Value computation in social and strategic contexts: evidence and theory[*]

Isabelle Brocas
USC and CEPR

Juan B. González
USC

Daniel Hermosillo Gutiérrez
USC

January 2024

Abstract

Humans' choices often reflect their considerations for the well-being of others. In some instances, these choices also showcase their capacity to gauge the intentions of others and utilize this insight to inform strategic decisions. This human ability to empathize and use this knowledge to navigate social situations hinges on intricate interactions among different brain networks. Recent research has unveiled compelling findings indicating the involvement of the central executive network and the default mode network in generating signals, which are subsequently combined to shape the subjective values of social choices. This review aims to explore this evidence within the context of both other-regarding concerns and strategic interactions, two paradigms that are frequently treated as distinct, yet they share common neural processes.

Keywords: Other regarding concerns; strategic behavior; social brain; theory of mind.

JEL Classification: D01, D87, D90.

[*] We thank Jerry Jiang, Erica Schug, Karthik Srinivasan and Ziwei Yu for their excellent research assistance on an early review published as a LABEL report (Brocas et al. (2022)). We also thank all the members of the Los Angeles Behavioral Economics Laboratory (LABEL) for their insights.

1 Introduction

Humans are inherently social animals. When they find themselves in scenarios where their decisions impact the well-being of others or are influenced by the decisions of others, their choices tend to demonstrate either a positive or negative concern for the welfare of those involved, as well as a concern for social perception. Simultaneously, their choices can also involve a more detached evaluation of the potential reactions and actions of others, aimed at making the most self-interested decisions. Economists categorize these two sets of considerations differently: social preferences delve into how individuals incorporate concerns for others into their choices, while game theory models selfish behavior through a rational assessment of the mental states of others. Social attitudes represent a multifaceted concept that encompasses a range of considerations pertinent to social contexts. These considerations include preferences regarding the fair distribution of wealth, tendencies towards reciprocating positive actions, the willingness to trust and be trusted by others, and one's preferences regarding adherence to social norms. On the other hand, strategic aptitude is another intricate construct that involves the capacity to discern the intentions of others and make reasoned deductions about their likely courses of action.

Yet, how does the brain carry out the necessary computations for assessing choices in social contexts and determining appropriate behavioral responses? Are the computational processes for social preferences and strategic interactions distinct? At first glance, both scenarios require individuals to factor in the presence of others and could possibly rely on shared mental algorithms. In this article, we discuss the current understanding of the neural circuitry involved in social valuation and social behavior. Earlier analyses, as outlined by Fehr and Camerer (2007) and Ruff and Fehr (2014), have primarily focused on social preferences.

Social preferences are gauged through a range of decision tasks designed to explore various facets of this intricate concept. Some of these tasks are presented as individual decision-making scenarios, where individuals must select between different allocations. These tasks can unveil an individual's concern for their own wealth in comparison to that of others and their inclination toward efficiency. In contrast, other tasks involve games, typically played in a one-time setting with anonymous partners and real stakes. These games assess how individuals place value on their relative wealth. Depending on the specific game, they also shed light on how these allocations may be influenced by our expectations of others or what we believe others expect from us. Game theoretic models predict outcomes driven by self-interest, which often diverge from what is empirically observed. In practice, more socially desirable choices are frequently made instead. In essence, all of these games provide insights into the intricate social attitudes that lead us to deviate from purely selfish decisions.

In most of the games mentioned earlier, an individual's motivation can be rooted in their inherent concern for the well-being of others and/or their ability to anticipate the actions of others. Put differently, social preferences are intertwined with strategic considerations. For example, a person may initially be willing to share with a partner who also appears inclined to share, but subsequently opt not to do so upon witnessing their partner's selfish behavior. This highlights the challenge of distinguishing between an individual's emotions and their capacity to understand the emotions of others – a dimension economists capture in preference representations – and their ability to accurately gauge the thoughts and strategies of others – a dimension economists associate with perspective-taking and strategic behavior.

From an evolutionary standpoint, advantageous adaptations that have shaped humans into social animals likely operated on both of these dimensions. In psychology, the concept of Theory of Mind (ToM) generally pertains to the ability to comprehend other individuals by attributing mental states to them, including beliefs, intentions, desires, emotions, or thoughts. Simultaneously, certain games appear to demand more than just the capacity to empathize with others; they necessitate grappling with intricate concepts, such as social norms, and applying logical reasoning that extends beyond social interactions. For instance, participants in a trust game must not only anticipate their partner's potential actions if they send a payment (requiring ToM) but also engage in a recursive thought process, employing backward induction arguments to arrive at a decision, and consider the expectations dictated by social norms (higher-order considerations). In summary, ToM and strategic thinking are inherently highly intertwined, and the cognitive functions supporting these intricate calculations are vital for making appropriate choices in social contexts.

From a neuroscientific standpoint, the aforementioned behavior is predictably underpinned by three primary networks that facilitate the following functions: (i) the representation of preferences and the evaluation of rewards, known as the *reward network*; (ii) the capacity for introspection and consideration of others, often referred to as the *social brain*; and (iii) the processing of higher order cognitive representations through the cognitive *executive network*. In contrast to a straightforward valuation task, such as selecting between two fruits, which primarily engages the goal value network, consisting of the reward system and specific regions within the ventromedial prefrontal cortex responsible for representing goal value, the process of social valuation involves a broader array of brain regions, each specializing in representing the additional dimensions inherent to social decision-making. As we will explore further, studies on social valuation reveal the involvement of brain regions that may not be engaged in straightforward valuation tasks, including the activation of the *salience network*. We shall note at the outset that we will disregard important details about brain functioning in order to provide a simple representation of the processes involved in social behavior.

The rest of the article is organized as follows. Section 2 describes the components that form the concept of social decision-making. Section 3 discusses how specific brain processes are involved in the representation of these components. Section 4 addresses how this information can be leveraged to model decision-making.

2 Building blocks of social decision-making

2.1 Other regarding concerns

Substantial evidence supports the notion that individuals are genuinely concerned about the welfare of others, indicating the presence of social preferences. Although various models exist to illustrate these social preferences, they all adhere to a common framework in which an individual's utility is directly tied to their own outcomes and the outcomes of others. Social preferences are measured via a variety of decision tasks that collectively tell us about how people perceive allocations of money and how perception varies across contexts.

Many games build on the classical Dictator Game. In the simplest version, a decision-maker possesses the authority to distribute a sum of money between themselves and the recipient. The recipient, in this context, holds no influence or decision-making power. Contrary to purely self-interested behavior, the typical allocation in this game is approximately 30% of the dictator's endowment (Engel, 2011). This underscores the prevalence of individuals who genuinely care about the well-being of others. The specific amount given to the recipient is subject to various influencing factors, including social proximity to the recipient, anonymity, gender, concerns related to fairness, and individual personality traits, among others. In the Ultimatum Game, the dictator takes on the role of a proposer, and the recipient holds the power to accept or reject the proposed allocation. Should the recipient decline the offer, neither party receives any reward. In this scenario, even a self-interested proposer must take into account the recipient's social preferences. If both parties were entirely self-interested, the proposer would logically offer nothing to the recipient, who, in turn, should accept the offer. However, empirical evidence reveals that offers below the 20% threshold are rejected nearly half the time. Additionally, on average, proposers anticipate these rejections and therefore extend offers amounting to approximately 40% of their endowment (Camerer, 2003; Cochard et al., 2021). This behavior suggests that proposers believe that recipients value fairness and strategically make offers aligned with these expectations. The Third Party Dictator Game introduces a third party who witnesses the dictator's decisions and has the option to penalize the dictator at a personal cost. Notably, approximately half of these observers opt to impose penalties when they observe selfish deviations from an equal split. This phenomenon underscores the willingness of individuals to make personal sacrifices to uphold the principles of fairness (Fehr and Fischbacher, 2004).

Social preferences also encompass the notion of trust, which is studied via the Trust Game. Here, two participants, A and B, each possess an initial endowment. A is given the option to send a transfer to B. B is privy to these transfers, which are subsequently multiplied by a factor known to both parties. B then faces the decision of retaining the entire sum or sending a portion of it back to A. Once again, A must anticipate and consider B's social preferences when making their strategic decision. In a scenario where both A and B exhibit purely self-interested behavior, A would abstain from sending any transfer, as they could reasonably expect B to retain the entire amount. However, empirical behavioral evidence paints a different picture: On average, A sends approximately 50% of their endowment, while B reciprocates by returning approximately 40% (Johnson and Mislin, 2011).

2.2 Collaborative behavior

Inherent within humans is their innate capacity for cooperative behavior. This propensity becomes particularly evident in situations where pursuing self-interest would result in unfavorable societal outcomes, or when multiple equilibria can be achieved, each with distinct social implications.

For instance, in the Public Good Game, each participant possess an initial endowment that they can collectively invest in a shared project, distributing their contributions in any proportion they choose. The optimal outcome for the group as a whole is to maximize their collective earnings through investment, but individually, each player stands to gain more by not investing at all. The Nash equilibrium in this game suggests that no one should invest. However, empirical evidence reveals that this equilibrium is uncommon, with players, on average, investing approximately half of their endowment. Over time, players tend to adjust their behavior and invest less in repeated iterations of the game (Ledyard, 1995; Gunnthorsdottir et al., 2007; Levitt and List, 2007). Similarly, in the Prisoner's Dilemma, both players face a choice between cooperation and defection. If both opt for defection, they end up in a worse position than if they were to both cooperate. However, choosing cooperation unilaterally results in the smallest individual payoff. Technically, the Prisoner's Dilemma can be seen as a variant of the Public Goods Game. In the one-shot game and in finitely repeated iterations, the Nash equilibrium is for both players to defect. However, when the game is played repeatedly over an indefinite period, cooperation can be upheld as an equilibrium through the use of threats to defect as a response to the partner's defection. In practical settings, people often choose to cooperate in all variations of this game (Camerer, 2003; Embrey et al., 2018).

Coordination games like the Battle of the Sexes or the Stag Hunt exhibit multiple equilibria in single-shot scenarios. However, when these games are played repeatedly, individuals tend to converge towards efficient and equitable outcomes. For example, in the Battle of the Sexes, where players A and B each favor distinct equilibria, we often witness people alternating between these equilibria as they interact over time.

Behavior in these situations do not necessarily require to directly care about others. It may emerge from the collective understanding that some outcomes are mutually beneficial and worth selecting even for self interest, hence becoming a norm that people comply with instinctively.

2.3 Mentalizing about others

Theory of Mind (ToM) refers to the ability to think about one's and others' mental states, and it is a key component in social interactions. It allows individuals to interpret and predict the behaviors of others. The ability to infer or even recognize the different mental states of others cannot be directly observed and therefore depends on responses to specially fabricated tasks and social interactions. A distinction is made between interpreting what an individual feels and what they think. In general, perspective-taking is categorized into affective and cognitive components. Researchers employ four main types of tasks to distinguish when a participant is engaging in affective Theory of Mind (aToM) or cognitive Theory of Mind (cToM). Tasks such as the false belief task and strategic games predominantly examine cToM, as they require participants to consider the thoughts and decision-making of the central character. In contrast, to assess aToM, researchers often employ the "mind in the eyes" task and cartoon vignettes, which prompt participants to provide insights into the emotional state of a depicted character.

False belief tasks test whether a person understands that another individual may possess beliefs that differ from reality. For example, a participant is shown a character observing the placement of an object and when the character is no longer observing, the object is displaced. The participant is then asked where he or she thinks the character believes the object is after starting to observe again. They should answer that the character will look for the object when he left it, but participants who have not developed cToM will believe the character knows what they themselves have observed. Experimental evidence suggests that cToM develops gradually but not fully. In particular, using the false belief task as a measure for cToM, there is evidence that children 3 years old cannot correctly respond to a first-order false belief task, while 4-5 years old can (Meinhardt et al., 2011). However, children 5-6 years old fail second-order false belief tasks, despite having first-order ToM (Arslan et al., 2017). A similar trend exists in a different age bracket with adolescents performing worse on third-order false belief tasks compared to young adults (Valle et al., 2015). Behavior in strategic games is very heterogeneous and differences often reflect logical difficulties attached to specific games and algorithms required to strategize correctly (Camerer, 2003). But they also reflect limitations in ToM. For instance, experimental subjects do better in games of complete information than in games of private information. Indeed, strategic settings require not only to form a mental model of an opponent but also a mental model of the opponent's mental model about oneself. In the presence of private information, a player must realize that their opponent does not have a piece of information that they have. This turns out to be very counterintuitive and requires precise use of ToM (Kolling et al., 2021). Therefore, a player tends to build mental models ignoring the asymmetry in information (Brocas et al., 2014).

Children develop the ability to recognize facial expressions as early as 12 months old, so, therefore, it is unsurprising that they develop the aToM component relatively early. Current evidence shows that children can perform fairly well in the mind in the eyes task even before adolescence (Manstead, 1995; van der Meulen et al., 2017; Moor et al., 2012). Typically developing children have a similar average score in the mind in the eyes task as typically developing adults (Vogindroukas et al., 2014; Moor et al., 2012).

A recent study conducted in our lab (Alfonso et al., 2023) reveals a strong interconnection between the abilities to assess others' emotions (aToM) and anticipate their behavior (cToM). The study also demonstrates that aToM and cToM develop together. This suggests that emotional components play a role as inputs in cognitive processes, and are processed by overlapping brain regions.

2.4 Creating logical models about others

As for any other decision-making problem, reasoning in the social domain requires to represent internal goals to influence behavior by selecting appropriate perceptual information, inhibiting inappropriate or automatic responses, and directing attention. Correlational studies have shown that performance in tests of executive function or math ability is associated with social behavior in social games, although not in a consistent way (Chen et al., 2013b; Ben-Ner et al., 2004; Benjamin et al., 2013). Game theoretic studies in adults and children also show that higher intelligence or better cognitive skills are associated with behavior closer to Nash predictions (Brañas-Garza et al., 2012; Gill and Prowse, 2016; Proto et al., 2019, 2022; Fe et al., 2022).

3 Brain regions and networks that support social evaluations

The neural correlates of social decision-making are observed by measuring neural responses associated with behavior in many of the tasks we have described in the previous section. Studies have helped mapping separate but very interrelated regions and networks. For each of these networks, we will restrict to the most fundamental regions taxed during social processing.

3.1 Reward network

The evaluation of rewards involves several crucial brain structures, with key players including the ventral tegmental area (VTA), the striatum, the thalamus, the amygdala, and the ventromedial prefrontal cortex (vmPFC). Reward value is encoded in several steps (see Figure 1 for a schematic diagram). Initially, the VTA responds to stimuli indicating that a reward is present (e.g. a food on display) and subsequently releases dopamine signals to the striatum. Within the striatum, dopaminergic neurons calculate the expected reward associated with the stimulus (Pessiglione et al., 2006; Schultz, 2016). The difference between the expected reward and the actual reward associated with a certain stimulus is known as reward prediction error, a key component of decision-making and reinforcement learning (Schultz, 2007).

Furthermore, the degree of uncertainty regarding the reward is also encoded by the dopaminergic neurons within the striatum, working in conjunction with the amygdala (Fiorillo et al., 2003). This fast process encodes low-dimensional information, conveying both the expected value and the variance of the reward to the vmPFC through the thalamus. The vmPFC accumulates and integrates this information with input from other brain regions (Levy and Glimcher, 2012).

It is now established that all types of rewards share common neural mechanisms (Bhanji and Delgado, 2014; Du and Chang, 2015; Ruff and Fehr, 2014). The anticipation of rewards stemming from both prosocial and selfish behaviors is encoded within the striatum, culminating in the accumulation of a unified value within the vmPFC (Hutcherson et al., 2015). However, the valuation of social interactions often necessitates consideration of more intricate attributes, such as social norms, perceived intentions, and expectations of reciprocity. These representations are typically provided by distinct neural networks.

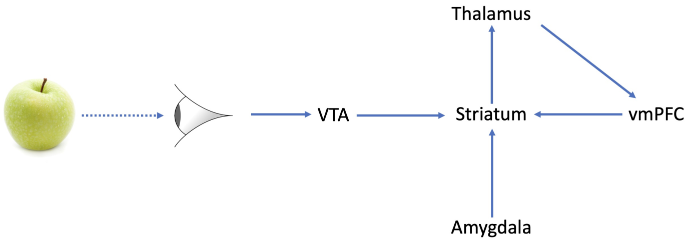

Figure 1: Simplified diagram of the **Reward System**. The VTA and Amygdala encode low-dimensional information contained in the stimulus (here an apple) and send it to the striatum. The striatum integrates this information and sends it to the vmPFC via the thalamus.

3.2 Social brain

The default mode network was initially recognized as the network that becomes active by default when individuals daydream or allow their minds to wander. However, it promptly deactivates when they engage in externally oriented tasks, which led to its classification as a task-negative network. Nonetheless, the functions of the default mode network are more intricate than this initial assessment suggests, as it also plays a role in tasks involving mentalizing or self-directed thought (Chiong et al., 2013).

The default mode network is distributed across the brain in a topographically dispersed manner, consisting of regions situated at a considerable distance from the somatosensory network. It is primarily composed of the posterior cingulate cortex, the medial prefrontal cortex and the angular gyrus (Figure 2). This spatial arrangement underscores the network's focus on internal thought processes (Smallwood et al., 2021). Furthermore, this dispersion is accompanied by a high degree of functional and connectivity diversity, with some researchers suggesting that the default mode network is the composite result of overlapping subnetworks (Andrews-Hanna et al., 2010; Buckner and DiNicola, 2019).

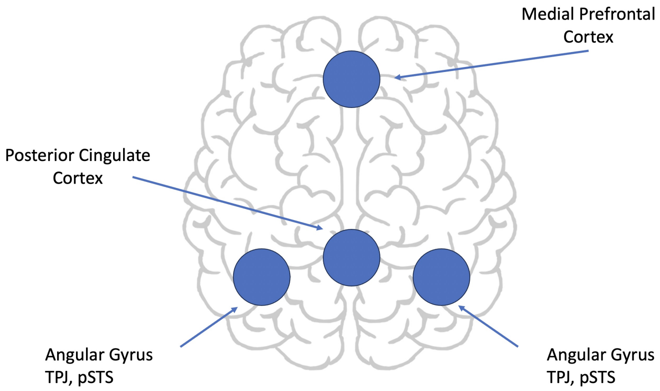

Figure 2: Simplified description of the **Social Brain**. The default mode network is primarily composed of the posterior cingulate cortex, the TPJ, the pSTS, the medial prefrontal cortex and the angular gyrus.

The default mode network functions as a system primarily engaged in self-referential processes, episodic memory, language, mentalizing about the self and others, and the ability to mentally travel through time (Smallwood et al., 2021). Its distinctive blend of rich self-referential and social cognitive processes positions it as a vital hub for integrating individual self-awareness and interactions with others (Yeshurun et al., 2021; Zadbood et al., 2017). Of these functions, mentalizing about oneself and others plays a pivotal role in social decision-making.

Key players in this network include the posterior superior temporal sulcus (pSTS) and the temporo-parietal junction (TPJ). The pSTS becomes active during social interactions, particularly when other individuals respond to the subject's actions, and it encodes essential information about these interactions, such as their emotional valence (Emonds et al., 2011; Isik et al., 2017; Varrier and Finn, 2022). On the other hand, the TPJ is associated with representing the perceived value of others (Crockett and Lockwood, 2018; Cutler and Campbell-Meiklejohn, 2019; Hutcherson et al., 2015), as well as their beliefs and intentions (Carter et al., 2012; Carter and Huettel, 2013; Schurz et al., 2014). Further it is activated in cooperation games (Sun et al., 2023; Thompson et al., 2021) and it reflects individual differences in social preferences in competitive settings (van den Bos et al., 2013).

The information processed within the default mode network is subsequently relayed to the reward network, where it is combined to form a unified choice value (Atique et al., 2011; Du and Chang, 2015; Hutcherson et al., 2015).

The default mode network collaborates with other networks to handle specific cognitive functions. Concerning theory of mind (ToM), the cognitive aspect (cToM) uniquely engages the dorsolateral prefrontal cortex (dlPFC) and the dorsal subnetwork of the default mode network (Bukowski, 2018; Schurz et al., 2014; Van Veluw and Chance, 2014; Young et al., 2010). In contrast, the affective aspect of ToM (aToM) uniquely involves the amygdala and the ventral subnetwork (Atique et al., 2011; Healey and Grossman, 2018). This aligns with the role of the dlPFC in the central executive network, which is active during various cognitive tasks. Likewise, the involvement of the ventromedial prefrontal cortex (vmPFC) and the amygdala in aToM corresponds to their significance in processing emotional stimuli.

Beyond mentalizing, the default mode network also plays a role in complex social decision-making, such as moral reasoning. It serves as the endpoint for moral decision processes, functioning as a platform that integrates multidimensional inputs from different brain regions. In the context of moral decision-making, the default mode network combines signals to generate a unified representation, which then informs the vmPFC (Chiong et al., 2013; Boccia et al., 2017; Obeso et al., 2018; Qu et al., 2022).

3.3 Central Executive Network

The central executive network, also known as the frontoparietal network, is anchored in the dlPFC and the posterior parietal cortex (PPC) (Menon, 2011). This network is needed to maintain and manipulate information in working memory, solve problems, and support complex cognition and goal-directed behavior (Constantinidis and Klingberg, 2016; Koechlin and Summerfield, 2007).

The central executive network, particularly the dlPFC, plays a central role in processes related to reward and subjective value evaluation. It provides critical input regarding the attributes of an item to the vmPFC (Levy and Glimcher, 2011; Rudorf and Hare, 2014). Notably, the dlPFC has been observed to encode the variability of multi-attribute objects (Kahnt et al., 2011) and exhibits heightened activity when trade-offs between different attributes are required (McFadden et al., 2015).

In the context of food choices, the dlPFC has been shown to modulate value (Camus et al., 2009; Hare et al., 2011a,b; Gluth et al., 2012; Sokol-Hessner et al., 2012; Chen et al.,

2018) and craving (Fregni et al., 2008; Hall et al., 2017). It is also involved in processes related to self-regulation and self-control (Hutcherson et al., 2012; Harris et al., 2013; Chen et al., 2018) and has been linked to conditions such as eating disorders (Brooks et al., 2011; Foerde et al., 2015; He et al., 2019; Lowe et al., 2019; Dalton et al., 2020) and addictive behavior (Koob and Volkow, 2010; Nakamura-Palacios et al., 2016).

Furthermore, the dlPFC has been found to have functional connections with regions responsible for value coding in self-control paradigms (Hare et al., 2009) and multi-attribute decision-making tasks (Rudorf and Hare, 2014). Collectively, this body of evidence underscores the pivotal role of the dlPFC in value-based decision-making, particularly when integrating multiple dimensions (such as taste, health, and future outcomes) to derive subjective value (Brocas and Carrillo, 2021). The dlPFC is notably more active when choices involve conflicts that need to be resolved (Baumgartner et al., 2011; de Wit et al., 2009), and it contributes to choice consistency (Zyuzin et al., 2023).

The role of the central executive network in social valuation is complex. The dlPFC serves as a crucial locus for encoding moral objectives and signaling breaches of social norms (Crockett et al., 2017; Qu et al., 2022; Zinchenko and Arsalidou, 2018). For example, the dlPFC detects instances of advantageous inequality (Gao et al., 2018) and identifies unfair offers in games like the ultimatum game, often predicting their rejection (Sanfey et al., 2003). This view is bolstered by causal evidence: inhibiting dlPFC activity has been shown to reduce rejections of unfair offers (Baumgartner et al., 2011).

Additionally, the dlPFC plays a pivotal role in strategic reasoning. Depending on the context, strategic thinking may result in either selfish or prosocial choices (Cutler and Campbell-Meiklejohn, 2019; McAuliffe et al., 2017; Steinbeis et al., 2012). Causal evidence supporting this notion is derived from studies employing anodal transcranial direct current stimulation (tDCS), which depolarizes neurons and heightens cortical excitability, or cathodal tDCS, which hyperpolarizes neurons and diminishes excitability. For instance, applying anodal tDCS to the dlPFC has been shown to increase strategic giving in the ultimatum game while reducing altruistic giving in the dictator game. By contrast, the application of cathodal tDCS yields the opposite effects (Ruff et al., 2013). Furthermore, the cortical thickness of the dlPFC is linked to more prosocial behavior in individuals who tend to be self-oriented, but it is associated with selfish behavior in pro-self individuals (Tanaka et al., 2023).

The dlPFC plays a pivotal role in integrating signals related to social norms and their breaches, as well as signals associated with the value of strategic interactions with partners. It then forwards these integrated inputs to the reward system for consideration in the valuation process (Baumgartner et al., 2011). Additionally, the dlPFC regulates the activity of the default mode network, often in conjunction with the salience network.

The dlPFC is also known to be engaged in strategic games like the prisoner's dilemma, where participants must exert cognitive control (Emonds et al., 2012; Gradin et al., 2016). Research examining differences in strategic thinking has revealed increased dlPFC activation in individuals with higher strategic thinking (Nagel et al., 2018). Furthermore, the dlPFC encodes the depth of recursion in cooperative games, reflecting the participants' executive sophistication (Yoshida et al., 2010).

3.4 Salience Network

The functions carried out by the central executive network consume a substantial amount of energy. As a result, the brain has evolved to minimize their default utilization and only engage this network when necessary. This decision is orchestrated by the salience network (Figure 3), an intrinsic connectivity network comprised of key hubs in the anterior insular cortex (AIC), the anterior cingulate cortex (ACC), with the participation of the amygdala (Menon and Uddin, 2010).

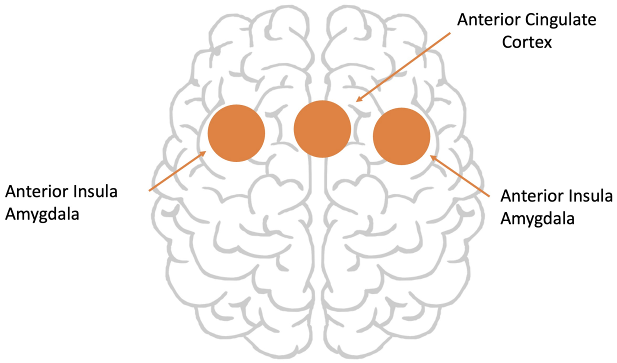

Figure 3: Simplified description of the **Salience Network**. The key players of the salience network within social paradigms are the AIC, the ACC and the Amygdala.

The AIC serves as the central hub of the salience network. Information from the posterior insular cortex, which receives inputs from the somatosensory network via the thalamus, is utilized to assess the salience of both internal and external stimuli (Namkung et al., 2017; Nguyen et al., 2016). This information is subsequently processed in the AIC to generate a multi-dimensional subjective assessment of the relevance of stimuli and to monitor the ongoing task (Uddin, 2015).

The right dorsal AIC plays a pivotal role in guiding this process by acting as a ”switch” from the default mode network to the more resource-intensive central executive network. When a task demands increased attention, this switch activates key regions within the central executive network (Jilka et al., 2014; Seeley et al., 2007; Sridharan et al., 2008). However, in tasks that involve self-referential processes and mentalizing, such as moral decision-making that requires empathy, the AIC doesn't deactivate the default mode network but rather enhances its activity (Chen et al., 2013a).

Given the flexibility required for this system to monitor very different tasks, there is some degree of functional specialization within the network. The more ventral regions of the ACC and the AIC, as well as the amygdala, primarily process emotional salience, while the more dorsal regions are responsible for cognitive salience (Salehinejad et al., 2021; Touroutoglou and Dickerson, 2019). While both subnetworks work together, this ventral-dorsal dissociation can be appreciated in various aspects, including their functionality, connectivity, and even structural characteristics (Touroutoglou et al., 2012; Klugah-Brown et al., 2023; Menon et al., 2020).

The salience network is often overlooked in simple valuation paradigms, yet it comes into play in various social decisions that involve cooperation, altruism, empathy, or aversion to inequality (Fehr and Krajbich, 2014). For instance, the AIC responds to unfair offers in games like the ultimatum, dictator, and impunity games, as well as to defections in the prisoner's dilemma (Du and Chang, 2015; Rilling et al., 2008; Takagishi et al., 2009). The amygdala reacts to unfair offers in dictator and ultimatum games (Fehr and Krajbich, 2014), and it is associated with selfish decisions in altruistic giving (Cutler and Campbell-Meiklejohn, 2019). Additionally, the amygdala is active in strategic giving (Fermin et al., 2016). The ACC plays a key role in signaling unfairness to the central executive network (Chen et al., 2021). It integrates both emotional and cognitive information (Salehinejad et al., 2021; Touroutoglou and Dickerson, 2019) and is involved in representing the intentions of others, often in collaboration with the default mode network (Du and Chang, 2015).

While early studies linked the salience network with the representation of fairness, more recent studies suggest that the salience network does not inherently represent a specific concern (Fehr and Krajbich, 2014). This is because the salience network primarily signals what is salient or unusual. When fairness is the established social norm and is thus expected, an uncommonly unfair offer in the ultimatum game becomes a relevant stimulus that triggers an emotional response. The salience network detects this stimulus and may redirect attention while involving the central executive network.

This interpretation implies that the role of the salience network can vary from person to person, contingent on individual expectations and behaviors. This is indeed what Van den Bos et al. (2009) discovered: the anterior insula is more active when prosocial individuals defect in a trust game, whereas for pro-self individuals, it activates when they reciprocate. Individual heterogeneity can be attributed to inherent characteristics. For example, connectivity between the amygdala and the ACC has been associated with more prosocial decision-making (Dal Monte et al., 2020; Scheggia et al., 2022), and the gray matter volume of the amygdala is linked to pro-social tendencies (Fermin et al., 2016).

Overall, the salience network is critical in modulating the activity of the central executive network and the default mode network and assigning weights on the information that is eventually routed to the vmPFC, which finalizes the valuation process.

4 Towards a neuroeconomic model of social valuation

4.1 Conceptual framework

The brain's computation of subjective value is essential for guiding decision-making. Simple valuation hinges on a sophisticated reward network. Various brain regions synchronize their activity, facilitating the accumulation of information in the vmPFC. This process is rapid and largely automatic. However, in certain situations, decision-making demands the construction and integration of more intricate inputs, such as interoception (often described as "gut feeling"), self-control, risk assessment, or temporal cognition. In such cases, the salience network assumes a pivotal role. It integrates both external and internal stimuli, determining whether to engage the central executive network to handle complex representations. If the central executive network is indeed recruited, it subsequently conveys its output to the vmPFC. The vmPFC, in turn, combines this input with information received from the reward system, culminating in the creation of a unified subjective value for effective decision-making.

Social contexts are inherently complex to evaluate (see Figure 4). They necessitate a delicate interplay among the reward system, central executive network, and salience network. Additionally, they demand specific computations involving various processes, often overlapping, predominantly situated within the default mode network. Emotional processing involves the recruitment of the limbic network, notably the amygdala. Mirroring the emotions of others and aspects of aToM similarly engage the amygdala and regions within the medial prefrontal cortex (mPFC). In contrast, cToM encompasses regions like the pSTS, TPJ, and mPFC. The processing of signals related to moral conflict and others' intentions involves the TPJ and the mPFC, while harm aversion and norm violations are handled by the salience network. More advanced levels of strategizing and mentalizing about the beliefs and intentions of others also implicate the central executive network. Also, contemplating social norms requires specific calculations within the default mode network and the dlPFC. The dlPFC is responsible for both moral and strategic reasoning (Qu et al., 2022; Li et al., 2022).

All these processes are interconnected, contributing to a dynamic integration of information. Ultimately, this information converges in the vmPFC, where it culminates in the computation of a single decision value. Recent evidence points to the integration of these signals in the vmPFC, with a posterior-to-anterior gradient (Cutler and Campbell-Meiklejohn, 2019; Hiser and Koenigs, 2018). Specifically, the posterior vmPFC exhibits higher activity in response to automatic values (such as negative emotions, simple stimuli, and pure altruism), while the anterior portion is associated with more complex and abstract values (including complex emotions, money, abstract items, and strategic altruism). This gradient in vmPFC activation reflects the increasing complexity of the values being processed.

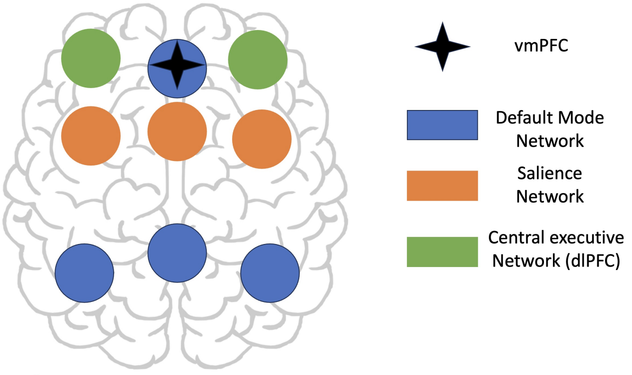

Figure 4: Simplified description of the 3 main actors involved in **Social Valuations**. The ventromedial prefrontal cortex (vmPFC) combines lower-order signals with higher-order signals processed by the default mode network and the central executive network, under the modulatory influence of the salience network.

This evidence shapes our approach to modeling decision-making in social contexts. It highlights the modularity of the brain, where distinct networks and regions handle various calculations separately. Moreover, it underscores the brain's preference for efficiency. Due to the cost associated with certain operations, the brain generally defaults to a simple process and activates a resource-intensive process only when necessary.

4.2 Modeling

The previously presented evidence offers insights into the way economic decision-making within a social context unfolds. To illustrate this, consider the following example. A decision-maker cares about rewards for themselves and, to a certain degree, rewards for others. We adopt a model similar to Brocas and Carrillo (2021) to represent the interplay between systems. Brocas and Carrillo (2021) focuses on self control and captures the interplay between the vmPFC and the dlPFC. In the model, the vmPFC computes the goal value based on low order information received by the reward system (e.g. taste) and high order information received by the dlPFC (e.g. health concerns). The dlPFC decides when to pass the high order information to the vmPFC, anticipating how this information will be integrated, and incurs a cost when it does. We call the decision to pass information *modulation*. In this model, we endow dlPFC with a central executive system function and put it in charge of deciding when to modulate.[1] The dlPFC cares only about optimizing the goal value and inform the vmPFC at the correct time.

[1] This is a modeling assumption. In practice, other systems such as the insula or the ACC may be involved in detecting when modulation may be beneficial and switch dlPFC on.

Here, we adopt a similar approach but incorporate specific modifications to accommodate the social context under consideration. Initially, an additional actor, the social network, conveys information to the vmPFC, signaling potential concerns for others. It is assumed that this information incurs no cost and is transmitted whenever the decision-maker is involved in a social scenario. This allows us to focus on the computations of the dlPFC. Additionally, we enhance the dlPFC with two advanced calculations—namely, equity concerns and strategic concerns—which can be relayed to the vmPFC based on the task requirements. The dlPFC has the flexibility to transmit either none, one or both types of information.

If the decision-maker is asked to distribute a sum of money, m, between themselves x and another individual $m - x$, these rewards are represented by an interplay between the reward system and the social brain network. The allocation's value can be expressed through a utility representation, denoted as $u(x) + \alpha v(m - x)$. Here, $u(x)$ and $v(m - x)$ are functions that capture the value associated with the respective rewards, and α is the weight attributed by the decision-maker to the other person, as evaluated by the social network. For the sake of simplifying the discussion, let's assume that both $u(x)$ and $v(m - x)$ are functions that increase and are concave in their arguments.

As previously mentioned, the dlPFC is responsible for modulating in two distinct ways. Firstly, for any chosen allocation, it calculates the cost associated with a deviation from a fair allocation, represented as $c(|x - \frac{m}{2}|)$, where $c()$ is, for simplification, an increasing and convex function in its argument. However, conveying this information to the vmPFC incurs a processing cost for the dlPFC, denoted as K_e. Secondly, the dlPFC predicts the potential behavior of the other individual. For example, if the partner has the authority to veto an allocation that provides the decision-maker with a higher reward, the dlPFC can relay information indicating that $x > \frac{m}{2}$ will yield a payoff of 0. Transmitting this information imposes a processing cost on the dlPFC, labeled as K_s.

Assuming there is no uncertainty on the value of the reward for self and for other, the interaction between the dlPFC and the vmPFC is captured by a simple stylized model. At stage 1, the dlPFC anticipates how the vmPFC will evaluate the options conditional on the information it receives and decides whether to supply information, considering the costs associated with this decision. At stage 2, the vmPFC calculates the value associated with each potential choice based on the information it receives.

At stage 2, four scenarios arise in terms of value representation and decision-making:

- The vmPFC only receives input from the reward network and the social network: choices are evaluated according to $u(x) + \alpha v(m - x)$ and comparisons over all pairs $(x, m - x)$ are performed to select the best option x^* that solves

$$u'(x^*) = \alpha v'(m - x^*)$$

The optimal solution is sensitive to α. By differentiating this equation with respect to

α, we can show that $\frac{\partial x^*}{\partial \alpha} < 0$ indicating that individuals who reflect higher concerns for others make more altruistic choices.

- The vmPFC also receives information from the dlPFC but about equity concerns only, the value of the allocation becomes $u(x) + \alpha v(m - x) - c(|x - \frac{m}{2}|)$ and the vmPFC selects the option $x^{**} < x^*$ (by construction) that solves

  $$u'(x^{**}) = \alpha v'(m - x^{**}) + c'(|x^{**} - m/2|)$$

  We also have $\frac{\partial x^{**}}{\partial \alpha} < 0$.

- The vmPFC also receives information from the dlPFC but about strategic concerns only, the value of the allocation becomes $u(x) + \alpha v(m - x)$ for all $x \leq \frac{m}{2}$ and 0 for $x > \frac{m}{2}$ and the optimal choice is $\min\{x^*, \frac{m}{2}\}$.

- The vmPFC also receives information from the dlPFC about both equity and strategic concerns, the value of the allocation becomes $u(x) + \alpha v(m - x) - c(|x - \frac{m}{2}|)$ for all $x \leq \frac{m}{2}$ and 0 for $x > \frac{m}{2}$ and the optimal choice is $\min\{x^{**}, \frac{m}{2}\}$.

At stage 1, the dlPFC decides what information to supply if any. In a dictator game, strategic concerns are irrelevant. Furthermore, by design, the payoff obtained when equity concerns are factored in is lower than the payoff obtained without such modulation. In this scenario, it is optimal to not modulate the decision of the vmPFC. As a consequence, the decision-maker offers x^*, and the value of the choice is expressed as $u(x^*) + \alpha v(m - x^*)$. Disparities among decision-makers stem from their inherent consideration for others, encapsulated by the parameter α. The calculations involve an interplay solely between the social network and the reward network.

In an ultimatum game, there is a risk of proposing an allocation that could be vetoed. This situation can be categorized into three cases, two of them are illustrated in Figure 5.

Case 1. If $x^{**} < x^* < \frac{m}{2}$, the decision-maker is altruistic enough to make an offer that satisfies the equity concern of the partner. There is no reason for the dlPFC to modulate the decision of the vmPFC. Eventually, x^* is chosen as in the dictator game.

Case 2. If $x^{**} < \frac{m}{2} < x^*$, not modulating the vmPFC results in 0 payoff. If the dlPFC sends information about equity concerns only, the vmPFC chooses x^{**}, which is accepted. If the dlPFC sends information about strategic concerns only, the vmPFC chooses $\frac{m}{2}$, which is also accepted, and yields a higher reward. If both concerns are passed, then the vmPFC chooses $\frac{m}{2}$ but the reward is now lowest. In that case, it is best to send only information about strategic concerns. This occurs if

$$u(\frac{m}{2}) + \alpha v(m - \frac{m}{2}) - K_s > 0$$

If modulation is too costly, the decision-maker offers x^* and the offer is rejected.

Case 3. Last, when $\frac{m}{2} < x^{**} < x^{*}$, not modulating the vmPFC again leads to 0 payoff. If the dlPFC sends information about equity concerns only, the vmPFC chooses x^{**}, which is not accepted. If the dlPFC sends information about strategic concerns only, the vmPFC chooses $\frac{m}{2}$, which is accepted. Sending both information also results in vmPFC choosing $\frac{m}{2}$. Overall, it is best to send only information about strategic concerns which occurs if

$$u(\frac{m}{2}) + \alpha v(m - \frac{m}{2}) - K_s > 0$$

Otherwise, the decision-maker offers x^{*} and the offer is rejected.

Hence, we observe that the decision-maker offers different amounts in the ultimatum game compared to the dictator game, reserving a higher sum for themselves in the dictator game (see Figure 5). This discrepancy doesn't arise from the decision-maker arbitrarily altering their preferences between games; rather, it stems from an optimization to represent preferences differently for efficiency reasons. Choices in the dictator game manifest as purely other-regarding concerns, while choices in the ultimatum game involve considerations of strategic concerns, and these considerations results in choosing the allocation that coincides with the social norm. However, this does not occur for equity concerns.

Various settings can trigger different calculations. For example, certain experimental manipulations, such as directing attention towards encouraging decision-makers to examine the consequences of their actions, have proven effective in activating dlPFC areas and altering behavior. In social contexts, like a dictator game, removing anonymity among players, scrutinizing gameplay, or encouraging dictators to consider the recipient has a similar impact. These manipulations eliminate the cost to represent equity concerns and compel players to take these concerns into account. In such cases, dictators optimally choose x^{**}. In an ultimatum game scenario, the dlPFC decides whether to factor in strategic concerns or not. When $x^{**} < \frac{m}{2}$, there is no incentive for the dlPFC to influence the vmPFC's decision, and x^{**} is chosen. In contrast, if $\frac{m}{2} < x^{**}$, not modulating the vmPFC results in a payoff of 0, and modulation occurs when K_s is not too high. These overall choices tend to be more advantageous for the recipients (see Figure 6). Still, individuals who are not sensitive to these manipulations should not change their behavior compared to the baseline example.

In summary, depending on their inherent levels of other-regarding concerns and the particular experimental context, decision-makers may display either purely strategic behavior or be guided by norms, either rooted in equity concerns or purely strategic considerations. Also, the engagement of the dlPFC is linked to the innate other-regarding concerns (captured by parameter α). For instance, altruistic individuals are expected to offer substantial amounts to their partners, irrespective of the game, with minimal involvement of the dlPFC. Conversely, selfish individuals might offer nothing in the dictator game but opt for a fair amount driven solely by strategic concerns in the ultimatum game. In that case, their choices should be accompanied by activation of the dlPFC.

Moreover, individuals with lower modulation costs would make optimal decisions at the appropriate moments. Unfair offers are expected to arise when modulation costs are high. This implies an additional source of variation: the costs of modulation are likely reflected in cognitive abilities or emotion regulation abilities. Consequently, we would expect to observe a higher incidence of unfair offers among individuals with lower performance in these dimensions. Alternatively, modulation may be more costly if the decision-maker is distracted or attending to several decisions at the same time. This also predicts that choices in these games can be manipulated by taxing cognitive functions.

Naturally, we can devise alternative games or scenarios where different moral considerations come into play, necessitating distinct inferences about partners. The model outlined here illustrates how social behavior can be effectively modeled within a unified framework based on brain function.

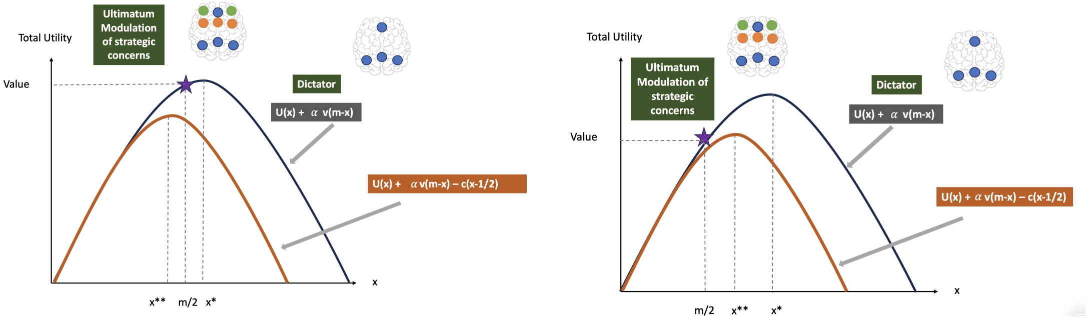

Figure 5: The dictator game requires the involvement of the social network and results in allocation x^* to self. In the ultimatum game, the left graph illustrates Case 2 and the right panel illustrates Case 3. Both result in the modulation of strategic concerns and the allocation of m/2 to self.

5 Concluding remarks

Understanding the brain's process of computing social values is essential for creating models that can accurately describe behavior within different contexts. This is crucial for explaining why individuals may exhibit self-centered behaviors in certain situations and generous actions in others. Additionally, the field of social behavior research stands to gain valuable insights by incorporating information about the evolution of brain structures that underpin social cognition and exploring the connection between variations in behavior and neural activity.

Indeed, the complexity of the valuation network likely reflects how evolution has shaped our brains and helped us adapt to environmental pressure. For instance, the dopamine

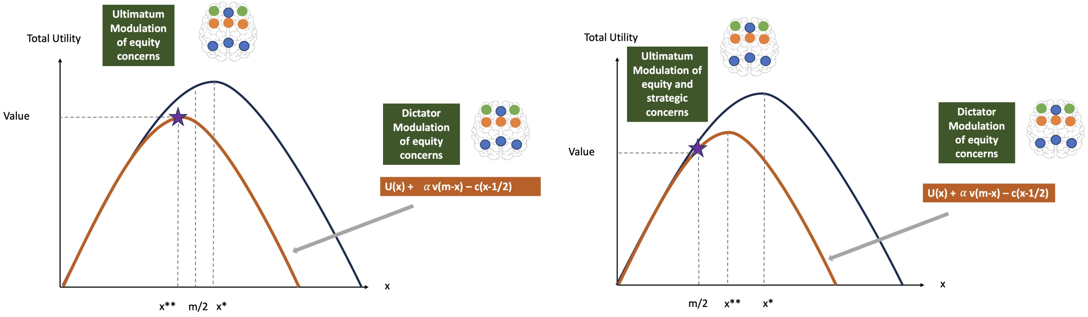

Figure 6: Under directed attention, the dictator game triggers the involvement of the social network and the representation of equity concerns and results in allocation x^{**} to self. In the ultimatum game, dlPFC may also represent strategic concerns (right) or not (left).

network is known to be an evolutionary old network shared by both vertebrates and invertebrates (Martínez-García and Lanuza, 2018; Barron et al., 2010; Waddell, 2013). Analogs of the default mode network exist in both primates and rodents (Lu et al., 2012). The PFC however represents a recent evolutionary specialization found only in anthropoid primates (Ma et al., 2022) while some of its functions are performed by other regions in other species such as birds (Diekamp et al., 2002; Güntürkün et al., 2020). In humans, the core decision-making network reflects a hierarchy of development. The earliest evolved regions specialize in processing rewards, fear, and salience. Subsequently, the default mode network emerged, enabling mentalization about the self and others. The most recent addition, the central executive network, supports intricate cognitive processing. Throughout the course of evolution, these networks have undergone reorganization, establishing connections between both new and ancient brain structures. A prime illustration of this reorganization is observed in the ancient salience network, which now regulates the activities of the more recent default mode and central executive networks.

Furthermore, while we have examined general trends in various valuation processes, it is crucial to recognize the substantial heterogeneity within living organisms. Variations in patterns of neural activation and brain structure are still not fully appreciated. Research into this heterogeneity often involves comparing populations that exhibit significantly different average behaviors, such as comparing neurotypical individuals with those who have behavioral disorders. These studies provide valuable insights into identifying core functions and understanding the mechanisms underlying these disorders.

As demonstrated, this information is valuable not only for reconciling diverse models of social preferences and behavior but also for integrating them into a unified framework. Individuals often exhibit varied objectives in different scenarios. The conventional approach in behavioral economics involves categorizing these situations and deducing utility functions from observed behaviors, yet it falls short in explaining how contextual factors influence outcomes.

An alternative perspective is to conceptualize decision-making as a form of information processing. This approach has already been exemplified in the context of discounting (Brocas and Carrillo, 2008), self control (Brocas and Carrillo, 2021), performance in cognitive tasks (Alonso et al., 2014) or memory (Brocas and Carrillo, 2016). The brain can be seen as a machine that encodes each component of a decision (a stimulus) and synthesizes this information into a single value, guiding a response. Distinct contexts elicit different representations due to the evolutionary mechanisms shaping the brain's functionality. This approach offers a deeper understanding of the impact of context on decision outcomes.

However, it is important to note that substantial heterogeneity exists even within these groups. Preferences, cognitive abilities, and social behaviors can vary significantly even among neurotypical individuals. Understanding the physiological underpinnings of such diversity is essential for predicting behavior accurately and constructing precise models of decision-making. This involves considering the diversity within each crucial component of social behavior, such as the capacity for mentalizing about others, the ability to strategize or collaborate. It requires detailing the interconnections between these factors and explaining how this information is combined to generate a spectrum of continuous variations in behavior.

Substantial evidence also underscores the tight link between behavioral outcomes and cognitive, affective, and personality traits. Moreover, an increasing body of research indicates a robust genetic foundation for these traits, influencing not only behavior but also variations in brain structure, volume, and connectivity. Concurrently, there is a growing recognition of the role of environmental factors in shaping behavior, emphasizing the intricate interplay between genes and the environment.

As our understanding of the biological mechanisms governing behavior deepens, it becomes evident that the fundamental drivers of decision-making are not mere preferences but rather the interplay of genes and experiences. These elements collaboratively shape our identity, preferences, and the neural processes engaged during decision-making. Modeling these brain mechanisms can offer a parsimonious explanation for the myriad sources of individual variability.

For instance, within the framework outlined in section 4, behavior is parameterized by α, representing an intrinsic attitude toward others. This intrinsic attitude might be a result of the intricate interplay between personality traits (which have a strong genetic component) and early life experiences, exemplifying the complex interweaving of genetic and environmental factors in shaping individual behavior. Isolating and understanding this parameter could prove instrumental in predicting behavior across diverse social environments. This is certainly a promising avenue for future research.

References

Alfonso, A., Brañas-Garza, P., Brocas, I., Carrillo, J. D., Gonzalez, J. B., and Vazquez, M. J. (2023). The development of rationality in games with hidden information. *mimeo.*

Alonso, R., Brocas, I., and Carrillo, J. D. (2014). Resource allocation in the brain. *Review of Economic Studies*, 81(2):501–534.

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., and Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. *Neuron*, 65(4).

Arslan, B., Taatgen, N. A., and Verbrugge, R. (2017). *Five-year-olds' systematic errors in second-order false belief tasks are due to first-order theory of mind strategy selection: A computational modeling study.* Frontiers in Psychology, 8(FEB).

Atique, B., Erb, M., Gharabaghi, A., Grodd, W., and Anders, S. (2011). Task-specific activity and connectivity within the mentalizing network during emotion and intention mentalizing. *Neuroimage*, 55(4):1899–1911.

Barron, A. B., Søvik, E., and Cornish, J. L. (2010). The roles of dopamine and related compounds in reward-seeking behavior across animal phyla. *Frontiers in behavioral neuroscience*, 4(163.).

Baumgartner, T., Knoch, D., Hotz, P., Eisenegger, C., and Fehr, E. (2011). Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. *Nature Neuroscience*, 14(11):11.

Ben-Ner, A., Kong, F., and Putterman, L. (2004). Share and share alike? Gender-pairing, personality, and cognitive ability as determinants of giving. *Journal of Economic Psychology*, 25(5):581–589.

Benjamin, D. J., Brown, S. A., and Shapiro, J. M. (2013). Who is 'behavioral'? Cognitive ability and anomalous preferences. *Journal of the European Economic Association*, 11(6):1231–1255.

Bhanji, J. P. and Delgado, M. R. (2014). The social brain and reward: social information processing in the human striatum. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(1).

Boccia, M., Dacquino, C., Piccardi, L., Cordellieri, P., Guariglia, C., Ferlazzo, F., and . . . Giannini, A. M. (2017). Neural foundation of human moral reasoning: an ALE meta-analysis about the role of personal perspective. *Brain Imaging and Behavior*, 11(1).

Brañas-Garza, P., García-Muñoz, T., and González, R. H. (2012). Cognitive effort in the beauty contest game. *Journal of Economic Behavior and Organization*, 83(2):254–260.

Brocas, I. and Carrillo, J. D. (2008). The brain as a hierarchical organization. *American Economic Review*, 98(4):1312–1346.

Brocas, I. and Carrillo, J. D. (2016). A neuroeconomic theory of memory retrieval. *Journal of Economic Behavior & Organization*, 130:198–205.

Brocas, I. and Carrillo, J. D. (2021). Value computation and modulation: a neuroeconomic theory of self-control as constrained optimization. *Journal of Economic Theory*, 198:105366.

Brocas, I., Carrillo, J. D., Wang, S. W., and Camerer, C. F. (2014). Imperfect choice or imperfect attention? Understanding strategic thinking in private information games. *Review of Economic Studies*, 81(3):944–970.

Brooks, S. J., O Daly, O. G., Uher, R., Friederich, H.-C., Giampietro, V., Brammer, M., Williams, S. C., Schiöth, H. B., Treasure, J., and Campbell, I. C. (2011). Differential neural responses to food images in women with bulimia versus anorexia nervosa. *PLoS One*, 6(7):e22259.

Buckner, R. L. and DiNicola, L. M. (2019). The brain's default network: updated anatomy, physiology and evolving insights. *Nature Reviews Neuroscience 2019*, 20(10):10.

Bukowski, H. (2018). The neural correlates of visual perspective taking: a critical review. *Current Behavioral Neuroscience Reports*, 5:189–197.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton University Press, Princeton.

Camus, M., Halelamien, N., Plassmann, H., Shimojo, S., O'Doherty, J., Camerer, C., and Rangel, A. (2009). Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex decreases valuations during food choices. *European Journal of Neuroscience*, 30(10):1980–1988.

Carter, R. M. K., Bowling, D. L., Reeck, C., and Huettel, S. A. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science*, 336(6090).

Carter, R. M. K. and Huettel, S. A. (2013). A nexus model of the temporal–parietal junction. *Trends in Cognitive Sciences*, 17(7).

Chen, A. C., Oathes, D. J., Chang, C., Bradley, T., Zhou, Z.-W., Williams, L. M., Glover, G. H., Deisseroth, K., and Etkin, A. (2013a). Causal interactions between fronto-parietal central executive and default-mode networks in humans. *Proceedings of the National Academy of Sciences*, 110(49):19944–19949.

Chen, C.-C., Chiu, I.-M., Smith, J., and Yamada, T. (2013b). Too smart to be selfish? measures of cognitive ability, social preferences, and consistency. *Journal of Economic Behavior & Organization*, 90:112–122.

Chen, F., He, Q., Han, Y., Zhang, Y., and Gao, X. (2018). Increased bold signals in dlpfc is associated with stronger self-control in food-related decision-making. *Frontiers in psychiatry*, 9:689.

Chen, M., Zhu, X., Zhang, J., Ma, G., and Wu, Y. (2021). Neural correlates of proposers' fairness perception in punishment and non-punishment economic games. *Current Psychology*, 40:1838–1849.

Chiong, W., Wilson, S. M., D'Esposito, M., Kayser, A. S., Grossman, S. N., Poorzand, P., and . . . Rankin, K. P. (2013). The salience network causally influences default mode network activity during moral reasoning. *Brain*, 136(6).

Cochard, F., Le Gallo, J., Georgantzis, N., and Tisserand, J.-C. (2021). Social preferences across different populations: Meta-analyses on the ultimatum game and dictator game. *Journal of Behavioral and Experimental Economics*, 90:101613.

Constantinidis, C. and Klingberg, T. (2016). The neuroscience of working memory capacity and training. *Nature Reviews Neuroscience*, 17(7).

Crockett, M. J. and Lockwood, P. L. (2018). Extraordinary altruism and transcending the self. *Trends in Cognitive Sciences*, 22(12).

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., and Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, 20(6):6.

Cutler, J. and Campbell-Meiklejohn, D. (2019). A comparative fMRI meta-analysis of altruistic and strategic decisions to give. *NeuroImage*, 184.

Dal Monte, O., Chu, C. C., Fagan, N. A., and Chang, S. W. (2020). Specialized medial prefrontal–amygdala coordination in other-regarding decision preference. *Nature neuroscience*, 23(4):565–574.

Dalton, B., Foerde, K., Bartholdy, S., McClelland, J., Kekic, M., Grycuk, L., Campbell, I. C., Schmidt, U., and Steinglass, J. E. (2020). The effect of repetitive transcranial magnetic stimulation on food choice-related self-control in patients with severe, enduring anorexia nervosa. *International Journal of Eating Disorders*, 53(8):1326–1336.

de Wit, S., Corlett, P. R., Aitken, M. R., Dickinson, A., and Fletcher, P. C. (2009). Differential engagement of the ventromedial prefrontal cortex by goal-directed and habitual behavior toward food pictures in humans. *Journal of Neuroscience*, 29(36):11330–11338.

Diekamp, B., Kalt, T., and Güntürkün, O. (2002). Working memory neurons in pigeons. *The Journal of Neuroscience*, 22(4):RC210.

Du, E. and Chang, S. W. C. (2015). Neural components of altruistic punishment. *Frontiers in Neuroscience, 9(FEB)*, 26.

Embrey, M., Fréchette, G. R., and Yuksel, S. (2018). Cooperation in the finitely repeated prisoner's dilemma. *The Quarterly Journal of Economics*, 133(1):509–551.

Emonds, G., Declerck, C. H., Boone, C., Vandervliet, E. J., and Parizel, P. M. (2011). Comparing the neural basis of decision making in social dilemmas of people with different social value orientations, a fmri study. *Journal of Neuroscience, Psychology, and Economics*, 4(1):11.

Emonds, G., Declerck, C. H., Boone, C., Vandervliet, E. J., and Parizel, P. M. (2012). The cognitive demands on cooperation in social dilemmas: an fmri study. *Social Neuroscience*, 7(5):494–509.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, 14(4).

Fe, E., Gill, D., and Prowse, V. (2022). Cognitive skills, strategic sophistication, and life outcomes. *Journal of Political Economy*, 130(10):2643–2704.

Fehr, E. and Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences*, 11(10).

Fehr, E. and Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and human behavior*, 25(2):63–87.

Fehr, E. and Krajbich, I. (2014). Social preferences and the brain. In *Neuroeconomics*, pages 193–218. Elsevier.

Fermin, A. S. R., Sakagami, M., Kiyonari, T., Li, Y., Matsumoto, Y., and Yamagishi, T. (2016). Representation of economic preferences in the structure and function of the amygdala and prefrontal cortex. *Scientific Reports*, 6(1):1.

Fiorillo, C. D., Tobler, P. N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614):1898–1902.

Foerde, K., Steinglass, J. E., Shohamy, D., and Walsh, B. T. (2015). Neural mechanisms supporting maladaptive food choices in anorexia nervosa. *Nature neuroscience*, 18(11):1571–1573.

Fregni, F., Orsati, F., Pedrosa, W., Fecteau, S., Tome, F. A., Nitsche, M. A., Mecca, T., Macedo, E. C., Pascual-Leone, A., and Boggio, P. S. (2008). Transcranial direct current stimulation of the prefrontal cortex modulates the desire for specific foods. *Appetite*, 51(1):34–41.

Gao, X., Yu, H., Sáez, I., Blue, P. R., Zhu, L., Hsu, M., and Zhou, X. (2018). Distinguishing neural correlates of context-dependent advantageous- and disadvantageous-inequity aversion. *Proceedings of the National Academy of Sciences of the United States of America*, 115(33).

Gill, D. and Prowse, V. (2016). Cognitive ability, character skills, and learning to play equilibrium: A level-k analysis. *Journal of Political Economy*, 124(6):1619–1676.

Gluth, S., Rieskamp, J., and Büchel, C. (2012). Deciding when to decide: time-variant sequential sampling models explain the emergence of value-based decisions in the human brain. *Journal of Neuroscience*, 32(31):10686–10698.

Gradin, V., Pérez, A., Macfarlane, J., Cavin, I., Waiter, G., Tone, E., Dritschel, B., Maiche, A., and Steele, J. (2016). Neural correlates of social exchanges during the prisoner's dilemma game in depression. *Psychological medicine*, 46(6):1289–1300.

Gunnthorsdottir, A. . H., A., D. . M., and K. (2007). Dispositions, history and contributions in public goods experiments. *Journal of Economic Behavior and Organization. 62*, 62.

Güntürkün, O., Stacho, M., and Ströckens, F. (2020). The brains of reptiles and birds. *Evolutionary neuroscience*, pages 159–212.

Hall, P. A., Lowe, C., and Vincent, C. (2017). Brain stimulation effects on food cravings and consumption: an update on lowe et al.(2017) and a response to generoso et al.(2017). *Psychosomatic Medicine*, 79(7):839–842.

Hare, T. A., Camerer, C. F., and Rangel, A. (2009). Self-control in decision-Making involves modulation of the vmPFC valuation system. *Science*, 324(5927).

Hare, T. A., Malmaud, J., and Rangel, A. (2011a). *Focusing attention on the health aspects of foods changes value signals in vmpfc and improves dietary choice.* Journal of Neuroscience, 31(30):11077–11087.

Hare, T. A., Schultz, W., Camerer, C. F., O'Doherty, J. P., and Rangel, A. (2011b). Transformation of stimulus value signals into motor commands during simple choice. *Proceedings of the National Academy of Sciences*, 108(44):18120–18125.

Harris, A., Hare, T., and Rangel, A. (2013). Temporally dissociable mechanisms of self-control: early attentional filtering versus late value modulation. *Journal of Neuroscience*, 33(48):18917–18931.

He, Q., Huang, X., Zhang, S., Turel, O., Ma, L., and Bechara, A. (2019). Dynamic causal modeling of insular, striatal, and prefrontal cortex activities during a food-specific go/nogo task. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 4(12):1080–1089.

Healey, M. L. and Grossman, M. (2018). Cognitive and affective perspective-taking: evidence for shared and dissociable anatomical substrates. *Frontiers in neurology*, 9:491.

Hiser, J. and Koenigs, M. (2018). The multifaceted role of the ventromedial prefrontal cortex in emotion, decision making, social cognition, and psychopathology. *Biological Psychiatry*, 83(8).

Hutcherson, C. A., Bushong, B., and Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, 87(2).

Hutcherson, C. A., Plassmann, H., Gross, J. J., and Rangel, A. (2012). Cognitive regulation during decision making shifts behavioral control between ventromedial and dorsolateral prefrontal value systems. *Journal of Neuroscience*, 32(39):13543–13554.

Isik, L., Koldewyn, K., Beeler, D., and Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences of the United States of America*, 114(43).

Jilka, S. R., Scott, G., Ham, T., Pickering, A., Bonnelle, V., Braga, R. M., and ... Sharp, D. J. (2014). Damage to the salience network and interactions with the default mode network. *Journal of Neuroscience*, 34(33).

Johnson, N. D. and Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5).

Kahnt, T., Heinzle, J., Park, S. Q., and Haynes, J.-D. (2011). Decoding different roles for vmpfc and dlpfc in multi-attribute decision making. *Neuroimage*, 56(2):709–715.

Klugah-Brown, B., Wang, P., Jiang, Y., Becker, B., Hu, P., Uddin, L. Q., and Biswal, B. (2023). Structural–functional connectivity mapping of the insular cortex: a combined data-driven and meta-analytic topic mapping. *Cerebral Cortex*, 33(5):1726–1738.

Koechlin, E. and Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6).

Kolling, N., Braunsdorf, M., Vijayakumar, S., Bekkering, H., Toni, I., and Mars, R. B. (2021). Constructing others' beliefs from one's own using medial frontal cortex. *Journal of Neuroscience*, 41(46):9571–9580.

Koob, G. F. and Volkow, N. D. (2010). Neurocircuitry of addiction. *Neuropsychopharmacology*, 35(1):217–238.

Ledyard, J. O. (1995). Is there a problem with public goods provision. *The handbook of experimental economics*, pages 111–194.

Levitt, S. D. and List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic perspectives*, 21(2):153–174.

Levy, D. J. and Glimcher, P. W. (2011). *Comparing apples and oranges: using reward-specific and reward-general subjective value representation in the brain.* Journal of Neuroscience, 31(41):14693–14707.

Levy, D. J. and Glimcher, P. W. (2012). The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology*, 22(6).

Li, Y., Hu, J., Ruff, C. C., and Zhou, X. (2022). Neurocomputational evidence that conflicting prosocial motives guide distributive justice. *Proceedings of the National Academy of Sciences*, 119(49):e2209078119.

Lowe, C. J., Reichelt, A. C., and Hall, P. A. (2019). The prefrontal cortex and obesity: a health neuroscience perspective. *Trends in cognitive sciences*, 23(4):349–361.

Lu, H., Zou, Q., Gu, H., Raichle, M. E., Stein, E. A., and Yang, Y. (2012). Rat brains also have a default mode network. *Proceedings of the National Academy of Sciences*, 109(10):3979–3984.

Ma, S., Skarica, M., Li, Q., Xu, C., Risgaard, R. D., Tebbenkamp, A. T., Mato-Blanco, X., Kovner, R., Krsnik, Ž., de Martin, X., et al. (2022). Molecular and cellular evolution of the primate dorsolateral prefrontal cortex. *Science*, 377(6614):eabo7257.

Manstead, A. S. (1995). Children's understanding of emotion. *Everyday Conceptions of Emotion: An Introduction to the Psychology, Anthropology and Linguistics of Emotion*, pages 315–331.

Martínez-García, F. and Lanuza, E. (2018). Evolution of vertebrate survival circuits. *Current Opinion in Behavioral Sciences*, 24:113–123.

McAuliffe, K., Blake, P. R., Steinbeis, N., and Warneken, F. (2017). *The developmental foundations of human fairness.* Nature Human Behaviour.

McFadden, B. R., Lusk, J. L., Crespi, J. M., Cherry, J. B. C., Martin, L. E., Aupperle, R. L., and Bruce, A. S. (2015). Can neural activation in dorsolateral prefrontal cortex predict responsiveness to information? an application to egg production systems and campaign advertising. *PloS one*, 10(5):e0125243.

Meinhardt, J., Sodian, B., Thoermer, C., Döhnel, K., and Sommer, M. (2011). True- and false-belief reasoning in children and adults: An event-related potential study of theory of mind. *Developmental Cognitive Neuroscience*, 1(1).

Menon, V. (2011). Large-scale brain networks and psychopathology: a unifying triple network model. *Trends in Cognitive Sciences*, 15(10).

Menon, V., Gallardo, G., Pinsk, M. A., Nguyen, V.-D., Li, J.-R., Cai, W., and Wassermann, D. (2020). Microstructural organization of human insula is linked to its macrofunctional circuitry and predicts cognitive control. *elife*, 9:e53470.

Menon, V. and Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain structure and function*, 214:655–667.

Moor, B. G., de Macks, Z. A. O., Güroğlu, B., Rombouts, S. A., Van der Molen, M. W., and Crone, E. A. (2012). Neurodevelopmental changes of reading the mind in the eyes. *Social cognitive and affective neuroscience*, 7(1):44.

Nagel, R., Brovelli, A., Heinemann, F., and Coricelli, G. (2018). Neural mechanisms mediating degrees of strategic uncertainty. *Social cognitive and affective neuroscience*, 13(1):52–62.

Nakamura-Palacios, E. M., Lopes, I. B. C., Souza, R. A., Klauss, J., Batista, E. K., Conti, C. L., Moscon, J. A., and de Souza, R. S. M. (2016). Ventral medial prefrontal cortex (vmpfc) as a target of the dorsolateral prefrontal modulation by transcranial direct current stimulation (tdcs) in drug addiction. *Journal of Neural Transmission*, 123:1179–1194.

Namkung, H., Kim, S. H., and Sawa, A. (2017). The insula: An underestimated brain area in clinical neuroscience, psychiatry, and neurology. *Trends in Neurosciences*, 40(4).

Nguyen, V. T., Breakspear, M., Hu, X., and Guo, C. C. (2016). The integration of the internal and external milieu in the insula during dynamic emotional experiences. *NeuroImage*, 124.

Obeso, I., Moisa, M., Ruff, C. C., and Dreher, J. C. (2018). A causal role for right temporo-parietal junction in signaling moral conflict. *ELife*, 7.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., and Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106).

Proto, E., Rustichini, A., and Sofianos, A. (2019). Intelligence, personality, and gains from cooperation in repeated interactions. *Journal of Political Economy*, 127(3):1351–1390.

Proto, E., Rustichini, A., and Sofianos, A. (2022). Intelligence, errors, and cooperation in repeated interactions. *The Review of Economic Studies*, 89(5):2723–2767.

Qu, C., Bénistant, J., and Dreher, J.-C. (2022). Neurocomputational mechanisms engaged in moral choices and moral learning. *Neuroscience & Biobehavioral Reviews*, 132:50–60.

Rilling, J. K., King-Casas, B., and Sanfey, A. G. (2008). The neurobiology of social decision-making. *Current Opinion in Neurobiology*, 18(2).

Rudorf, S. and Hare, T. A. (2014). Interactions between Dorsolateral and Ventromedial Prefrontal Cortex Underlie Context-Dependent Stimulus Valuation in Goal-Directed Choice. *Journal of Neuroscience*, 34(48).

Ruff, C. C. and Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, 15(8):8.

Ruff, C. C., Ugazio, G., and Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science*, 342(6157).

Salehinejad, M. A., Ghanavati, E., Rashid, M. H. A., and Nitsche, M. A. (2021). Hot and cold executive functions in the brain: A prefrontal-cingular network. *Brain and Neuroscience Advances*, 5:23982128211007769.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, 300(5626).

Scheggia, D., La Greca, F., Maltese, F., Chiacchierini, G., Italia, M., Molent, C., Bernardi, F., Coccia, G., Carrano, N., Zianni, E., et al. (2022). Reciprocal cortico-amygdala connections regulate prosocial and selfish choices in mice. *Nature Neuroscience*, 25(11):1505–1518.

Schultz, W. (2007). Behavioral dopamine signals. *Trends in Neurosciences*, 30(5).

Schultz, W. (2016). Dopamine reward prediction-error signalling: a two-component response. *Nature Reviews Neuroscience*, 17(3):3.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., and Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, 42.

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., and . . . Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, 27(9).

Smallwood, J., Bernhardt, B. C., Leech, R., Bzdok, D., Jefferies, E., and Margulies, D. S. (2021). The default mode network in cognition: a topographical perspective. *Nature reviews neuroscience*, 22(8):503–513.

Sokol-Hessner, P., Hutcherson, C., Hare, T., and Rangel, A. (2012). Decision value computation in dlpfc and vmpfc adjusts to the available decision time. *European Journal of Neuroscience*, 35(7):1065–1074.

Sridharan, D., Levitin, D. J., and Menon, V. (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proceedings of the National Academy of Sciences of the United States of America*, 105(34).

Steinbeis, N., Bernhardt, B. C., and Singer, T. (2012). Impulse control and underlying functions of the left DLPFC mediate age-related and age-independent individual differences in strategic social behavior. *Neuron*, 73(5).

Sun, P., Zheng, L., Zhang, Q., Cheng, X., Li, L., and Ling, X. (2023). Neural basis of cooperation in the prisoner's dilemma in a loss context. *Social Behavior and Personality: an international journal*, 51(4):1–13.

Takagishi, H., Takahashi, T., Toyomura, A., Takashino, N., Koizumi, M., and Yamagishi, T. (2009). Neural correlates of the rejection of unfair offers in the impunity game. *Neuroendocrinology Letters*, 30(4).

Tanaka, H., Shou, Q., Kiyonari, T., Matsuda, T., Sakagami, M., and Takagishi, H. (2023). Right dorsolateral prefrontal cortex regulates default prosociality preference. *Cerebral Cortex*, 33(9):5420–5425.

Thompson, K., Nahmias, E., Fani, N., Kvaran, T., Turner, J., and Tone, E. (2021). The prisoner's dilemma paradigm provides a neurobiological framework for the social decision cascade. *Plos one*, 16(3):e0248006.

Touroutoglou, A. and Dickerson, B. C. (2019). Cingulate-centered large-scale networks: Normal functions, aging, and neurodegenerative disease. *Handbook of Clinical Neurology*, 166.

Touroutoglou, A., Hollenbeck, M., Dickerson, B. C., and Barrett, L. F. (2012). Dissociable large-scale networks anchored in the right anterior insula subserve affective experience and attention. *Neuroimage*, 60(4):1947–1958.

Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. *Nature Reviews Neuroscience*, 16(1).

Valle, A., Massaro, D., Castelli, I., and Marchetti, A. (2015). Theory of mind development in adolescence and early adulthood: The growing complexity of recursive thinking ability. *Europe's Journal of Psychology*, 11(1).

van den Bos, W., Talwar, A., and McClure, S. M. (2013). Neural correlates of reinforcement learning and social preferences in competitive bidding. *Journal of Neuroscience*, 33(5):2137–2146.

Van den Bos, W., Van Dijk, E., Westenberg, M., Rombouts, S. A., and Crone, E. A. (2009). What motivates repayment? neural correlates of reciprocity in the trust game. *Social cognitive and affective neuroscience*, 4(3):294–304.

van der Meulen, A., Roerig, S., de Ruyter, D., van Lier, P., and Krabbendam, L. (2017). A comparison of children's ability to read children's and adults' mental states in an adaptation of the reading the mind in the eyes task. *Frontiers in psychology*, 8:594.

Van Veluw, S. J. and Chance, S. A. (2014). Differentiating between self and others: an ale meta-analysis of fmri studies of self-recognition and theory of mind. *Brain imaging and behavior*, 8:24–38.

Varrier, R. S. and Finn, E. S. (2022). Seeing social: A neural signature for conscious perception of social interactions. *Journal of Neuroscience*, 42(49):9211–9226.

Vogindroukas, I., Chelas, E.-N., and Petridis, N. E. (2014). Reading the mind in the eyes test (children's version): a comparison study between children with typical development, children with high-functioning autism and typically developed adults. *Folia Phoniatrica et Logopaedica*, 66(1-2):18–24.

Waddell, S. (2013). Reinforcement signalling in drosophila; dopamine does it all after all. *Current opinion in neurobiology*, 23(3):324–329.

Yeshurun, Y., Nguyen, M., and Hasson, U. (2021). The default mode network: where the idiosyncratic self meets the shared social world. *Nature Reviews Neuroscience*, 22(3):181–192.

Yoshida, W., Seymour, B., Friston, K. J., and Dolan, R. J. (2010). Neural mechanisms of belief inference during cooperative games. *Journal of Neuroscience*, 30(32):10744–10751.

Young, L., Dodell-Feder, D., and Saxe, R. (2010). What gets the attention of the temporo-parietal junction? an fmri investigation of attention and theory of mind. *Neuropsychologia*, 48(9):2658–2664.

Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit memories to other brains: constructing shared neural representations via communication. *Cerebral cortex*, 27(10):4988–5000.

Zinchenko, O. and Arsalidou, M. (2018). Brain responses to social norms: Meta-analyses of f mri studies. *Human brain mapping*, 39(2):955–970.

Zyuzin, J., Combs, D., Monterosso, J. R., and Brocas, I. (2023). The neural correlates of value representation: From single items to bundles. *Human Brain Mapping*, 44(4):1476–1495.