Opinion

The Psychology of Motivated versus Rational Impression Updating

Minjae Kim,[1,*] BoKyung Park,[1] and Liane Young[1]

People's beliefs about others are often impervious to new evidence: we continue to cooperate with ingroup defectors and refuse to see outgroup enemies as rehabilitated. Resistance to updating beliefs with new information has historically been interpreted as reflecting bias or motivated cognition, but recent work in Bayesian inference suggests that belief maintenance can be compatible with procedural rationality. We propose a mentalizing account of belief maintenance, which holds that protecting strong priors by generating alternative explanations for surprising information involves more mentalizing about the target than nonrational discounting. We review the neuroscientific evidence supporting this approach, and discuss how both types of processing can lead to fitness benefits.

Highlights

Recent theoretical work suggests that instances of seemingly motivated belief maintenance may be compatible with Bayesian-rational inference. When we have strong prior beliefs, protecting these beliefs from revision by generating an *ad hoc* auxiliary hypothesis is procedurally rational.

Computational neuroimaging work shows that brain regions in and outside of the ToM network are involved in probabilistic belief updating. The medial prefrontal cortex in particular is implicated in updating beliefs about social agents, and updating value representations with information from social sources.

We propose that, given its role in both mental state inference and belief updating, the ToM network will be recruited for both rational belief updating and rational belief maintenance, but less so for nonrational discounting of information.

Belief Maintenance in Person Perception

Once we have made up our minds about people, it can be hard for us to change our beliefs. This is especially so when our beliefs are desirable in some way – we often hold onto positive beliefs about people we care for and negative beliefs about those we dislike. Much of the extant literature in social psychology reveals that beliefs about others persist in the face of counter-attitudinal information. We dismiss stereotype disconfirmers as outliers [1], invest in friends who reciprocate only half the time [2], and overblame outgroup members for unintentional harms [3]. Such findings indicate that motivational factors [4], which are shaped by social distance and group membership, can inhibit flexible **belief updating** (see Glossary) in light of new evidence. This resistance to belief updating is a puzzle, as up-to-date inferences about others' moral intent (e.g., helpful vs. harmful) and character (e.g., trustworthy vs. untrustworthy) allow us to make informed decisions about when and whom to trust, reciprocate, punish, and forgive. However, failure to update beliefs is not necessarily nonrational; seemingly motivated belief maintenance can instead result from a rational process. When we have strong **prior beliefs** about others, inferring alternate causes for their unexpected behavior can be compatible with **Bayesian rationality**. A given instance of belief maintenance, then, can be either procedurally rational – the result of Bayesian reasoning over strong priors, or procedurally nonrational – the result of discounting of counter-attitudinal information.

The key question here is this: How can we adjudicate between belief maintenance that is compatible with the rational incorporation of priors, and belief maintenance that is nonrational? To answer this question, we propose a **mentalizing** account of belief maintenance, which posits that rational belief maintenance involves mentalizing about the target to come up with alternative explanations, while nonrational discounting is characterized by the absence of mentalizing. As the behavioral signatures of rational belief maintenance and nonrational discounting are indistinguishable, we need to examine mentalizing-related neural activity, in conjunction with behavioral data, to infer which mechanism resulted in belief maintenance. This proposal thus invites a unique contribution for neuroscientific evidence, and calls for a re-examination of multiple studies that have documented belief maintenance: ingroup favoritism [5–7], stereotyping [1,8,9], impression

[1] Department of Psychology, Boston College, Chestnut Hill, MA 02467, USA

*Correspondence:
minjae.kim@bc.edu (M. Kim).

formation and updating [10–12], and ultimate attribution error [13–15]. We deploy new neuroimaging evidence and revisit phenomena that have typically been interpreted as reflecting biased cognition, considering cases that are compatible with rational belief preservation, and cases that are not. We further argue that the two forms of processing (i.e., procedurally rational vs. nonrational belief maintenance) may serve distinct functions.

Belief Maintenance Can Be Rational

Recent theoretical work has explored how seemingly motivated belief maintenance can be compatible with Bayesian reasoning over strong priors. One account holds that observers with strong prior beliefs can generate *ad hoc* **auxiliary hypotheses** (claims designed to accommodate conflicting information) to explain unpredicted events, and that this is a form of rational inference [16]. This mechanism adheres to the probabilistic tenets of the Bayesian framework: auxiliaries are likelier to be invoked when they are highly consistent with the new information, and when the central belief has a relatively high prior probability. Credit for the new observation is thus distributed between central and auxiliary hypotheses according to their posterior probabilities [16].

For example: imagine that you observe someone take money from a tip jar. If a trustworthy friend were performing this action, you might generate an auxiliary hypothesis about her innocent intent (e.g., in fact, she was intending to make change for a dollar), because you have stronger (more certain) prior beliefs about her trustworthiness. In invoking an auxiliary, in this case a situational explanation, you give less credit to the hypothesis that her character or stable disposition is responsible for the observed outcome. By contrast, if a stranger performed the same action, you might be less likely to make such a situational attribution, as you have weaker (less certain) prior beliefs about her trustworthiness. You give more credit to the hypothesis that her character produced the observed outcome.

The likelihood of invoking an auxiliary hypothesis rests on more than the mere existence of a relationship history; rather, it depends probabilistically on the certainty of prior beliefs (although this is often a function of relationship history). From this perspective, cases of belief maintenance that have been construed as motivated are theoretically compatible with a Bayesian-rational mechanism, wherein our strong prior beliefs warrant alternative explanations of inconsistent information.

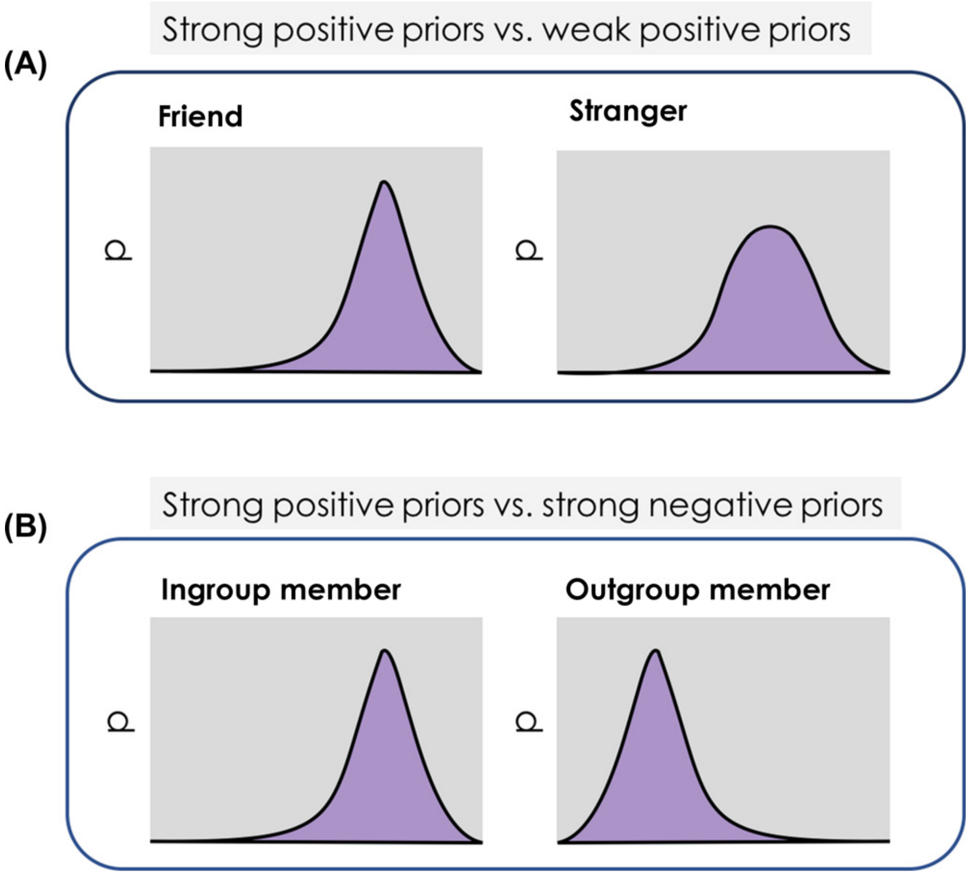
Forms of Strong Prior Beliefs

A key feature of the Bayesian account is that new information is weighed against the strength of our priors. In what contexts do observers have stronger versus weaker priors? **Belief distributions** can vary across social distance (e.g., friend vs. stranger), and in valenced contexts (e.g., friend vs. enemy; ingroup vs. outgroup) (Figure 1). In the tip jar example, closeness was the operant dimension: we have stronger prior beliefs about our friend because we have built them up through repeated interaction. This is an instance where strong, positive priors about a friend support an auxiliary explanation for negative behavior, whereas weak, neutral priors about a stranger do not favor such an inference. Another notable feature of this framework is that it predicts the observer will infer an auxiliary explanation for a friend's, but not a stranger's, extremely positive behavior. Evidence that contradicts – from either direction – strong priors can potentially be explained by an auxiliary hypothesis.

In other situations, group membership may supplant experience, serving as shorthand for moral character. This is evident in cases where observers have positive prior beliefs about ingroup strangers, and negative prior beliefs about outgroup strangers. In contrast to friend–stranger contexts, in intergroup contexts, the observer may hold comparably strong beliefs

Glossary

Auxiliary hypothesis: claim designed to accommodate conflicting information.
Belief distribution: the probability of each possible value of a belief.
Belief updating: revising a belief in light of new evidence.
Bayesian rationality: a form of procedural rationality wherein beliefs are revised probabilistically according to Bayes' theorem.
Kullback–Leibler divergence: a measure of the difference between two belief distributions (e.g., between a prior distribution and a posterior distribution).
Prediction error: the difference between what was expected (per a belief distribution) and what was observed.
Prior belief/distribution: a belief/distribution before new evidence is observed.
Mentalizing: also known as ToM; the ability to attribute and reason about agents' mental states, such as beliefs, desires, goals, and intentions.
Mentalizing network: brain regions implicated in mentalizing/ToM, including DMPFC, ventromedial prefrontal cortex, RTPJ, LTPJ, and precuneus.



Trends in Cognitive Sciences

Figure 1. Examples of Prior Belief Distributions. Belief distributions can vary across social distance (e.g., friend vs. stranger), and in valenced contexts (e.g., friend vs. enemy; ingroup vs. outgroup). (A) We have strong prior beliefs about our friends because we have built them up through repeated interactions with them, while we lack strong priors for strangers. This means that our positive prior beliefs about a friend can support an auxiliary explanation for negative behavior, whereas weak, distributed priors about a stranger do not favor such an inference. (B) In intergroup contexts, group membership may supplant experience, serving as shorthand for moral character. This is evident in cases where observers have positive prior beliefs about ingroup strangers, and negative prior beliefs about outgroup strangers. In contrast to friend–stranger contexts, in intergroup contexts, the observer may hold comparably strong beliefs about each target, but the content of their beliefs will often be opposite in valence. The current framework predicts that an observer may account for negative information about an ingroup member but not an outgroup member; conversely, an observer may account for positive information about an outgroup member but not an ingroup member.

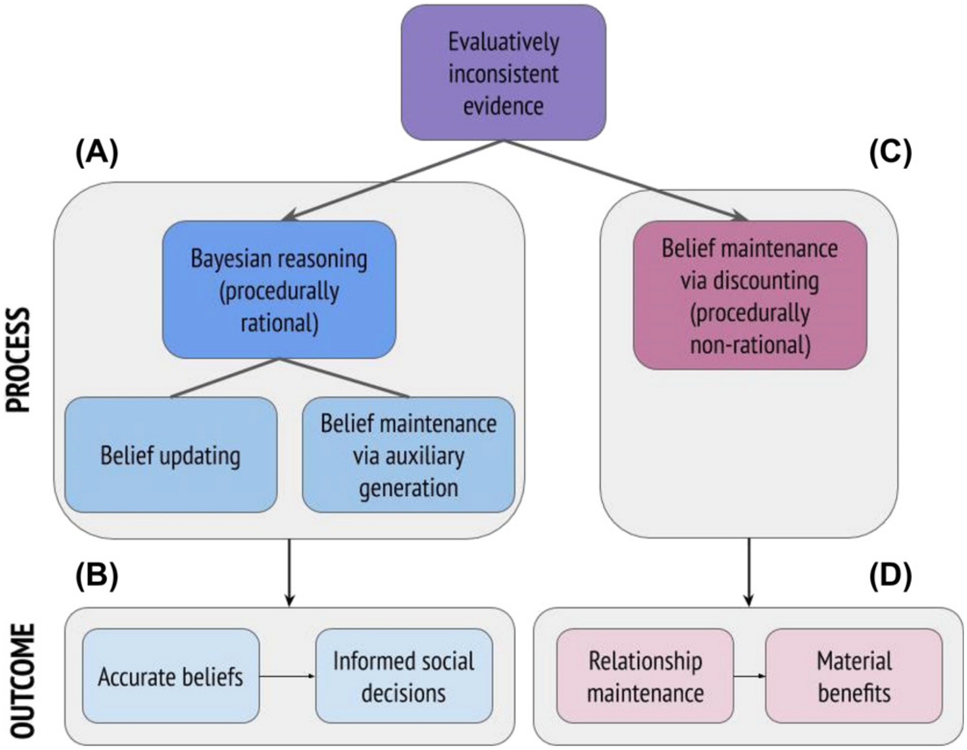
about each target, but the content of their beliefs will often be opposite in valence. The current framework predicts that an observer may account for negative information about an ingroup member but not an outgroup member; conversely, an observer may account for positive information about an outgroup member but not an ingroup member (but see [17] for an investigation on how beliefs about initially bad agents are more amenable to Bayesian impression updating).

A Mentalizing Account of Rational Belief Maintenance

When an observer is faced with new information that is inconsistent with strong prior beliefs, they can take one of three paths: (i) update their prior beliefs about the target; (ii) produce an auxiliary hypothesis to accommodate the information; or (iii) discount the information (Figure 2). The first two paths are compatible with procedural rationality, while discounting is not; additionally, the first two paths are expected to involve more mentalizing – reasoning about the mental states of others – compared to discounting. In the updating case, the observer may draw a straightforward mental state inference (e.g., of harmful intent given a harmful outcome), and use this inference to update their beliefs about the target's moral character. In the case of auxiliary generation, the observer may infer a mental state that is concordant with their prior beliefs about the target's character, but is a likely cause for the observed outcome (e.g., intent to make change from a tip jar).

The nonrational route to belief maintenance involves discounting the new information and disengaging from further mentalizing, precluding proper updating. In the case of a previously trustworthy person performing an untrustworthy action, the non-Bayesian observer may preserve their prior beliefs about her by opting out of considering the new evidence, thus pre-empting an unfavorable mental state inference. This is a procedurally nonrational form of belief maintenance, in that the observer fails to engage with or account for new evidence in any way.

This account presents a novel opportunity to leverage neural indices of mentalizing, to distinguish between belief preservation that is compatible with the rational incorporation of priors, and


Evaluatively inconsistent evidence
(A)
(C)
PROCESS
Bayesian reasoning (procedurally rational)
Belief updating
Belief maintenance via auxiliary generation
Belief maintenance via discounting (procedurally non-rational)
(B)
(D)
OUTCOME
Accurate beliefs
Informed social decisions
Relationship maintenance
Material benefits

Trends in Cognitive Sciences

Figure 2. Rational and Nonrational Processing and Their Outcomes. When an observer is faced with new information that is inconsistent with strong prior beliefs, they can take one of three paths: update their prior beliefs about the target, produce an auxiliary hypothesis to accommodate the information, or discount the information. (A, B) Both belief updating and auxiliary generation are procedurally rational, and allow the observer to maintain accurate representations of reality, which in turn inform social decisions. (C, D) Discounting is procedurally nonrational, but may help strengthen and maintain social relationships, which in turn can bring social or material benefits.

belief preservation that is nonrational. A large body of neuroimaging work has shown that the **mentalizing network**, including the dorsomedial prefrontal cortex (DMPFC), right temporoparietal junction (RTPJ), left temporoparietal junction (LTPJ), and precuneus [18–21], are involved in: computing and representing mental states [22,23]; incorporating mental states into moral judgments [24–26]; tracking violations of social predictions [27–30]; and updating beliefs about moral intent and character [31,32] [see Box 1 for a review of functions of the Theory of Mind (ToM) network]. We thus propose that reduced neural activity in the ToM network in light of prior-inconsistent information indicates nonrational discounting: a failure to process and draw inferences from evaluatively meaningful evidence. We further propose that enhanced activity in the ToM network following prior-inconsistent information is compatible with auxiliary generation in the service of rational belief maintenance.

Inferring Auxiliary Hypothesis Generation from Mentalizing Activity

We view ToM activity that accompanies belief maintenance as indicating auxiliary generation, given: (i) the role of the ToM network in inferring and representing mental states (Box 1); and (ii) the mentalistic nature of explanations for others' behavior. Crucially, we do not interpret activity in mentalizing regions to indicate domain-general auxiliary generation *per se*; the neural mechanisms for generating explanations may vary by domain (e.g., generating auxiliaries to explain surprising observations of rigid bodies in motion may elicit activity in a wholly different neural network). In the moral domain, however, explanations of others' behavior are dominated by mental state information. Past work has shown that observers tend to explain others' actions by referencing transient mental states such as desires and beliefs [33]. Nondispositional explanations for others' behavior thus involve an interaction between external situations and agents' mental states, rather than situations alone. For example, the auxiliary hypothesis for the tip jar scenario involves both a situation (my friend is out of change) and the relevant mental states (desire and intent to make change).

Box 1. Mentalizing and Bayesian Processing in the Brain

A large body of neuroimaging work has shown that brain regions for ToM, such as RTPJ, are recruited for incorporating mental state information into moral judgments, for example, when forgiving accidents, condemning attempted harms, and withholding praise for accidental help [24–26]. Additionally, ToM regions have been found to encode, in their spatial patterns of activation, distinct kinds of moral intent, such as harmful versus innocent [22], and competitive versus cooperative [23]. Furthermore, recent studies have explored a role for mentalizing regions in processing social PE. These regions show greater activity to behaviors that violate (vs. confirm) prior expectations that are based on: past behavior [27,28], instructed trait knowledge [29], and reward feedback in economic games [30]. Finally, RTPJ activity is associated with behavioral measures of belief updating: increased activity in RTPJ accompanies negative moral judgments of previously fair social partners [31], and worsened impressions of ingroup targets following negative information [32].

Other studies have found updating-related activity outside of typical ToM regions. One fMRI study found that bilateral TPJ was preferentially recruited for tracking mundane changes in behavior, whereas, for diagnostic changes, left ventrolateral prefrontal cortex (VLPFC) and left inferior frontal gyrus (IFG) were preferentially recruited instead [47]. Another study using a reinforcement learning paradigm found that PEs related to the generosity of human and slot machine targets correlated with activity in left VLPFC, in addition to bilateral inferior parietal lobule, posterior cingulate cortex, precuneus, and RTPJ [30].

Computational neuroimaging work has implicated regions in and outside the ToM network in Bayesian processing. One study showed that observers track the volatility (uncertainty) of an advisor's trustworthiness in accordance with estimates from a Bayesian reinforcement learning model, and that these estimates correlate with activity in the anterior cingulate cortex gyrus [48]. Additionally, PE signals in this paradigm correlated with activity in right middle temporal gyrus, right superior temporal sulcus/TPJ, and DMPFC. Another study found that DMPFC parametrically tracks trial-by-trial updating of value judgments in response to social consensus information, where updating was indexed by **Kullback–Leibler (K–L) divergence** [49]. Other work looking at updating of social power hierarchies has implicated the hippocampus and amygdala, in addition to MPFC [50]. Specifically, activity in these regions was correlated with the degree to which participants had to update their power estimates for targets after receiving feedback, as indexed by K–L divergence.

Based on the prior work on behavior explanation [33], we expect any explanation of behavior – not just an auxiliary explanation – to involve mental state inference. Mentalistic explanations can then lead to either an impression update or belief maintenance. We thus expect to observe enhanced mentalizing activity both when belief updating occurs (e.g., we infer that our friend intended to steal the money), and when belief maintenance occurs (e.g., we infer that she intended to make change), compared to when discounting occurs. For this reason, we interpret mentalizing activity that accompanies belief maintenance as likely reflecting the generation of mentalistic alternative causes for surprising information.

In addition, mentalistic auxiliaries can target different agents associated with the event. The observer may reason about the mental states of the protagonist or perpetrator, or the person who recounted the event to the observer (the source). For example, given strong prior beliefs about a friend's trustworthiness, if we learn about contradictory evidence secondhand, we may question the reliability of the source: perhaps they were misinformed, or they intended to mislead us. Evaluating surprising information as less reliable in light of strong prior beliefs can be consistent with Bayesian reasoning, especially under the assumption that the information people typically receive is not totally reliable [34]. Thus strong priors can impact the sufficiency of new information for belief updating, both through inferences about the target, and inferences about the source.

When the source is an identifiable agent, it is likely that producing auxiliary hypotheses concerning reliability will involve mentalizing. When the source is not an agent (e.g., a newspaper), reasoning about the reliability of the source may or may not involve mentalizing, perhaps depending on the extent to which the source is viewed as having moral agency [35]. An important topic for future research is the nature and prevalence of auxiliary hypotheses in the domain of moral updating that do not involve mental state inference (see Outstanding Questions).

Motivated Rationalization and Rational Discounting

A motivated observer can use many different strategies to reach their desired conclusion. They can not only selectively ignore incoming evidence, but also conduct a biased memory search, subjectively select statistical heuristics, and importantly, generate alternative explanations for the observation [4]. Kunda's seminal proposal holds that motivated cognition is characterized by selecting an information processing strategy that will uphold the desired belief. Under this broad construal of motivated cognition, there are two cases where coming up with alternative explanations may be described as motivated. In one case, an individual may deviate substantially from Bayesian rationality (perhaps as described by a computational model) and generate an auxiliary explanation despite it not being warranted by their priors. This raises the question of what degrees of deviation from Bayesian rationality should count as motivated, given that observers are not always optimal Bayesian agents [36]. It is important that future work investigates the extent to which individuals update beliefs in contexts where priors and likelihoods are well defined, to establish descriptive norms for procedural rationality. At present, however, it is difficult to infer precise application of Bayes' rule from ToM activation alone. Our proposal instead affords stronger inferences about nonrational discounting given the absence of ToM activity (but see Box 2 for a discussion of ways to test for rationality).

In the second case, an individual may generate an auxiliary explanation to support a foregone conclusion (i.e., *post hoc* rationalization). This generates a time-course hypothesis: a motivated individual will likely finalize their judgment before coming up with an alternative explanation. We can use functional magnetic resonance imaging (fMRI) to examine the time-course of neural activity for nonrational and rational auxiliary generation: we expect that motivated individuals will exhibit a peak in mentalizing activity after submitting their behavioral response (Box 2).

Box 2. Testing for Rationality

The present proposal affords stronger inferences of nonrational discounting given the absence of ToM activity, than inferences of rational auxiliary generation given the presence of ToM activity. While enhanced mentalizing activity in light of prior-inconsistent information is compatible with rational processing, it is also compatible with motivated rationalization. We propose several paradigms that can help test for rationality.

Motivation may hijack the auxiliary generation process, such that an observer endorses an auxiliary hypothesis despite it not being Bayes optimal. Future work can measure participants' priors and motivations, and take advantage of cases where the two diverge (see [51] for an application of this paradigm to political beliefs). Enhanced ToM activity in response to prior-inconsistent information, but not motivation-inconsistent information, would suggest a specific role for ToM in supporting rational auxiliary generation. In addition, an impression updating task combined with a minimal group manipulation will allow us to test whether ToM is recruited in the absence of the experience or meaningful information that comes with social groups.

A complementary approach is to examine the time-course of neural activity in the ToM network during impression updating. By definition, an individual engaging in *post hoc* rationalization commits to maintaining their belief, prior to coming up with a plausible explanation. We hypothesize that rational auxiliary generation and *post hoc* rationalization will be temporally differentiable in their neural time-course: a rational observer will exhibit a peak in mentalizing activity prior to submitting their behavioral response, while a *post hoc* rationalizer will exhibit the peak after the behavioral response.

On the flip side, a rational updater may discount evidence before considering auxiliary explanations. If prior-inconsistent information about the target is coming from a third-party source (agentic or not) that is known to be highly unreliable (e.g., *National Enquirer*), a rational observer may dismiss that information out of hand. This form of discounting should occur regardless of the content and strength of the prior belief. In cases of extreme source unreliability, the absence of mentalizing activity in response to the prior-inconsistent information cannot be used to infer that nonrational discounting has occurred.

Boundaries of the Present Account

The proposed framework examines the process by which beliefs are updated; here, we note the boundaries of our proposal. First, the designations of rational and nonrational are agnostic with respect to the source of the prior belief. Both priors that are evidence based, and priors that are heavily sourced from affective value can be subjected to Bayesian processing; procedural rationality is orthogonal to the source of the prior. Our mentalizing account does not adjudicate between different possible sources of the prior. Second, the mechanics of how observers generally evaluate evidence, independent of updating, will not be addressed by the current proposal. Third, we limit our account to the updating of personally held beliefs, rather than contexts where one has to publicly defend or sway public opinion on the moral character of an associate (a domain where strategic rationalization is highly expected to occur). Finally, it is important to note that the designations of rational and nonrational describe how the observer processes new evidence during belief updating, rather than whether or not the observer is maximizing their utility as a rational actor. For example, an observer may preserve their prior belief because they have calculated that acting according to an updated belief would be costlier than acting according to a potentially incorrect belief; this can be rational if the metric of rationality is achieving the best trade-off between costs and rewards given a particular set of goals (see [37] for a discussion of cognitive resource considerations for rationality). In the present discussion, however, we focus on procedural rationality during belief updating, and later return to different ways in which belief preservation may be rational.

Belief Maintenance Compatible with Rational Processing

Neural evidence can be harnessed to determine whether ingroup–outgroup asymmetries in belief updating, and subsequent discrepancies in social decisions, are compatible with Bayesian reasoning over different **prior distributions**. In an fMRI investigation of such an asymmetry, participants had the opportunity to punish ingroup and outgroup members who defected against another person in the Prisoner's Dilemma Game [38]. Participants showed increased activity in DMPFC and bilateral TPJ – prominent nodes in the mentalizing network – when deciding to punish an ingroup versus outgroup defector; furthermore, increased connectivity between DMPFC and LTPJ was associated with weaker punishment of ingroup defectors. Here, greater ToM involvement when faced with an ingroup member may reflect the inference of auxiliaries that are consistent with strong positive priors about the ingroup (e.g., perhaps the ingroup member did not intend to defect). Moreover, disrupting RTPJ activity using transcranial magnetic stimulation (TMS) reduced relative forgiveness for ingroup members in the same paradigm, pointing to a causal role for RTPJ – and perhaps mentalizing – in parochial punishment and forgiveness [39].

We tend to have stronger beliefs about the positive traits of close others than of strangers. In a recent study using the Ultimatum Game, unfair offers from romantic partners elicited greater activity in MPFC compared to unfair offers from strangers, and participants who exhibited lower levels of MPFC–DACC (dorsal anterior cingulate cortex) functional connectivity were likelier to accept unfair offers from their partner than from a stranger [40]. The authors suggest that: (i) participants likely engaged in greater mentalizing to make sense of the intentions behind their partners' surprising actions; and (ii) decreased coupling between conflict-related signals in DACC and mentalizing-related activity in MPFC may enhance prosocial responses toward close others.

Overall, these findings indicate that, when observers hold strong positive prior beliefs about an agent, they engage in greater mentalizing in light of negative information, potentially to generate a nondispositional explanation of the surprising behavior. This is the case whether the beliefs are formed through direct experience, as in the case of close relationships, or are suggested by group membership.

Belief Maintenance Following Nonrational Discounting

Disengagement from mentalizing can shield the observer from an unfavorable character inference. One study found that decreased bilateral TPJ activity was associated with a failure to worsen impressions of ingroup members in response to negative information [32]. If the participants in this study were invoking auxiliaries to account for negative information, they would have mentalized more about ingroup targets; instead, they may have disengaged from mentalizing to preserve their positive prior beliefs. Participants in this study who were able to overcome ingroup bias in updating engaged the lateral prefrontal cortex and dorsal anterior cingulate cortex, in addition to TPJ and precuneus. This suggests that the mentalizing and executive control networks can work in concert to overcome the tendency to discount unfavorable information (see [41] for more on exerting control to overcome effortless group bias).

Another study tested the effect of prior record on moral intent inference [31]. Participants first played an economic game with partners who were fair or unfair, before learning about the players' harmful actions in new contexts of ambiguous intent. When a partner who was previously fair (vs. unfair) was described as performing a harmful action, the action was judged as both less intentional and less blameworthy. Crucially, these judgments were associated with decreased activation in RTPJ. Here, participants may have discounted the negative outcomes of actions performed by fair partners by disengaging from mentalizing.

In a similar vein, a recent fMRI study explored moral impression updating for friends versus strangers during an economic game. Scan participants witnessed their close friend or a stranger give money to or take money from them over multiple rounds, and rated each partner on trustworthiness and closeness after each round [42]. Overall, participants updated their partner ratings based on both initial ratings, and trial-by-trial valuations derived from a computational model. When comparing the friend-taking and stranger-taking conditions, we found: (i) greater negative **prediction error** (PE) for friends than strangers; (ii) reduced RTPJ activity in response to friends than strangers; and (iii) less negative updating for friends than strangers. We also found that, within the friend-taking condition, increased RTPJ activity was associated with both greater negative PE and increased negative updating. These results suggest that, across conditions, a failure to mentalize in light of negative PE about a friend leads to the maintenance of positive beliefs. To the extent that RTPJ was activated on friend-taking trials; however, participants were able to combine the new information with initial impressions to effectively update. Thus, in this paradigm, participants on average engaged in nonrational discounting for their friends, but also engaged in a limited form of updating in response to PE.

The evidence indicates that, in light of information that affords unfavorable belief updates for ingroup members or previously prosocial partners, observers sometimes discount information about these targets to avoid drawing any unfavorable inferences. This is a nonrational process, in that it fails to account for potentially meaningful new evidence that would have otherwise prompted either a belief update or the generation of an auxiliary hypothesis.

Ultimate Rationality

What are the ultimate benefits of rational and nonrational processing? The chief function of rational belief updating is that an observer can use their beliefs to make informed decisions about social partners and interactions. Nonrational belief updating, by contrast, can lead to less accurate beliefs, although we suggest that it may still confer fitness benefits outside the realm of calculated decision-making. Rational and nonrational belief updating, then, can be seen as two paths to fitness.

Predictive Value of Rational Belief Maintenance

While both Bayesian auxiliary generation and nonrational discounting result in maintenance of the prior belief, the former allows the observer to hold more accurate beliefs about reality. The act of inferring an auxiliary hypothesis entails the possibility of adjusting the prior belief; that is, a rational observer must first weigh the prior probabilities and likelihoods of both the central belief and candidate auxiliary beliefs, before invoking an auxiliary explanation. If a candidate auxiliary hypothesis cannot sufficiently account for the new observation, rational inference holds that it be discarded in favor of adjusting the prior belief. If an auxiliary hypothesis is invoked at the end of such a process, the observer then possesses a set of beliefs that is likelier to predict new information well. By contrast, if an observer opts out of auxiliary generation, they also bypass a checkpoint for re-evaluating and potentially adjusting their prior belief. Deactivating the rational route impairs the ability to make judicious social decisions regarding whom to approach and befriend, and with whom to cooperate.

Bayesian belief processing, however, does not guarantee accurate beliefs. As mentioned above, procedural rationality of updating is orthogonal to the source of the prior belief, which may be more affect based than evidence based. For instance, someone who is strongly prejudiced against outgroup members may hold such beliefs largely because there are perceived affective benefits to interacting with similar others [43]. Such strong negative priors about the outgroup may then undergo Bayesian processing in light of disconfirmatory evidence, but procedural rationality in this case is unlikely to yield a set of beliefs with predictive accuracy, since the priors were not evidence based to begin with. There may be some limited predictive value in this instance, in that the observer's predictive model of other similar observers may facilitate interactions with them, but their model of agents will still be inaccurate.

Socioaffective Value of Nonrational Discounting

Nonrational discounting can still be ultimately rational, in that it can increase social fitness and actual economic benefits (for a related discussion on the adaptive function of nonrational influences on behavior, see [44]). In contexts where one's social identity is made salient, such as in hyperpartisan environments, a desire for benefits conferred by maintaining good standing in the social group, such as a sense of belonging and status, can override accuracy goals, promoting the adoption of false beliefs held by the group [43]. Additionally, social reward signals in the brain may support the maintenance of close relationships, even at the cost of rational economic decision-making. For example, one study using the Trust Game found that participants trusted their friend more than a stranger, despite an equal reciprocation rate of 50% from both friend and stranger [2]. Computational modeling of neural data indicated that trust decisions were predicted by a social value reward signal in ventral striatum (a reward-implicated region) in response to reciprocation from friends. Although participants' repeated decisions to trust their friend were financially suboptimal, these decisions were associated with a social reward signal, which may encourage the maintenance of close relationships in real life. Moreover, seemingly nonrational decisions do not always lead to suboptimal economic outcomes. A recent study found that, in an economic game, when given the opportunity to calculate the future costs of cooperation, players who decided to cooperate without considering the costs were reciprocated more often [45]. Finally, another recent study showed that people who were more reluctant to update their impressions of a friend after learning negative information about them, and those who showed relatively less RTPJ activity in response to their friend's negative behavior, reported having more friends in real life [46]. These effects persisted even after controlling for participants' prior experiences with their friend, including how much they communicate with their friend and how long they have known their friend. In addition, prior experience did not explain the degree of impression updating, suggesting that motivated impression updating accounted for better social outcomes, above and beyond the effect of prior knowledge. These findings indicate that nonrational processing may promote the maintenance of broad social networks, and the material benefits of having close collaborators that trust us.

Concluding Remarks and Future Directions

We propose a mentalizing account of belief maintenance, which posits that protecting strong priors by invoking an auxiliary hypothesis requires more mentalizing about the target than nonrational discounting. This framework invites a unique contribution for neural evidence in adjudicating between belief maintenance that is nonrational, and belief maintenance that is compatible with Bayesian reasoning over strong priors. The Bayesian route is likelier to promote beliefs with greater predictive utility, as it provides a potential checkpoint for adjusting prior beliefs; nonrational discounting can help preserve positive social relationships, which may come with material benefits.

Numerous interesting questions remain (see Outstanding Questions). Overall, the proposed framework calls for the use of neural evidence to revisit phenomena that have been taken for granted as motivated, and a more nuanced examination of how different forms of belief maintenance can be proximately or ultimately adaptive.

Acknowledgments

The authors thank Stefano Anzellotti, Adam Bear, Fiery Cushman, Simon Karg, Josh Hirschfeld-Kroen, Josh Knobe, Justin Martin, Ryan McManus, Shaun Nichols, Peter Railton, Rebecca Saxe, Ben Tappin, attendees of the 2019 Reason Workshop at Boston College, and members of the Boston College Morality Laboratory for their helpful comments and feedback; Josh Hirschfeld-Kroen for assistance with figure creation; and Jordyn Mason for assistance with manuscript preparation. The authors acknowledge the support of the John Templeton Foundation.

Outstanding Questions

Individual and contextual features. What are individual differences (e.g., in mentalizing ability, cognitive reflection) that predict Bayes-compatible versus nonrational belief maintenance? How does motivational context (e.g., prediction and affiliation) affect belief updating? How do manipulated priors for unfamiliar targets differ from priors (based on experience) for familiar targets? How do observers update impressions of minimal groups? Are there comparable mechanisms for belief updating outside the social domain?

Boundaries of the mentalizing account. Are there auxiliary hypotheses in the domain of moral updating that do not involve mental state inference? How are nonagentic sources of social information treated? Do motivated rationalizers ever generate auxiliaries before arriving at a judgment?

Updating in the brain. What can the neural time-course of mentalizing activity tell us about *post hoc* rationalization? Are there distinct roles for each ToM region during updating, or can some regions take on multiple roles (e.g., mentalizing, trait updating, and PE tracking)? Are there univariate/multivariate differences in ToM activity for different types of auxiliaries, such as inferring a trait-concordant intent versus doubting the reliability of the source? How does the ToM network interact with other PE regions, and with other brain networks, such as the reward learning and cognitive control networks, during updating?

Mechanics of updating. When do we update the uncertainty versus the mean value of belief distributions? How are multiple instances of the same information handled, especially if the first instance did not lead to an update in mean value? To what degrees are uncertainty and mean value influenced by the diagnosticity of each observation, and the number of concordant observations over time?

References

1. Richards, Z. and Hewstone, M. (2001) Subtyping and subgrouping: processes for the prevention and promotion of stereotype change. *Personal. Soc. Psychol. Rev.* 5, 52–73
2. Fareri, D. *et al.* (2015) Computational substrates of social value in interpersonal collaboration. *J. Neurosci.* 35, 8170–8180
3. Monroe, A.E. and Malle, B.F. (2019) People systematically update moral judgments of blame. *J. Pers. Soc. Psychol.* 116, 215–236
4. Kunda, Z. (1990) The case for motivated reasoning. *Psychol. Bull.* 108, 480
5. Tajfel, H. and Turner, J. (1979) An integrative theory of intergroup conflict. In *The Social Psychology of Intergroup Relations* (Austin, W.G. and Worchel, S., eds), pp. 33–37, Brooks/Cole
6. Crocker, J. and Luhtanen, R. (1990) Collective self-esteem and ingroup bias. *J. Pers. Soc. Psychol.* 58, 60–67
7. Greenwald, A.G. and Pettigrew, T.F. (2014) With malice toward none and charity for some: ingroup favoritism enables discrimination. *Am. Psychol.* 69, 645–655
8. Kunda, Z. and Oleson, K.C. (1995) Maintaining stereotypes in the face of disconfirmation: constructing grounds for subtyping deviants. *J. Pers. Soc. Psychol.* 68, 565–579
9. Sidanius, J. and Pratto, F. (1999) *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression*, Cambridge University Press
10. Asch, S.E. (1946) Forming impressions of personality. *J. Abnorm. Soc. Psych.* 41, 258–290
11. Reeder, G.D. and Spores, J.M. (1983) The attribution of morality. *J. Pers. Soc. Psychol.* 44, 736
12. Reeder, G.D. and Brewer, M.B. (1979) A schematic model of dispositional attribution in interpersonal perception. *Psychol. Rev.* 86, 61–79
13. Pettrigrew, T.F. (1979) The ultimate attribution error: extending Allport's cognitive analysis of prejudice. *Pers. Soc. Psychol. B.* 5, 461–476
14. Storms, M.D. (1973) Videotape and the attribution process: reversing actors' and observers' points of view. *J. Pers. Soc. Psychol.* 27, 165–175
15. Sagar, H.A. and Schofield, J.W. (1980) Racial and behavioral cues in Black and White children's perceptions of ambiguously aggressive acts. *J. Pers. Soc. Psychol.* 39, 590–598
16. Gershman, S.J. (2019) How to never be wrong. *Psychon. Bull. Rev.* 26, 13–28
17. Siegel, J.Z. *et al.* (2018) Beliefs about bad people are volatile. *Nat. Hum. Behav.* 2, 750
18. Saxe, R. and Kanwisher, N. (2003) People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". *Neuroimage.* 19, 1835–1842
19. Saxe, R. *et al.* (2004) Understanding other minds: linking developmental psychology and functional neuroimaging. *Annu. Rev. Psychol.* 55, 87–124
20. Saxe, R. and Wexler, A. (2005) Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia* 43, 1391–1399
21. Molenberghs, P. *et al.* (2016) Understanding the minds of others: a neuroimaging meta-analysis. *Neurosci. Biobehav. Rev.* 65, 276–291
22. Koster-Hale, J. *et al.* (2013) Decoding moral judgments from neural representations of intentions. *Proc. Natl. Acad. Sci. U. S. A.* 110, 5648–5653
23. Tsoi, L. *et al.* (2016) Distinct neural patterns of social cognition for cooperation versus competition. *Neuroimage* 137, 86–96
24. Young, L. and Saxe, R. (2009) An fMRI investigation of spontaneous mental state inference for moral judgment. *J. Cogn. Neurosci.* 21, 1396–1405
25. Young, L. *et al.* (2010) Investigating the neural and cognitive basis of moral luck: it's not what you do but what you know. *Rev. Philos. Psychol.* 1, 333–349
26. Young, L. *et al.* (2011) Neural evidence for "intuitive prosecution": the use of mental state information for negative moral verdicts. *Soc. Neurosci.* 6, 302–315
27. Mende-Siedlecki, P. *et al.* (2013) Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *J. Neurosci.* 33, 19406–19415
28. Dungan, J.A. *et al.* (2016) Theory of mind for processing unexpected events across contexts. *Soc. Cogn. Affect. Neurosci.* 11, 1183–1192
29. Heil, L. *et al.* (2019) Processing of prediction errors in mentalizing areas. *J. Cogn. Neurosci.* 31, 900–912
30. Hackel, L.M. *et al.* (2015) Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nat. Neurosci.* 18, 1233–1235
31. Kliemann, D. *et al.* (2008) The influence of prior record on moral judgment. *Neuropsychologia* 46, 2949–2957
32. Hughes, B.L. *et al.* (2017) Motivation alters impression formation and related neural systems. *Soc. Cogn. Affect. Neurosci.* 12, 49–60
33. Malle, B.F. (2001) Folk explanations of intentional action. In *Intentions and Intentionality: Foundations of Social Cognition*, pp. 265–286, MIT Press
34. Tappin, B.M. and Gadsby, S. (2019) Biased belief in the Bayesian brain: a deeper look at the evidence. *Conscious. Cogn.* 68, 107–114
35. Gray, K. *et al.* (2012) Mind perception is the essence of morality. *Psychol. Inq.* 23, 101–124
36. Hackel, L.M. and Amodio, D.M. (2018) Computational neuroscience approaches to social cognition. *Curr. Opin. Psychol.* 24, 92–97
37. Lieder, F. and Griffiths, T.L. (2019) Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* Published online February 4, 2019. https://doi.org/10.1017/S0140525X1900061X
38. Baumgartner, T. *et al.* (2012) The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Hum. Brain Mapp.* 33, 1452–1469
39. Baumgartner, T. *et al.* (2014) Diminishing parochialism in intergroup conflict by disrupting the right temporo-parietal junction. *Soc. Cogn. Affect. Neurosci.* 9, 653–660
40. Fatfouta, R. *et al.* (2018) Accepting unfairness by a significant other is associated with reduced connectivity between medial prefrontal and dorsal anterior cingulate cortex. *Soc. Neurosci.* 13, 61–73
41. Hughes, B.L. and Zaki, J. (2015) The neuroscience of motivated cognition. *Trends Cogn. Sci.* 19, 62–64
42. Park, B. *et al*. How theory-of-mind brain regions process prediction error across relationship contexts. *Soc. Cogn. Affect. Neurosci.* (invited revision)
43. Van Bavel, J.J. and Pereira, A. (2018) The partisan brain: an identity-based model of political belief. *Trends Cogn. Sci.* 22, 213–224
44. Cushman, F. (2018) Rationalization is rational. *Behav. Brain Sci.* 28, 1–69
45. Jordan, J.J. *et al.* (2016) Uncalculating cooperation is used to signal trustworthiness. *Proc. Natl. Acad. Sci. U. S. A.* 113, 8658–8663
46. Park, B. and Young, L. (2020) An association between biased impression updating and relationship facilitation: a behavioral and fMRI investigation. *J. Exp. Soc. Psychol.* 87. Published online November 23, 2019. https://doi.org/10.1016/j.jesp.2019.103916
47. Mende-Siedlecki, P. and Todorov, A. (2016) Neural dissociations between meaningful and mere inconsistency in impression updating. *Soc. Cogn. Affect. Neurosci.* 11, 1489–1500
48. Behrens, T.E. *et al.* (2008) Associative learning of social value. *Nature* 456, 245
49. De Martino, B. *et al.* (2017) Social information is integrated into value and confidence judgments according to its reliability. *J. Neurosci.* 37, 6066–6074
50. Kumaran, D. *et al.* (2016) Computations underlying social hierarchy learning: distinct neural mechanisms for updating and representing self-relevant information. *Neuron* 92, 1135–1147
51. Tappin, B.M. *et al.* (2017) The heart trumps the head: desirability bias in political belief revision. *J. Exp. Psychol. Gen.* 146, 1143