An examination of accurate versus “biased” mentalizing in moral and economic decision-making

BoKyung Park, Minjae Kim, and Liane Young

Boston College

KEY WORDS: Mentalizing, Moral judgment, Economic decision-making

Correspondence should be addressed to BoKyung Park (parkanj@bc.edu) and Minjae Kim (minjae.kim@bc.edu).

Abstract

In this chapter, we review evidence that supports the engagement of mentalizing for optimal and suboptimal decision-making associated with social and/or non-social rewards. Mentalizing can support optimal decisions in the contexts of moral judgment and social interaction. However, mentalizing can also be influenced by apparently irrelevant factors, including agents' moral history and group status. We suggest that, while these factors can "bias" mentalizing, resulting in suboptimal decisions, some seemingly biased decisions may in fact be procedurally rational. For example, when people discount new negative evidence about someone they have a strong positive impression of, they may nevertheless be engaging in procedurally rational Bayesian updating. We suggest that Bayesian-rational updating involves generating ad hoc auxiliary hypotheses to explain away inconsistent evidence, which may be supported by mentalizing-related neural activity.

Value-based decision-making, especially in social contexts, depends critically on the ability to think about agents' mental states, i.e., mentalizing or Theory of Mind (ToM). Although the involvement of mentalizing in different decision-making processes, such as moral judgment and economic exchange, is generally acknowledged, whether mentalizing leads to optimal or suboptimal decisions is a relatively open question. Accurate mentalizing leads to optimal decisions that maximize immediate or future benefits, including learning from others, defeating others, evaluating others, and predicting others. Yet, mentalizing is also vulnerable to "bias"; mentalizing is affected by a number of ostensibly irrelevant factors, including the identity and group status of the interacting agents, the mentalizer's own beliefs and values, and other contextual factors. We suggest that, in these cases, mentalizing can lead to suboptimal decisions. In the last section of this chapter, we revisit cases of mentalizing that appear to be biased, taking ingroup bias as a case study, and we suggest that a subset of these cases may be compatible with rational Bayesian reasoning.

Thus, in this chapter, we review cases in which mentalizing supports both optimal and suboptimal value-based decisions, in the domains of moral judgment and economic exchange. We will also examine how seemingly biased mentalizing and subsequent suboptimal decisions may in fact arise from a rational procedure.

Mentalizing network

Before we discuss the role of mentalizing in moral and economic decision-making, we briefly summarize research on the network of brain regions that support mentalizing, also known as the ToM network. Decades of work, using functional magnetic resonance imaging (fMRI) and event-related potential (ERP) methods, points to several key nodes: the right temporo-parietal junction (rTPJ; often labeled as posterior superior temporal cortex, inferior parietal lobule, or brodmann area 39), left temporo-parietal junction (lTPJ), posterior superior temporal sulcus (pSTS), dorsomedial prefrontal cortex (dmPFC) , and precuneus (Decety & Cacioppo, 2012;

Saxe, 2009; Saxe & Powell, 2006; Saxe, Carey, & Kanwisher, 2004; Saxe & Kanwisher, 2003; Saxe, Xiao, Kovacs, Perrett, & Kanwisher, 2004; Saxe & Wexler, 2005). Recent empirical and theoretical evidence suggests that these regions support mentalizing by, to some extent, encoding social prediction error, i.e., they respond preferentially to unexpected agent behaviors (Koster-Hale & Saxe, 2013). Other work reveals that other sub-regions of the medial prefrontal cortex (mPFC), including ventromedial prefrontal cortex (vmPFC), anterior cingulate cortex (ACC), and adjacent paracingulate cortex are also recruited for mentalizing (Amodio & Frith, 2006; Frith & Frith, 2006; Krueger, Grafman, & McCabe, 2008; Lombardo et al., 2009; Walter et al., 2004). Of these regions, the MPFC and bilateral TPJ emerged as consistently activated across ToM tasks in a massive activation likelihood estimation (ALE) meta-analyses of 144 datasets (3150 participants) (Molenberghs, Johnson, Henry, & Mattingley, 2016).

Accurate mentalizing leads to optimal decisions

Evidence demonstrates that accurate mentalizing can result in immediate rewards, such as earning money, or more distant rewards, such as identifying future cooperators versus competitors.

First, given the primacy of moral signals in impression updating compared to other trait information (Brambilla, Carraro, Castelli, & Sacchi, 2019; Goodwin, 2015), moral judgment–evaluating whether an agent's behavior is right or wrong–is crucial for identifying potential friend versus foe, and maximizing future social benefits. A large body of previous research has identified the key role of mentalizing regions, and specifically the rTPJ, in the formation and revision of moral judgments (e.g., Decety & Cacioppo, 2012; Young, Cushman, Hauser, & Saxe, 2007). Specifically, rTPJ activity is consistently recruited for intent-based moral judgments, including: forgiving accidents (innocent intent), condemning failed attempts to harm (malicious intent) (Young, Nichols, & Saxe, 2010; Young & Saxe, 2009), and even withholding praise for unintentionally helpful behaviors (Young, Scholz, & Saxe, 2011). Spatial patterns of activity in

rTPJ discriminate between intentional and accidental harms, and also correlate with moral judgments, though this pattern discrimination is absent in high-functioning adults with autism (Koster-Hale, Saxe, Dungan, & Young, 2013). Moreover, mentalizing supports the integration of mitigating intent information even for extreme harms (e.g., killing one's wife to relieve her suffering); reduced punishment was associated with increased rTPJ activity (Yamada et al., 2012). Other work has suggested that forgiving accidents may involve suppressing emotional responses to negative outcomes, indexed by greater coupling between the mentalizing network activity and amygdala activity in response to unintended harms (Treadway et al., 2014).

Convergent transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS) studies have shown that modulating rTPJ activity leads to systematically different moral judgments, establishing a causal role for the rTPJ in mentalizing for moral judgment. Disrupting rTPJ activity using TMS leads to more outcome-based moral judgments (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010), whereas increasing the excitability of the rTPJ with tDCS leads to more intent-based moral judgments (Sellaro et al., 2015).

Developmental work reveals that young children aged 3 to 4 years, who lack mature mentalizing capacities, fail to incorporate the intention information and make more outcome-based moral judgments as well (Killen, Mulvey, Richardson, Jampol, & Woodward, 2011; see also Cushman, Sheketoff, Wharton, & Carey, 2013). The ability to integrate mental state information with other task-relevant information for moral judgment is supported by developing neural circuitry, including the rTPJ, for mentalizing (Gweon, Dodell-Feder, Bedny, & Saxe, 2012; Richardson, Lisandrelli, Riobueno-Naylor, & Saxe, 2018).[1] Thus, accurately inferring the mental

[1] Other work has focused not on the inference of mental states but the inference of moral traits, including generosity. In one study, trait generosity (i.e., proportion of money an agent offered) was encoded separately from total reward provided by the agent, in the rTPJ (Hackel, Doll, & Amodio, 2015). Partner choice decisions relied primarily on trait generosity.

states of others helps people to appropriately assign blame and praise to agents, contributing to future cooperative relationships.

In addition to its role in identifying social partners with whom it is a good idea to cooperate, mentalizing can also be a useful tool in competitive contexts – situations in which we have to figure out what other agents are thinking in order to predict and outsmart them (Singer & Fehr, 2005). In one study, when participants were asked to play a game against strategic human partners, greater engagement of MPFC was associated with better game performance (Coricelli & Nagel, 2008). RTPJ was also recruited when participants defected against their partners to earn greater profits (Bhatt, Lohrenz, Camerer, & Montague, 2010), suggesting a role for mental state inference in strategizing against other agentic opponents.

Meanwhile, when people engage in repeated interactions with the same partner, they must track their partner's actions and select optimal interaction strategies – processes that also depend on mentalizing (Lee, 2006). Ample evidence shows that the mentalizing network supports the processes by which people update their representations of others' personality traits when their behaviors change meaningfully over time (Baron, Gobbini, Eengell, & Todorov, 2011; Mende-Siedlecki, Baron, & Todorov, 2013; Mende-Siedlecki, Cai, & Todorov, 2013; Thornton & Mitchell, 2018; see also Mende-Siedlecki & Todorov, 2016).

Additional research has investigated more complex contexts in which people had to interactively revise their own behavior in response to the behavior of other agents who could impact their outcomes (Hampton, Bossaerts, & O'Doherty, 2008). Participants were paired with a partner and took turns playing as the "employee," who could either work or shirk, or the "employer," who could either inspect or not inspect ("inspection game"). The employee earned money when she shirked and the employer did not inspect, or when she worked and the employer did inspect. In contrast, the employer earned a reward when she did not inspect and the employee worked, or when she inspected and the employee shirked. Thus, accurate prediction of the partner's next action should be based on the history of her choices, as well as

the fact that the participant's own action can in turn modify the partner's behavior. The researchers found that activity in the STS tracked updating of the partner's strategy based on this computation. Moreover, MPFC and ventral striatum were conjointly recruited and highly correlated with STS; these two regions encoded different components of expected reward from the interactions. These findings suggest that this interactive network supports the revision of decisions based on action valuation and mentalizing. By coordinating these different networks, this process potentially generates more fine-grained representations of opponents' future actions. A recent study using a similar inspection game paradigm probed the causal impact of mentalizing on interactive updating using TMS (Hill et al., 2017). When participants whose rTPJ was disrupted by TMS played as the employee, they failed to consider the causal link between their own actions and the employer's future behavior.[2]

Accurately forming and revising representations of another agent's mental states is also important in cases where that agent has privileged access to useful information. One example is when a decision-maker considers advice from others ("advisors"). Decisions to follow the advice or not depend on the advisor's mental states, including her intention and expertise (Harvey & Fischer, 1997; Klucharev, Smidths, & Fernández, 2008; Schilbach, Eickhoff, Schultze, Mojzisch & Vogeley, 2013; Van Swol, 2009; Yaniv & Kleinberger, 2000). In one study, participants were asked to choose one of two fixed options that would likely return greater scores (Behrens, Hunt, Woolrich & Rushworth, 2008). An ostensible advisor gave advice to participants across trials, but the advisor's goal was to ensure that participants score within a certain limited range, not to maximize the score. Thus, participants needed to keep track of two different elements based on each outcome: the predicted scores associated with each of the two choice options, and the

[2] However, another body of research failed to find that repeated interaction with partners necessarily involves mentalizing. These studies focused on the involvement of the reward-processing circuitry, or interpreted activity of regions that constitute both mentalizing and reward valuation networks (e.g., MPFC) in the light of reward computation. See Delgado, Frank, & Phelps, 2005; Fareri, Chang, & Delgado, 2015; Izuma, Saito, & Sadato, 2008; King-Casas et al., 2005; and Phan, Sripada, Angstadt, & McCabe, 2010, as examples.

current intention of the advisor as a function of their current score. While participants' reward computations (in a reinforcement learning model) were tracked by reward-processing regions (ventral striatum and vmPFC), critically, computations of the advisor's intention were tracked by nodes in the mentalizing network (dmPFC and rTPJ). The two information sources were combined in vmPFC. Thus, to maximize personal value, participants recruited different brain regions, including the mentalizing network, to update representations of the reward and the advisor. In another study, participants made predictions about whether an asset value would increase or decrease on some trials (Boorman, O'Doherty, Adolphs, & Rangel, 2013). More importantly, on other trials, there were advisors who made the predictions about asset value, and participants had to bet for or against the advisor before the prediction was revealed; they earned a bonus for correct bets. Thus, tracking the expertise of the advisors was important. Over the course of the game, participants could use the feedback to form and update beliefs about the expertise of the advisors. When participants learned about the expertise of the advisor, based on whether the advisor's choice agreed with their own prediction or not, they recruited the mentalizing network, including rTPJ and dmPFC. Thus, mentalizing regions may have supported participants' capacity to generate accurate representations of the advisor's expertise.

Similarly, people can reflect on others' choices to infer relative values of available options to maximize their own rewards. An interesting case of this is how people navigate fluctuating stock markets. A certain stock price rises, and people have to infer what other traders think. Does this pattern reflect the stock's real value, or arbitrary noise from other sources? Bruguier and colleagues (2010) found that mentalizing can support optimal decisions in these contexts, especially when other traders in the market were known to have better access to critical information ("insiders"). When there were insiders who knew the specific dividends of stocks in the market, price changes of stocks in these markets were a diagnostic tool to estimate their dividends. Participants were informed whether there were insiders in the current

market or not, and chose the number of shares of different types of stocks. Of key relevance, mentalizing regions, including the medial paracingulate cortex, were recruited more when insiders were present than absent. Participants' mentalizing ability was independently measured in two separate tasks: where they (1) predicted agentic movement of shapes and (2) inferred mental states from eye gazes. Participants who were better at mentalizing performed better in forecasting market trades, supporting the argument that accurate mentalizing for inferring insider strategies helps participants navigate the financial market more successfully (Bruguier, Quartz, & Bossaerts, 2010).

To conclude this section, we reviewed evidence that mentalizing is critical for evaluating moral agents, including harmful agents, for predicting how competitors will behave, and for learning from those who have special information about a shared environment (e.g., stock markets). Thus, mentalizing can allow decision-makers to maximize profits in direct economic interactions, and to identify future cooperative partners through third-party observations. While mentalizing is essential for social decision-making, mentalizing can also go astray, as we will see in the next section.

Inaccurate mentalizing leads to suboptimal decisions

Although mentalizing supports successful social interaction, mentalizing is also susceptible to influence by factors that may be irrelevant to the decision at hand, including the moral character or group status of the target. Consequently, people may assign more or less blame than is warranted, leading to inaccurate identification of future friend or foe.

Prior research has manipulated participants' prior moral impressions of targets to investigate the impact on mentalizing. In one study, participants first interacted with fair and unfair agents; they then read vignettes describing good and bad actions presented as performed by the agents (Kliemann, Young, Scholz, & Saxe, 2008). Participants showed reduced rTPJ activity when a previously *fair* agent caused a negative outcome, compared to

when a previously *unfair* agent caused a negative outcome. Furthermore, fair agents were judged as less blameworthy for causing negative outcomes, and these actions were judged as less intentional, compared the same negative outcomes caused by unfair agents. Together, these behavioral and neural patterns suggest that participants disengaged from intent attribution for previously fair agents. Consistent with this idea, a recent behavioral study revealed that when people initially had optimistic impressions of a financial advisor's expertise, they preferentially incorporated positive information about the advisor's accuracy and took the advisor's advice more than they should have given actual feedback (Leong & Zaki, 2017). Critically, when participants' initial impressions were directly manipulated to be more well-calibrated, the optimism bias went away. These findings suggest that when people make initially optimistic judgments about experts, they preferentially discount new evidence that is inconsistent with these judgments.

Another body of research suggests that salient negative outcomes of an agent's behavior can distort mentalizing processes. In one set of studies, participants were presented with vignettes describing a CEO causing environmental damage as a side effect of a new business policy (Knobe, 2003). Importantly, the CEO stated explicitly that he did not intend to cause environmental harm; however, participants who treated the environment as "sacred" perceived the harm as more intentional compared to participants who did not (Tannenbaum, Ditto, & Pizzaro, 2008; Ditto, Pizarro, & Tannenbaum, 2009). Thus, morally unacceptable outcomes might lead participants to overestimate harmful intent. In broadly consistent studies, people blame agents who benefit from uncontrollable negative events, i.e., agents who bet on natural disasters, or agents who are forced to harm an enemy (Inbar, Pizarro & Cushman, 2012; Woolfolk, Doris, & Darley, 2006; see also Pizarro, Uhlmann, & Bloom, 2003; Pizarro, Uhlmann, & Salovey, 2003).

Similar findings have emerged when participants themselves are actually impacted by bad or good behaviors. In one study, participants responded to offers from a partner in the

Ultimatum Game, who either was forced to make an unfair offer or could choose between fair and unfair distributions (Güroğlu, van den Bos, Rombouts, & Crone, 2010). The researchers found that participants engaged in greater mentalizing, as indexed by greater rTPJ responses, when they rejected forced unfair offers compared to intended unfair offers. This finding suggests that people might justify their blame of faultless others by over-attributing harmful intent, while no mentalizing effort was required to reject the unambiguously intentional unfair offers.

Convergent evidence comes from studies examining the influence of group membership on mentalizing. Specifically, research has found that participants discount ingroup members' negative behaviors (e.g., taking money from the participants; heckling a speaker during a talk) and thus fail to negatively update their impressions. In recent work, overcoming this bias to negatively evaluate a close friend (Park, Fareri, Delgado, & Young, in prep) or an ingroup member (Hughes, Zaki, & Ambady, 2017) was accompanied by recruitment of brain regions associated with mentalizing, including bilateral TPJ and ACC. A similar study examined this effect behaviorally. When participants were presented with an outgroup member's negative behavior first, and intention information later, they increased blame for intentional harms to a greater degree than they reduced blame for unintentional harms. However, for ingroup members, participants used exacerbating and mitigating intent information symmetrically to assign blame (Monroe & Malle, 2019), suggesting that participants might engage in mentalizing more readily when they encounter the negative behavior of outgroup members versus ingroup members. Thus, group membership across diverse contexts influences when and how people engage in mentalizing, leading to occasionally inaccurate moral judgments.

The evidence reviewed thus far shows that people can disengage from mentalizing about targets for whom they have positive prior impressions, resulting in mitigated blame and reduced impression updating. However, another body of work indicates that *greater* mentalizing can also facilitate forgiveness and cooperation (Hare, Camerer, Knoepfle, O'Doherty, & Rangel, 2010; Krueger et al., 2007; Strang, Utikal, Fischbacher, Weber, & Falk, 2014; Will, Crone, &

Güroğlu, 2015). Indeed, some studies found that greater mentalizing for ingroup members was associated with blame mitigation. Specifically, in one study, participants had the opportunity to punish ingroup and outgroup members who defected against another person in the Prisoner's Dilemma Game (Baumgartner, Götte, Gügler, & Fehr, 2012). When participants were presented with an ingroup defector, they showed increased activity in dmPFC and bilateral TPJ, reflecting an attempt to infer the intentions behind the defection. Moreover, increased connectivity among the nodes of the mentalizing network was associated with weaker punishment of ingroup members. Furthermore, disrupting rTPJ activity using TMS reduced forgiveness for ingroup members (Baumgartner, Schiller, Rieskamp, Gianotti, & Knoch, 2014). Another study found that the greater dmPFC activity participants showed when they played the Prisoner's Dilemma Game with ingroup compared to outgroup members, the likelier they were to cooperate more with ingroup than outgroup members in the game (Rilling, Dagenais, Goldsmith, Glenn, & Pagnoni, 2008).

Thus, the act of mentalizing can lead to seemingly opposite consequences: exacerbating and mitigating blame (or decreasing and increasing cooperation). We will revisit this puzzle in the final section, where we explore whether these processes reflect rational versus motivated cognition. But, in either case, mentalizing serves the same purpose of preserving pre-existing impressions. For now, we note that mentalizing is vulnerable to the influence of irrelevant factors, which can lead to biased judgments and perhaps inaccurate action predictions.

Finally, inaccurate, biased mentalizing can also result in concrete financial losses. People often rely on others' mental states to infer potential reward from future decisions, such as seeing other customers' response to their food in a restaurant. Depending on the accuracy of the mental state inference, the value represented in one's mind may not reflect the real intrinsic value. A group of researchers tested this possibility in a paradigm that extended the target of mental state inference to a whole group of agents. Participants viewed experimental asset prices, some of which were inflated beyond their intrinsic value by crowds in the market, i.e.,

financial bubbles (De Martino, O'Doherty, Ray, Bossaerts, & Camerer, 2013). The researchers found that, compared to the non-bubble markets, in the bubble markets where participants had to infer intentions of other traders, the computed values of participants' current possession--reflecting the inflated value of their assets--were parametrically tracked by increased dmPFC activity as well as vmPFC activity. Moreover, there was greater functional coupling between dmPFC and vmPFC in the bubble market, and greater vmPFC activity was ultimately associated with greater likelihood of following the crowd in bubble markets. This pattern suggests that the computed intentions of other traders, reflected in dmPFC, was projected to vmPFC, a region associated with reward computation, perhaps leading participants to overestimate the role of intent in the rise of prices. Consequently, these participants purchased assets at high prices and ultimately earned less. Thus, observers who engage in excessive mentalizing for crowds may follow suboptimal trends and incur a financial loss.

Social interaction requires mentalizing; yet, as we have reviewed in this section, mentalizing is vulnerable to bias and can lead to suboptimal decisions. Prior moral impressions, which may be built through direct feedback, or implicitly signaled by group membership, can bias mentalizing, increasing the possibility of inaccurate mental state inferences. However, as we will explore in the final section, these biases may not reflect truly "irrational" processes. Although the resultant decisions, such as favorable judgments about ingroup members (and the discounting of negative information about ingroup members), may appear biased, the underlying *processes* may nevertheless be rational[3]. This idea may be the key to explaining why people sometimes engage in greater mentalizing, and other times less mentalizing, in order to protect positive impressions of ingroup members (and negative impressions of immoral agents). In the

[3] Here we focus on procedural rationality, which may produce either accurate or inaccurate judgment. By contrast, see Cushman, 2019, for a theoretical account of how people ultimately benefit from rationalization.

final section, we will discuss this puzzle and explore the possibility that seemingly biased social and moral judgments may actually reflect rational decision-making.

Motivated mentalizing or rational updating?

Our prior knowledge of a person influences how we evaluate their behavior. Consider a close friend who you know to be trustworthy. One day, you see her take a quarter from a tip jar. Would you then judge her to be an untrustworthy person? Or–given your prior knowledge of her trustworthiness–would you consider this observation a noisy data point, and reattribute her behavior to situational factors? For instance: perhaps she was trying to make change for a dollar. By contrast, seeing a stranger take a quarter from a tip jar often leads to the inference that they are untrustworthy.

This asymmetry in our trait evaluations of friend versus stranger appears to be an instance of the well-known bias to positively evaluate close others or ingroup members. A key proposal of this chapter, however, is that the asymmetry can be accounted for by differences in the strength of prior knowledge. In the case of the stranger, we have no prior knowledge of their trustworthiness, so a single bad behavior is highly diagnostic of their character. But in the case of our friend, we have ample prior knowledge of her trustworthiness, so entirely revising our impression of her based on a single action may not be optimal. The confusing feature here is that strong prior knowledge of close others often co-occurs with factors that typically contribute to motivated decision-making, such as congenial affect, a long relationship history, and attachment. It is likely the case that both prior knowledge and socio-affective factors contribute to reduced belief updating in response to negative feedback; the relative contributions of these factors across contexts is a difficult but important empirical question. Here we highlight cases of seemingly motivated judgments that may instead be compatible with a rational updating process.

Bayesian updating provides a normative framework for how beliefs about others should be updated when new information is acquired. Bayes' rule holds that the probability of a belief being true given new evidence–e.g., *P*(*my friend is trustworthy*|*she stole a quarter*)–is equal to the likelihood of the evidence being acquired given the prior belief, *P*(*she stole a quarter*|*she is trustworthy*), multiplied by the probability of the prior belief being true before receiving the new evidence, *P*(*she is trustworthy*), scaled by the probability of the new evidence being acquired, *P*(*she stole a quarter*). This process factors the strength of the prior belief into updating; it follows that new information that contradicts strong prior beliefs may be discounted. While Bayesian updating does not necessarily guarantee accurate mental state inference, it confers *procedural* rationality on the inference process (Hahn & Harris, 2014), and serves as a normative criterion for assessing deviations from rational belief updating (Hackel & Amodio, 2018). Why adopt Bayesian processing in particular as a criterion for rationality? According to a set of epistemological accounts called the 'Dutch Book' arguments, when an agent possesses degrees of belief that violate the axioms of probability theory, they are vulnerable to logically-ensured losses when acting on their beliefs (e.g., accepting a wager that will lead to a sure loss, regardless of outcome), and to internally inconsistent evaluations (see Hájek, 2008 for a review). By this account, adhering to the axioms of probability theory can protect us from holding beliefs that are logically guaranteed to be false, and which would impair utility maximization.

How can a Bayesian framework be used to understand the robustness of prior beliefs to contradictory evidence, especially in the case of moral updating? A theoretical account suggests that people can generate *ad hoc* auxiliary hypotheses to explain away evidence that contradicts prior beliefs, and that this process is Bayesian-rational (Gershman, 2019; see Lakatos, 1976 for discussion on the role of auxiliaries in science). This process adheres to probability theory: auxiliary hypotheses are likelier to be invoked when they are highly consistent with the new information, and when the prior belief has a relatively high probability. For instance, given your

strong prior belief in the trustworthiness of your coin-taking friend, you may generate the auxiliary hypothesis that your friend was making change for a dollar. That is, the unexpected event is attributed to a situational cause, instead of a dispositional cause. While the tendency to invoke situational explanations for close others or ingroup members has been described as a cognitive bias, situational attributions can be procedurally rational if warranted by the strength of prior beliefs. Additionally, to return to the epistemological arguments for Bayesian rationality, invoking an auxiliary hypothesis in a graded manner allows the observer to retain a coherent set of beliefs that takes new evidence into account.[4]

How can we discern whether a case of reduced belief updating is the result of Bayesian-rational updating over strong priors, rather than non-rational discounting of contradictory evidence? Our novel proposal is that, the rational route to belief preservation will recruit more mentalizing activity than the non-rational route. When a Bayesian observer is faced with new, meaningful information that is inconsistent with their prior evaluations, they can account for the discrepancy by updating their prior beliefs, or by generating an auxiliary hypothesis to explain away the information. We speculate that, at least in the domain of moral judgment and character evaluation, both of these processes will recruit the mentalizing network, in particular, rTPJ, given its role in supporting mental state-based moral judgment. Thus an association between increased rTPJ activity and increased updating may suggest Bayesian updating of prior beliefs, and an association between increased rTPJ activity and reduced updating may suggest the generation of auxiliary hypotheses. An association between *decreased* rTPJ activity and reduced updating, however, may suggest motivated discounting of new evidence.

We now apply this logic to several studies discussed above. Recall that Baumgartner and colleagues (2012) found increased mentalizing network activity in response to ingroup vs.

[4] We also note that procedural rationality is orthogonal to the source of the prior belief: both priors that are evidence-based and priors that are derived largely from socio-affective value (e.g., positive beliefs about the ingroup in minimal group contexts) can undergo Bayesian processing.

outgroup defectors, and that greater connectivity in this network was associated with forgiveness of ingroup members. Increased mentalizing activity in this case can be reinterpreted as supporting the generation of auxiliary hypotheses that are consistent with strong positive beliefs about the ingroup. For example, perhaps the ingroup member did not intend to defect, or had a good reason to do so.

Turning to cases of motivated discounting, Kliemann and colleagues (2008) had found that, when a previously fair (vs. unfair) social partner was described as performing a harmful action, participants judged the action to be less intentional, and this judgment was associated with reduced rTPJ activity. If participants were taking a Bayesian route to belief maintenance, they would have engaged in more mentalizing for fair partners, in order to explain away the evaluatively inconsistent information. We hypothesize that participants took a motivated route instead: they may have opted out of explaining the discrepancy by disengaging from mentalizing about fair partners, resulting in decreased inferences of harmful intent. The function of this selective disengagement may be to preserve a historically cooperative relationship. Further, disengagement from mentalizing can be seen for ingroup members as well. Generally, group membership may serve as a proxy for moral character, such that in the absence of direct evidence, ingroup members are viewed as good moral agents. Hughes and colleagues (2017) found that decreased rTPJ activity was associated with reduced impression updating in response to negative feedback for ingroup members, consistent with what the researchers termed “an effortless bias” account. In this case, participants who disengaged from mentalizing were able to maintain desirable beliefs about ingroup members, by failing to incorporate evidence that would have led to a negative character inference. These studies suggest that, in the face of evidence that affords disfavorable trait inferences about ingroup members or previously moral agents, people may opt out of rational updating by mentalizing less about these agents altogether. Future work should examine the role of decreased mentalizing in other contexts, such as economic games, in which people are resistant to updating in response to

feedback about moral or ingroup targets (see Evans, Fleming, Dolan, & Averbeck, 2011; Fareri, Chang, & Delgado, 2015; Hackel, Doll, & Amodio, 2015).

Maintaining beliefs by discounting new information is not rational in the Bayesian sense, but it may be *adaptively* beneficial, in that in can increase social fitness and affective well-being. Specifically, it can be beneficial to maintain relationships with potential cooperative partners. For example, a group of researchers found that participants trusted their friends more than strangers in the Trust Game, even though reciprocation rates were equal for friend and stranger (Fareri, Chang, & Delgado, 2015). Neural and computational evidence indicated that trust decisions were driven by a striatal reward response to reciprocation from close friends. There can thus be affective benefits to non-rational processing of feedback about close others. Moreover, in recent research, we found that individuals who were more resistant to negatively updating their evaluations about a friend also reported having more friends in real life (Park & Young, under review). These are cases in which reduced belief updating leads to inaccurate predictions and financially suboptimal decisions, but may ultimately maximize the affective and social benefits of interacting with and maintaining close friends.

Within a given context, individuals may vary in whether they take a procedurally rational or irrational path to belief maintenance. For example, one study examined the public's impressions of Bill Clinton eight months before and three days after the Lewinsky story broke (Fischle, 2000). Respondents were interviewed on various aspects of the scandal, including the credibility and importance of the allegations, and attitudes towards the president's resignation. The study found that perceived importance of the scandal increased support for resignation by 57% for Clinton detractors, but only by 19% for Clinton supporters. The author argued that a Bayesian framework cannot account for such moderated effects, while a motivated reasoning process can capture affect-dependent weighting of factors like perceived importance. While this argument holds for those supporters who thought the allegations were important but did not support resignation, there were also supporters who exhibited–per our interpretation–the

Bayesian response. In particular, this study also found that supporters were likelier than detractors to view the scandal as a conspiracy, and this reduced supporters' certainty of impropriety and their endorsement of resignation. Given supporters' robust prior beliefs about Clinton, this set of respondents may have generated the auxiliary hypothesis that the scandal was a conspiracy planted by the president's opponents. More generally, in studies that find motivation-derived evaluations, there may be individual differences in whether participants take a Bayesian route to belief maintenance, or deviate from Bayesian reasoning in order to maintain prior beliefs.

Comparing participants' behavioral belief updates with predictions from a Bayesian model of inference can reveal the contexts in which people engage in probabilistic belief updating. One recent study examined how people learn factual political statements based on noisy feedback from a computer, and found that participants closely followed Bayesian updating, but not perfectly (Hill, 2017). Specifically, participants evaluated the same factual statement across multiple rounds; in some rounds, the computer signaled, with 75% accuracy, whether the statement was true or not. Comparing participants' initial responses (prior beliefs) with their final responses (posterior beliefs), the author found that when the signal was consistent with their prior beliefs, participants did not deviate from what was expected by a Bayesian model. When the signal was inconsistent with their prior beliefs, however, participants updated less than expected by the Bayesian model, suggesting motivational influences. Importantly, a growing body of work has used a computational approach to investigate how people update their evaluations of others, such as when making repeated judgments of whether advisors are trustworthy and accurate. Some studies have found that observers are biased towards learning from evaluatively consistent information (e.g., Leong & Zaki, 2017; Fareri, Chang, & Delgado, 2012); others have found that participants derive inferences in a Bayesian-rational manner (Behrens, Hunt, Woolrich, & Rushworth, 2008; Cao, Kleiman-Weiner, & Banaji, 2019; Diaconescu et al., 2014; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015; see

Hackel & Amodio, 2018 for a review of the literature). These findings highlight the utility of Bayesian models for systematically investigating when people adhere to, and when they deviate from, procedurally rational updating when learning about others' traits.

Neuroimaging studies, combined with a computational modeling approach, will provide an important window into the link between mentalizing network activity and motivated vs. Bayesian decision-making. Given that the rTPJ has been found to be engaged for probabilistic belief updating (Mengotii, Dombert, Fink, & Vossel, 2017), future research should combine fMRI and computational methods to further explore how the rTPJ may support Bayesian reasoning. Additionally, if there is a role for the rTPJ in Bayesian reasoning, it may be context-dependent. When the observer acquires evidence that does not warrant revision of a strong prior belief, the rTPJ may support the generation of auxiliary hypotheses (e.g., blame-mitigating mental states, appeal to situational factors); this may be the process underlying reduced punishment for ingroup defectors (Baumgartner et al., 2012). When the new evidence does warrant belief updating, the rTPJ may support revision of the strong prior belief (e.g., by attributing harmful intent). However, if the observer is motivated to maintain desired prior beliefs in the face of exceedingly strong contradictory evidence, their departure from Bayesian updating may be indexed by decreased rTPJ activity; this may underlie reduced negative updates for ingroup members (Hughes, Zaki, & Ambady, 2017). Further work will be needed to characterize the conditions under which observers who have strong prior beliefs about targets will mentalize about them *less* upon receiving contradictory evidence–therefore opting out of drawing any inferences that would prompt belief updating–versus mentalize about them *more*– therefore generating alternative hypotheses to accommodate the surprising behavior.

To summarize, our proposal is that instances of social and moral decision-making that appear to be motivated may instead be compatible with Bayesian-rational reasoning. Our strong prior beliefs–e.g., positive beliefs about the ingroup–are often protected from revision through the generation of auxiliary hypotheses (Gershman, 2019). Further work is needed to

differentiate between procedurally rational updating that appears irrational, and motivated updating that is driven by social and affective considerations (e.g., attachment to ingroup members). Finally, we call upon future work to examine the proximate and ultimate costs and benefits of motivated updating, above and beyond those of Bayesian updating.

Conclusion

We reviewed evidence that supports the engagement of mentalizing for optimal and suboptimal decision-making, in the contexts of moral judgment and economic exchange. People engage in accurate mentalizing, leading to social and non-social rewards, i.e., beating the competition, learning from others' strategies, identifying cooperative partners. But, people can also engage in "biased" mentalizing, with the aim of protecting their positive impressions of close others (friends, ingroup members), leading to direct and indirect losses. Even so, as we discussed in the final section, seemingly "biased" decisions, i.e., discounting negative feedback about close others, may in fact be Bayesian-rational, stemming from differences in people's prior beliefs and knowledge. Determining which kinds of decisions reflect rational updating versus motivated reasoning will be an important question to address going forward. We look forward to future research, which will continue to enhance our understanding of how mentalizing contributes to value-based decision-making.

Acknowledgements

We thank Josh Hirschfeld-Kroen, Emma Alai, and Simon Karg for their thoughtful feedback, and the John Templeton Foundation for support.

References

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature reviews neuroscience, 7*, 268-277.

Baron, S. G., Gobbini, M. I., Engell, A. D., & Todorov, A. (2011). Amygdala and dorsomedial prefrontal cortex responses to appearance-based and behavior-based person impressions. *Social Cognitive and Affective Neuroscience, 6*(5), 572-581.

Baumgartner, T., Götte, L., Gügler, R., & Fehr, E. (2012). The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Human Brain Mapping, 33*, 1452-1469.

Baumgartner, T., Schiller, B., Rieskamp, J., Gianotti, L. R. R., & Knoch, D. (2014). Diminishing parochialism in intergroup conflict by disrupting the right temporo-parietal junction. *Social Cognitive and Affective Neuroscience, 9*(5), 653-660.

Behrens, T. K. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature, 456*, 245-249.

Bhatt, M. A., Lohrenz, T., Camerer, C. F., & Montague, R. (2010). Neural signatures of strategic types in a two-person bargaining game. *Proceedings of the National Academy of Sciences of United States of America. 107*(46), 19720-19725.

Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron, 80*, 1558-1571.

Brambilla, M., Carraro, L., Castelli, L., & Sacchi, S. (2019). Changing impressions: Moral character dominates impression updating. *Journal of Experimental Social Psychology, 82*, 64-73.

Bruguier, A. J., Quartz, S. R., & Bossaerts, P. (2010). Exploring the nature of "trader intuition." *The Journal of Finance, 65*(5), 1703-1723.

Cao, J., Kleiman-Weiner, M., & Banaji, M. R. (2019). People make the same Bayesian judgment

they criticize in others. *Psychological Science, 30*(1), 20-31.

Coricelli, G., & Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences of United States of America. 106*(23). 9163-9168.

Cushman, F. A., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition, 127*(1), 6-21.

Decety, J., & Cacioppo, S. (2012). The speed of morality: A high-density electrical neuroimaging study. *Journal of Neurophysiology, 108*, 3068-3072.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience, 8*(11), 1611-1618.

De Martino, B., O'Doherty, J. P., Ray, D., Bossaerts, P., & Camerer, C. (2013). In the mind of the market: Theory of mind biases value computation during financial bubbles. *Neuron, 79*, 1222-1231.

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., Fehr, E., & Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Computational Biology*, doi: https://doi.org/10.1371/journal.pcbi.1003810

Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. In Bartels, D. M., Bauman, C. W., Skitka, L. J., & Burlington, D. (Eds). *The Psychology of Learning and Motivation.* Burlington: Academic Press.

Evans, S., Fleming, S., Dolan, R. J., & Averbeck, B. B. (2011). Effects of emotional preferences on value-based decision-making are mediated by mentalizing and not reward networks. *Journal of Cognitive Neuroscience, 23*(9), 2197-2210.

Fareri, D. S., Chang, L. J., & Delgado, M. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience, 6*, 148.

Fareri, D. S., Chang, L. J., & Delgado, M. (2015). Computational substrates of social value in interpersonal collaboration. *The Journal of Neuroscience, 35*(21), 8170-8180.

Fischle, M. (2000). Mass response to the Lewinsky Scandal: Motivated reasoning or Bayesian updating? *Political Psychology, 21*(1), 135-159.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*, 531-534.

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin and Review*, doi: https://doi.org/10.3758/s13423-018-1488-8

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science*, 24(1), 38-44.

Güroğlu, B., van den Bos, W., Rombouts, S. A. R. B., & Crone, E. A. (2010). Unfair? It depends: Neural correlates of fairness in social context. *Social Cognitive and Affective Neuroscience, 5*(4), 414-423.

Gweon, H., Dodell-Feder D., Bedny M., & Saxe R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. *Child Development*. 1853-1868.

Hackel, L. M. & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology, 24*, 92-97.

Hackel, L. M, Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature Neuroscience, 18*(9), 1233-1235.

Hahn, U. & Harris, A. J. L. (2014). What does it mean to be biased: Motivated reasoning and rationality. In Ross, B. H. (Ed.), *Psychology of learning and motivation,* San Diego, CA: Academic Press.

Hájek, A. (2008). Dutch book arguments. In Anand, P., Pattanaik, P., & Puppe, C. (Eds.), *The Handbook of Rational and Social Choice*, New York: Oxford University Press.

Hampton, A. N., Bossaerts, P. & O'Doherty, J. P. (2008). Neural correlates of mentalizing-

related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of United States of America. 105*(18), 6741-6746.

Hare, T., Camerer, C. F., Knoepfle, D. T., O'Doherty, J. P., & Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *The Journal of Neuroscience, 30*(2), 583-590.

Harvey, N. & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes. 70*(2), 117-133.

Hill, S. J. (2017). Learning together slowly: Bayesian learning about political facts. *The Journal of Politics, 79*(4), 1403-1418.

Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J. P., & Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience, 20*, 1142-1149.

Hughes, B. L., Zaki, J., & Ambady, N. (2017). Motivation alters impression formation and related neural systems. *Social Cognitive and Affective Neuroscience, 12*(1), 49-60.

Inbar, Y., Pizarro, D. A., & Cushman, F. (2012). Benefiting from misfortune: When harmless actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin, 38*(1), 52-62.

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron, 58*, 284-294.

Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition, 119*, 197-215.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science, 308*(5718), 78-83.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). *Inference of*

intention and permissibility in moral decision making. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.), Proceedings of the 37th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia, 46*, 2949-2957.

Klucharev, V., Smidts, A., & Fernández, G. (2008). Brain mechanisms of persuasion: how 'expert power' modulates memory and attitudes. *Social Cognitive and Affective Neuroscience, 3*(4), 353-366.

Knobe, J. (2003). Intentional action and side-effects in ordinary language. *Analysis*, 63, 190-193.

Koster-Hale, J., & Saxe R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*, 836-848.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences of United States of America. 110*(14), 5648-5653.

Krueger, F., Grafman, J., & McCabe, K. (2008). Neural correlates of economic game playing. *Philosophical Transactions of The Royal Society. 363*, 3859-3874.

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., Heinecke, A., & Grafman, J. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences of United States of America. 104*(50), 20084-20089.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In *Can Theories be Refuted?,* (pp. 205-259): Springer.

Lee, D. (2006). Neural basis of quasi-rational decision making. *Current Opinion in Neurobiology, 16*(2), 191-198.

Leong, Y. C. & Zaki, J. (2017). Unrealistic optimism in advice taking: A computational account.

Journal of Experimental Psychology: General, 147(2), 170.

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Wheelright, S. J., Sadek, S. A., Suckling, J., MRC AIMSConsortium, & Baron-Cohen, S. (2009). Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience, 22*(7), 1623-1635.

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *The Journal of Neuroscience, 33*(50), 19406-19415.

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience, 8*(6), 623-631.

Mende-Siedlecki & Todorov, A. (2016). Neural dissociations between meaningful and mere inconsistency in impression updating. *Social Cognitive and Affective Neuroscience, 11*(9), 1489-1500.

Mengotii, P., Dombert, P. L., Fink, G. R., & Vossel, S. (2017). Disruption of the right temporoparietal junction impairs probabilistic belief updating. *The Journal of Neuroscience, 37*(22), 5419-5428.

Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience and Biobehavioral Reviews, 65*, 276-291.

Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology, 116*(2), 215-236.

Park, B., Fareri, D. S., Delgado, M. R., & Young, L. *How theory-of-mind brain regions process prediction error across relationship contexts.* Manuscript in preparation.

Park, B., & Young, L. *Does biased impression updating facilitate close relationships? A behavioral and fMRI investigation.* Manuscript under review.

Phan, K. L., Sripada, C. S., Angstadt, M., & McCabe, K. (2010). Reputation for reciprocity

engages the brain reward center. *Proceedings of the National Academy of Sciences of United States of America. 107*(29), 13099-13104.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology, 39*, 653-660.

Pizarro, D. Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science, 14*(3), 267-272.

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications, 9*, doi: 10.1038/s41467-018-03399-2

Rilling, J. K., Dagenais, J. E., Goldsmith, D. R., Glenn, A. L., & Pagnoni, G. (2008). Social cognitive neural networks during in-group and out-group interactions. *Neuroimage, 41*, 1447-1461.

Saxe, R. (2009). Theory of mind (neural basis). In Banks, W. (Ed.), *Encyclopedia of Consciousness*. Cambridge, MA: MIT Press.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology, 55*, 87-124.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in “theory of mind.” *Neuroimaging, 19*, 1835-1842.

Saxe, R., & Powell, L. J. (2006). It’s the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science, 17*(8), 692-699.

Saxe, R., & Wexler A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia. 43*(10), 1391-1399.

Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., & Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia, 42*, 1435-1446.

Schilbach, L., Eickhoff, S. B., Schultze, T., Mojzisch, A., & Vogeley, K. (2013). To you I am

listening: Perceived competence of advisors influences judgment and decision-making via recruitment of the amygdala. *Social Neuroscience, 8*(3), 189-202.

Sellaro, R., Güroğlu, B., Nitsche, M. A., van den Wildenberg, W. P. M., Massaro, V., Durieux, J., Hommel, B., & Colzato, L. S. (2015). Increasing the role of belief information in moral judgments by stimulating the right temporoparietal junction. *Neuropsychologia 77*, 400–408.

Singer, T., & Fehr, E. (2005). The neuroeconomics of mind reading and empathy. *American Economic Review, 95*(2), 340-345.

Strang, S., Utikal, V., Fischbacher, U., Weber, B., & Falk, A. (2014). Neural correlates of receiving an apology and active forgiveness: An fMRI study. PLoS ONE 9:e87654. doi: 10.1371/journal.pone.0087654

Tannenbaum, D., Ditto, P. H. & Pizarro, D. A. (2008). *Different moral values produce different judgments of intentional action.* Poster presented at the annual meeting of the Society for Personality and Social Psychology, Albuquerque, NM.

Thornton, M. A. & Mitchell, J. P. (2018). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex, 28*, 3505-3520.

Treadway, M. T., Buckholtz, J. W., Martin, J. W., Jan, K., Asplund, C. L., Ginther, M. R., Jones, O. D., & Marois, R. (2014). Corticolimbic gating of emotion-driven punishment. *Nature Neuroscience, 17*(9), 1270-1275.

Van Swol, L. M. (2009). The effects of confidence and advisor motives on advice utilization. *Communication Research, 36*(6), 857-873.

Walter, H., Adenzato, M., Ciaramidaro, A., Enrici, I., Pia, L., & Bara, B. G. (2004). Understanding intentions in social interaction: The role of the anterior paracingulate cortex. *Journal of Cognitive Neuroscience, 16*(10), 1854-1863.

Will, G.-J., Crone, E. A., & Güroğlu, B. (2015). Acting on social exclusion: neural correlates of

punishment and forgiveness of excluders. *Social Cognitive and Affective Neuroscience, 10*(2), 209-218.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition, 100*, 283-301.

Yamada, M., Camerer, C. F., Fujie, S., Kato, M., Matsuda, T., Takano, H., Ito, H., Suhara, T., & Takahashi, H. (2012). Neural circuits in the brain that are activated when mitigating criminal sentences. *Nature Communications, 3*, 759. doi:10.1038/ncomms1757

Yaniv, I. & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes. 83*(2), 260-281.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of United States of America. 107*(15), 6753-6758.

Young, L., Cushman, F., Hauser, M, & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of United States of America. 104*(20), 8235-8240.

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology, 1*, 333-349.

Young, L. & Saxe, R. (2009). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience, 21*(7), 1396-1405.

Young, L., Scholz, J., & Saxe, R. (2011). Neural evidence for "intuitive prosecution": The use of mental state information for negative moral verdicts. *Social Neuroscience, 6*(3), 302-315.