Self-Other Prediction Error: How Individuals Process Others' Behavior Based on Perceived Alignment

Paul C. Bogdan[1]

[1] Psychology & Neuroscience, Duke University, Durham, 27708, NC, USA.

Running head: *Self-Other Prediction Error and Perceived Alignment*

Author Note

Correspondence: Paul Bogdan, email: paulcbogdan@gmail.com

Acknowledgments: The author thanks Daryl Cameron and his research group, Florin Dolcos, Kyle Mou, Sanda Dolcos, Sean M. Laurent and his research group, and Suraiya Allidina for their feedback on an earlier version of this manuscript.

Abstract

The present report argues that the pursuit of social alignment drives all aspects of how individuals evaluate and respond to another person's behavior. Social/moral judgment – whether you see another person as good and their behavior as appropriate – can be framed as an evaluation of whether the person's values align with yours and whether they acted the same way you would. Following negative judgments, your responses can be understood as attempts to reestablish alignment, such as by conforming to this person's behavior or punishing it. From this perspective, the present research proposes the Self-Other Prediction Error (SOPE) model, which conceptualizes these mechanisms within frameworks of predictive coding, Theory of Mind, and dual-process decision-making. Recent studies experimentally testing this model validate its core arguments and show that SOPE predicts participants' reactions to others' behavior more accurately than earlier theories focusing on norms or utility. Furthermore, reinterpreting older studies suggests that the proposed model successfully reconciles a wide array of phenomena concerning moral psychology, behavioral economics, and other social cognition topics. Hence, the proposed SOPE account is posed as a foundation for understanding social information processing generally and its computational underpinnings.

Keywords: Behavioral economics, moral judgment, reinforcement learning, prediction error, reciprocity.

1. Self-other prediction error: How individuals process others' behavior based on perceived alignment

Imagine you return to your desk, and your favorite pen is missing. You immediately suspect your colleague took it. This is unfair and wrong. You would not take their pen, after all. You think about confronting them. You consider even asking for the pen's return in earshot of others, so the embarrassment teaches them not to steal again. You also wonder if you should take their pens in the future, especially if you are in a rush. Before this, you scour your desk to ensure you did not simply misplace the pen. It does not pop up. Nonetheless, you take a step back. This person is your friend, and you do not want trouble. Perhaps they desperately needed to sign a document, which required them to take the pen. The explanation seems implausible, but you settle on it for now.

I will argue that the common mechanism guiding these thoughts and responses is the drive for *social alignment*. The pen's theft is seen as bad because you perceived misalignment: You saw your colleague treat you differently than how you would treat them (action-misalignment), and you inferred that your colleague cares about you less than you care about them (goal-misalignment). Further, your subsequent cognitive and behavioral responses are attempts to reestablish alignment: You thought about shaming your colleague to be more like you (punishment), you pondered becoming more like them (reciprocity/conformity), and you tried reinterpreting your colleague's behavior (dismissing evidence of misalignment).

Earlier theories conceptualized alignment as an innate drive to eliminate dissimilarity between one's own values and another person's perceived values (Constant et al., 2019; Shamay-Tsoory et al., 2019; Theriault et al., 2020; Veissière et al., 2020). These earlier frameworks derive from predictive coding theory and define discrepancies between one's values and others' values as *self-other prediction errors* (SOPE), which individuals seek to minimize. However,

these earlier papers exclusively considered social conformity, arguing that SOPE causes people to adjust their behaviors and goals to be more like others. The present research drastically expands this idea.

To develop the SOPE model, the drive for alignment is considered alongside frameworks of decision-making and Theory of Mind. Decision-making research examines how individuals represent their personal values, and Theory of Mind research describes how individuals draw inferences about other people. Together, these fields lay out the computational mechanisms necessary for understanding how SOPE can arise from the interactions between individuals' personal values and their inferences about others' values. Probing this interplay between SOPE, decision-making, and Theory of Mind is the main innovation here compared to earlier papers on social alignment.

The proposed model has a broad scope, putting forth that social alignment underlies all ways in which an individual reacts negatively or positively to another person's behavior. Such reactions are usually studied in terms of moral judgments or social-economic decisions, and given this rich experimental and theoretical literature, moral and economic studies are the core of the present review. However, the proposed SOPE account is also meant to capture other reactions, such as taking offense to minute actions or anything that confers information about another person. Accordingly, rather than focusing on specific constructs such as "liking", "righteousness", etc., the outcomes of judgment are described in terms of general positivity or negativity with emphasis on the specific and concrete behaviors following judgment. This scope widens the pool of phenomena to be explained and the possibility for counterarguments, but the scope limits the need to segment social cognition across constructs that may be seen as arbitrary.

While reviewing this extant research, the proposed model is contrasted to earlier theories focusing of social judgment, including (1) expectation-violation/deontological theories, (2)

consequentialist/utilitarian theories, and (3) Theory-of-Mind/virtue-ethics theories. As will be described, different aspects of the SOPE model bear key parallels to each of these earlier ideas. The SOPE account overall is meant to be compatible with their supporting evidence, while synthesizing and expanding them and generally providing more accurate predictions. The penultimate section of this report summarizes two prior empirical reports testing these ideas, the core SOPE predictions, and how SOPE relates to these earlier views (Bogdan et al., 2023, 2024). Altogether, the present research aims to combine conceptual explanations with experimental evidence to put forth a perspective that synthesizes psychological topics and seeks to clarify underlying mechanisms.

2. Social alignment and self-other prediction error

2.1. Adjusting oneself to match others

Earlier research on alignment has principally focused on individuals' adjusting their goals and behaviors to match others' goals and behaviors (Asch, 1951; Bandura & Walters, 1977; Charness et al., 2019; Gray et al., 2014; Jung et al., 2020; Nowak & Sigmund, 2005). This manifests in multiple ways. For example, suppose one person steals from another person. The victim will want revenge and be motivated to steal back from the original thief (reciprocity). To a lesser degree, the victim will also "pay it forward," becoming more likely to steal from other people in general (upstream reciprocity). Any third-party onlookers will tend to similarly feel justified to steal from the original thief (downstream reciprocity) and overall feel more comfortable stealing in general (conformity). Overall, alignment is multifaceted, influencing both second parties and third parties along with both person-specific and generalized values.

Along with being complex, there are several reasons why the drive for alignment is foundational to social interaction: (1) The drive to align is universal, given ethnographic research on conformity and reciprocity across cultures (Bond & Smith, 1996; Curry et al., 2019). (2) The

drive to align is innate, given that infants conform to and reciprocate others' behavior (Buttelmann et al., 2013; Hamlin et al., 2011). (3) The drive to align is strong, and individuals tend to reciprocate others' generous or selfish behavior, even when it means worse personal outcomes and even when stakes are high (E. Fehr et al., 1993, 2002). (4) The drive to align may be an evolutionarily preserved dimension of social cognition, as conformity and reciprocity are seen across some non-human primates (Schweinfurth & Call, 2019; Whiten et al., 2022). (5) Further, the drive to align is adaptive. Simulations show that agents using reciprocity-based strategies find success and are more cooperative (Axelrod, 1980; Traulsen & Nowak, 2006). These many properties put alignment in a special place within the social cognition landscape.

Along with the social alignment guiding individuals' decision-making, the *expectation of alignment* characterizes how individuals perceive others' decisions: (1) Across the world, individuals tend to punish behavior that is more selfish or overly generous relative to their own (Herrmann et al., 2008). (2) This expectation is innate, as even infants negatively view others who do not reciprocate their behavior, and infants predict pairs of third parties will reciprocate one another (Jin et al., 2024). (3) The drive to respond to deviations from reciprocity is potent, and individuals tend to individuals retributory punish others even at their own expense and sometimes even when there is no opportunity for the punishment to deter future deviations (Tan & Xiao, 2018). (4) Beyond humans, primates also expect reciprocity, evidenced by monkeys and apes acting more prosocially toward others who have the ability to reciprocate (Benozio et al., 2023; Suchak & de Waal, 2012). (5) Finally, simulations show that punishing or avoiding others more selfish than oneself leads to fair and cooperative outcomes (Page & Nowak, 2001), and moral homogeneity is argued to be adaptive because it deters conflicts (DeScioli & Kurzban, 2013). Each of these different points gives further weight to the general premise that the expectation of alignment is an intrinsic and core pillar of social judgment.

2.2. Alignment as prediction error minimization

Mechanistically, the drive for alignment can be conceptualized in terms of prediction error. “Prediction error” is essentially a discrepancy between two mental representations, and this is a general notion applicable beyond social cognition. For instance, if a participant presses the incorrect button in a flanker task, this is believed to elicit a prediction error between their representations of the performed action versus the action they wish they performed (Di Gregorio et al., 2016; Yeung et al., 2004). After an incorrect response, the prediction error may provoke the recruitment of cognitive resources to increase accuracy in upcoming trials (Cavanagh & Frank, 2014). An innate drive to minimize prediction error is posed to derive from a living entity’s homeostatic needs. For instance, humans pursue a 98 °F body temperature, which may be achieved by decisions that minimize prediction error between this “predicted” temperature and their actual body temperature. Achieving these biological states involves attaining intermediaries: Maintaining a healthy body temperature requires a house, which requires a job, which requires cordial coworker relations, etc. Hence, prediction error at an intermediatory can propagate toward those deeper goals, compelling individuals to minimize violations of those intermediaries themselves. Per this organization, information is always processed relative to expectations, which is thought to aid in the efficient use of cognitive resources (Keller & Mrsic-Flogel, 2018; Schütt et al., 2024). Accordingly, arguments have been made that all neurocognitive processes should be understood and unified in terms of prediction error (Barrett, 2017; Friston, 2010; Köster et al., 2020).

Social information about others’ actions and values may elicit prediction errors with respect to one’s own decision-making. In turn, several papers have argued conformity is the minimization of prediction error (Constant et al., 2019; Shamay-Tsoory et al., 2019; Theriault et al., 2020; Veissière et al., 2020). This rationale builds upon older cognitive dissonance theories,

describing how discrepancies between beliefs about oneself and others produce negative responses because they defy assumptions of similarity (Festinger, 1954). Compared to the prediction-error-conformity claims, cognitive dissonance accounts have been more elaborate in describing how individuals reduce interpersonal discrepancies. Rather than just confirming, individuals may persuade others to match their beliefs or move to more aligned groups (Matz & Wood, 2005). Further, to minimize cognitive dissonance elicited by peers' inappropriate behavior, individuals may draw convenient assumptions about their motivations and dismiss problematic evidence (Barkan et al., 2015). Overall, many outcomes of dissonance minimization have been proposed

The present work builds on prediction error theories and their computational formulations which focus on how social information interacts with personal values linked to decision-making systems. In addition, the present research takes and expands upon the explanatory breadth of cognitive dissonance work, seeking to develop it into a general view covering all types of reactions, judgments, and responses to social information. Later descriptions will focus on linking these conceptual scales, specifying computational pathways whereby decision-making systems are leveraged for social information processing and interact with Theory of Mind mechanisms. However, before this, the next section reviews psychological evidence to illustrate at a high level how different aspects of social interaction can be viewed under the lens of social alignment and the minimization of self-other discrepancies.

3. The SOPE model

3.1. SOPE in action

The proposed view is the "*self-other prediction error*" (SOPE) model. Negative reactions to social information are argued to derive from the presence of SOPE, and subsequent cognitive and behavioral responses are posed to attempts at minimizing SOPE. For example, suppose

Adam finds evidence that Beth stole money from him, even though he has never stolen her money. The SOPE causes Adam to react negatively. Per earlier conformity/reciprocity accounts, he can minimize SOPE by conforming and stealing from other people or by reciprocating and stealing from Beth. Such responses involve Adam adjusting himself to match Beth.

However, there are ways to minimize SOPE beyond conformity and reciprocity: (1) Adam can punish Beth. Punishment is often a form of communication meant to change another person's behavior (Ho et al., 2019; Sarin et al., 2021). Through punishment, Adam encourages Beth to change her behavior and values to match his. (2) Adam can avoid Both, which decreases SOPE by making her deviancy less salient. In a strict sense, minimizing SOPE is not increasing alignment *per se* but rather decreasing misalignment. (3) Adam can justify the behavior and update his beliefs, concluding that if he was in Beth's context, then he would do the same. Adam may assume Beth quickly needed cash for a bus fare to visit his parents. Reinterpreting Beth's behavior allows Adam to maintain that he and Beth still hold similar goals in valued areas – e.g., both value family time. (4) Adam can reinterpret the evidence, so he no longer believes any theft occurred. These third and fourth points concern uncertainty during Theory of Mind. Uncertainty often enables individuals to judge transgressions by in-group members less harshly by permitting lenient interpretations of the behavior (Kim et al., 2020). The drive to eliminate SOPE may motivate these irrational justifications. (5) Beth's later behavior can minimize Adam's SOPE. She may apologize and convince Adam that she will not steal again. Understanding apologies as SOPE minimization speaks to evidence that apologies are most effective when they communicate that the perpetrator agrees their behavior was inappropriate and has changed their views to align with the victim's (R. Fehr & Gelfand, 2010; Lewicki et al., 2016). Overall, this example lays out the potential for SOPE to capture numerous elements of social behavior, which the following subsection elaborates upon with more experimental evidence.

3.2. SOPE sensitivity and minimization

Earlier research indeed shows that people like others similar to themselves, particularly if they are alike in highly valued (moral) domains (Barnby et al., 2022; Earle & Siegrist, 2006; Siegrist et al., 2000). When judging misdeeds, individuals tend to be less harsh if they are guilty of those misdeeds too (Alicke, 1993). In contrast, people who hold themselves to high moral standards tend to impose these on others. For instance, participants who just behaved generously tend to evaluate and punish selfish behavior most harshly (Irwin & Horne, 2013; Volk et al., 2019). Such reactions are consistent with the idea that people's responses depend on comparisons between others and themselves. The implications of this can manifest in peculiar ways, such as participants with siblings judging incest more harshly than participants without siblings (R. M. Miller et al., 2014). Such findings have led to theories consistent with the SOPE model stating that moral judgment of another person's action depends on the evaluator's aversion to behaving the same way (R. M. Miller & Cushman, 2013).

Further intriguing evidence for the role of SOPE in social evaluation comes from studies showing that selfish groups generally dislike and exclude generous people (Irwin & Horne, 2013; Parks & Stone, 2010). Such disapproval occurs even when another person's generosity directly benefits the selfish critics. These effects also emerge cross-culturally. Across both low- and high-income nations, individuals deciding whether to punish another person's behavior consider whether the person's behavior is either more generous or more selfish than their own (Herrmann et al., 2008). Beyond the laboratory, most people likewise dislike behavior that is far more moral than their own. Individuals usually do not praise unexpectedly large charitable donations (Klein & Epley, 2014). Likewise, most people dislike vegans who avoid all animal products and climate advocates who avoid all airplane travel (De Groeve & Rosenfeld, 2022; Sparkman & Attari, 2020). This relationship appears causal: Participants instructed to commit an immoral act tend to

dislike confederates who refused to do it (Monin et al., 2008). Given the curious nature of these results, they have received much attention, and other explanations have been presented, such as these being rooted in norm violations (Irwin & Horne, 2013; Kawamura & Kusumi, 2020). These alternative accounts will be discussed and compared to the view proposed here, although as it stands, show how SOPE addresses several intuitive social phenomena.

The SOPE model also finds interesting support from research on observing mimicry – e.g., seeing bodily movements that parallel one's own or listening to people speak with similar word usage. Encountering this type of mimicry generates positive impressions (Bocian et al., 2018; Fischer-Lokou et al., 2011; Kulesza et al., 2014, 2022; Quiros et al., 2021). Note that these studies specifically show that observing mimicry causes positive impressions and not just that mimicry is correlated with positive views. This link between mimicry and evaluations is robust, emerging in real-world settings (Ireland et al., 2011; Otterbacher et al., 2017; Rains, 2016). Furthermore, these effects arise in primates: Macaques who most mimic their peers' postures tend to be treated better by others (Anderson & Kinnally, 2021). Hence, evaluating others based on alignment may be an evolutionarily preserved dimension of social cognition.

3.3. Interim summary and upcoming directions

The studies reviewed in this section focused specifically on alignment or closely related issues. These are taken alongside initial motivation about the universality, adaptiveness, and biological conservation of the drive for reciprocity and its expectation from others. Altogether, the SOPE model finds a plausible foundation. However, justifying SOPE's relevance to all aspects of social information processing and judgment requires a much more thorough review, specifically focusing on how studies covering other topics – e.g., social norms, harms, etc. – can be reframed in terms of SOPE. Later sections do this, but beforehand, a finer definition of SOPE mechanisms is necessary to make specific predictions that can be linked to these other concepts.

4. The computation of SOPE

4.1. Model-free and model-based decision-making

SOPE is posed to emerge from discrepancies between incoming social information and their own values. “Value” can be defined as that which drives decision-making, and the decision-making literature describes how values can exist in different forms. One prominent perspective is that values should be categorized based on whether they support fast/heuristic mechanisms or slow/deliberative mechanisms (Evans, 2008). This distinction can be formalized as model-free and model-based processing (Dolan & Dayan, 2013; K. J. Miller et al., 2019; Pauli et al., 2018). Model-free processes encourage individuals to act habitually, based on their *context* and without deliberation on specific goals. *Habits* constitute values assigned to actions within a context. Model-based processes involve individuals bringing *outcomes* to mind before making a choice and selecting the action whose expected outcomes best achieve their goals. *Goals* constitute values assigned to outcomes. This model-free/based division improves behavioral predictions and tracks neural signatures (Daw et al., 2011; Herd et al., 2021).

Moral psychology research often analogously suggests that social evaluation is split across heuristic and deliberative processes. For example, high cognitive load encourages heuristic rule- or norm-based reasoning, whereas low load promotes outcome- or intentionality-based reasoning (Buon et al., 2013; Greene et al., 2008; Martin et al., 2021). Neuroscientific research endorses this division, showing how reliance on norms and heuristics during judgment recruits different brain regions than deliberation about outcomes (Greene et al., 2004; Hutcherson et al., 2015; Koenigs et al., 2007). To explain how dual processes arise in moral processing, Cushman (2013) proposed an account based on the aforementioned model-free versus model-based mechanisms. Model-free processes were argued to support rule-based judgments, such that individuals negatively evaluate behavior they would be uncomfortable

performing themselves. Slower model-based processes were argued to support consequentialist judgments, such that individuals negatively evaluate behavior yielding undesired outcomes. These definitions begin to pin down elusive notions into specific computations and describing social evaluative mechanisms as repurposing established "non-social" decision-making pathways gives credence to their plausibility (Lockwood et al., 2020; Parkinson & Wheatley, 2015). SOPE is likewise divided into dual processes, but with key differences relative to earlier views.

4.2. Fast judgments

Fast judgments are posed to be due to $SOPE_{Action}$. The degree of $SOPE_{Action}$ depends on whether an *Actor's* behavior aligns with how an *Observer* would act in their context, per the Observer's habits. If $SOPE_{Action}$ arises, the magnitude of the negative reaction is modulated by the degree $SOPE_{Action}$ interferes with the Observer's goals, which depends on the action's expected outcomes. Roughly, this explanation describes fast judgment in terms of two components: the emergence of $SOPE_{Action}$ and its modulation by expectations. These two elements are discussed in turn.

4.2.1. $SOPE_{Action}$ reflects habit-based comparisons

The idea that Observers use their habits to judge self-other similarity bears resemblance to the model-free pathway by Cushman (2013). However, unlike this earlier view, $SOPE_{Action}$ supposes that Observers attempt to suppress their own perspective and consider habits from the perspective of an Actor's context. By extension, Observers' prior beliefs about the action and the Actor come into play because these shape perspective-taking and the Observer's perception of the other person's context.

Illustrating the $SOPE_{Action}$ computation, suppose Anne flees a store with an unknown product, and Bob sees this. Bob believes there is an 80% chance she is stealing food for her family and a 20% chance she is stealing goods for herself. Bob personally would steal if his

family needed food, but he would not steal if he wanted goods for himself. Hence, Bob perceives an 80% chance that he would act the same way in Anne's shoes. Bob would judge Anne less harshly than Charles, who would not steal under either context. Additionally, Bob judges Anne less harshly than Dave, who would steal for his family but believes this explanation is unlikely because he already dislikes Anne, and he is confident she is stealing for herself.

The social reasoning across this example is presumed to be fast, operating as a heuristic. This may seem unintuitive, as the example appears to involve Theory of Mind, which is often conceptualized as a slow and deliberate process (Ho et al., 2022). However, $SOPE_{Action}$ is instead posed to leverage "Theory-of-Mind-like" social beliefs (Schneider et al., 2012, 2014). Apperly and Butterfill (2009) discuss this distinction and how, as heuristics developed over previous social experiences, the fast and implicit application of social knowledge becomes possible. For instance, one can sometimes immediately recognize when a close friend deviates from their usual behavior. Heuristically, even probabilistic reasoning may become embedded into an Observer's representation of an Actor's context and influence the application of their habits. For example, if Bob frequently hears news about people stealing food, then when he encounters theft, he will not need to deliberate but will judge theft softer than someone who does not encounter such news.

4.2.2. The impact of $SOPE_{Action}$ is modulated by outcomes

The effect of $SOPE_{Action}$ on Observers' reactions is posed to be modulated by the extent dissimilar behavior is expected to produce outcomes that interfere with the Observer's goals. As above, this mechanism can be illustrated through examples: (i) Seeing someone drink soup with a straw will elicit some negative reaction due to action-dissimilarity but because such behavior is not expected to cause meaningful harm, the negative response is muted. (ii) By contrast, observing violence toward a helpless victim will be strict because the behavior is not just deviant but also produces undesired outcomes. (iii) Yet, if a violent Actor's behavior matches how the

Observer believes they would behave in the Actor's context – e.g., the violence is toward someone deserving it – then there is presumed to be no negative judgment at all. These cases illustrate how $SOPE_{Action}$ is posed to be necessary for a negative judgment, but a strongly negative one also requires that the action's expected outcome diverge from Observers' goals (i.e., their outcome valuations).

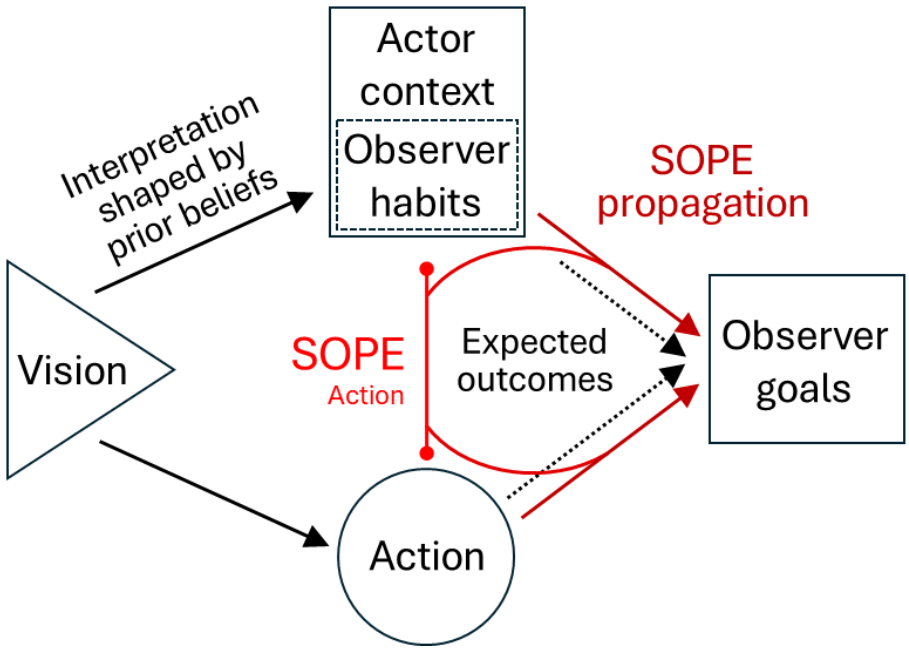
Critically, for fast judgments, the dissimilar action's *expected* outcome is the modulator rather than its actual outcome. This focus on expected outcomes carries two key features. First, this accounts for judgments where negative outcomes are felt but are difficult to see – e.g., negative judgment of gay behavior is linked to the perception that it harms families (Royzman et al., 2015; Schein et al., 2016). Second, operating upon expected outcomes would hasten the judgment. As prior work has shown, evaluating actual outcomes and associating them with an Actor's behavior is a slow process (Greene, 2009). Yet, these meticulous assessments are not necessary for fast judgments if Observers can leverage existing action-outcome and habit-goal associations. Available psychophysiological evidence indeed shows that fast judgments can distinguish morally loaded from non-morally loaded instances of dissimilar behavior (Lahat et al., 2013; Yucel et al., 2020). Hence, judgments based on habits must be somehow imbued with another form of Observers' values. Simply claiming that these concerns are baked into habits themselves seems inadequate – e.g., an Observer may have strong habits for self-preservation, but observing another person commit suicide will elicit a less negative judgment than observing a murder. The present account, whereby expected outcomes modulate $SOPE_{Action}$ attempts to reconcile these many phenomena (Kelly et al., 2007; Nichols, 2002; Tisak & Turiel, 1988).

This proposal is broadly consistent with prediction error theory, although there are different possibilities regarding the exact computations at play. Habit representations may predict both how the Actor will behave and how the habits relate to goals (Kruglanski & Szumowska,

2020). Hence, dissimilar behavior may elicit $SOPE_{Action}$ that propagates through the habit predictions and then elicits a prediction error with respect to the Observer's outcome values. Alternatively, the representation of the action itself may predict the outcome, so $SOPE_{Action}$ may propagate through the preexisting action-outcome association; Figure 1 illustrates both possible propagation trajectories. Admittedly, these computational claims about prediction-error propagation are speculative. Decisive answers on such topics generally require neuronal recording or intracranial electroencephalography (Chao et al., 2018), and these techniques remain relatively rare in social cognition research. Nonetheless, using purely behavioral data, Section 6 will describe studies that formally modeled $SOPE_{Action}$ and tested the core propositions here: (i) Observers' judgments depend on the likelihood that they would likewise perform a given action in the Actor's context, and (ii) the impact of this effect on judgments is modulated by an action's outcomes.

Figure 1

Self-other prediction error linked to habit-action discrepancies



Note. Social perceptual information is parsed into representations of an Actor's context and their action; not necessarily simultaneously, a contextual representation may have been formed before seeing the action. $SOPE_{Action}$ is computed as the extent of discrepancy between an observed action and the Observer's habits associated with the Actor's context. If $SOPE_{Action}$ arises, it is posed to be propagated from the habit and/or action representations to the Observer's goals (i.e., the values they ascribe to outcomes), which modulates the negative reaction due to $SOPE_{Action}$.

4.3. Slow judgments

4.3.1. Theory of Mind

Slower judgments are posed to involve proper Theory of Mind (ToM), whereby inferences about an Actor enter awareness – unlike the ToM-like heuristics noted above for fast judgments. ToM is the process of an Observer reasoning about an Actor by seeing them as an entity that behaves with intentionality (Dennett, 1989; Gergely & Csibra, 2003). For example, reasoning about why a student skipped class engages ToM because it requires consideration of the student's goals. This contrasts, for instance, discerning why a car engine failed. During ToM, Observers update their beliefs about an Actor and the Actor's context based on an observed action. For instance, after observing a student skip class, one may infer that the class is poorly taught (context-updating) and/or that the student does not care about their grade (goal-updating); also referred to as "situational" or "dispositional" attribution, respectively (Jones & Harris, 1967; Ross, 1977, 2018).

An Observer can infer an Actor's goals and context by assuming their action sought to maximize their expected value – sometimes called "inverse reinforcement learning" (Collette et al., 2017; Jara-Ettinger, 2019). Formally, these types of inferences can be modeled as Bayesian processes, which account for the uncertainty surrounding any given inference (Baker et al., 2017;

FeldmanHall & Shenhav, 2019; Saxe & Houlihan, 2017). Many aspects of social judgment concern how Observers deal with uncertainty, what assumptions they draw, and how these are influenced by prior beliefs. Bayesian ToM models grapple with this, and this is the interpretation of ToM adhered to in the present report. However, ultimately, the present thesis is compatible with any ToM framework that presumes Observers draw inferences about an Actor's goals and the Actor's context in some way.

4.3.2. Slow SOPE$_{Action}$

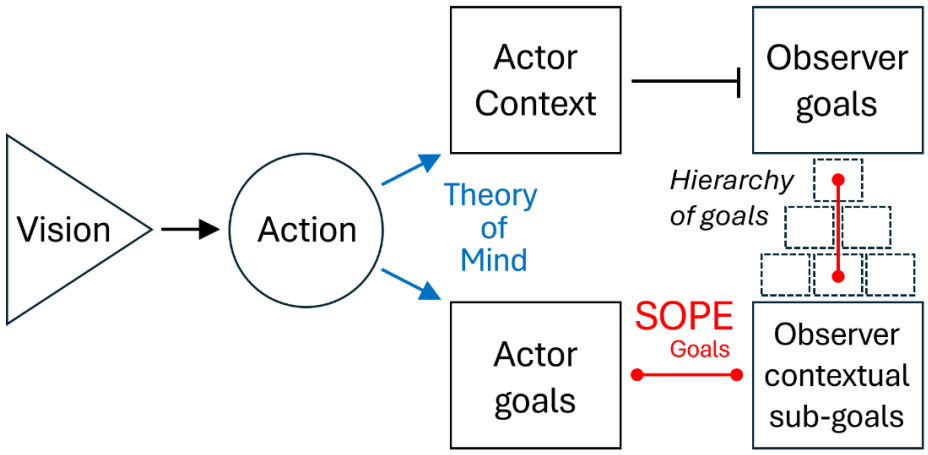
ToM is involved in slow social judgments, and the next subsection will discuss how ToM generates a second form of SOPE based on self-other goal dissimilarity. However, it must be also briefly noted that, during slow judgments, SOPE$_{Action}$ presumably does not entirely disappear. Instead, it seems most sensible that SOPE$_{Action}$ is repeatedly computed and continuously influences cognition, driving an Observer to minimize it. However, over time, the degree of SOPE$_{Action}$ may shift somewhat. Rather than just being modulated by an action's expected outcomes, over a slower time scale, the action's actual outcomes have time to be assessed and may begin to modulate SOPE$_{Action}$. Such processes fuse the model-free and model-based pathways for social judgment posed by Cushman (2013). Also over time, ToM-based inferences would help an Observer construct a more accurate representation of the Actor's context and should thus inform which habit representations should be used to compute SOPE$_{Action}$. This is analogous to decision-making accounts of how model-based reasoning can map the effects of an action, and then model-free processes help clarify the valuations of those effects and suggest later behaviors (Keramati et al., 2016; Kool et al., 2018). Overall, several implications stem just from the premises laid out thus far, but there remain several propositions left to make on the core mechanisms underlying social judgment.

4.3.3. SOPE$_{Goals}$ reflects goal comparisons

ToM is posed to be the backbone for a second SOPE computation: SOPE$_{Goals}$, which represents discrepancies between an Observer's goals and their perception of an Actor's goals. The exact goals selected for comparison depend on an Actor's context and an Observer's prior beliefs about how context-specific goals relate to more universal goals (Figure 2). Compared to model-free/based views that focus just on habits (deontology) and outcomes (consequentialism), invoking ToM and goal comparison is poised to enhance explanatory power.

Figure 2

Self-other prediction error linked to inferred goal discrepancies



Note. Visual information is processed as an observed action. Then, via Theory of Mind, an Observer infers the Actor's context and goals; although not illustrated, these inferences are influenced by the Observer's prior beliefs along with any other relevant visual information. SOPE$_{Goals}$ is computed as the extent of dissimilarity between the Actor's goals and the Observer's relevant goals. The relevant goals are determined via the Actor's perceived context modulating a hierarchy of universal goals and contextual sub-goals. Although this diagram focuses on how observed Actions prompt Theory of Mind, SOPE$_{Goals}$ would also emerge from any social information leading to inferences about an Actor's goals (e.g., due to hearing gossip).

Illustrating these processes, consider again the example where Bob sees Anne steal from the store. Bob cares for the store's well-being and values its harm as -3 utility. Because he saw Anne steal, Bob infers that she cares less about harming the store than him. Specifically, Bob infers that Anne values store harm at only -1 utility (difference of 2). Per $SOPE_{Goals}$, Bob judges Anne less harshly than Charles who also infers Anne values the store at -1 utility, but Charles knows the storeowner and values the store's harm at -5 utility (difference of 4). Bob also judges Anne less harshly than Dave. Dave values store harm at -3 utility like Bob, but Dave infers Anne takes sadistic pleasure from harming the store and she values it at +1 utility (difference of 4). Note that this example focuses on visible harm to another entity, but one's goals can be abstract. For instance, individuals may value being loyal or being pure and perceive violations of these states as being harmful (Graham et al., 2011; Royzman et al., 2015; Schein et al., 2016). Misalignment of any such goals would contribute to $SOPE_{Goals}$, causing a negative reaction.

4.3.4. Goal hierarchies

Goal comparisons and the generation of $SOPE_{Goals}$ are presumed to occur at different levels of a goal hierarchy, and this expands the explanatory power of the $SOPE_{Goals}$ mechanism. Like many claims here, this proposition is based on decision-making research, which shows that the brain organizes value structures hierarchically, such that sub-goals are pursued to achieve higher goals (Merel et al., 2019; Ribas-Fernandes et al., 2019). Leveraging this organization to make a decision relies on one's context regulating which sub-goals are currently relevant (Hunter & Daw, 2021; Palminteri & Lebreton, 2021). Applied to social judgment, when an Observer considers the Actor's context, goal comparison would specifically concern the context's upregulated sub-goals. This allows for "asymmetric" but harmonious relationships. For instance, consider roommates Isaac and Jada, who assign chores so Isaac cooks and Jada cleans. Even though Jada never cooks herself, if Isaac neglects cooking, then Jada will be annoyed. At a

surface level, if Isaac devalues cooking, this makes his goals more like Jada's. However, *cleaning* is specifically Jada's sub-goal for the higher aim of *being a good roommate*, and when she takes Isaac's perspective, *cooking* is the most relevant sub-goal. Hence, when Isaac fails to cook, the discrepancy in higher goals creates SOPE and explains Jada's irritation. These types of relationships, wherein individuals assume different roles, are ubiquitous in the social world, and thus it is critical that a theory of judgment accounts for them (Rai & Fiske, 2011).

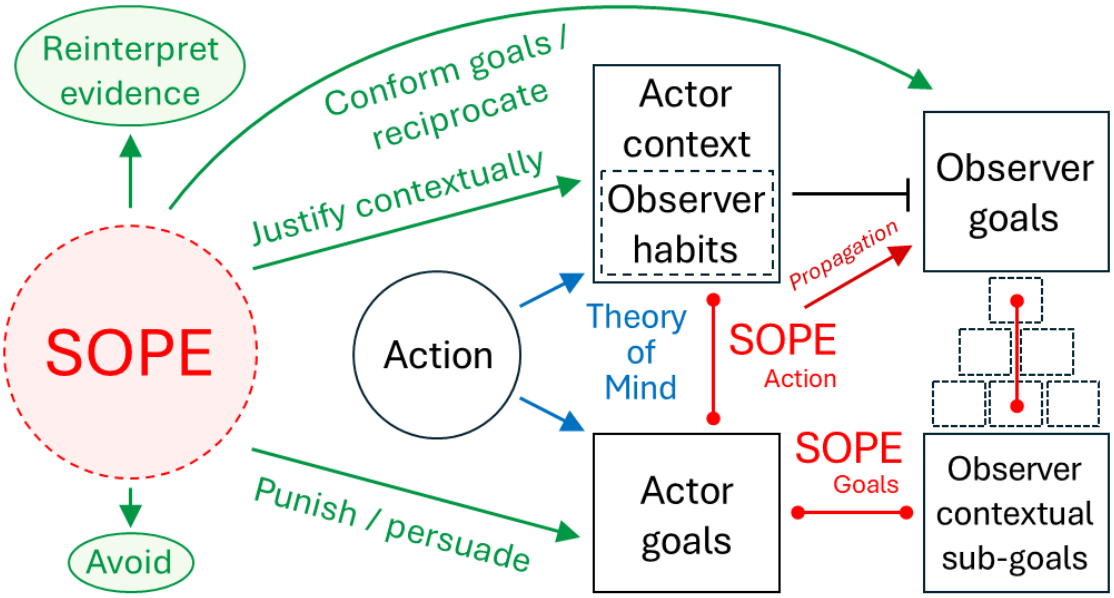
In engaging ToM and goal hierarchies, $SOPE_{Goals}$ elevates the concepts of self-other similarity and reciprocity into a framework that supports complex human coordination and the emergence of phenomena like contractualism – the idea that society is awash with cooperative agreements and moral violations constitute actions that defy these agreements. That is, SOPE allows stable and efficient social interactions with respect to factors such as relationship models, bargaining power, and comparative advantage (Le Pargneux & Cushman, 2024; Rai & Fiske, 2011). Consider, for instance, the relationship between a principal investigator and their trainee. Even if the trainee values a project sufficiently to invest 20 hours/week into it, they will not expect the same from their principal investigator. Instead, the trainee may expect just 5 hours of investment but in the form of wise feedback. Despite the trainee investing 4x more time, this is a surface-level difference influenced by each person's context. Deeper down, the trainee and investigator could actually be perfectly reciprocal in their valuations for each other – i.e., no $SOPE_{Goals}$ with respect to fundamental prosocial values like welfare trade-off ratios or inequity aversion levels (E. Fehr & Schmidt, 1999; Nowak, 2006). Experimental work indeed shows that such contextual differences are accounted for during judgment – e.g., individuals with greater power or with less to gain from cooperation are held to looser standards (Le Pargneux & Cushman, 2024). This is consistent with contractualism and the present arguments.

4.4. Action-trait unity

The two presented mechanisms for judging self-other similarity may appear to define judgment differently: $SOPE_{Action}$ may seem to describe action evaluations (e.g., fairness or wrongness) whereas $SOPE_{Goals}$ may seem to concern character evaluations (e.g., trust or respect). Although some studies may separately investigate each, at a deep level, action and character judgments are inseparable – e.g., work on the Side-Effect Effect shows how participants asked to judge a person's action may base this judgment on inferences about the person's character (Knobe, 2003; Laurent et al., 2019). Hence, rather than seeing $SOPE_{Action}$ and $SOPE_{Goals}$ as targeting different constructs, it is more informative to view $SOPE_{Action}$ as a heuristic for $SOPE_{Goals}$. Analogously to habits being heuristics for achieving goals, the compatibility of an Actor's actions with the Observer's habits will usually track the compatibility of the Actor's goals with the Observer's goals. Technically, an action could prompt low $SOPE_{Action}$ but high $SOPE_{Goals}$ – e.g., if a conservative vegan encounters vegan behavior, they may agree with it but then infer that the Actor is liberal, which they dislike. However, this arrangement would be exceptional, and these two forms of SOPE should usually yield converging conclusions but at different time scales. Adding to this idea about the unity of action and character judgments, both high $SOPE_{Action}$ and high $SOPE_{Goals}$ are offered to lead the same subsequent behavior: Both prompt Observers to respond in a way that minimizes SOPE (Figure 3).

Figure 3

Responses to minimize self-other prediction error



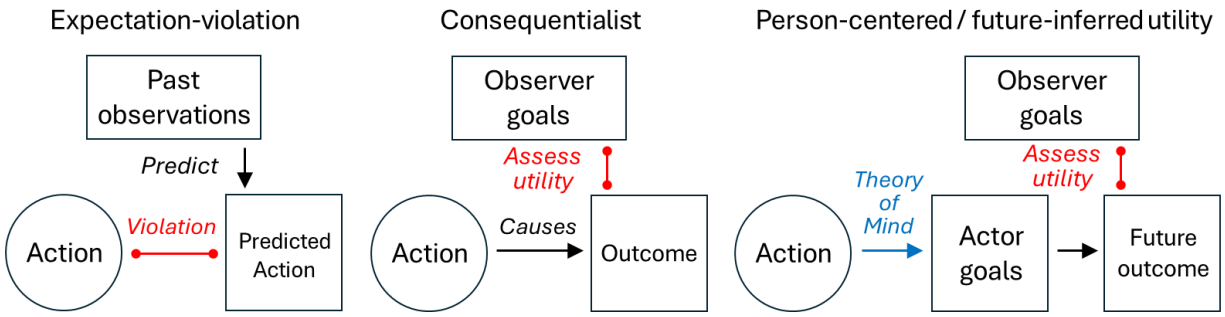
Note. The diagram summarizes the $SOPE_{Action}$ and $SOPE_{Goals}$ pathways shown in Figures 1 and 2 associated with a negative reaction to another person's action. Additionally, the present diagram shows how SOPE – coming from either of the two sources – may influence the Observer's subsequent cognitive and/or behavioral responses. SOPE may encourage contextual justification (updating the Actor's perceived context), punishment/persuasion (convincing the Actor to change their goals), conformity/reciprocity (modifying oneself to align with the Actor), or other possible responses. The diagram also notes a "reinterpret evidence" response, which is not directed toward any box because this is a general concept that may influence every aspect of processing – e.g., the Observer may conclude that an inappropriate action did not occur at all or that it did not actually elicit any unwanted outcomes. The diagram also notes an "Avoid' response, which is also not directed to other boxes, as it does not represent the Observer minimizing their discrepancy with the Actor but rather making the discrepancy less salient, which may involve other mechanisms.

5. Standard models of social judgment

With the complete set of SOPE pathways laid out, the focus pivots to discussing how the SOPE model is compatible with results from the moral judgment and behavioral economics literature. These research areas are developed and diverse, making their review ideal for evaluating the SOPE models' explanatory scope. In addition, this literature provides several alternative theories of social judgment, which can be compared to the SOPE model. Earlier theories can be largely categorized as (1) expectation-violation views that focus on whether another person's action violates social norms, (2) consequentialist views that hinge on the benefits/harms of another person's action, and (3) person-centered views that depend on how an observed action motivates inferences about the Actor's character (Figure 4). These perspectives are summarized below, alongside experimentally tested predictions about how these existing ideas relate to SOPE.

Figure 4.

Three standard models of moral judgment



Note. For each diagram, the red lines indicate the conflicts targeted by the corresponding theory. **Left.** Per the expectation-violation view, negative judgments arise from discrepancies between an observed action and a predicted action. Said predictions are based on social norms or beliefs about a specific Actor. **Middle.** Per the consequentialist view, negative judgments are rooted in an action's outcomes deviating from an Observer's desired outcomes (i.e., their goals). **Right.** Per the person-centered/future-inferred utility view, negative judgments arise from actions that

imply an Actor's future behavior will cause undesired outcomes. This can be understood as a discrepancy between inferred future outcomes and Observers' goals.

5.1. Expectation-violation view

Many moral judgment studies focus on social expectations and argue that Observers negatively judge behavior that violates these expectations. Often, these expectations are defined as descriptive norms, meaning that Observers are presumed to disapprove of atypical behavior. Indeed, both children and adults usually see typical behavior as being morally correct (Eriksson et al., 2021; Lindström et al., 2018; Roberts et al., 2019), and atypically may explain why disgust/purity violations are viewed negatively despite being ostensibly harmless (Gray & Keeney, 2015). Further, manipulating perceived typicality causes shifts in judgments (Hetu et al., 2017; Vavra et al., 2018; Walker et al., 2021; Xiang et al., 2013). Hence, social expectations have been posed to drive judgments or, at least, serve as heuristics for fast evaluations.

The expectation-violation view parallels $SOPE_{Action}$, both focusing on how individuals apply rules to quickly perceive actions. Accordingly, the two accounts produce somewhat similar explanations. For instance, per the expectation-violation view, hitting another person with a fish is seen as worse than punching them because the former deviates more from norms (Walker et al., 2021). However, harsher judgments could also be explained by Observers themselves being unlikely to attack someone with a fish. Experimental comparisons between these accounts are necessary, and later, Section 6 will present studies modeling SOPE and expectation-violation to assess which better predicts behavioral data. The studies also tested whether both views are correct, in a sense, but the expectation-violation effect on judgment is mediated by $SOPE_{Action}$. For example, in economic games, frequently observing selfish behavior discourages punishment of said selfish behavior (Hetu et al., 2017; Vavra et al., 2018; Xiang et al., 2013). This may be because the observations encourage individuals to conform and behave selfishly themselves,

which, in turn, makes them accept selfish acts (norm → self → judgment). However, per $SOPE_{Action}$, if Observers do not conform themselves, then norms should have no impact on judgment. The Section 6 studies would test this mediation and do so across several different ways of conceptualizing social expectations – e.g., as descriptive norms across a large population, as norms within a group of peers, or as person-specific predictions based on one Actor's personal history. In every case, $SOPE_{Action}$ is predicted to mediate all effects of expectations on judgment.

5.2. Consequentialist view

This second branch of literature focuses on how Observers' judgments reflect their perceptions of an action's outcomes. For instance, an action that harms a human rather than an animal will generally be seen as worse, and this may reflect Observers' subjective valuations of human versus animal lives (Cohen & Ahn, 2016; Engelmann & Waldmann, 2022). Some consequentialist theories argue that harm is a universal feature of moral violations (Gray et al., 2022; Schein & Gray, 2018). Available evidence indeed shows that negative judgments of even seemingly victimless acts involve perceived harm – e.g., gay marriage elicits the strongest disapproval among those who believe it hurts families (Royzman et al., 2015; Schein et al., 2016). Further, studies on "moral luck" have isolated and manipulated harm's effects, showing that participants will tend to judge an action as worse if it leads to a bad outcome even if just by chance (Nagel, 1979). This causal demonstration makes a strong case for outcomes being a necessary element of a complete social judgment theory.

For the SOPE model, harms are predicted to modulate the effect of $SOPE_{Action}$ on judgments, such that harmful dissimilar behavior leads to more negative reactions than minimally harmful dissimilar behavior. However, unlike pure consequentialist theories, the SOPE argument is that negative judgment will only occur if the harm is caused by behavior that

the Observer believes they would not do themselves. This explanation fills some explanatory gaps in harm-based perspectives. For instance, a theory must distinguish intentionally versus unintentionally harmful acts, as the former is judged more harshly (Ames & Fiske, 2013). Further, a theory must capture why some behaviors are harmful but not at all judged negatively by third parties – e.g., a domestic violence victim will not be judged negatively for ending a relationship, even if the breakup hurts their former partner (Royzman & Borislow, 2022). The factors at play in these cases are baked into $SOPE_{Action}$. That is, when processing the domestic violence victim's behavior, the Observer perceives that they would act the same way. These conceptual arguments would be formally tested in the Section 6 experiments on SOPE, specifically assessing whether: (i) outcomes indeed modulate the effect of $SOPE_{Action}$ on judgments and (ii) outcomes otherwise do not significantly influence judgments.

5.3. Person-centered view and Future Inferred Utility model

This third body of work focuses on ToM, and associated theories posit that judgment depends on an Observer's inferences about an Actor, such that actions most informative about an Actor's character beget the strongest evaluations (Carlson et al., 2022; Uhlmann et al., 2015). Formalizing this idea, "future-inferred utility" models predict that actions will elicit a negative judgment if they cause an Observer to update their beliefs about an Actor's goals in a way that implies the Observer will lose utility in the future (Gerstenberg et al., 2018; Krasnow et al., 2016). For example, seeing a colleague steal a pen may elicit harsh judgments, as even though theft is minor, it suggests the colleague may steal again, potentially in a major situation.

ToM theories cover several unique aspects of judgment. For instance, ToM views neatly capture findings showing that intentional acts beget stronger judgments, as observing intentional behavior induces stronger inferences about an Actor's character (Ames & Fiske, 2013). More generally, several studies have shown how inferences about an Actor mediate Observers' moral

judgments (Johnson & Ahn, 2021; Johnson & Park, 2021; Siegel et al., 2017; Tannenbaum et al., 2011). ToM-based perspectives also unpack some potentially unintuitive phenomena related to uncertainty and group membership: Inappropriate behavior by in-group members is sometimes judged laxly (Hewstone, 1990), which may be because Observers more readily imagine contextual justifications during ToM to justify inappropriate behavior. Yet at other times, members are judged more harshly (McLeish & Oxoby, 2007; Mendoza et al., 2014; Shinada et al., 2004), which may occur when contextual explanations are unlikely and updates about the Actor are larger relative to prior impressions. These cases illustrate how invoking ToM mechanisms comes with strong explanatory power.

Person-centered theories most parallel $SOPE_{Goals}$. Both focus on inferences about an Actor and describe judgment as proportional to the degree an action is diagnostic about the Actor's goals. However, the two perspectives diverge on exactly how inferences lead to negative judgments. For example, learning that an executive does not care about the environment leads to them being viewed negatively (Knobe, 2003; Laurent et al., 2019). The future-inferred utility view posits that negative judgment occurs because inferences about anti-environmental goals imply the executive's future behavior will be harmful. On the other hand, $SOPE_{Goals}$ posits that anti-environmental goals are judged negatively because they deviate from an Observer's goals (presuming the Observer cares about the environment). As with the case of $SOPE_{Action}$ and expectation-violation, discerning which explanation is most valid here too requires direct comparisons between these theories, and the Section 6 studies would do this. The studies were predicted to show that inferences about an Actor's goals entirely influence judgments based on the inferred goals' similarities to the Observer's own goals.

5.4. Comparisons

These three theoretical branches are each specialized for explaining different aspects of social interaction. Yet, these theories are also quite flexible. Table 1 shows this, summarizing ten of the reviewed psychological findings, how the three existing views each attempt to explain them, and how the findings can alternatively be explained by $SOPE_{Action}$ and/or $SOPE_{Goals}$. Although parsimony is subjective, the SOPE explanations seem, at minimum, plausible. However, one critical open question about many of the listed effects is whether they unfold over fast or slow timescales. For instance, it is known that violations by in-group members are sometimes judged more harshly than violations by out-group member (McLeish & Oxoby, 2007; Mendoza et al., 2014; Shinada et al., 2004), but it is unclear how quickly this group-membership effect emerges. If it only occurs slowly, then limited explanations by fast mechanisms (expectation-violation and/or $SOPE_{Action}$) would not challenge those mechanisms' validities. Thus, it is difficult to definitively compare these ideas based on earlier evidence, and hence, several original studies were performed that modeled and formally compared these mechanisms' powers in predicting judgment across different settings.

Table 1

Ten phenomena in social judgment

Result	Expectation-violation	Consequences (harm)	Trait inference (future utility)	$SOPE_{Action}$ (habit similarity)	$SOPE_{Goals}$ (goal similarity)
(1) Judgment of ostensibly harm-less (e.g., unpure) actions can be harsh	Such acts are atypical	Such acts are actually harmful	Such acts predict future harm	Such acts deviate from one's own behavior, which norms inform	Such acts prompt inferences about dissimilar goals
(2) Weird harms judged harsher than non-weird harms	Weird harms are more atypical	Typicality influences harm perceptions	Weirdness influences inferences about the Actor	Weird acts deviate more from own behavior	Weirdness influences inferences about the Actor
(3) Some harmful and intentional behavior is not judged negatively	Atypicality is required for negative judgment		Such acts do not cause negative inferences about the Actor	Such acts are not inconsistent with own behavior	Such acts do not cause negative inferences about the Actor
(4) Chance harms impact judgment (moral luck)	Harm biases the application of norms	Harm is the basis of judgment	Harm biases Observers toward harsher trait inferences	Effect of habit dissimilarity is modulated by outcomes/harms	Harm biases Observers toward harsher trait inferences
(5) Moral violations judged harsher than unconventional behavior		Harm is the basis of judgment	Moral violations beget stronger inferences about the Actor	Effect of habit dissimilarity is modulated by outcomes/harms	Moral violations beget stronger inferences about the Actor
(6) Diagnostic behavior (e.g., intentional acts) are judged harsher	Diagnostic acts more so violate norms		Diagnostic acts prompt larger inferences about others	Diagnostic acts deviate more from own behavior	Larger inferences increase perceived self-other dissimilarity
(7) Violations by in-group members are judged harsher	Some norms are stricter for in-group members		Larger change between prior and posterior beliefs		Membership raises the salience of dissimilarity
(8) Violations by in-group members are judged softer	Some norms are looser for in-group members	Membership biases perceptions of outcomes	Membership biases inference about others	Membership biases ToM-like act/context interpretation	Membership biases inferences about the context
(9) Innocuous similarity causes slight positive judgments			Similarity causes prosocial inferences	Similarity drives judgment	Similarity drives judgment
(10) Excessively altruistic behavior is viewed negatively	Such behavior breaks norms	Such behavior implies hidden harm	Such behavior implies ulterior motives	Such behavior deviates from observers' habits	Implies goals that deviate from observers' goals

Note. Each of the ten results are mentioned in the report. Blank spaces indicate no reasonable explanation to my knowledge.

6. Testing the proposed model

In eight studies across two reports, fifteen predictions were tested concerning SOPE and its relationship to existing views (Bogdan et al., 2023, 2024) (Table 2). The first report focused on modeling self-other comparisons generally and their effect on judgment. Then, the second report tested $SOPE_{Action}$ and $SOPE_{Goals}$ precisely. Overall, converging evidence supported the core claims of the SOPE argument, finding the SOPE models to be more predictive than expectation-violation, consequentialist, or person-centered views. In addition, the specific predictions about how these constructs interact (e.g., via mediation or moderation) were confirmed. The studies supporting these claims had high statistical power and consistently yielded robust results, which were often replicated between studies. These experiments, their methodology, and the results supporting the predictions are detailed below.

Table 2

Fourteen predictions about SOPE

General SOPE predictions		
Magnitude	**(A)**	Greater self-other dissimilarity (SOPE) predicts harsher judgments
Ergodic	**(B)**	SOPE predicts both within- and across-subject differences in judgment
Universal	**(C)**	SOPE's impact on judgment emerges cross-culturally
Bidirectional	**(D)**	Both dissimilarly prosocial and dissimilarly antisocial acts are judged negatively; albeit antisocial dissimilarity more so due to harm modulation
Causal	**(E)**	Manipulating reciprocity increases judgments even if outcomes are constant
Response-general	**(F)**	SOPE guides action judgments, character judgments, and later behavior
Task-general	**(G)**	SOPE effects emerge across moral vignettes and different economic games
Probabilistic	**(H)**	Bayesian ToM-reasoning guides the assessment of self-other similarity
Comparison predictions		
Expectations	**(1)**	When SOPE is accounted for, any effect of social expectations on judgment is fully mediated by Observers conforming themselves to the expectations
	(2)	Full mediation occurs regardless of expectations' content (generous/selfish)
	(3)	SOPE remains more predictive than expectation-violation regardless of how expectations are defined (e.g., social norms or person-specific expectations)
	(4)	Rather than being the basis of judgment, social expectations inform ToM
Outcomes	**(5)**	Outcomes modulate the effect of $SOPE_{Action}$ on judgment
	(6)	Outcomes otherwise have no direct impact on judgment
Person-centered	**(7)**	Goal differences ($SOPE_{Goals}$) better predict judgments than inferred goals themselves

Note. The Section 6 studies endorse every prediction, generally each one several times.

6.1. Self-other similarity and judgment

6.1.1. Magnitude, ergodicity and norm-violation comparisons

The first report consisted of four studies, modeling self-other comparisons (Bogdan et al., 2023). Study 1 and replication Study 2 used the Ultimatum Game, which is a two-player task wherein each trial one player proposes how to share a pool of money (e.g., $10), and their partner chooses whether to accept or reject the offer. Acceptance causes the money to be split as proposed whereas rejection causes neither player to receive any money (i.e., costly punishment). After each trial, participants switched between the two roles, meaning that analyses could link participants' proposed offers to their judgments of received offers. Also after each trial, participants were told that they would be paired with a new partner, among a large group of other human participants. Thus, participants could develop expectations about the offers typical players proposed, and the effects of these perceived descriptive norms could be studied.

Analyses confirmed several predictions: After proposing generous offers, participants became more likely to punish selfish offers, which suggests self-other comparisons (*magnitude prediction A*). In addition, participants who proposed most generously, on average, punishd selfishness the most (*ergodicity prediction B*). By contrast, participants' expectations of others' behavior, modeled as the means of their previously received offers, only indirectly impacted judgment. That is, receiving selfish offers only increased acceptance of selfishness if participants conformed and began proposing selfish offers themselves (norm → self → judgment) (*expectations prediction 1*). The mediation was reproduced in a condition where participants observed overwhelmingly selfish behavior in others, meaning that the present conclusions generalize across different norm settings (*expectations prediction 2*).

6.1.2. Universality, bidirectionality, and other expectations

Study 3 analyzed Public Goods Game data collected from 16 cities across the globe (Herrmann et al., 2008). Participants were organized into four-player groups. In each trial, players could contribute to a pot of money, which would be multiplied and then distributed equally among everyone. After the contribution phases, participants saw each other player's contribution amount and could choose to punish them. This game was repeatedly played with the same group. In each trial, SOPE was modeled from the perspective of each of the four players toward each of the other three people. The SOPE elicited by a given player's decision was defined as the absolute difference (|*other – self*|) between the player's contribution (*other*) and an evaluator's averaged contribution in past trials (*self*). As a comparison, perceived expectation-violation was also modeled as the absolute difference (|*other – norm*|) between a player's contribution and the other two player's mean contribution in past trials (*norm*).

As hypothesized, in 94% of cities, SOPE better predicted participants' punishments than did social-norm violation (*universality prediction C*). SOPE's advantage here is striking as social norms would have been prominent, given that this was a multi-player game, and each decision impacted everyone. Also in 94% of cities, SOPE was more predictive than the signed difference between another player's behavior and the evaluator's average (*other – self*). That is, the best model posited participants would punish behavior that was either more selfish or more generous than their own, validating this potentially surprising prediction (*bidirectional prediction D*).

6.1.3. Causality, dyadic expectations, and response-generality

Study 4 returned to the role-changing Ultimatum Game design, but now, participants were told that they would play with the same person repeatedly in multi-trial blocks, only changing partners between blocks. In reality, across blocks, participants played with various computer agents, each programmed with a different playstyle. The computers included: (i)

reciprocity agents that adjusted their behavior each trial to be more similar to the participants', (ii) generous agents that proposed high amounts and accepted nearly all offers, and (iii) control-condition agents that make choices by sampling from the decision probabilities of participants in the first two studies. Following each block, participants' overall impressions of their partners were assessed using the Trust Game (Berg et al., 1995).

Analyzing the Trust Game data showed that the reciprocity agent elicited similar levels of trust as the generosity agent (insignificant difference), and the reciprocity agent produced far higher trust than the control-condition agent despite the Ultimatum Game payouts being fairly similar. Given that this design manipulates perceived social alignment, this causally demonstrates its effects (*causal prediction E*). Additionally, whereas the first three studies investigated action fairness, by demonstrating these points using trust, SOPE's relevancy generalizes to character-level evaluations (*response-general prediction F*). As a separate branch of analysis, the study also examined participants' Ultimatum Game behavior, attempting to replicate the first three study's results while now modeling person/dyad-specific social expectations – i.e., an expectation that one's partner will treat the participant similarly to how they did in the past. As before, SOPE better tracked participants' judgments, confirming its predictive over different conceptualizations of expectation-violation (*expectation prediction 3*).

6.2. Dual-process and probabilistic SOPE

6.2.1. SOPE$_{Action}$ and task-generality

The second report investigated self-other similarity while modeling SOPE$_{Action}$ and SOPE$_{Goals}$ specifically and comparing them to all three of the existing theories of judgment presented (Bogdan et al., 2024). Study 1 began this investigation with a moral vignette design. Participants evaluated characters' trustworthiness after learning about their inappropriate behavior in an ambiguous context (e.g., seeing another person steal but being unsure what was

stolen). Afterward, participants reported how they themselves would behave in related specific contexts (e.g., [i] would they steal food for their family, [ii] would they steal a luxury item for themselves, etc.). Participants also reported how they expected most people to behave in each specific context. Analyses modeled a simplified form of $SOPE_{Action}$ as the percentage of specific contexts where participants said they too would behave inappropriately. As hypothesized $SOPE_{Action}$ better predicted participants' trust evaluations than did violations of participants' expectations about typical behavior. Hence, the SOPE conclusions generalize to a vignette design (*task-general prediction G*).

6.2.2. $SOPE_{Action}$ and probabilistic ToM

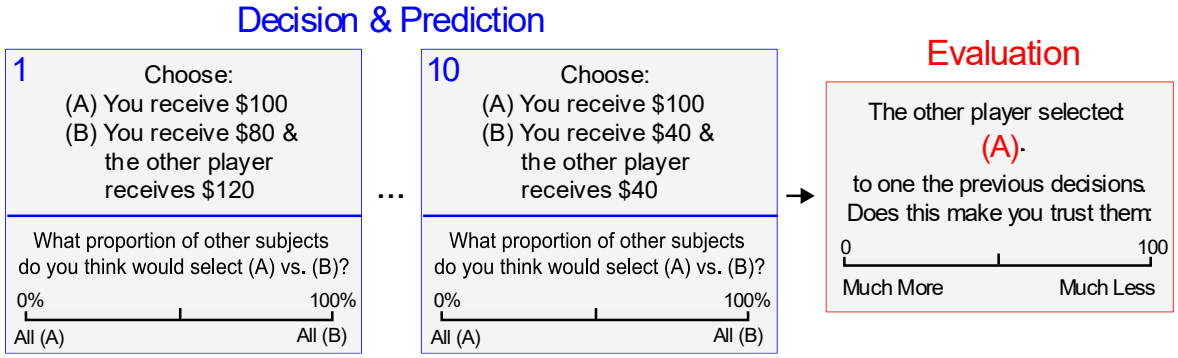
Study 2 investigated $SOPE_{Action}$ using an economic game. The game was analogous to the Study 1 vignette design, but the new task's design also allowed modeling whether participants used Bayesian inference to parse uncertainty about an Actor's context. Hence, the study permitted stricter tests of $SOPE_{Action}$ and specifically the idea that participants compute the likelihood that they would behave the same as an Actor.

Study 2 involved a two-player economic game, wherein each trial, participants chose whether to share money with a partner across ten different *contexts* (Figure 5). For instance, one context asked participants whether they would sacrifice $60 so their partner receives $40, whereas another context asked whether they would sacrifice $20 so their partner receives $100. For each context, participants also predicted what percentage of people would choose to sacrifice. After making decisions and predictions for all ten contexts, participants were told of another player's choice (sacrifice or not) for one of the ten contexts. Critically, participants were not told which context the choice was for. Thus, participants would recognize whether the player behaved selfishly or generously but not to what extent. For instance, being unwilling to sacrifice $60 so the other player receives $40 may be seen as reasonable whereas being unwilling to

sacrifice $20 so the other player receives $100 may be seen as highly selfish. Based on just knowing their partner's decision in one unknown context, participants were asked to judge the player's trustworthiness using a 0-to-100 scale. Participants were told they would change partners between trials.

Figure 5

Task diagram of Study 2 from the second report



Note. This diagram represents one trial of the economic game, containing a Decision & Prediction phase followed by a Trust Evaluation phase. Participants were told that the other player's decision (here, A) was directed to one of the ten contexts for which they just made decisions and predictions (here, two contexts are shown). Figure is from Bogdan et al. (2024).

The first set of analyses modeled $SOPE_{Action}$ as the percentage of contexts wherein participants made the same choice as in the decision being evaluated. Consistent with the first study's results, participants who behaved selfishly in more contexts trusted others who behaved selfishly. Analogously, participants who behaved generously in more contexts trusted others who behaved generously. These predictions were more accurate than those by a competing expectation-violation model.

Next, new analyses tested $SOPE_{Action}$ more precisely, examining whether participants used Bayesian inference to infer the overall likelihood that they made the same choice as their

partner. The analyses involved modeling participants' perceptions of each context's likelihood of being the one selected. For example, suppose a participant predicts 50% of people will sacrifice in context 1 and 25% of people will sacrifice in context 2. If the participant is told that their partner indeed sacrificed, then the participant will perceive context 1 as being twice as likely to have been the context for which the decision was made. In turn, for assessing the overall likelihood that they acted the same way ($SOPE_{Action}$), participants' own decision-making in context 1 contributes twice as much as their decision-making in context 2. The behavioral data indeed showed that this occurred and influenced judgments. That is, when participants evaluated a selfish action, their judgment was most influenced by their own decision-making in contexts where they predicted most people would behave selfishly. Likewise, when participants evaluated a generous action, their judgment was most influenced by their own decision-making in contexts where they predicted most people would behave generously. Hence, participants seem to use Bayesian reasoning when judging self-other similarity (*probabilistic prediction*). Social expectations – i.e., predictions about how others will behave – inform this reasoning but are not the standard for judgment themselves (*expectations prediction 4*).

6.2.3. $SOPE_{Action}$, outcome modulation, and consequentialist-view comparisons

Studies 3A and 3B were similar to Study 2, but instead of using ten contexts per trial, there were just two. This lowered the informational load, and participants could better gauge the likely outcomes of their partner's decision. Hence, a consequentialist model could be tested – defined based on the monetary effects of their partner's choice, averaged plainly across the two contexts; preliminary tests also attempted weighing contexts by their Bayesian-inferred likelihood, but this had little impact on the consequentialist model's predictions.

As hypothesized, the consequentialist model was weaker than $SOPE_{Action}$ for predicting perceptions of others' trustworthiness (Study 3A). $SOPE_{Action}$'s predictive advantage also

emerged when the design instead asked participants to judge the fairness of their partner's choice (Study 3B). Yet also as hypothesized, the consequentialist variable modulated the effect of $SOPE_{Action}$ on judgment, such that high $SOPE_{Action}$ and larger outcomes led to stronger judgments across both Study 3A and 3B (*consequentialist prediction 5*). When the $SOPE_{Action}$ effect was accounted for, outcomes otherwise did not influence evaluations (*consequentialist prediction 6*).[1]

6.2.4. $SOPE_{Goals}$, dual processes, and person-centered-view comparisons

Finally, Studies 3A and 3B were also used to investigate $SOPE_{Goals}$ and compare it to a person-centered model. The studies' money-sharing design allowed quantifying player's prosocial/antisocial goals (outcome valuations), which was done using the Fehr-Schmidt model (E. Fehr & Schmidt, 1999). This is a social decision-making model that defines a person's perception of an action's utility as its payout minus the utility lost from inequity. This inequity utility loss is weighed by the person's personal inequity aversion; defined by one parameter representing distaste for receiving less than another person (α) and one parameter representing distaste for receiving more than another person (β). The Fehr-Schmidt model – combined with a standard logistic/softmax function for converting utility to decision probabilities – was fit in two contexts: First, this framework was used to quantify each participant's inequity concerns (α_{self} and β_{self}), which involved fitting the model for each participant to identify the parameters that best predicted their decisions to sacrifice or not across the experiment. Second, the framework was used to measure each participant's perceptions of their partner's inequity aversion after learning about their partner's decision. For this, the framework was fit many times, estimating α_{other} and β_{other} for each trial; given the uncertainty regarding the two possible contexts, modeling assumed participants used Bayesian inference to discern the likelihood of each one.

[1] The expectation-violation model was also tested and, replicating the Study 2 results, it was less predictive than $SOPE_{Action}$. Both comparisons are provided in the report but note that the consequentialist modulation findings are only provided in the report's associated OSF repository.

After measuring participants' inequity aversions and their perceptions of others' inequity aversions, $SOPE_{Goals}$ was measured for each trial. $SOPE_{Goals}$ was defined as the absolute difference between the participant's inequity aversion and their perception of the other player's aversion: $SOPE_{Goals,\alpha} = |\alpha_{self} - \alpha_{other}|$ and $SOPE_{Goals,\beta} = |\beta_{self} - \beta_{other}|$. For both Study 3A (trust) and Study 3B (fairness), $SOPE_{Goals,\beta}$ best predicted judgments. By contrast, participants' raw perceptions of the other player (β_{other}) represent the person-centered view, and this quantity was starkly less predictive than in $SOPE_{Goals,\beta}$ in both studies. Hence, social judgment is sensitive to self-other differences in goals, not inferences about the goals themselves (*person-centered prediction 7*).[2,3] Beyond this comparison, this finding also provides further general evidence that individuals probabilistically reason during judgment.

6.3. Conclusions and outstanding questions

In sum, the targeted studies endorse the SOPE model and show its predictive power relative to three previously proposed mechanisms of how individuals evaluate others' behavior. To be sure, there also exist theories of judgment that combine these three established mechanisms. For instance, the Theory of Dyadic Morality states that moral condemnation is directed toward non-normative behavior that causes harm, and perceived harm exists as either visible immediate negative outcomes or implied negative outcomes (Gray et al., 2022; Schein & Gray, 2018). Multi-faceted theories like these are challenging to definitively specify and compare: How exactly should an action's unexpectedness be weighed against its harmfulness when making an overall prediction? The SOPE model faces similar difficulties: If judgment involves both $SOPE_{Action}$ and $SOPE_{Goals}$, then precisely how much does each contribute?

[2] None of the alpha inequity measures ($SOPE_{Goals,\alpha}$ and α_{other}) were predictive.
[3] A model representing the signed difference in betas ($\beta_{self} - \beta_{other}$) was also tested and was less predictive than the absolute difference ($SOPE_{Goals,\beta} = |\beta_{self} - \beta_{other}|$). This reiterates the earlier conclusions about SOPE's bidirectionality, effectively predicting negative reactions to both antisocial and overly prosocial behavior.

Nevertheless, by demonstrating that these two SOPE computations are more predictive than all three alternative core social-judgment mechanisms, this intends to be a strong demonstration.

Aside from this experimental evidence, the proposed SOPE view also carries several conceptual advantages for understanding social cognition relative to earlier theories of moral judgment. *First*, because SOPE is not just a theory of moral judgment but rather concerns social information processing generally, the present rhetoric does require defining "morality" – an arguably socially constructed category that may allude to cognitive mechanisms but precisely delineate them (McHugh et al., 2022). *Second*, the SOPE view also represents behavior and cognition following judgment: Namely, after prediction error, participants will act and think to minimize SOPE, such as via conformity, reciprocity, punishment, justification, avoidance, persuasion, etc. This logic about subsequent behavior follows from prediction error theory. However, extending alternative accounts to describe behavior following judgment leads to challenges – e.g., a consequentialist view may explain that Observers will reciprocate or punish to improve future outcomes, but it is unclear how such a view could capture a drive to justify misdeeds. *Third*, because the SOPE account has been shaped by the idea that social judgment repurposes decision-making systems, this adds to the computational plausibility of the arguments and mechanisms put forth. Additionally, given the richness of the decision-making literature, including on the neuroscientific side, parallels to social cognition may inspire further hypotheses concerning the nature of judgment.

Finally, the studies in this section tested the most critical predictions about SOPE. However, in describing how the SOPE view covers diverse social phenomena (Section 4), several claims were made but remain to be tested. For instance, Section 4.3.4 discusses asymmetric relationships, wherein two individuals have different roles, and how judgment in this setting involves contextual modulation to determine which goals are most relevant for the

comparison. Testing this type of claim may require psychophysiological evidence. Another open topic for research is examining whether the effects of innocuous similarity/dissimilarity (i.e., mimicry) indeed invoke the same mechanisms as those for moral judgment. Although the behavioral evidence reviewed shows that these both influence participants' impressions of others, these ideas about common computational pathways still need to be confirmed.

7. Conclusion

The Golden Rule puts forth that you should treat others how you wish to be treated. The SOPE model agrees but also inverts the rule, stating that you likewise expect others to treat you how you treat them. Further, the model proposes that your subsequent behavior constitutes attempts to reestablish the perception of reciprocal treatment. These ideas share key similarities with earlier accounts of moral judgment and are compatible with their supporting evidence. However, as the studies testing SOPE suggest, the SOPE predictions are more precise in predicting individuals' reactions to others' behavior. These conclusions, along with helping to synthesize different aspects of social cognition, notably also carry practical value: The everyday quarrels that pop up in social life are misalignments. However, unlike in laboratory studies, these real-world misalignments can often be resolved through communication. The present research emphasizes the benefits of communication in disputes and learning about other parties, which may ultimately show how each party actually agrees on what behavior is appropriate for what contexts.

References

Alicke, M. D. (1993). Egocentric standards of conduct evaluation. *Basic and Applied Social Psychology*, *14*(2), 171–192.

Ames, D. L., & Fiske, S. T. (2013). Intentional harms are worse, even when they're not. *Psychological Science*, *24*(9), 1755–1762.

Anderson, J. A., & Kinnally, E. L. (2021). Behavioral mimicry predicts social favor in adolescent rhesus macaques (Macaca mulatta). *Primates*, *62*(1), 123–131.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In *Organizational influence processes* (pp. 295–303). Routledge.

Axelrod, R. (1980). Effective choice in the prisoner's dilemma. *Journal of Conflict Resolution*, *24*(1), 3–25.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Bandura, A., & Walters, R. H. (1977). *Social Learning Theory* (Vol. 1). Prentice-hall Englewood Cliffs, NJ.

Barkan, R., Ayal, S., & Ariely, D. (2015). Ethical dissonance, justifications, and moral behavior. *Current Opinion in Psychology*, *6*(Dec), 157–161.

Barnby, J., Raihani, N., & Dayan, P. (2022). Knowing me, knowing you: Interpersonal similarity improves predictive accuracy and reduces attributions of harmful intent. *Cognition*, *225*, 105098.

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization (vol 12, pg 17, 2017). *Social Cognitive and Affective Neuroscience*, *12*(11), 1833–1833. https://doi.org/10.1093/scan/nsx060

Benozio, A., House, B. R., & Tomasello, M. (2023). Apes reciprocate food positively and negatively. *Proceedings of the Royal Society B*, *290*(1998), 20222541.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142.

Bocian, K., Baryla, W., Kulesza, W. M., Schnall, S., & Wojciszke, B. (2018). The mere liking effect: Attitudinal influences on attributions of moral character. *Journal of Experimental Social Psychology*, *79*, 9–20.

Bogdan, P. C., Dolcos, F., Moore, M., Culpepper, S., Kuznetsov, I., & Dolcos, S. (2023). Social Expectations are Primarily Rooted in Reciprocity: An Investigation of Fairness, Cooperation, and Trustworthiness. *Cognitive Science*, *47*(8), e13326.

Bogdan, P. C., Dolcos, S., & Dolcos, F. (2024). How Likely Is it that I Would Act the Same Way: Modeling Moral Judgment During Uncertainty. *Cognitive Science*, *48*(11), e70010.

Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin*, *119*(1), 111.

Buon, M., Jacob, P., Loissel, E., & Dupoux, E. (2013). A non-mentalistic cause-based heuristic in human social evaluations. *Cognition*, *126*(2), 149–155.

Buttelmann, D., Zmyj, N., Daum, M., & Carpenter, M. (2013). Selective imitation of in-group over out-group members in 14-month-old infants. *Child Development*, *84*(2), 422–428.

Carlson, R. W., Bigman, Y. E., Gray, K., Ferguson, M. J., & Crockett, M. (2022). How inferred motives shape moral judgements. *Nature Reviews Psychology*, *1*(8), 468–478.

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, *18*(8), 414–421. https://doi.org/10.1016/j.tics.2014.04.012

Chao, Z. C., Takaura, K., Wang, L., Fujii, N., & Dehaene, S. (2018). Large-scale cortical networks for hierarchical prediction and prediction error in the primate brain. *Neuron*, *100*(5), 1252-1266. e3.

Charness, G., Naef, M., & Sontuoso, A. (2019). Opportunistic conformism. *Journal of Economic Theory*, *180*, 100–134.

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General*, *145*(10), 1359.

Collette, S., Pauli, W. M., Bossaerts, P., & O'Doherty, J. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *Elife*, *6*, e29718.

Constant, A., Ramstead, M. J., Veissière, S. P., & Friston, K. J. (2019). Regimes of expectations: An active inference model of social conformity and human decision making. *Frontiers in Psychology*, *10*, 679.

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*(1), 47–69.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–292.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

De Groeve, B., & Rosenfeld, D. L. (2022). Morally admirable or moralistically deplorable? A theoretical framework for understanding character judgments of vegan advocates. *Appetite*, *168*, 105693.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin*, *139*(2), 477.

Di Gregorio, F., Steinhauser, M., & Maier, M. E. (2016). Error-related brain activity and error awareness in an error classification paradigm. *Neuroimage*, *139*, 202–210.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Earle, T. C., & Siegrist, M. (2006). Morality information, performance information, and the distinction between trust and confidence 1. *Journal of Applied Social Psychology*, *36*(2), 383–416.

Engelmann, N., & Waldmann, M. R. (2022). How to weigh lives. A computational model of moral judgment in multiple-outcome structures. *Cognition*, *218*, 104910.

Eriksson, K., Vartanova, I., Ornstein, P., & Strimling, P. (2021). The common-is-moral association is stronger among less religious people. *Humanities and Social Sciences Communications*, *8*(1), 1–8.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, *59*, 255–278.

Fehr, E., Fischbacher, U., & Tougareva, E. (2002). Do high stakes and competition undermine fairness? Evidence from Russia. *Evidence from Russia (July 2002)*.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *The Quarterly Journal of Economics*, *108*(2), 437–459.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868.

Fehr, R., & Gelfand, M. J. (2010). When apologies work: How matching apology components to victims' self-construals facilitates forgiveness. *Organizational Behavior and Human Decision Processes*, *113*(1), 37–50.

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour*, *3*(5), 426–435.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140.

Fischer-Lokou, J., Martin, A., Guéguen, N., & Lamy, L. (2011). Mimicry and propagation of prosocial behavior in a natural setting. *Psychological Reports*, *108*(2), 599–605.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? From expectations to responsibility judgments. *Cognition*, *177*, 122–141.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366.

Gray, K., & Keeney, J. E. (2015). Impure or just weird? Scenario sampling bias raises questions about the foundation of morality. *Social Psychological and Personality Science*, *6*(8), 859–868.

Gray, K., MacCormack, J. K., Henry, T., Banks, E., Schein, C., Armstrong-Carter, E., Abrams, S., & Muscatell, K. A. (2022). The affective harm account (AHA) of moral judgment: Reconciling cognition and affect, dyadic morality and disgust, harm and purity. *Journal of Personality and Social Psychology*, *123*(6), 1199.

Gray, K., Ward, A. F., & Norton, M. I. (2014). Paying it forward: Generalized reciprocity and the limits of generosity. *Journal of Experimental Psychology: General*, *143*(1), 247.

Greene, J. D. (2009). The cognitive neuroscience of moral judgment. *The Cognitive Neurosciences*, *4*, 1–48.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144–1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400.

Hamlin, J. K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proceedings of the National Academy of Sciences*, *108*(50), 19931–19936.

Herd, S., Krueger, K., Nair, A., Mollick, J., & O'Reilly, R. (2021). Neural mechanisms of human decision-making. *Cognitive, Affective, & Behavioral Neuroscience*, *21*(1), 35–57.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367. https://doi.org/10.1126/science.1153808

Hetu, S., Luo, Y., D'Ardenne, K., Lohrenz, T., & Montague, P. R. (2017). Human substantia nigra and ventral tegmental area involvement in computing social error signals during the ultimatum game. *Social Cognitive and Affective Neuroscience*, *12*(12), 1972–1982. https://doi.org/10.1093/scan/nsx097

Hewstone, M. (1990). The 'ultimate attribution error'? A review of the literature on intergroup causal attribution. *European Journal of Social Psychology*, *20*(4), 311–335.

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019). People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General*, *148*(3), 520.

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*, *26*(11), 959–971.

Hunter, L. E., & Daw, N. D. (2021). Context-sensitive valuation and learning. *Current Opinion in Behavioral Sciences*, *41*, 122–127.

Hutcherson, C. A., Montaser-Kouhsari, L., Woodward, J., & Rangel, A. (2015). Emotional and utilitarian appraisals of moral dilemmas are encoded in separate areas and integrated in ventromedial prefrontal cortex. *Journal of Neuroscience*, *35*(36), 12593–12605.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, *22*(1), 39–44.

Irwin, K., & Horne, C. (2013). A normative explanation of antisocial punishment. *Social Science Research*, *42*(2), 562–570.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110.

Jin, K., Ting, F., He, Z., & Baillargeon, R. (2024). Infants expect some degree of positive and negative reciprocity between strangers. *Nature Communications*, *15*(1), 7742.

Johnson, S. G., & Ahn, J. (2021). Principles of moral accounting: How our intuitive moral sense balances rights and wrongs. *Cognition*, *206*, 104467.

Johnson, S. G., & Park, S. Y. (2021). Moral signaling through donations of money and time. *Organizational Behavior and Human Decision Processes*, *165*, 183–196.

Jung, H., Seo, E., Han, E., Henderson, M. D., & Patall, E. A. (2020). Prosocial modeling: A meta-analytic review and synthesis. *Psychological Bulletin*, *146*(8), 635.

Kawamura, Y., & Kusumi, T. (2020). Altruism does not always lead to a good reputation: A normative explanation. *Journal of Experimental Social Psychology*, *90*, 104021.

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron*, *100*(2), 424–435.

Kelly, D., Stich, S., Haley, K. J., Eng, S. J., & Fessler, D. M. (2007). Harm, affect, and the moral/conventional distinction. *Mind & Language*, *22*(2), 117–131.

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences*, *113*(45), 12868–12873.

Kim, M., Park, B., & Young, L. (2020). The psychology of motivated versus rational impression updating. *Trends in Cognitive Sciences*, *24*(2), 101–111.

Klein, N., & Epley, N. (2014). The topography of generosity: Asymmetric evaluations of prosocial actions. *Journal of Experimental Psychology: General*, *143*(6), 2366.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908–911.

Kool, W., Cushman, F. A., & Gershman, S. J. (2018). Competition and cooperation between multiple reinforcement learning systems. *Goal-Directed Decision Making*, 153–178.

Köster, M., Kayhan, E., Langeloh, M., & Hoehl, S. (2020). Making sense of the world: Infant learning from a predictive processing perspective. *Perspectives on Psychological Science*, *15*(3), 562–571.

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological Science*, *27*(3), 405–418.

Kruglanski, A. W., & Szumowska, E. (2020). Habitual behavior is goal-driven. *Perspectives on Psychological Science*, *15*(5), 1256–1271.

Kulesza, W., Chrobot, N., Dolinski, D., Muniak, P., Bińkowska, D., Grzyb, T., & Genschow, O. (2022). Imagining is Not Observing: The Role of Simulation Processes Within the Mimicry-Liking Expressway. *Journal of Nonverbal Behavior*, 1–14.

Kulesza, W., Dolinski, D., Huisman, A., & Majewski, R. (2014). The echo effect: The power of verbal mimicry to influence prosocial behavior. *Journal of Language and Social Psychology*, *33*(2), 183–201.

Lahat, A., Helwig, C. C., & Zelazo, P. D. (2013). An event-related potential study of adolescents' and young adults' judgments of moral and social conventional violations. *Child Development*, *84*(3), 955–969.

Laurent, S. M., Reich, B. J., & Skorinko, J. L. (2019). Reconstructing the side-effect effect: A new way of understanding how moral considerations drive intentionality asymmetries. *Journal of Experimental Psychology: General*, *148*(10), 1747.

Le Pargneux, A., & Cushman, F. (2024). Moral judgment is sensitive to bargaining power. *Journal of Experimental Psychology: General*.

Lewicki, R. J., Polin, B., & Lount Jr, R. B. (2016). An exploration of the structure of effective apologies. *Negotiation and Conflict Management Research*, *9*(2), 177–196.

Lindström, B., Jangard, S., Selbing, I., & Olsson, A. (2018). The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology: General*, *147*(2), 228.

Lockwood, P. L., Apps, M. A., & Chang, S. W. (2020). Is there a "social" brain? Implementations and algorithms. *Trends in Cognitive Sciences*.

Martin, J. W., Buon, M., & Cushman, F. (2021). The Effect of Cognitive Load on Intent-Based Moral Judgment. *Cognitive Science*, *45*(4), e12965.

Matz, D. C., & Wood, W. (2005). Cognitive dissonance in groups: The consequences of disagreement. *Journal of Personality and Social Psychology*, *88*(1), 22.

McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. L. (2022). Moral judgment as categorization (MJAC). *Perspectives on Psychological Science*, *17*(1), 131–152.

McLeish, K. N., & Oxoby, R. J. (2007). *Identity, cooperation, and punishment*.

Mendoza, S. A., Lane, S. P., & Amodio, D. M. (2014). For members only: Ingroup punishment of fairness norm violations in the ultimatum game. *Social Psychological and Personality Science*, *5*(6), 662–670.

Merel, J., Botvinick, M., & Wayne, G. (2019). Hierarchical motor control in mammals and machines. *Nature Communications*, *10*(1), 1–12.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological Review*, *126*(2), 292.

Miller, R. M., & Cushman, F. (2013). Aversive for me, wrong for you: First-person behavioral aversions underlie the moral condemnation of harm. *Social and Personality Psychology Compass*, *7*(10), 707–718.

Miller, R. M., Hannikainen, I. A., & Cushman, F. A. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion*, *14*(3), 573.

Monin, B., Sawyer, P. J., & Marquez, M. J. (2008). The rejection of moral rebels: Resenting those who do the right thing. *Journal of Personality and Social Psychology*, *95*(1), 76.

Nagel, T. (1979). Moral luck. *Mortal Questions [New York: Cambridge University Press, 1979]*, 31–32.

Nichols, S. (2002). Norms with feeling: Towards a psychological account of moral judgment. *Cognition*, *84*(2), 221–236.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560–1563. https://doi.org/10.1126/science.1133755

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291–1298.

Otterbacher, J., Ang, C. S., Litvak, M., & Atkins, D. (2017). Show me you care: Trait empathy, linguistic style, and mimicry on Facebook. *ACM Transactions on Internet Technology (TOIT)*, *17*(1), 1–22.

Page, K. M., & Nowak, M. A. (2001). A generalized adaptive dynamics framework can describe the evolutionary Ultimatum Game. *Journal of Theoretical Biology*, *209*(2), 173–179.

Palminteri, S., & Lebreton, M. (2021). Context-dependent outcome encoding in human reinforcement learning. *Current Opinion in Behavioral Sciences*, *41*, 144–151.

Parkinson, C., & Wheatley, T. (2015). The repurposed social brain. *Trends in Cognitive Sciences*, *19*(3), 133–141.

Parks, C. D., & Stone, A. B. (2010). The desire to expel unselfish members from the group. *Journal of Personality and Social Psychology*, *99*(2), 303.

Pauli, W. M., Cockburn, J., Pool, E. R., Perez, O. D., & O'Doherty, J. P. (2018). Computational approaches to habits in a model-free world. *Current Opinion in Behavioral Sciences*, *20*, 104–109.

Quiros, J. D. V., Kapcak, O., Hung, H., & Cabrera-Quiros, L. (2021). Individual and joint body movement assessed by wearable sensing as a predictor of attraction in speed dates. *IEEE Transactions on Affective Computing*.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*(1), 57.

Rains, S. A. (2016). Language style matching as a predictor of perceived social support in computer-mediated interaction among individuals coping with illness. *Communication Research*, *43*(5), 694–712.

Ribas-Fernandes, J. J., Shahnazian, D., Holroyd, C. B., & Botvinick, M. M. (2019). Subgoal-and goal-related reward prediction errors in medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *31*(1), 8–23.

Roberts, S. O., Ho, A. K., & Gelman, S. A. (2019). The role of group norms in evaluating uncommon and negative behaviors. *Journal of Experimental Psychology: General*, *148*(2), 374.

Royzman, E. B., & Borislow, S. H. (2022). The puzzle of wrongless harms: Some potential concerns for dyadic morality and related accounts. *Cognition*, *220*, 104980.

Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of Julie and Mark: Unraveling the moral dumbfounding effect. *Judgment & Decision Making*, *10*(4).

Sarin, A., Ho, M. K., Martin, J. W., & Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. *Cognition*, *208*, 104544.

Saxe, R., & Houlihan, S. D. (2017). Formalizing emotion concepts within a Bayesian model of theory of mind. *Current Opinion in Psychology*, *17*, 15–21.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32–70.

Schein, C., Ritter, R. S., & Gray, K. (2016). Harm mediates the disgust-immorality link. *Emotion*, *16*(6), 862.

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology: General*, *141*(3), 433.

Schneider, D., Slaughter, V. P., Becker, S. I., & Dux, P. E. (2014). Implicit false-belief processing in the human brain. *NeuroImage*, *101*, 268–275.

Schütt, H. H., Kim, D., & Ma, W. J. (2024). Reward prediction error neurons implement an efficient code for reward. *Nature Neuroscience*, 1–7.

Schweinfurth, M. K., & Call, J. (2019). Revisiting the possibility of reciprocal help in non-human primates. *Neuroscience & Biobehavioral Reviews*, *104*, 73–86.

Shamay-Tsoory, S. G., Saporta, N., Marton-Alper, I. Z., & Gvirts, H. Z. (2019). Herding brains: A core neural mechanism for social alignment. *Trends in Cognitive Sciences*.

Shinada, M., Yamagishi, T., & Ohmura, Y. (2004). False friends are worse than bitter enemies:"Altruistic" punishment of in-group members. *Evolution and Human Behavior*, *25*(6), 379–393.

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, *167*, 201–211.

Siegrist, M., Cvetkovich, G., & Roth, C. (2000). Salient value similarity, social trust, and risk/benefit perception. *Risk Analysis*, *20*(3), 353–362.

Sparkman, G., & Attari, S. Z. (2020). Credibility, communication, and climate change: How lifestyle inconsistency and do-gooder derogation impact decarbonization advocacy. *Energy Research & Social Science*, *59*, 101290.

Suchak, M., & de Waal, F. B. (2012). Monkeys benefit from reciprocity without the cognitive burden. *Proceedings of the National Academy of Sciences*, *109*(38), 15191–15196.

Tan, F., & Xiao, E. (2018). Third-party punishment: Retribution or deterrence? *Journal of Economic Psychology*, *67*, 34–46.

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology*, *47*(6), 1249–1254.

Theriault, J. E., Young, L., & Barrett, L. F. (2020). The sense of should: A biologically-based framework for modeling social pressure. *Physics of Life Reviews*.

Tisak, M. S., & Turiel, E. (1988). Variation in seriousness of transgressions and children's moral and conventional concepts. *Developmental Psychology*, *24*(3), 352.

Traulsen, A., & Nowak, M. A. (2006). Evolution of cooperation by multilevel selection. *Proceedings of the National Academy of Sciences*, *103*(29), 10952–10955.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81.

Vavra, P., Chang, L. J., & Sanfey, A. G. (2018). Expectations in the Ultimatum Game: Distinct Effects of Mean and Variance of Expected Offers. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00992

Veissière, S. P., Constant, A., Ramstead, M. J., Friston, K. J., & Kirmayer, L. J. (2020). Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences*, *43*.

Volk, S., Nguyen, H., & Thöni, C. (2019). Punishment under threat: The role of personality in costly punishment. *Journal of Research in Personality*, *81*, 47–55.

Walker, A. C., Turpin, M. H., Fugelsang, J. A., & Białek, M. (2021). Better the two devils you know, than the one you don't: Predictability influences moral judgments of immoral actors. *Journal of Experimental Social Psychology*, *97*, 104220.

Whiten, A., Harrison, R. A., McGuigan, N., Vale, G. L., & Watson, S. K. (2022). Collective knowledge and the dynamics of culture in chimpanzees. *Philosophical Transactions of the Royal Society B*, *377*(1843), 20200321.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational Substrates of Norms and Their Violations during Social Exchange. *Journal of Neuroscience*, *33*(3), 1099–1108. https://doi.org/10.1523/jneurosci.1642-12.2013

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*, *111*(4), 931.

Yucel, M., Hepach, R., & Vaish, A. (2020). Young children and adults show differential arousal to moral and conventional transgressions. *Frontiers in Psychology*, *11*, 548.