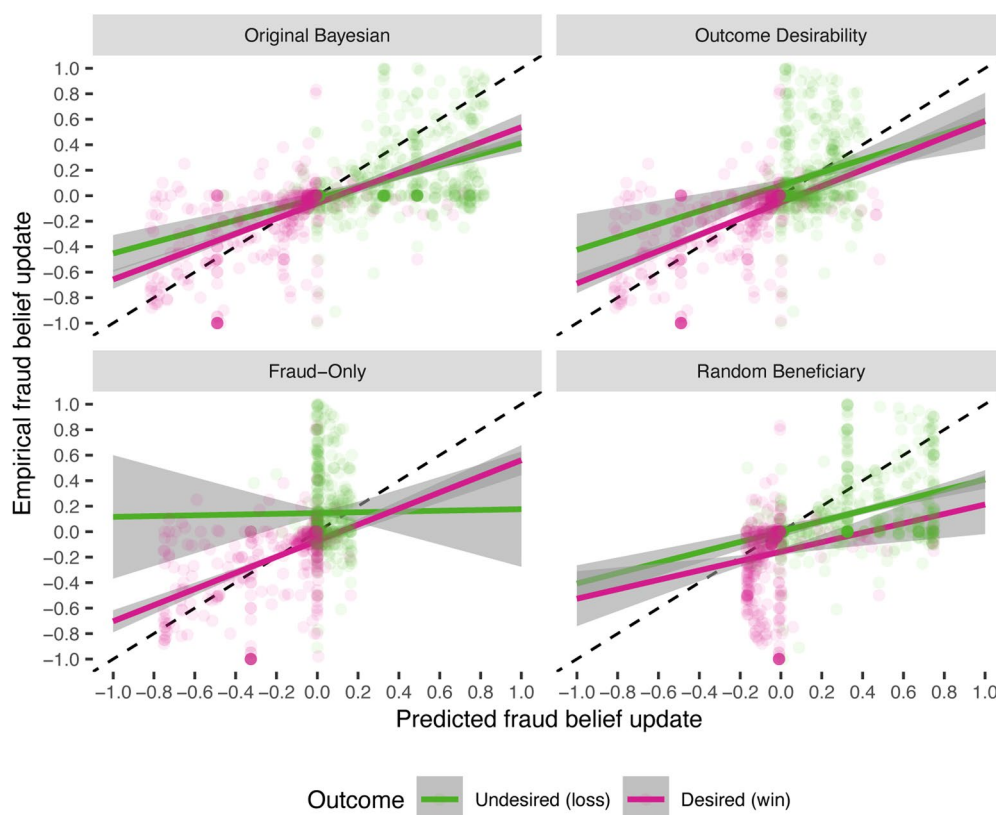
Extended Data Fig. 3 | Simulations of predicted fraud update across the prior belief space for all models. Predictions are shown for each of the four models: (a) The original Bayesian model (note that the plot for the Bayesian model is also the plot for the Hypothesis Desirability model, because fitting the latter to the

data yields $\alpha = 0$); (b) the Outcome Desirability model, with predictions based on the mean preference strength in the sample ($u = .95$) and the strength of bias that best fits the data ($\alpha = .3$); (c) the Fraud-Only model; and (d) the Random Beneficiary model.



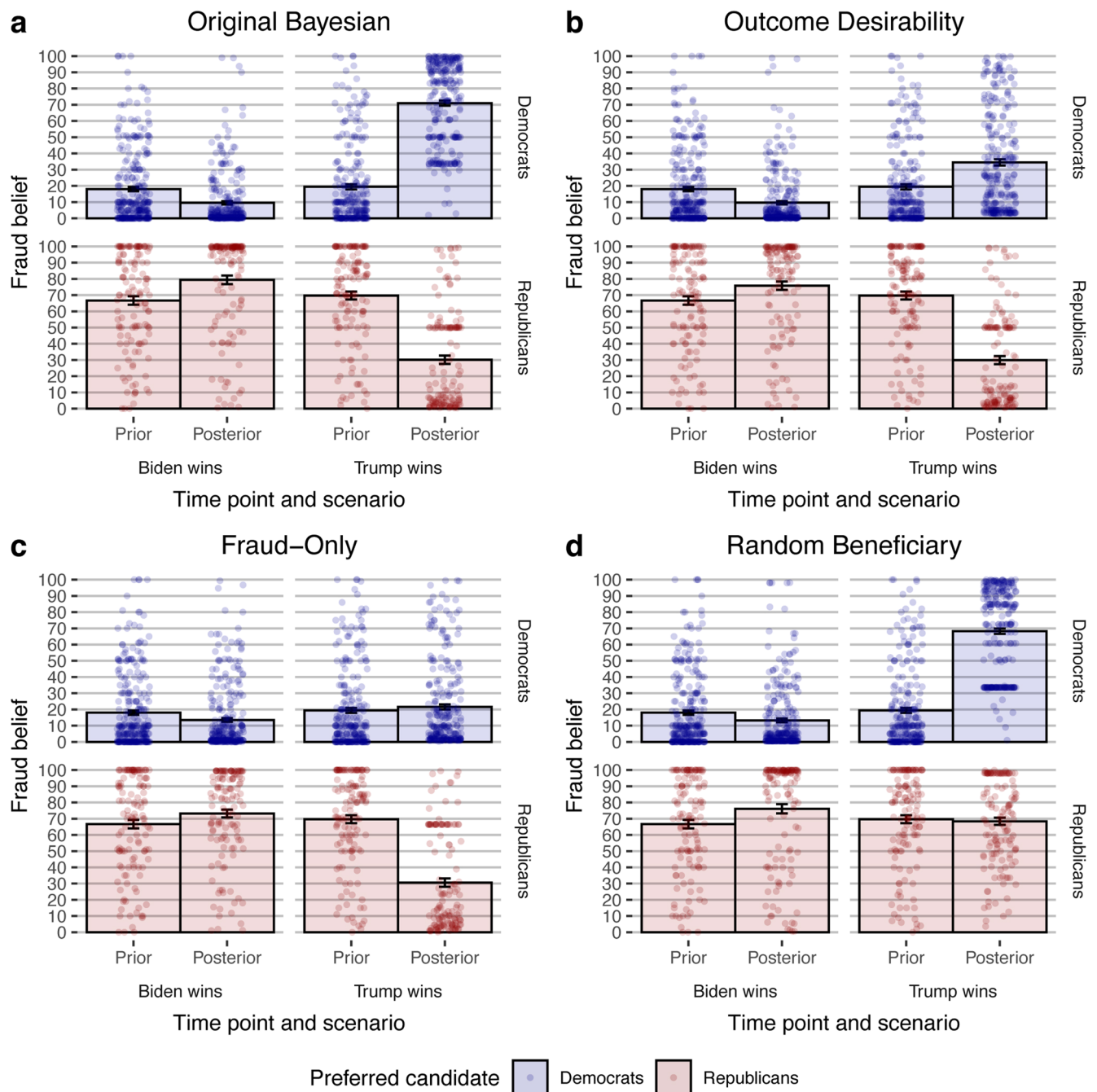
Extended Data Fig. 4 | Update of true vote belief. Simulations of predicted true vote belief update across the prior belief space for the Bayesian model and the Outcome Desirability model. Outcome Desirability model predictions assume

the strength of bias that best fits the data ($\alpha = .3$) and are shown for the mean preference strength in the sample ($u = .95$) as well as a lower value ($u = .55$) for comparison.



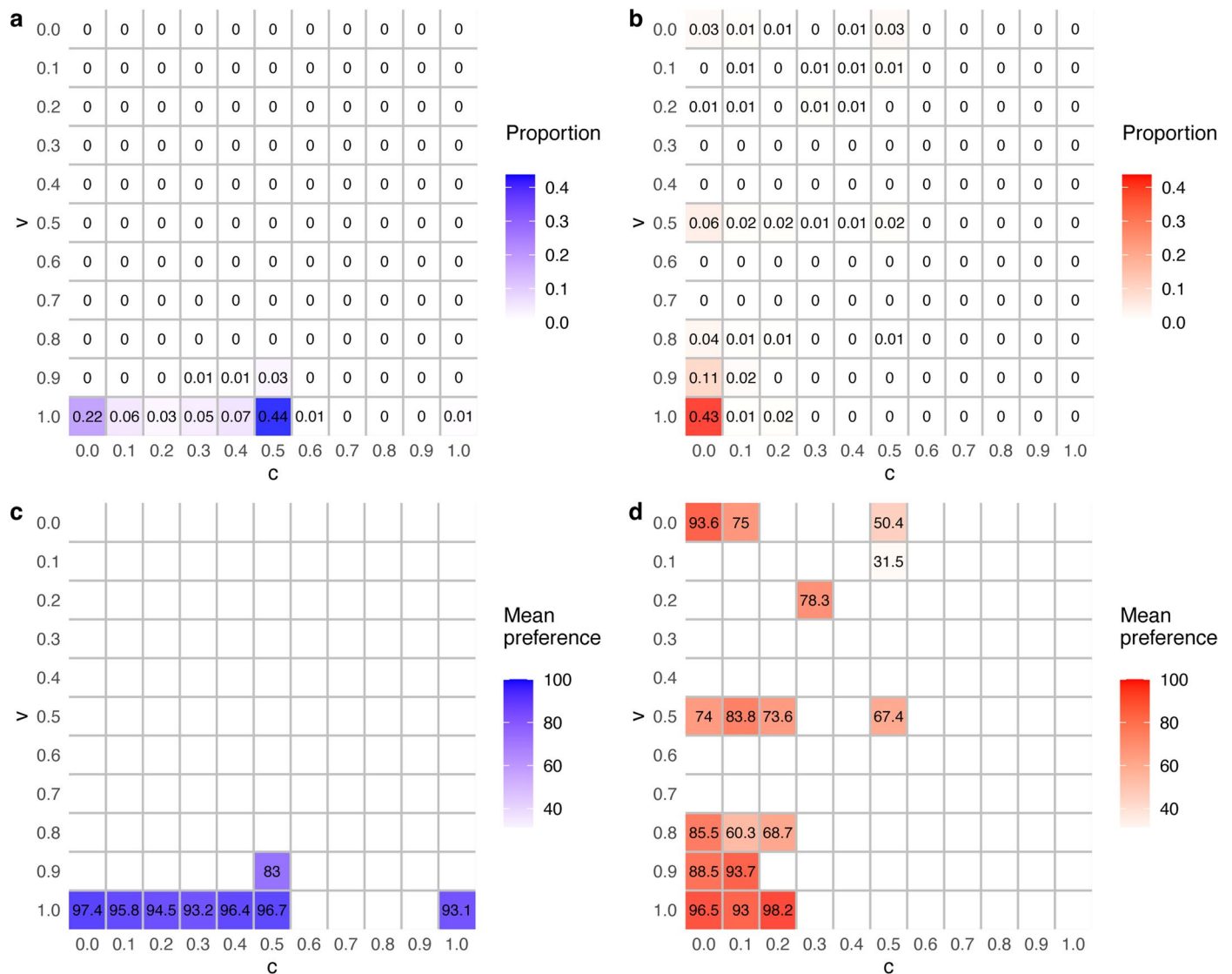
Extended Data Fig. 5 | Empirical and predicted fraud belief updates across models. Empirical fraud belief updates as a function of the predicted fraud belief updates from each model (based on empirically measured priors). Each point represents a single participant. The 95% confidence regions are for the linear

approximation of the regression of empirical upon predicted fraud belief updates (the lines have slopes less than unity in part because of regression effects). Note that the plot for the Bayesian model is also the plot for the Hypothesis Desirability model, because fitting the latter to the data yields $\alpha = 0$.



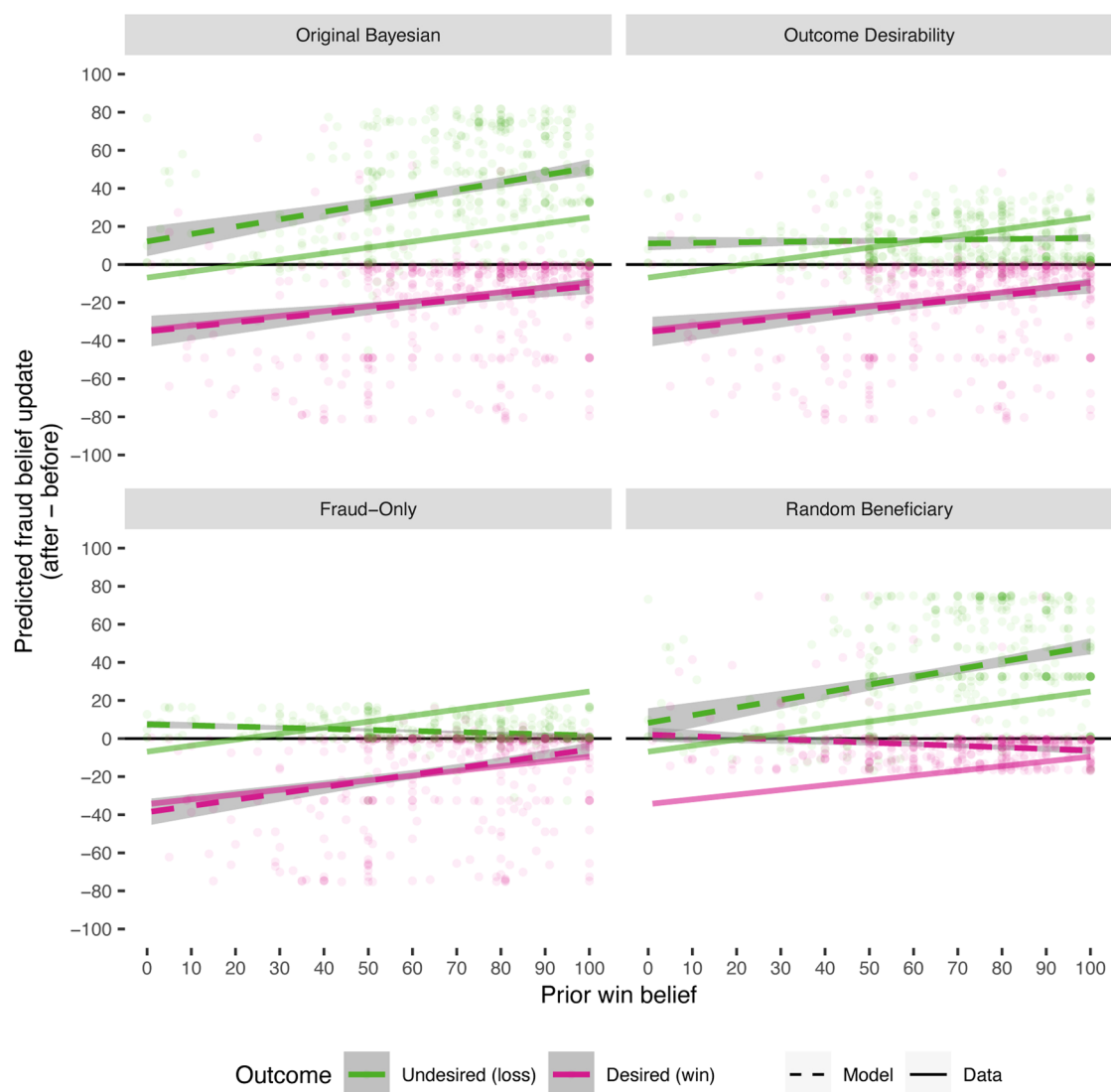
Extended Data Fig. 6 | Predicted posterior fraud belief for each scenario across models. Paralleling the empirical results from Fig. 1c, the model-based predictions of the posterior fraud belief for each participant ($N = 828$) are presented as a function of the map scenario (hypothetical winner), for each of the four models: (a) The original Bayesian model (note that the plot for the Bayesian model is also the plot for the Hypothesis Desirability model, because fitting the

latter to the data yields $\alpha = 0$); (b) the Outcome Desirability model; (c) the Fraud-Only model; and (d) the Random Beneficiary model. In all panels, the prior values are based on the empirically obtained prior fraud beliefs from the original survey, only for participants who completed the follow-up survey. Points represent single participants, and error bars represent standard error of the mean across participants.



Extended Data Fig. 7 | Empirically measured priors from the follow-up survey. (a–b) Heat maps for the proportion of participants with each combination of ν (probability of the preferred candidate winning the true votes) and c (probability of fraud, if present, favoring the preferred candidate), based on the follow-up

sample, for (a) Democratic participants and (b) Republican participants. (c–d) Heat map of the mean preference strength of participants for each combination of ν and c , for (c) Democratic participants and (d) Republican participants. For panels C and D, mean preference is shown only for cells with $>1\%$ of participants.



Extended Data Fig. 8 | Predicted fraud belief update as a function of prior win belief across models. Paralleling the empirical results from Fig. 1e, the model-based predictions of the fraud belief update for each participant are presented as a function of the empirically measured prior probability of the preferred candidate's win. Dashed lines show linear fits to models' predictions,

with 95% confidence regions. For comparison, solid lines show the linear fits to the observed empirical patterns (Fig. 1e). Note that the plot for the Bayesian model is also the plot for the Hypothesis Desirability model, because fitting the latter to the data yields $\alpha = 0$.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection No code was used in data collection

Data analysis Analysis code are publicly available at <https://doi.org/10.5281/zenodo.5730630> (release v5.0.0), and were run with R version 3.6.3.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All data are publicly available at <https://doi.org/10.5281/zenodo.5730630> (release v5.0.0).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	A quantitative experimental design
Research sample	Online sample of N = 1760 (1642 after exclusions, see below) American citizens / permanent residents, ages 18-84 years, at least one participant from each state. Since the beliefs tested were related to the U.S. election, the rationale was to obtain a large sample across the U.S. We did not attempt to assess average beliefs or belief updates across the population (like election polls often do), and thus did not aim for a strictly representative sample of the U.S. population.
Sampling strategy	Online convenience sample via Cloud Research. Sample size was chosen based on available funds aiming for a large sample across the U.S.
Data collection	Online sample collected from Amazon's Mechanical Turk (MTurk) via the Cloud Research platform. Participants completed the experiment via their computer, on the Qualtrics platform. Since this is an online experiment, the researcher was not present during the experiment (i.e., blindness of the researcher to the experimental condition is not applicable here).
Timing	Original sample: November 4-5, 2020. Follow-up sample (sub-sample of the same participants): January 19-20, 2021.
Data exclusions	Thirty-two participants were excluded due to failure to correctly answer a simple attention check included in the survey and one due to not meeting the participation requirement of American citizenship or permanent residency. Eighty-five additional participants were excluded from analyses because they chose "Other" for their preferred Presidential candidate (and not "Donald Trump" or "Joe Biden").
Non-participation	18 participants started the online survey and provided informed consent, but did not complete the survey. Since participants could choose to stop responding at any point without providing explanation, we do not know the reasons for their non-participation.
Randomization	Participants were randomly assigned to one of two experimental conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Participants' age ranged from 18-84 years, with 1-138 participants per state. We did not collect other demographic information such as gender, race, or ethnicity.
Recruitment	Participants were recruited online via Cloud Research (on Amazon's Mechanical Turk). The user pool of MTurk is known to be more liberal, young and educated compared to the general population. These biases are of limited consequence in our study, which is not trying to estimate the beliefs of a larger population, but rather utilizes an experimental manipulation to examine how people update their beliefs given new evidence when multiple causal explanations are competing to explain the

evidence. Moreover, previous studies have shown that results tend to replicate well between representative and MTurk samples. In addition, democratic and older participants from the original sample were more likely to complete the follow-up survey. However, sensitivity analyses revealed that our findings were robust to the sample characteristics available in our data.

Ethics oversight

Dartmouth College's Committee for the Protection of Human Subjects.

Note that full information on the approval of the study protocol must also be provided in the manuscript.