Bayesianism and Wishful Thinking are Compatible

David E. Melnikoff[1,*] **and Nina Strohminger**[2,3]

[1]Department of Psychology, Northeastern University
[2]Department of Legal Studies and Business Ethics, The Wharton School, University of Pennsylvania
[3]Department of Psychology, University of Pennsylvania
[*]davidemelnikoff@gmail.com

ABSTRACT

Bayesian principles show up across many domains of human cognition, but wishful thinking—where beliefs are updated in the direction of desired outcomes rather than what the evidence implies—seems to threaten the universality of Bayesian approaches to the mind. In this paper, we show that Bayesian optimality and wishful thinking are, despite first appearances, compatible. The setting of opposing goals can cause two groups of people with identical prior beliefs to reach opposite conclusions about the same evidence through fully Bayesian calculations. We show that this is possible because, when people set goals, they receive privileged information in the form of affective experiences, and this information systematically supports goal-consistent conclusions. We ground this idea in a formal, Bayesian model in which affective prediction errors drive wishful thinking. We obtain empirical support for our model across four studies.

Introduction

Among the most influential ideas in the cognitive sciences is that the mind emerges from approximations to normatively rational Bayesian calculations[1–4]. This idea, commonly called the "Bayesian brain hypothesis," has been applied successfully to a wide range of psychological processes, from low-level perception to categorization, reasoning, and emotion[5–9]. However, some phenomena appear deeply incompatible with Bayesian principles, raising the question of how fundamental these principles really are to mental processing.

Arguably the greatest threat to the Bayesian brain hypothesis is the phenomenon of *belief polarization*[10–12], which occurs when two people update their beliefs in opposing directions on the basis of the same information[13]. For example, someone may conclude that their preferred political candidate has an even greater chance of winning after seeing the latest polls, whereas a detractor may conclude from the same data that the candidate's odds have gone down even further[14]. Because Bayesian norms demand that beliefs be updated in the direction of the evidence, it seems that belief polarization cannot be the result of Bayesian inference. Instead, belief polarization appears to emerge from "wishful thinking," defined as belief updating that is driven by the directional motivation to reach a particular conclusion (as opposed to the non-directional motivation to be accurate)[15–17]. On this account, when people who encounter the same data shift their beliefs in opposing directions, they do so because they want to reach opposing conclusions, not because of any rational Bayesian calculus.

Defenders of the Bayesian brain hypothesis are acutely aware of the existential threat posed by belief polarization. In response, they have sought to establish that belief polarization is not driven by wishful thinking, and that sufficiently sophisticated Bayesian models can explain it without appealing to directional motivation. They have taken an *eliminative* stance toward wishful thinking, pitting it against Bayesian accounts of belief polarization in a zero-sum game in which the latter must emerge victorious for the Bayesian brain hypothesis to survive.

The eliminative approach has had some success. Bayesian models have proven capable of reproducing many examples of belief polarization in the absence of wishful thinking. But as we will see, the eliminative approach cannot save Bayesianism on its own.

To counter claims that wishful thinking drives belief polarization, eliminativists have appealed to the fact that belief polarization can be Bayes-rational, and not directionally motivated, when it occurs among people with different prior beliefs that bear on the meaning of the evidence[11,18–20]. Consider again the supporter and detractor whose beliefs about a candidate's odds of winning diverged on the basis of the same poll. These partisans were not randomly assigned to their positions; before encountering the poll they likely had different prior beliefs. The detractor may have believed that polls systematically overestimate support for the candidate in question, and the supporter may have believed the opposite. In this scenario, a neutral polling result would give the detractor reason for pessimism about the candidate's prospects, and the supporter reason for optimism. Thus, Bayesian

inference could lead the supporter and detractor to draw opposing conclusions from the same poll in the absence of directional motivation. This principle is applicable to the majority of demonstrations of belief polarization, especially within the motivated reasoning literature, where the typical study design involves presenting evidence to two groups of people with different prior beliefs[11,13,15].

The problem for the eliminativist project is that belief polarization occurs even when priors are held constant. In one set of studies, participants were randomly assigned to prosecute or defend a person in a mock criminal trial in order to achieve the goal of earning money[21]. Participants did not advocate, nor did they agree to advocate—they merely learned that they could earn money by doing so. Despite receiving identical evidence, participants assigned to defend came to believe that the defendant was less likely to be guilty, and participants assigned to prosecute came to believe that the defendant was more likely to be guilty.

This result does not depend on people possessing different priors—priors were held constant through random assignment. Rather, this is a form of belief polarization driven by the motivation to defend opposing positions—a clear case of *wishful belief polarization*, distinct from belief polarization that is attributable to group differences in prior beliefs rather directional motivation.

Wishful belief polarization represents a huge and unignorable challenge to claims that the mind is fully pervaded by Bayesian reasoning. We take up this challenge in the current paper by introducing a novel approach to reconciling Bayesianism with belief polarization. Ours is a compatibilist approach: Rather than pitting Bayesian inference and wishful thinking against each other, we contend that the two can coexist. We defend this claim by developing and validating a fully Bayesian model of wishful belief polarization—one in which Bayesian calculations implement, rather than replace, a causal path from directional motivation to diverging beliefs. In so doing, we reconcile one of the most influential accounts of human mental function with one of its greatest empirical threats.

Reconciling Bayesianism with Wishful Belief Polarization

In the aforementioned study, in which goal attainment required defending one of two randomly assigned positions, the experimenters provided all participants with identical information about the trial. But maybe the information provided by experimenters was not the only information participants used—maybe participants had access to some less obvious, "hidden" information.

Past research indicates that the brain takes a rather ecumenical approach to what counts as information. From the brain's perspective, information is not limited to the sort of things that would be presented in a court of law (e.g. finger prints recovered from a crime scene) but includes internal evidence, such as emotions, mood, and other features of affective processing. Affect is frequently used to build out a full model of "what one knows" about a situation, and is used to shape judgments, decisions, and other cognitive processes[9,22–25]. If I feel nostalgia while recalling an ambiguous childhood memory, I may infer this was a good experience; if I feel uneasy about a business deal, I may infer something about it must be amiss. The major point to take away from this literature is that people treat affect as a form of information. Affective information, therefore, is potentially used along with other information to perform Bayesian belief updating[9,26–29].

The novel contribution of this paper is the idea that affective information could be the key to reconciling the Bayesian brain hypothesis with wishful belief polarization. We formalize this idea in a fully Bayesian model of wishful belief polarization (see Methods), which we interrogate across four empirical studies.

From Affective Prediction Errors to Wishful Belief Polarization

How, exactly, is affect used to update beliefs? The Bayesian brain hypothesis makes specific and unique predictions, which are instantiated in our model. One prediction is that people should not update their beliefs based on affect per se. Instead, people should update their beliefs based on *affective prediction errors*: how much better or worse they feel relative to expectations[25,30]. The logic behind this is straightforward. When a person observes exactly what they expect to observe, this implies that the beliefs underlying that person's expectations are correct and should be maintained. Conversely, when expectations are violated, this suggests that the beliefs underlying those expectations are wrong and should be updated. A large body of work from computational neuroscience shows that the brain constantly predicts upcoming sensory events, including affective experiences, and uses deviations from these predictions to update its model of the external world[4,7,9,26]. This process unfolds automatically, regardless of explicit intentions.

Suppose a person feels better than expected about the prospect of defending a position. Under a Bayesian model, this positive affective prediction error should lead the person to update their belief about the to-be-defended position. It should lead the person to infer that the to-be-defended position is more valid than initially thought, since people typically believe that the more valid a position is, the better they will feel about defending it[31,32].

Informally, the line of reasoning would look something like this: "The more valid a position is, the better I feel about defending it; I feel better than expected about the prospect of defending this position; therefore, this position is more valid than I initially thought." (For the formal treatment, see Methods.) Now suppose that someone feels worse than expected about defending a position—they encode a negative affective prediction error. This person may use the same Bayesian reasoning process infer that the to-be-defended position is less valid than initially thought. Should a person feel exactly as they expected to feel about defending a position, the absence of an affective prediction error should, under a Bayesian model, lead to an absence of belief updating.

Affective prediction errors play a central role in our Bayesian model of wishful belief polarization. To see how our model works, imagine someone is given an incentive to defend a position—for instance, they will earn money by doing so. In this situation, at least two changes would take place. For one, the person would come to feel better about the prospect of defending the position. By making the act of defending a position instrumental to a goal attainment, the valence of the act itself should become more positive[33–35].

But the person would not merely feel better about the prospect of defending the position—they would feel better than expected.[†] People systematically underestimate how much their feelings will shift as a result of changing incentives due to a general tendency to overestimate the degree to which future affective states will resemble current affective states ("projection bias")[36,37]. For example, people who have just eaten tend to underestimate how much they will enjoy food when they are hungry again[38]. This suggests that people will underestimate the degree to which increasing the goal-conduciveness of defending a position will increase the subjective pleasantness of doing so, resulting in positive affective prediction errors. Because our model stipulates that the subjective validity of a position increases whenever people feel better than expected about the prospect of defending it, such prediction errors should increase the subjective validity of the to-be-defended position.

Our model potentially explains how the directional motivation to defend opposing positions could lead two people to update their beliefs in opposing directions despite receiving the same external evidence and holding identical priors. Consider two people, one who is incentivized to defend capital punishment, and one who is incentivized to oppose capital punishment. These incentives lead both people to feel better about the prospect of defending their respective positions than predicted. In response to these affective prediction errors, Bayesian calculations lead the pro-capital punishment individual to assign greater validity to capital punishment, and the anti-capital punishment individual to assign less validity to capital punishment. This pattern of polarization is Bayesian and, at the same time, directionally motivated: the affective prediction errors that drove it reflect changes in people's desire to defend specific positions.

To summarize, we contend that when people are incentivized to defend a position, they encode positive affective prediction errors, a kind of "hidden" information that supports the position's validity. This information is used to update beliefs in a normatively rational Bayesian fashion, and can account for a phenomenon—wishful belief polarization—long touted as a threat to the Bayesian brain hypothesis. Three hypotheses follow from our model, which we describe below, and then test across four studies.

Hypotheses

One hypothesis of our model is the *underestimation hypothesis*: Making the act of defending a position goal-conducive systematically induces a positive affective prediction error. That is, the prospect of defending the position feels better than predicted.

The second hypothesis is the *error-based updating hypothesis*: People update their beliefs in proportion to the size of the affective prediction error. This is important for distinguishing our model from non-Bayesian models. For example, a non-Bayesian model may predict that, regardless of how people expect to feel about the prospect of defending a position, when they feel good they assign greater validity to the position, and when they feel bad they assign less validity to that position. So, if people update their beliefs based only on how they feel, rather than how they feel relative to expectations, this would rule out a Bayesian account.

The third hypothesis is also important for distinguishing Bayesian from non-Bayesian inference. Bayesian inference stipulates that belief updating is inversely proportional to subjective observation noise—a person's belief about the inherent randomness, or unpredictability, of the observations they use to update their beliefs[30]. Suppose that someone thinks their own affective experiences are very noisy and therefore inherently unpredictable. According to Bayes' rule, this person should not update their beliefs very much in response to affective prediction errors; if affect is inherently noisy, then an affective prediction error may not be attributable to an erroneous belief—it may instead be attributable to random noise. Conversely, suppose that someone thinks their own affective experiences

[†]While on the whole most people make this prediction error, not everyone will do so every time. This is an important detail to which we will soon return.

are not noisy at all and therefore highly predictable. This person, according to Bayes' rule, should update their beliefs substantially in response to affective prediction errors. If affect is highly predictable, then affective prediction errors are unlikely to result from random noise. Instead, they probably result from erroneous beliefs that ought to be changed. In short, Bayesian principles dictate that the noisier a person thinks their affective experiences are, the less they should update their beliefs in response to affective prediction errors (see Methods). This is the *noise hypothesis*. Evidence for the noise hypothesis would distinguish our model from alternatives in which belief updating is error-based but non-Bayesian, such as Rescorla-Wagner[39] and other delta-rule models[25,40–42].

Study 1

According to the underestimation hypothesis, people tend to underestimate how good they would feel about the prospect of defending a position if they were incentivized to do so. This hypothesis was supported by the results of Study 1.

To begin Study 1, participants were informed that, later in the experiment, they would be incentivized to defend either the plaintiff or defendant in a mock trial. Participants did not learn *which* side they would be incentivized to defend—only that, if they secured a victory for their client, they would win a cash prize. Next, participants read one of six randomly selected court briefs based on actual cases from the United States legal system (for full briefs, see Supporting Information). Participants predicted (i) how good they would feel about defending the plaintiff's side and (ii) how good they would feel about defending the defendant's side. Immediately after, we revealed to participants which side they needed to defend in order to win the cash prize. Participants then reported how good they actually felt about the prospect of defending their assigned position.

The only relevant change that occurred between the reporting of predicted and actual affect was the degree of incentive to defend the assigned positions. This feature of the study design is crucial; it ensures that any affective prediction errors reflect a change in motivation to defend the assigned position, and are therefore are capable, in principle, of driving directionally motivated belief change.

The underestimation hypothesis says that regardless of which position they are incentivized to defend, participants should feel better about doing so than predicted. This is what we found (Figure 1). On average, affective prediction errors (actual affect minus predicted affect) were significantly greater than zero for all 12 combinations of legal case and assigned position.

The six cases that made up our stimulus set covered a diverse range of issues. Among them were a free speech case involving a lawsuit brought by a neo-Nazi group alleging a First Amendment violation after their pro-Nazi demonstration was blocked; a healthcare case involving a lawsuit brought against the United States aiming to repeal the Affordable Care Act; a custody case involving a lawsuit brought against the mother of an 8-year-old girl alleging that she posed a threat to her child and should be denied custody. Other topics dealt with the issues of teaching evolution in public schools, voting rights, and a police shooting. Across all these topics, participants on average felt better than predicted about the prospect of defending their assigned position, no matter which side they were incentivized to defend. Consistent with prior research[36,37], participants failed to fully account for the influence of incentives when generating affective predictions. This finding represents the first piece of our Bayesian account of wishful belief polarization, which suggests that the affective prediction errors observed here are used, in a Bayesian manner, to update beliefs.

Study 2

Study 2 serves as an initial test of the error-based updating and noise hypotheses. In this study, participants read a vignette about a fictional woman, Harriet, described as the Chief Financial Officer of a manufacturing company on trial for financial fraud and embezzlement. After reading the vignette, participants learned that, in the next few minutes, they would be assigned to one of two roles in Harriet's trial: prosecuting attorney (which entails defending the position that Harriet is guilty) or defense attorney (which entails defending the position that Harriet is innocent). Participants also learned that if they succeeded in defending their assigned position, they would win a cash prize. After providing this information, but before assigning participants to defend a particular position, we collected self-report measures of (i) the subjective probability of Harriet's guilt, (ii) predicted affect conditional on prosecuting Harriet, (iii) predicted affect conditional on defending Harriet, and (iv) subjective observation noise (i.e., the subjective noisiness of affect) (see Methods).

Next, all participants were assigned to the role of prosecuting attorney, introducing an incentive to defend the position that Harriet is guilty. Immediately after, we measured how participants felt about the prospect of prosecuting Harriet. As in Study 1, the introduction of the incentive was the only relevant change that occurred between

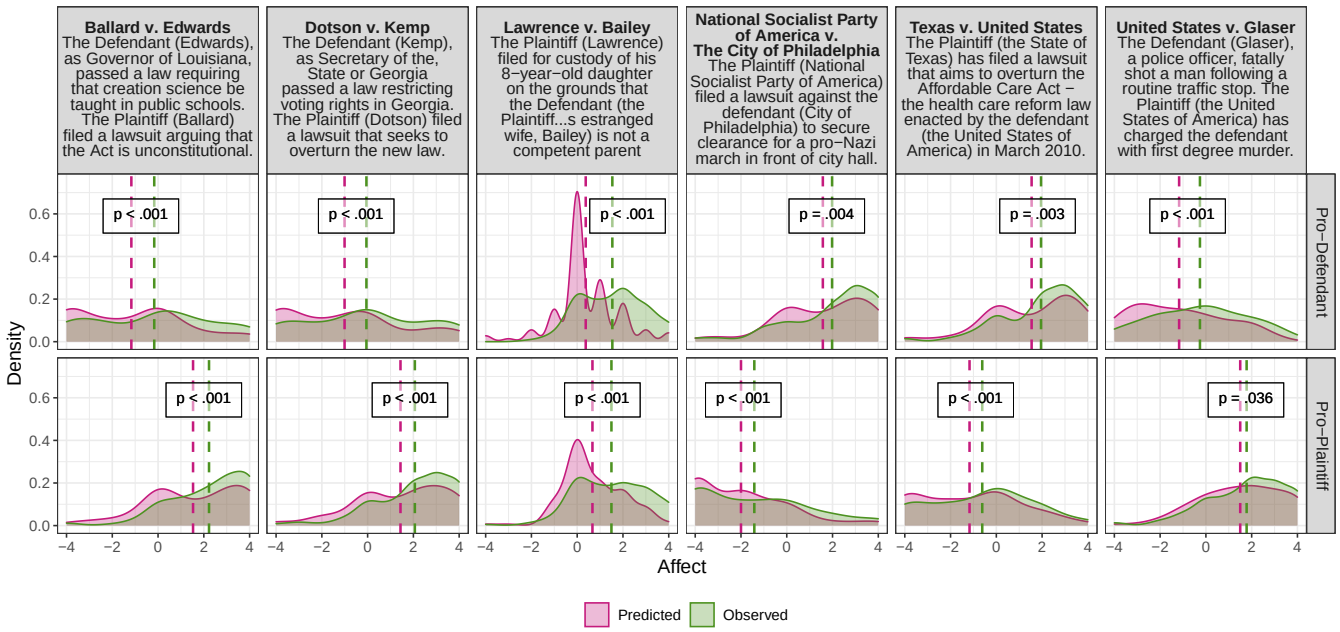

Figure 1. Predicted versus observed affect as a function of case and whether participants were incentivized to defend the plaintiff's position (Pro-Plaintiff) or the defendant's position (Pro-Defendant). Distributions of predicted affect are in pink, and distributions of observed affect are in green. Pink and green dashed lines represent average predicted and observed affect, respectively. P-values correspond to the contrasts between predicted and observed affect for each combination of case and position. Descriptions of each case are abbreviated. For full descriptions, see Supporting Information.

the reporting of predicted and actual affect. Accordingly, any affective prediction errors reflect a change in motivation to defend the assigned position, and can therefore drive directionally motivated belief change. Consistent with the underestimation hypothesis, these affective prediction errors were systematically positive (b = .53, SE = .07, $t(568) = 8.05$, $p < .001$; Figure 2A).

After measuring actual affect, we gave participants a description of what Harriet's trial would entail: during the trial, evidence would appear on their screen, and their job would be to present incriminating evidence to the jury by pressing their spacebar whenever the evidence on their screen made Harriet seem guilty. This task was never actually performed; the purpose of describing the task was to obscure the true purpose of the study by bolstering our cover story, which was that the study was intended to examine ability to recall information presented during criminal trials.

After learning what Harriet's trial would involve, participants reported, for the second time, the subjective probability of Harriet's guilt. Wishful belief updating was operationalized as the change in the subjective probability of Harriet's guilt from time 1 (before prosecuting was incentivized) to time 2 (after prosecuting was incentivized). Participants exhibited a significant amount of wishful belief updating: Harriet was judged more likely to be guilty once participants were incentivized to prosecute Harriet (b = .04, SE = .006, $t(568) = 6.97$, $p < .001$).

We found support for the error-based updating and noise hypotheses in the form of a significant interaction between affective prediction error and subjective observation noise on wishful belief updating (b = .01, SE = .004, $t(565) = 2.56$, p = .01; Figure 2B). When subjective observation noise was low, there was a main effect of affective prediction error on wishful belief updating such that larger affective prediction errors were associated with greater increases in the subjective probability of Harriet's guilt (b = .03, SE = −.005, $t(565) = 6.48$, $p < .001$). Conversely, when subjective observation noise was high, the effect of affective prediction error on wishful belief updating was not significant (b = .01, SE = −.005, $t(565) = 1.92$, p = .055).

To confirm that wishful belief updating was a function of the difference between the predicted and observed affect, and not just one of these factors on its own, we ran another model of wishful belief updating with predicted and observed affect simultaneously included as separate regressors (Figure 2C). When subjective observation noise

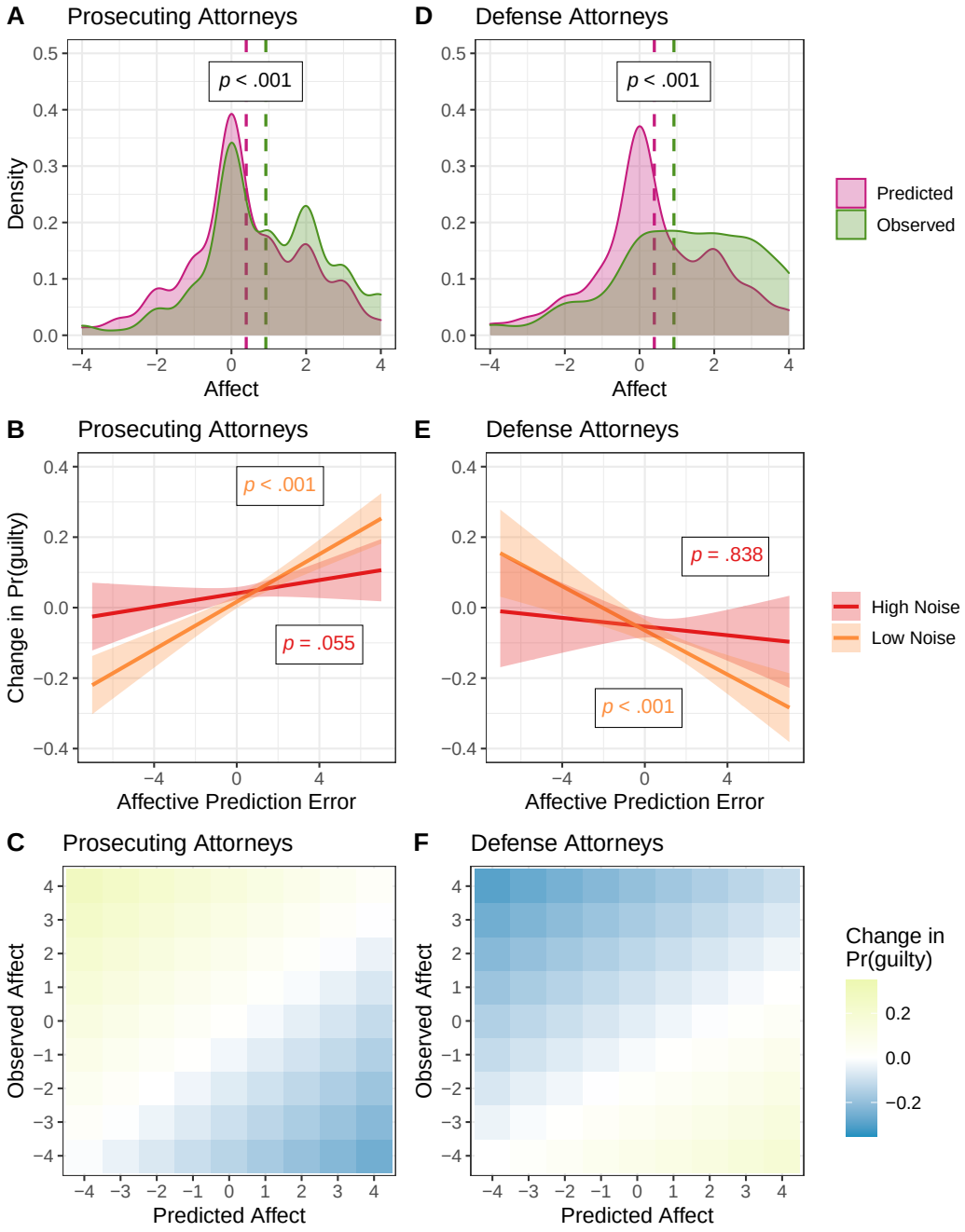

Figure 2. **A.** Predicted versus observed affect regarding the defending of the position that Harriet is guilty. **B.** Updating of the subjective probability of Harriet's guilt as a function of affective prediction errors (actual minus predicted affect) and the subjective noisiness of affect (high noise vs. low noise). High and low noise correspond to 1 standard deviation (SD) above and below mean noise, respectively. **C.** Independent effects of predicted and actual affect on updating of the subjective probability of Harriet's guilt when noise is low. **D.** Predicted versus actual affect regarding the defending of the position that Harriet is innocent. **E.** Updating of the subjective probability of Harriet's guilt as a function of affective prediction errors (actual minus observed affect) and subjective noisiness of affect (high noise vs. low noise). High and low noise correspond to 1 SD above and below mean noise, respectively. **F.** Independent effects of predicted and observed values of affect on updating of the subjective probability of Harriet's guilt when noise is low.

was low, we found a significant, positive effect of observed affect (b = .04, SE = .005, $t(563)$ = 6.71, $p < .001$) and a significant, negative effect of predicted affect ($b = -.03$, SE = .006, $t(563)$ = 4.89, $p < .001$), consistent with the hypothesis that it is the difference between these variables, rather than either variable on its own, that drives wishful belief updating.

An alternative account of these findings is that affective prediction errors are outcomes, rather than causes, of wishful belief updating. On this account, participants shifted their beliefs, which led them to feel better than predicted about the prospect of defending their assigned position. This account can be ruled out however, as it does not explain why the subjective noisiness of affect would attenuate the relationship between affective prediction errors and wishful belief updating. Unless affective prediction errors are used as information to update beliefs, there is no reason why their subjective noisiness (which determines how diagnostic they are about the state of the world) should make their relationship with belief updating weaker.

One may wonder if an a different Bayesian model can account for the results of Study 2. Cognitive dissonance theory[43] can be used to formulate a Bayesian account of wishful belief polarization that differs from our own. Dissonance theory suggests that when someone agrees to defend a position they believe is invalid, they experience negative affect, which they reduce by updating their belief to be more consistent with their behavior. This dissonance reduction process can be formalized by a Bayesian model in which a person's own behavior is treated as evidence[44,45]. According to such a model, when people agree to defend a position that they consider invalid, their negative affect motivates them to compute, using Bayes' rule, a posterior belief about the position's validity given the evidence furnished by their own behavior[44,46]. This process creates a posterior belief that is more aligned with the behavior, reducing negative affect. Of course, this process would also generate belief polarization among two individuals who agree to defend opposing positions and, like our model, would do so through Bayesian calculations.

So, can a dissonance-inspired Bayesian model account for the results of Study 2? The answer is no. Dissonance theory predicts that people should form more negative beliefs about Harriet to the extent that they experience negative affect[47]. This is not what we found. Rather, participants formed more negative beliefs about Harriet to the extent that they experienced more *positive affect* than anticipated.

It should come as no surprise that, in the present study, wishful belief updating was unrelated to cognitive dissonance. According to dissonance theory, the psychological discomfort responsible for belief updating occurs only when the choice to defend a subjectively invalid position is made freely[31]. Support for this idea comes from decades of research using the induced compliance paradigm, in which participants are asked to defend a position, and either explicitly informed that they can opt out of doing so (high choice condition), or not (low choice condition). Dissonance effects emerge in the high choice condition, and not the low choice condition. Our paradigm should be considered a low choice condition, as participants are not given the option to opt out of defending their assigned position without foregoing payment.

For these reasons it is doubtful that cognitive dissonance can account for our findings. Nonetheless, we provide even stronger evidence against a dissonance-inspired Bayesian model in the next study. In Study 3, we measure perceptions of free choice, allowing us to test our claims that perceptions of free choice in our paradigm are low to non-existent, and unrelated to wishful belief updating.

Study 3

Study 3 was similar to the previous study with one important exception. Instead of prosecuting Harriet, participants were told that, in order to win a cash prize, they would have to advocate for the position that Harriet is innocent by playing the role of defense attorney. In addition, at the end of this study, we asked participants to rate how free they were to opt out of defending their assigned position (see Methods).

As in Study 2, we observed a significant amount of wishful belief updating. After being incentivized to defend Harriet, participants considered Harriet's guilt less probable (b = -.08, SE = .01, $t(310)$ = 6.9, $p < .001$). Affective prediction errors were positive overall (b = .82, SE = .09, $t(310)$ = 8.76, $p < .001$; Figure 2D), consistent with the underestimation hypothesis.

We obtained support for the error-based updating and noise hypotheses in the form of a significant interaction between affective prediction error and subjective observation noise on wishful belief updating (b = .01, SE = .005, $t(307)$ = 2.33, p = .02; Figure 2E). When observation noise was low, there was a significant effect of affective prediction error on wishful belief updating such that larger affective prediction errors were associated with greater decreases in the subjective probability of Harriet's guilt ($b = -.04$, SE = .009, $t(307)$ = 4.32, $p < .001$). Conversely, when the subjective noisiness of affect was high, the effect of affective prediction error on wishful belief updating was not significant ($b = -.002$, SE = .011, $t(307)$ = .21, p = .838).

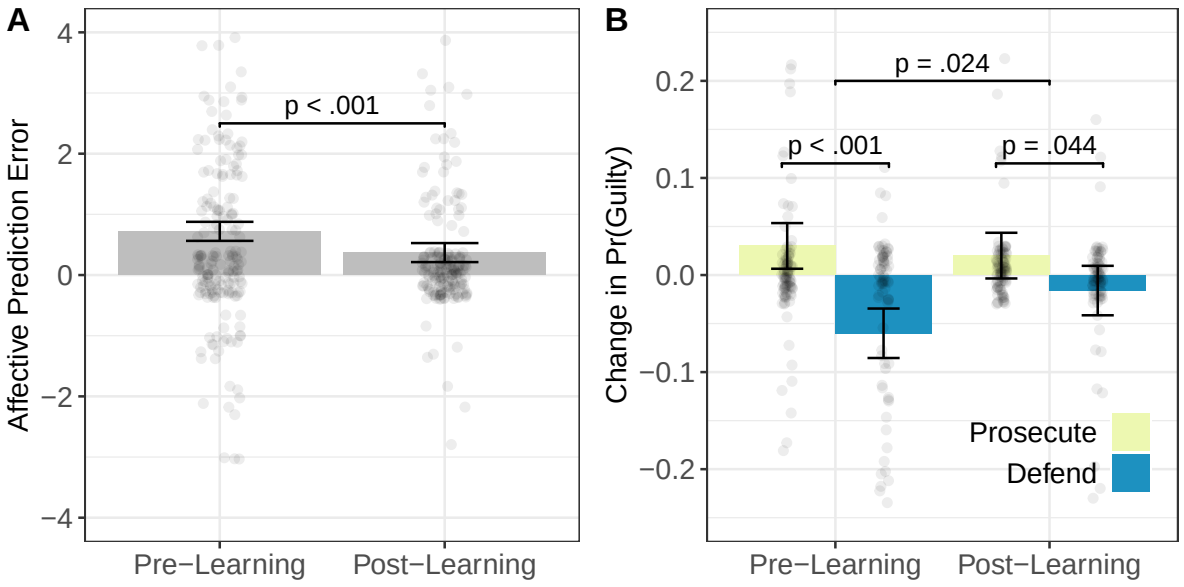

Figure 3. **A.** Affective prediction errors at time 1 (Pre-Learning) and time 2 (Post-Learning). **B.** Belief updating as a function of goal-conducive action (Prosecute vs. Defend) at time 1 (Pre-Learning) and time 2 (Post-Learning).

Wishful belief updating was a function of affective prediction errors per se—not just predicted or observed affect. When predicted and observed affect were entered simultaneously into a regression model predicting wishful belief updating, we observed a significant, negative effect of observed affect ($b = -.04$, $SE = .009$, $t(305) = 4.62$, $p < .001$), and a significant, positive effect of predicted affect ($b = .03$, $SE = .01$, $t(305) = 2.84$, $p = .005$) when subjective observation noise was low.

Consistent with our claim that our paradigm induces low perceptions of free choice, the modal response on our free choice scale was zero ($Mdn = 1$, $M = 1.74$, $SD = 1.95$), corresponding to the perception of no freedom of choice. In addition, perceptions of free choice had no effect on wishful belief updating ($b = -.003$, $SE = -.006$, $t(304) = .57$, $p = .569$). These results confirm that cognitive dissonance—which occurs only under conditions of high perceived choice—cannot account for wishful belief updating in our paradigm.

Study 4

If affective prediction errors drive wishful belief polarization, then wishful belief polarization should be reduced by manipulations that reduce affective prediction errors. We tested this hypothesis our pre-registered Study 4 (https://aspredicted.org/NL3_RJW) by giving participants an opportunity to learn how good they actually feel about the prospect of defending a particular position once doing so is incentivized. As in Studies 2 and 3, Study 4 began with a vignette about a person accused of a crime. Participants learned that, in the next few minutes, they would be assigned to defend one of two positions (that the defendant is innocent or guilty) in exchange for a cash prize. We measured how participants thought they would feel about defending both positions, and then randomly incentivized participants to defend one of them. Next, we measured how participants actually felt about the prospect of defending the assigned position, and told participants what the trial would follow the same procedure from Studies 2 and 3.

Participants then learned that they would participate in not one, but two separate trials, giving them an opportunity to win two cash prizes. In preparation for the second trial, participants received a new vignette about a second defendant, and reported predictions of how they would feel about prosecuting as well as defending them. When participants reported their predictions for the second time, they were unsure of which action would be incentivized in the second trial. However, we always incentivized the same action in both trials. As a result, affective predictions should be more accurate the second time around; participants assigned to prosecute in the second trial would have just observed how they felt about the prospect of prosecuting another defendant, and participants assigned to defend the in the second trial would have just observed how they feel about the prospect of defending

another defendant. Such experiences should result in more accurate affective predictions, and this hypothesis was borne out by our analyses. Regardless of which action was incentivized, affective prediction errors were smaller at time 2 relative to time 1 (b = .35, SE = .08, t(374.33) = 4.33, p < .001; Figure 3A) — though, at both time points, they was significantly greater than zero (time 1: b = .72, SE = .08, t(633.32) = 9.57, p < .001; time 2: b = .37, SE = .08, t(633.32) = 4.88, p < .001). We also found that affective prediction errors were positively associated with wishful belief polarization: the degree to which participants shifted their belief into alignment with their cause increased with the positivity of their affective prediction error (b = .04, SE = .004, t(371) = 8.73, p < .001). This effect was unmoderated by the particular position participants were assigned to defend (b = .002, SE = .008, t(370) = .24, p = .808).

According to our model, reducing affective prediction errors should reduce wishful belief polarization. To test this hypothesis, we measured belief updating either at time 1 (by measuring the subjective probability of the first defendant's guilt before, and immediately after, the incentive was introduced in the first trial) or at time 2 (by measuring the subjective probability of the second defendant's guilt before, and immediately after, the incentive was introduced in the second trial). Since affective prediction errors were reduced at time 2 relative to time 1, wishful belief polarization should be reduced at time 2 relative to time 1 as well. This is what we found (b = .06, SE = .03, t(370) = 2.27, p = .024; Figure 3B). At time 1, we observed a significant difference in wishful belief updating as a function of which position was goal-conducive to defend (b = .09, SE = .02, t(370) = 5.32, p < .001): the subjective probability of the defendant's guilt increased after participants were incentivized to prosecute (b = .03, SE = .01, t(370) = 2.53, p = .012) and decreased after participants were incentivized to defend (b = .06, SE = .01, t(370) = 4.85, p < .001). At time 2, wishful belief updating (b = .04, SE = .02, t(370) = 2.02, p = .044) was attenuated: the incentive to prosecute failed to increase the subjective probability of the defendant's guilt (b = .02, SE = .01, t(370) = 1.58, p = .116), and the incentive to defend failed to decrease the subjective probability of the defendant's guilt (b = .02, SE = .01, t(370) = 1.25, p = .211). Consistent with the idea that affective prediction errors mediate the effect of time on belief updating, we found that the effect of time on wishful belief polarization was reduced to non-significance after statistically controlling for affective prediction errors (b = .02, SE = .01, t(370) = 1.52, p = .129), whereas the effect of affective prediction errors on wishful belief polarization remained significant (b = .04, SE = .004, t(370) = 8.56, p < .001).

Discussion

On the face of it, the human capacity for wishful thinking seems incompatible with the Bayesian brain hypothesis. This is why defenses of Bayesianism have taken an eliminative stance toward wishful thinking, showing that many apparent instances of wishful thinking are not wishful after all[11,18–20]. This strategy has succeeded in defeating many challenges to the Bayesian brain hypothesis, with at least one major exception: the phenomenon of wishful belief polarization.

Instead of recasting wishful belief polarization in non-wishful terms, we have established that wishful belief polarization and Bayesianism are compatible. Our data support a fully Bayesian account of wishful belief polarization. In our model, when defending a position is incentivized, the motivation to obtain the incentive leads to better-than-predicted feelings about defending the position. The resulting affective prediction error is used to shift beliefs into alignment with the to-be-defended position—a process that is at once directionally motivated and Bayes-rational.

In addition to reconciling the Bayesian brain hypothesis with wishful thinking, our model advances theories of how people use their feelings as information about the external world. That people use feelings as information is well-established[9,22–24], but accounts of the computational mechanisms underlying this process are in their infancy. Recent work suggests that feelings are used to update beliefs according to Bayesian principles[26,48,49]. Our work supports this idea by providing evidence that affect-based belief updating adheres to two predictions of Bayesian optimality: (i) it is a function not of affect per se, but of affective prediction errors, and (ii) the noisier people think their affect is, the weaker the relationship between affective prediction errors and belief updating.

Wishful belief polarization probably does not emerge from exact Bayesian inference, which is computationally expensive and, in many cases, intractable. Rather, our current best evidence suggests that Bayesian behaviors emerge from cognitive processes that approximate Bayesian inference[50,51]. Researchers have developed various models of belief updating that are both efficient and roughly Bayesian, including variational methods[52] and sampling algorithms[53]. Identifying which of these approximations best describes the cognitive computations underlying wishful belief polarization will be an important direction for future research.

One point that our results highlight is that Bayesian brains, though rational in one sense of the word, are perfectly capable of producing irrational output. This is because the rationality of a Bayesian brain is limited to

the structure of its learning process: It updates its beliefs rationally given the data it encounters and its current internal model. When fed false data, or when using a flawed internal model, Bayesian logic can easily produce irrational outcomes[18]. Here, the irrational outcome is wishful belief polarization, and the modeling flaw that seems to produce it is an overly pessimistic belief about how good it will feel to defend a position in the service of a desired outcome.

Where does this overly pessimistic belief come from? Part of an answer may be found in research on projection bias, which shows that people overestimate the degree to which their future feelings will resemble their current feelings[37]. In Bayesian terms, people have an erroneous prior belief that affect is more autocorrelated than it really is. The question then becomes: Why does this erroneous prior persist in the face of disconfirming evidence? Why has a lifetime of learning not eliminated the projection bias? One possibility is that people fail to generalize appropriately. If a person feels better-than-expected about defending a position in one specific context, that person may infer that their feelings are less autocorrelated than previously thought *in that particular context*, but not in general. This would lead to more accurate affective predictions within the context in which the initial prediction error was encoded (as we show in Study 4) but not in other contexts. If this is the case, then it may be possible to reduce wishful thinking by encouraging people to generalize more broadly when learning from affective prediction errors.

Perhaps the deepest question for future work is the one we started with: is the brain Bayesian? The current investigation does not settle the issue, but it does lend support to proponents of the Bayesian approach by defending it against one of its greatest empirical challenges.

Materials and Methods

All studies were approved by the Institutional Review Board of the University of Pennsylvania. All participants gave informed consent and were compensated for their time. Data and analysis scripts are available on the OSF repository at: https://osf.io/59dmr/?view_only=b8ea1a66b5e84d1e8d67391662b60d82. Surveys were administered, and data collected, with Qualtrics. Using RStudio version 1.3 and R version 3.6, all between-subjects analyses were conducted using linear regression and fit using the 'stats' package, and all within-subjects and mixed analyses were conducted using linear mixed models with subject-level random intercepts and fit using the 'lmer' package.

In each study, participants had a goal to prosecute or defend someone in a criminal trial, but never actually participated in a trial. Instead, all participants learned that that survey would end early and that they would receive the financial bonus they had expected to compete for during the trial. Throughout each study, participants were asked to summarize key information to promote attention and comprehension. In addition, we included several questions in each study to confirm that participants had read and understood pertinent facts about the case and their task (see Supplementary Information).

In each study, participants had a goal to prosecute or defend someone in a criminal trial, but never actually participated in a trial. Instead, all participants learned that that survey would end early and that they would receive the financial bonus they had expected to compete for during the trial. Throughout each study, participants were asked to summarize key information to promote attention and comprehension. In addition, we included several questions in each study to confirm that participants had read and understood pertinent facts about the case and their task (see Supplementary Information).

Power

Data for establishing empirical estimates of effect sizes were unavailable, so we adopted a conservative approach of assuming, in all studies, that all effect of interest would be small-to-medium in size (d = .3, f^2 = .03). Under this assumption we computed the sample size necessary to achieve 80% power ($p < .05$, two-sided). In Study 1, the effects of interest were within-subject differences between predicted and observed affect across 12 between-subjects conditions, so N = 1070 was required to achieve sufficient power. In Studies 2 and 3, the effect of interest was a two-way interaction between affective prediction errors and observation noise, so N = 262 was required to achieve sufficient power. In Study 4, the effect of interest was a between-subject effect of learning, so N = 351 was required to achieve sufficient power.

Participants

We recruited participants using Prolific for Studies 1, 2, and 4, and CloudResearch for Study 3. We note that the exclusion rate (see below) was substantially higher in Study 3 relative to the other studies, which is likely due to the CloudResearch subject pool being less reliable than the Prolific subject pool, as all other relevant features of

the studies were held constant. Our initial sample sizes were N = 1502 in Study 1, N = 603 in Study 2, N = 571 in Study 3, and N = 451 in Study 4. Our final samples after exclusions were N = 1370 in Study 1 (59% female; Mdn_{age} = 32), N = 569 in Study 2 (47% female; Mdn_{age} = 33), N = 311 in Study 3 (46% female; Mdn_{age} = 35), and N = 375 in Study 4 (52% female; Mdn_{age} = 30).

Attrition did not vary significantly by condition in any study. For instance, in Study 1, where statistical power to detect such an effect was greatest, only 1.8% of participants who were assigned to defend a particular position failed to complete the survey, and the likelihood of abandoning the study prematurely was unrelated to (i) the predicted affect of defending the assigned position (b = .54, SE = .45, $t(1528)$ = 1.19, p = .232), and (ii) whether participants were assigned to their preferred position (b = .54, SE = .44, z = 1.22, p = .224).

Model

Let a_K be the act of defending, or otherwise acting in favor of, some position K, and let $V(a_K)$ be the subjective valence of this action—that is, how positive or negative a person feels about the prospect of defending K. We assume that people regard $V(a_K)$ as a function of N different variables including, but not limited to, the validity of K, denoted as x_1, and the goal-conduciveness of a_K, denoted as x_2. Let all N variables that people use to predict $V(a_K)$ be elements of the vector $\mathbf{x} = [x_1, \ldots, x_N]$. Though $\mathbf{x}$ is predictive of $V(a_K)$, it is not perfectly so; $V(a_K)$ is inherently noisy. We assume that people represent this uncertainty. Formally, we assume that people represent $V(a_K)$ as a random draw from a normal distribution:

$$V(a_K) \sim \mathcal{N}\left(\boldsymbol{\beta}\mathbf{x}^{\mathrm{T}},\, \sigma_v^2\right) \tag{1}$$

$\boldsymbol{\beta} = [\beta_1, \ldots, \beta_N]$ is a vector of weights corresponding to beliefs about the magnitude and direction of the relationship between each element of $\mathbf{x}$ and $V(a_K)$. β_1 denotes a person's belief about the relationship between x_1 (the validity of K) and $V(a_K)$, and β_2 denotes a person's belief about the relationship between x_2 (the goal-conduciveness of a_K) and $V(a_K)$. We take β_1 and β_2 to be positive. The variance, σ_v^2, denotes a person's belief about observation noise, with greater values of σ_v^2 corresponding to greater noise.

$\mathbf{x}$ comprises variables whose values cannot be known with certainty. For instance, a person cannot know for sure how valid a position is. People can only form *beliefs* about the elements of $\mathbf{x}$—beliefs that may be wrong. Let a person's beliefs about the elements of $\mathbf{x}$ be elements of the vector $\hat{\mathbf{x}} = [\hat{x}_1, \ldots, \hat{x}_N]$. $\hat{x}_1$ is a person's belief about the validity of K (a person's belief about the value of x_1) and $\hat{x}_2$ is a person's belief about the goal-conduciveness of a_K (a person's belief about the value of x_2). We assume that people represent $\mathbf{x}$ as normally distributed around $\hat{\mathbf{x}}$:

$$\mathbf{x} \sim \mathcal{N}(\hat{\mathbf{x}}, \mathbf{Q}) \tag{2}$$

where $\mathbf{Q}$ =diag$(q_1, \ldots, q_N)$ is a covariance matrix.

Given the generative model defined by equations 1 and 2, Bayes' rule stipulates the optimal way to update $\hat{x}_1$ (the subjective validity of K) upon observing $V(a_K)$. The update involves three steps. In the first step, $\hat{\mathbf{x}}$ is used to predict the value of $V(a_K)$. This prediction, denoted a $\hat{V}(a_K)$, quantifies how a person expects to feel about the prospect of defending K. It is computed as follows:

$$\hat{V}(a_K) = \boldsymbol{\beta}\hat{\mathbf{x}}^{\mathrm{T}} \tag{3}$$

The next step involves computing the affective prediction error, δ_K, which quantifies how much better or worse a person feels about the prospect of defending K than expected:

$$\delta_K = V(a_K) - \hat{V}(a_K) \tag{4}$$

Note that δ_K is goal-dependent when it results from an overly small value of β_2 (i.e., an underestimation of the degree to which $V(a_K)$ increases with the goal-conduciveness of a_K). Finally, δ_K is used to update $\hat{x}_1$:

$$\hat{x}_1' = \hat{x}_1 + \delta_K \beta_1 \alpha \tag{5}$$

$\hat{x}_1'$ is the new belief about the validity of K, and α is a non-negative weighting factor that controls the amount of updating:

$$\alpha = \frac{q_1}{\boldsymbol{\beta}\mathbf{Q}\boldsymbol{\beta}^{\mathrm{T}} + \sigma_v} \tag{6}$$

where q_1 is the element of covariance matrix $\mathbf{Q}$ corresponding to the variance of x_1. Notice that α is a decreasing function of observation noise, σ_v^2. Accordingly, equations 5 and 6 say that as observation noise increases, the amount of updating in response to δ_K decreases, in line with the noise hypothesis. Also notice that, according to equation 5, if $\beta_1 > 0$ (i.e., if people expect $V(a_K)$ to increase with the validity of K), then the subjective validity of K should increase as a positive function of δ_K, in line with the error-based updating hypothesis.

Goal-Dependent Prediction Errors

We quantified goal-dependent prediction errors as the difference between the predicted and observed values of $V(a_K)$. We measured predictions and observations of $V(a_K)$ by asking participants how positive or negative they thought they would feel—or how positive or negative they did feel—about (i) representing the plaintiff or defendant (Study 1), or (ii) prosecuting or defending the defendant (Studies 2–4). Participants responded on a 9-point scale from 1 (extremely negative) to 9 (extremely positive), with 5 (neutral) in the middle.

Observation Noise

In Study 2, we measured affective noisiness in two steps. First, we asked participants four questions of the following form: "If you knew for sure that Harriet was [innocent / guilty], how do you think you would feel about the prospect of [defending / prosecuting] her?" Next, we asked participants to rate their confidence in each answer. Participants responded to four questions of the following form: "You predicted that if you knew for sure that Harriet was [innocent / guilty], you'd feel X about [defending / prosecuting] her. How certain are you that this is how you'd feel?" where X is the relevant affective prediction. Participants responded on 9-point scales from 0 ("Not at all certain") to 9 ("Completely certain"). For participants assigned to defend a particular position (i.e., participants assigned to prosecute or defend Harriet), observation noise was computed as the average degree of uncertainty about the subjective valence of defending that position (i) if Harriet was known to be guilty and (ii) if Harriet was known to be innocent. The logic underlying this approach is the following: If someone is highly uncertain of how they would feel about prosecuting or defending someone they know to be guilty or innocent, this uncertainty must reflect the belief that observation noise is high (i.e., the belief that the subjective valence of prosecuting or defending is noisy). Conversely, if someone is highly certain of how they would feel about prosecuting or defending someone they know they to guilty or innocent, this certainty must reflect the belief that observation noise is low (i.e., the belief that the subjective valence of prosecuting or defending is predictable).

In Study 3, we adopted a simpler, more straightforward approach to measuring the subjective noisiness of affect. After participants reported their initial affective predictions, we asked participants to rate their confidence in their predictions. We used the following item: "You predicted that you'd feel X about [defending / prosecuting] Harriet. How certain are you that this is how you'd feel?" where X is the relevant affective prediction. Participants responded on 9-point scales from 0 ("Not at all certain") to 9 ("Completely certain").

Freedom of Choice

In Study 3, we measured perceived freedom of choice using two self-report items asking participants to rate how much they agree or disagree with the following statements: "Had I wanted to, I could have declined to perform the role of [prosecuting attorney / defense attorney]" and "I could not have declined to perform the role of [prosecuting attorney / defense attorney] even if I had wanted to" (reverse coded). Participants responded on 7-point scales from 0 ("Not at all accurate") to 6 ("Completely accurate").

References

1. Knill, D. C. & Pouget, A. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).
2. Doya, K., Ishii, S., Pouget, A. & Rao, R. P. *Bayesian brain: Probabilistic approaches to neural coding* (MIT press, Cambridge, MA, 2007).
3. Hohwy, J. *The predictive mind* (Oxford University Press, Oxford, UK, 2013).

4. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).
5. Lee, T. S. & Mumford, D. Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A* **20**, 1434–1448 (2003).
6. Oaksford, M. & Chater, N. *Bayesian rationality: The probabilistic approach to human reasoning* (Oxford University Press, Oxford, UK, 2007).
7. Friston, K. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
8. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: Statistics, structure, and abstraction. *Science* **331**, 1279–1285 (2011).
9. Barrett, L. F. The theory of constructed emotion: An active inference account of interoception and categorization. *Soc. Cogn. Affect. Neurosci.* **12**, 1–23 (2017).
10. Williams, D. Hierarchical Bayesian models of delusion. *Conscious. Cogn.* **61**, 129–147 (2018).
11. Tappin, B. M. & Gadsby, S. Biased belief in the Bayesian brain: A deeper look at the evidence. *Conscious. Cogn.* **68**, 107–114 (2019).
12. Mandelbaum, E. Troubles with Bayesianism: An introduction to the psychological immune system. *Mind & Lang.* **34**, 141–157 (2019).
13. 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).
14. Madson, G. J. & Hillygus, D. S. All the best polls agree with me: Bias in evaluations of political polling. *Polit. Behav.* **42**, 1055–1072 (2020).
15. Kunda, Z. The case for motivated reasoning. *Psychol. Bull.* **108**, 480–498 (1990).
16. Weber, E. U. & Stern, P. C. Public understanding of climate change in the United States. *Am. Psychol.* **66**, 315–328 (2011).
17. Kahan, D. M. The politically motivated reasoning paradigm. *Emerg. trends social behavioral sciences* 1–15 (2015).
18. Gershman, S. J. How to never be wrong. *Psychon. Bull. & Rev.* **26**, 13–28 (2019).
19. Jern, A., Chang, K.-M. K. & Kemp, C. Belief polarization is not always irrational. *Psychol. Rev.* **121**, 206–224 (2014).
20. Cook, J. & Lewandowsky, S. Rational irrationality: Modeling climate change belief polarization using bayesian networks. *Top. Cogn. Sci.* **8**, 160–179 (2016).
21. Melnikoff, D. E. & Strohminger, N. The automatic influence of advocacy on lawyers and novices. *Nat. Hum. Behav.* 1–7 (2020).
22. Loewenstein, G. & Lerner, J. S. The role of affect in decision making. In Davidson, R. J., Scherer, K. R. & H., G. H. (eds.) *Handbook of affective sciences*, 619–642 (Oxford University Press, Oxford, UK, 2003).
23. Schwarz, N. & Clore, G. L. Mood as information: 20 years later. *Psychol. Inq.* **14**, 296–303 (2003).
24. Clore, G. L. & Huntsinger, J. R. How emotions inform judgment and regulate thought. *Trends Cogn. Sci.* **11**, 393–399 (2007).
25. Heffner, J., Son, J.-Y. & FeldmanHall, O. Emotion prediction errors guide socially adaptive behaviour. *Nat. Hum. Behav.* **5**, 1391–1401 (2021).
26. Barrett, L. F. & Simmons, W. K. Interoceptive predictions in the brain. *Nat. Rev. Neurosci.* **16**, 419–429 (2015).
27. Seth, A. K. & Friston, K. J. Active interoceptive inference and the emotional brain. *Philos. Transactions Royal Soc. B: Biol. Sci.* **371**, 20160007 (2016).
28. Atzil, S., Gao, W., Fradkin, I. & Barrett, L. F. Growing a social brain. *Nat. Hum. Behav.* **2**, 624–636 (2018).
29. Hoemann, K., Xu, F. & Barrett, L. F. Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Dev. Psychol.* **55**, 1830–1849 (2019).
30. Mathys, C. D. *et al.* Uncertainty in perception and the hierarchical gaussian filter. *Front. Hum. Neurosci.* **8**, 825 (2014).

31. Festinger, L. & Carlsmith, J. M. Cognitive consequences of forced compliance. *The J. Abnorm. Soc. Psychol.* **58**, 203–210 (1959).

32. Bem, D. J. Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychol. Rev.* **74**, 183–200 (1967).

33. Robinson, M. J. & Berridge, K. C. Instant transformation of learned repulsion into motivational "wanting". *Curr. Biol.* **23**, 282–289 (2013).

34. Dayan, P. & Berridge, K. C. Model-based and model-free pavlovian reward learning: Revaluation, revision, and revelation. *Cogn. Affect. & Behav. Neurosci.* **14**, 473–492 (2014).

35. Melnikoff, D. E. & Bailey, A. H. Preferences for moral vs. immoral traits in others are conditional. *Proc. Natl. Acad. Sci.* **115**, E592–E600 (2018).

36. Loewenstein, G. Out of control: Visceral influences on behavior. *Organ. Behav. Hum. Decis. Process.* **65**, 272–292 (1996).

37. Loewenstein, G., O'Donoghue, T. & Rabin, M. Projection bias in predicting future utility. *The Q. J. Econ.* **118**, 1209–1248 (2003).

38. Read, D. & Van Leeuwen, B. Predicting hunger: The effects of appetite and delay on choice. *Organ. Behav. Hum. Decis. Process.* **76**, 189–205 (1998).

39. Rescorla, R. A. & Wagner, A. R. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. & Prokasy, W. F. (eds.) *Classical Conditioning II: Current Research and Theory*, 64–99 (Appleton-Century-Crofts, New York, NY, 1972).

40. Bush, R. R. & Mosteller, F. *Stochastic models for learning.* (John Wiley & Sons, Inc., New York, NY, 1955).

41. Sutton, R. S. Learning to predict by the methods of temporal differences. *Mach. Learn.* **3**, 9–44 (1988).

42. Gershman, S. J. A unifying probabilistic view of associative learning. *PLoS Comput. Biol.* **11**, e1004567 (2015).

43. Festinger, L. *A theory of cognitive dissonance* (Stanford University Press, Palo Alto, CA, 1957).

44. Kruglanski, A. W. *et al.* Cognitive consistency theory in social psychology: A paradigm reconsidered. *Psychol. Inq.* **29**, 45–59 (2018).

45. Luu, L. & Stocker, A. A. Post-decision biases reveal a self-consistency principle in perceptual inference. *eLife* **7**, e33334 (2018).

46. Cushman, F. Rationalization is rational. *Behav. Brain Sci.* **43**, 1–59 (2020).

47. Elliot, A. J. & Devine, P. G. On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *J. Pers. Soc. Psychol.* **67**, 382–394 (1994).

48. Seth, A. K. Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.* **17**, 565–573 (2013).

49. Smith, R. *et al.* A Bayesian computational model reveals a failure to adapt interoceptive precision estimates across depression, anxiety, eating, and substance use disorders. *PLoS Comput. Biol.* **16**, e1008484 (2020).

50. Gershman, S. J., Horvitz, E. J. & Tenenbaum, J. B. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* **349**, 273–278 (2015).

51. Lieder, F. & Griffiths, T. L. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* **43**, E1 (2020).

52. Mathys, C., Daunizeau, J., Friston, K. J. & Stephan, K. E. A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* **5**, 39 (2011).

53. Sanborn, A. N. & Chater, N. Bayesian brains without probabilities. *Trends Cogn. Sci.* **20**, 883–893 (2016).

Acknowledgements

This research was supported by the NIMH under award F32MH124430 (to D.E.M.), and the Wolpow Family Faculty Scholar Fund, the Wharton Dean's Research Fund, and the Wharton Behavioral Lab.

Author contributions statement

D.E.M. and N.S. jointly designed the studies, analyzed the data, and wrote the paper. D.E.M. is responsible for the formal model and, in an equally weighty intellectual achievement, N.S. is responsible for typesetting that model in LaTeX.