



Universiteit
Leiden

The Netherlands

Preferences and beliefs in behavior and the brain

Farina, A.

Citation

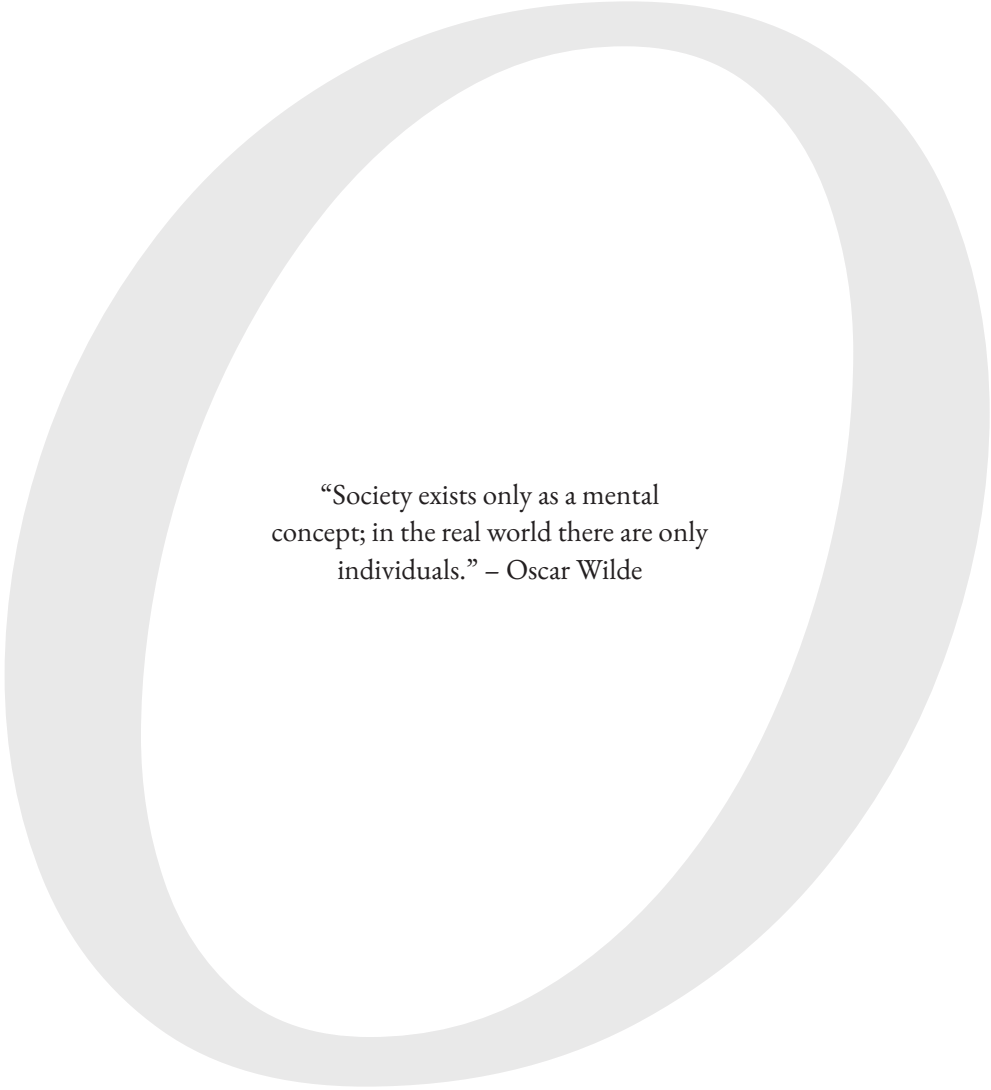
Farina, A. (2024, January 10). *Preferences and beliefs in behavior and the brain*. Retrieved from <https://hdl.handle.net/1887/3677340>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3677340>

Note: To cite this publication please use the final published version (if applicable).



“Society exists only as a mental
concept; in the real world there are only
individuals.” – Oscar Wilde

Chapter One _____

Introduction

me

As inherently social animals, humans are continuously engaged and cooperate with others. Consequently, our social preferences regularly influence the decisions we make throughout our day. As prominent as they are, it is no surprise that research from diverse fields such as evolutionary biology, psychology and economics have developed different theories about the origins of social preferences. Although cooperation has been a subject of research for over a century, relatively little is known about the basis of social preferences and whether these can vary across the lifespan or the environment. In the current thesis, I address this gap by bringing into question the stability of social preference and investigating the neural underpinnings of parochial cooperation.

Social Preferences

Charles Darwin is perhaps one of the first prominent scientists to contemplate the idea of social preferences, in describing what would later be known as kin selection in his 1859 book *On the Origin of Species*. Later formalized as Hamilton's rule (1964), kin selection extends Darwin's theory of natural selection to genes. The principal rationale is that agents will help related others, even if the behavior incurs a fitness cost to themselves, so long as it increases the probability of their shared gene to survive and reproduce. Formally, genes increase in frequency when:

$$rB > C$$

where r is the genetic relatedness of the helper and the recipient, B is the added benefit of the behavior gained by the recipient, and C is the cost to the helper. The closer the genetic relationship between the two actors, the more willing an individual should be to help another. Thus, this stylized rule offers an explanation as to why social preferences exist: humans care about more than just their own outcomes, so long as the outcomes of genetically related others are also affected. Given that our genes do not change after we are born, kin selection posits that these social preferences are stable throughout our lifetime.

Yet, much of our lives are spent alongside unrelated strangers. How can we make sense of social preferences that extend outside of the family nucleus? A more recent theoretical framework suggests that the degree of interdependence (Balliet et al., 2017; Gross & De Dreu, 2019b), or the likelihood to interact again in the future, makes our preferences social. This idea rests on the principle of reciprocity – that is, people will cooperate with someone today because it is an investment towards (expected)

reciprocal altruistic behavior from this same person in the future. For our social preferences to be influenced by reciprocal concerns, it need not be limited to repeated interactions with the same person (i.e., direct reciprocity). Most often in fact, social decisions are based on indirect reciprocity. When one person helps another, they help the recipient but also may gain a reputation for being cooperative. In the future, this person seen as cooperative may be more likely to receive help by another (Roberts et al., 2021). Accordingly, one of the most robust findings in social psychology is that of ingroup favoritism, also known as parochialism. This behavior was first documented in the 1970s by Henri Tajfel, noting that individuals cooperate more with ingroup compared to outgroup members (Tajfel, 1970), even when the groups are demarcated along inherently meaningless lines. One explanation for parochialism is that actors are more concerned with their in-group reputation and directly or indirectly reciprocate kind acts with members of their in-group, but do not care or care much less about what out-group members think of them.

Economic Games in Social Science

We all have preferences that guide our decisions, whether we are aware of them or not. These preferences are shaped by our environment, our experience and our biology. Ask someone to disclose their preferences, and you will surely be met with descriptions of an idealized version of themselves, a blank stare, or some version of “it depends”. In contrast, put someone in the actual position to choose, and their preferences are revealed.

The notion of revealed preferences is precisely what underlies the utility of standardized economic games. By observing choices made under given constraints, one can directly reveal the variable of interest – a preference – without the need for broad assumptions or reconstructions based on self-reports or introspection. Unlike approximating a bias such as prejudice from differences in response times to stimuli (as in the Implicit Association Test, Greenwald et al., 1998) or generosity from self-reports on questionnaire items (e.g. the Prosocial Tendencies Measure, Carlo & Randall, 2002), choices in economic games divulge the specific preference one aims to explain.

Economic games performed in the laboratory strip away the superfluous context of real-life decision-making, in order to avoid implicit norms and role expectations. For instance, experimental instructions avoid alluding to real world analogues such as overfishing or public education, give decision options neutral labels (i.e. “option A or option B”), and outcomes are set in concrete monetary amounts. This aims to free

participants from pre-conceived notions about what their choice *should* be, and instead focus on what they value to guide their behavior. As a result, responses are less prone to social desirability bias, the desire to *appear* rather than *actually be* prosocial. Though this method abstracts away from perceived realism, it nonetheless creates scenarios of true consequence by providing monetary incentives. Behaving in a way that is not aligned with one's true preference becomes subjectively costly and suboptimal. Thus, economic games allow for tractable modeling of the basic mechanisms underlying human cooperation and competition.

The use of economic games to study human behavior has its origins in Game Theory, first introduced in 1944 by von Neumann and Morgenstern. In their *Theory of Games and Economic Behavior*, they provided the first formal analysis of strategic interaction. They set up the three necessary elements of a game: (i) the interacting participants (players), (ii) their sets of available actions (sets of strategies), and (iii) their preferences for all possible combinations of these strategies and the resulting outcomes (payoffs or utilities). Subsequent research in experimental economics showed that theorizing, albeit mathematically rigorous, was not enough to arrive at useful predictions of human behavior – these predictions needed to be tested against actual human decisions. Over the past few decades, economic games have become a mainstay of many research disciplines trying to understand human behavior - from economics, to psychology, neuroscience, evolutionary biology, and political science, among others.

While there are several subcategories of economic games, I focus here on games that can be used to model social preferences. Social preferences - also referred to as other-regarding preferences in the broader economics literature - are preferences which arise in a social context, as opposed to risk or temporal preferences that can guide decisions in non-social contexts. The distinct characteristic of social preferences is that agents make decisions that simultaneously affect their own outcome, as well as others' welfare. Past theoretical and empirical research has shown that humans not only place value on their own outcomes, but they also place value on the outcome of others, and the difference between the two (Fehr & Schmidt, 1999). Social preferences differ from person to person and depend on the context. They run on a continuous scale from least to most prosocial: from preferring to harm others even at your own expense (i.e., sadomasochistic), to giving greater value to others' outcomes over your own (i.e., altruistic). Chapter 2 of this thesis provides an illustration of the range of social preferences.

Games that are used to study social preferences inherently model situations where individuals are interdependent. These are mixed-motives situations, or social dilemmas, so called because the players face a choice with competing motives: to choose what is best for the individual, or what is best for the collective. There are four

main standard economic games used to study the sources of cooperation, coordination, prosocial behavior and free-riding: the dictator game, ultimatum bargaining, the trust game, and the public goods game (see Box 1 for detailed descriptions).

Experimental modifications within each of these standard games, such as starting endowments, including punishment options, possibilities for partner switching, or revealing certain aspects of the partner, allow for more nuanced understandings of what pushes people to sustain cooperation or free-ride, or when social norms are created and enforced. In addition to these four archetypal games, another experimental set-up called the Intergroup Attacker-Defender Game (IADC; De Dreu & Gross, 2019; De Dreu et al., 2020) allows us to look at cooperation not only as contribution to the public good, but also as coordination in a conflict situation. In this context, two groups compete for a limited resource, but must coordinate within each group to win the contest.

As explored in chapter 3 of this thesis, the IADC models a non-symmetric conflict situation. Although both groups begin with the same endowment, only one group (the attackers) can seek to gain the resources of the other (the defenders). Those in the attacker group can invest in attack, while those in the defender group invest in defense. If the attack investment is larger than that of defense, the attackers gain all the money the defenders did not spend on defense (i.e., they gain the spoils of war). Otherwise, both teams simply lose everything they have not invested in attack or defense (i.e., the defense successfully defend their assets). Decisions made under this framework can tell us more about how cooperation can also lead to detrimental outcomes, be it in terms of wasted resources or unnecessary harm to others. Studying these scenarios helps to gain a broader understanding of intergroup dynamics, as well as within group pressures to cooperate or defect.

In addition to using economic games to study social preferences, they can also be a useful tool for probing beliefs. By asking people how they think others will act in the same scenario, or in response to their actions, we can gain a clearer picture of their expectations and ideas regarding social norms. In chapters 3 and 4 of this thesis, we measure beliefs in the IADC and the trust game to explain how beliefs influence behavior.

Box 1 – Social economic games

The dictator game (Forsythe et al., 1994) is the simplest game to study prosocial preferences. It consists of 2 anonymized players: player 1 (the dictator) is endowed with a certain amount of money, say for example 10 monetary units (MU), and player 2 (the receiver) who does not receive any endowment from the experimenter. The dictator simply decides if they wish to donate any money, if at all, to their partner. The receiver is passive in this game and can only accept what the first player decides to do, at which point the game is complete. While classic economic theory would predict that the dictator will maximize payouts for themselves and thus never give any money to their partner, experimental studies have consistently shown this is not the case, and a large part of participants do in fact donate something nontrivial (Hoffman et al., 2008). Because there is no strategic advantage to donating since player 2 cannot reciprocate, or punish player 1's behavior, how much the dictator donates is used as a measure of prosociality. Variations of this game, such as the slider measure used in Chapter 2 of this thesis, allow allocators to choose between a fixed set of possible monetary splits between themselves and another in order to reveal a particular social value orientation, such as a preference for profit-maximization, minimizing differences in payoffs or collective efficiency.

The ultimatum bargaining game (Güth et al., 1982) builds on the dictator game, by making player 2 an active participant. Here, player 1 is again given a starting endowment. They make an offer to their partner, and only if this offer is accepted, the money is split as agreed. If, however, the responder does not accept the donor's offer, both parties leave with nothing. In this scenario, player 1 is incentivized to make an offer large enough so that it is accepted by player 2, but small enough for the split to still be attractive. This game has been used extensively to study concepts of fairness, coordination and social norms (Murphy et al., 2011).

Similarly, the trust game involves 2 players. First introduced in 1995 as the Investment Game (Berg et al., 1995), the first player (trustor) is given a certain endowment, which s/he can transfer to their partner (trustee). Before it reaches them, the investment is tripled. The trustee can then decide whether they want to send back any money to the trustor. The more the first player invests, the more they stand to make. The final outcome of course, depends on the trustworthiness of the second player. This setup can be used to investigate concepts like trust, reciprocity, fairness, and betrayal aversion.

→

In a public goods game (Ledyard, 1995), two or more individuals begin with the same endowment from which they can make a contribution to a common pool. All contributions are then multiplied by some factor k , and then divided evenly among all group members. All individuals receive the same return from the pool, regardless of whether and how much they invested. This scenario opens up the possibility of free-riding: leaving others to provide to the common pool, thereby earning profit while not providing any of their own funds. As a result, this experimental set-up is ideal to model real-world public goods scenarios such as tax evasion, littering, and fighting climate change.

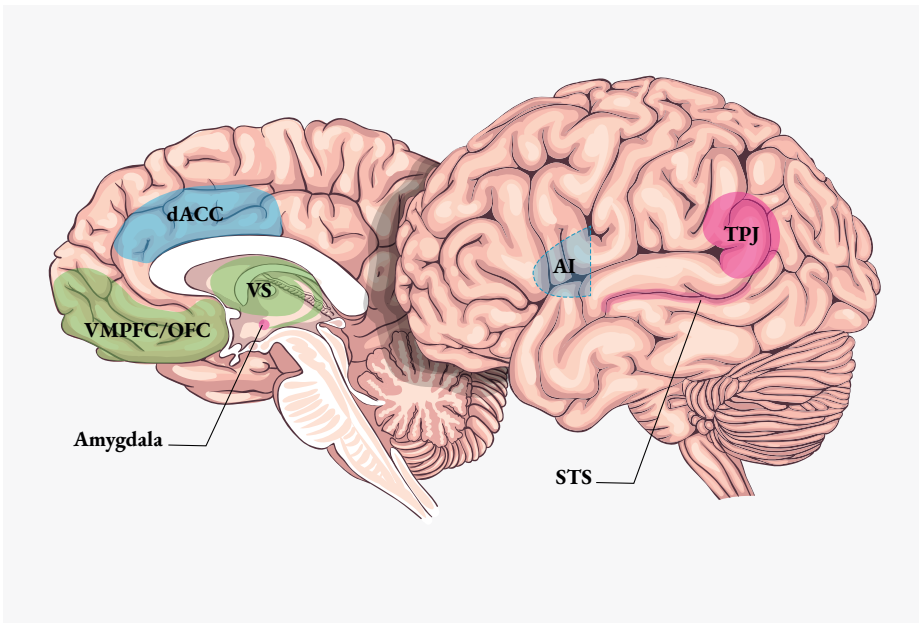
Neural correlates of social preferences and beliefs

Theories about the foundations of social preferences are, of course, not mutually exclusive. We can have social preferences that encompass a desire for genetically related others to thrive, as well as a need to make strategic investments in our more immediate future. In fact, both imply that our choices are made on the basis of value-based computations. Humans interpret the possibilities they are faced with, compare them based on their subjectively perceived value, in order to ultimately select the option with the greatest value for them.

Neuroeconomics is grounded in these theories of social decision-making as a starting point and attempts to formulate models that are also cognitively and neurobiologically plausible. There is evidence that value-based decisions, whether social or not, recruit the same neural circuitry in the ventral striatum, ventromedial prefrontal cortex (VMPFC), orbitofrontal cortex (OFC) and amygdala, and support the idea of a “common neural currency” (for a review see Ruff & Fehr, 2014); A concept that economists have worked with as an assumption for utility models, but for which neuroscientific studies offer substantiation. As it pertains to more specific social information such as face recognition and learning from others, studies on rodents, non-human primates and in humans have shown that the amygdala, the prefrontal cortex (especially subregions in the OFC and medial prefrontal cortex) and their interactions (Gangopadhyay et al., 2021) (Figure 1) play an important role.

Figure 1

Brain regions involved in social decision-making. In green: areas recruited during value-based decisions; in pink: areas recruited during theory of mind; in blue: regions representing ingroup/outgroup distinctions. AI: anterior insula; dACC: dorsal anterior cingulate cortex; STS: superior temporal sulcus; TPJ: temporal parietal junction; VMPFC/OFC: ventromedial prefrontal cortex/orbitofrontal cortex; VS: ventral striatum.



In addition to determining our own subjective value, social decision-making often involves incorporating our beliefs about others, as well as our estimates of how others perceive the situation. This ability to reason on the perspective of others – known as Theory of Mind – has been shown to involve the temporoparietal junction (TPJ) and the superior temporal sulcus (STS) (Saxe & Powell, 2016) (Figure 1).

The role of social preferences becomes particularly relevant when interacting across group lines. Research specifically examining intergroup conflict has found consistent neural activation in reward-related areas such as the ventral striatum (Hackel et al., 2017; Telzer et al., 2015), and orbitofrontal cortex (OFC) (Molenberghs et al 2016). Yet even in the absence of competition, people are sensitive to cues about functional relations and categorize others into coalitions based on perceived ease of coordination (Cikara, 2021).

Indeed, generalized representations of “us” versus “them” have been found in the dorsal anterior cingulate cortex and the anterior insula, irrespective of how group membership was characterized – whether these were arbitrary or based on political affiliations (Cikara et al., 2017).

Outline of this thesis

Despite over 150 years since Darwin first postulated kin selection, much about our social preferences still remains poorly understood. How are these preferences formed? Are they stable, or context-dependent? Can they be manipulated? The present dissertation contributes to our understanding of human social preferences by weighing in on these questions in the three following chapters. Chapter 2 introduces the idea of fixed social preferences and concludes that individual differences in social preferences only weakly relate to structural differences in the brain. Chapter 3 continues this line of reasoning by examining to what degree the social context modulates other-regarding preferences and provides evidence that prosocial behavior can also be used as a cue for the value as an interaction partner. Chapter 4 concludes by investigating the neural mechanisms underlying parochialism and puts forward the notion that our social preferences rely on distinct sets of processes whereby the left TPJ and right DLPFC are causally involved in reducing distrust in the outgroup. Below I present a brief overview of each chapter.

In chapter 2, we measured the social preferences of 194 individuals, using the Social Value Orientation (SVO) Ring Measure. This measure consists of 24 incentivized decomposed dictator games, where the participant must choose between pairs of own-other monetary outcomes, constraining participants to systematically trade-off their own economic welfare with that of an anonymous partner. While social preferences predict trust, public goods provision and mutual gains bargaining, the permanence of said preferences has not been thoroughly tested on a neural level. To fill this gap, we performed a comprehensive whole-brain analysis on the relationship between general social preferences and anatomical differences. We tested whether these preferences, as measured by SVO, correlated with brain structure in 74 distinct bilateral brain areas with identifiable functionalities for human cognition and behavior. Neither concerns for personal outcomes nor concerns for the outcomes of others in isolation were related to anatomical differences. Yet, social preferences positively scaled with cortical thickness in the left olfactory sulcus, a structure in the orbital frontal cortex previously shown to be involved in value-based decision-making. Consistent with work showing that heavier usage corresponds to larger brain

volume, findings suggest that prosocial preferences relate to cortical thickness in the left olfactory sulcus because of heavier reliance on the orbital frontal cortex during social decision making. This study covered the whole brain in a relatively large sample of healthy participants. All these analyses revealed a single unique area with a significant relationship to social preferences that survived statistical procedures aimed to reduce spurious statistical associations (permutation testing). However, the amount of variance explained was rather small, suggesting that brain anatomy contributes little to the direct prediction of social preferences. Though people have consistent social preferences, behavior is often influenced by the environment and the (lack of) history of social interactions, and hence may explain why no strong relationship exists in neural structure.

Chapter 3 explores how partner choice influences one's cooperative behavior in intergroup conflicts. The possibility that human cooperation depends on reputation and partner selection is well-supported in theoretical, experimental, and ethnographic work. What remains unclear, however, is what partners people prefer during intergroup conflict and how partner selection during intergroup conflict modulates decision-making. We fill this void with experiments in which individuals could form coalitions with others to attack and exploit out-group rivals or, alternatively, to collectively defend against possible out-group aggression. After being randomly assigned to a green or yellow group, participants across three online studies ($N = 750$) performed three tasks. Task 1 elicited pre-conflict social preferences using a helping task. Task 2 elicited conflict participation and partner choices in an attacker-defender contest, and Task 3 elicited post-conflict social preferences again with the helping task. The decisions made in the helping task were used to classify them as either universal cooperators, parochial cooperators or selfish types. Mixed regressions revealed that participants in both attacker and defender groups preferred selfish partners less than parochial or universal partners. When comparing the distribution of pre-conflict preferences to those elicited post-conflict (when preferences could be used as signal for possible inclusion in the intergroup contest by other participants), we find a significant shift. In all experiments, post-conflict preferences signal less universal cooperation and selfishness and more parochial cooperation. In line with previous findings that individuals cooperate more when seen by others, we find a robust effect of visibility on people's cooperative behavior. Participants alter their helping decisions to become more parochial when these decisions can be seen by potential future partners.

The phenomenon of parochialism is further examined in Chapter 4. By disrupting the TPJ via Transcranial Magnetic Stimulation (TMS), we can study this region's role in intergroup trust. Previous work in cognitive neuroscience has indeed shown

that reduced perspective-taking ability increased the difference in trusting in-group versus out-group members, and revealed a link between perspective taking and BOLD response in the TPJ. Here we tested the hypothesis that disrupting the functionality of the TPJ reduces trust. 90 right-handed participants played an incentivized Trust Game in the role of the trustor with ingroup and outgroup members (manipulated within-subjects) while in an fMRI scanner immediately after receiving inhibitory Transcranial Magnetic Stimulation (TMS) on their TPJ (left, right, sham; manipulated between subjects). As expected, we found trust to be lower when paired with outgroup partners. Trust in the outgroup was further reduced when (left) TPJ functionality was disrupted. At the whole brain level, all trust decisions reliably associated with neural activity in areas involved in mentalizing (inferior frontal gyrus, insula, TPJ and cerebellum), and cognitive control (anterior cingulate cortex, and dorsolateral prefrontal cortex (DLPFC)). ROI analyses revealed a partner \times TMS-treatment interaction on neural activity in the right DLPFC. Participants with a disrupted left TPJ showed less activity in the right DLPFC in ingroup compared to outgroup trials. Results support the possibility that the (left) TPJ is causally involved in trust, in particular by reducing distrust in out-group members. These findings also suggest that cognitive control and mentalizing work in concert when deciding whom to trust, and whom to discriminate against.

Limitations and Directions for Future Research

Across the three chapters, this thesis contributes to the understanding of human social behavior, the underlying preferences, and its neural foundations. All of the studies presented employ large sample sizes to test theories on the nature of prosocial behavior and be able to detect effect sizes observed in past research. Taken together, we can conclude that while social preferences are stable, and even reflected in brain structure, they are also dependent on situational factors. Interacting with an ingroup or outgroup member, whether behavior is public or not, contributing to a public good or competing to win a conflict, all affect revealed social preferences. Manipulating these contextual features critically affect prosocial behavior.

The methodologies presented in this dissertation offer an interdisciplinary approach to the study of social preferences: borrowing from neuroscience, experimental economics and psychology. On the one hand, this approach benefits from relying on real people making incentivized choices from which we can deduce causal links. On the other hand, the ecological validity of these experiments has been criticized; arguing that behavior collected in an artificial laboratory setting cannot be gener-

alized to the “real world”. While it is true that experiments use simplified scenarios to study complex social behavior, they also allow for controlled observation and measurement. With the use of anonymous partners in a clearly defined game, we can ensure that changes in our dependent variable (e.g., economic choices, prosocial behavior) are in fact due to the manipulation of an independent variable of interest (e.g., interaction partners’ group affiliation, visibility of behavior), and not the myriad factors that are impossible to disentangle in the real world.

Though interesting in its own right, the use of neuroscience to study value-based decision-making has also been criticized as irrelevant to advancing our understanding of behavior (Gul & Pesendorfer, 2011). Some have argued that knowing what areas of the brain are involved in decision-making does not add to the predictive power of existing psychological or economic theories. However, both of these fields suffer from a tension between descriptive and prescriptive approaches. Is it more useful to attempt to capture optimal human decision-making under a single paradigm? Or is science better served by a more accurate detailing of how humans actually behave in given contexts? With the help of neuroscience, researchers can rely on the physical mechanisms and constraints of the human brain to reconcile this tension. As such, we can create predictive and parsimonious models of behavior based on the actual computations performed by the brain – thereby avoiding theories that offer erroneous divisions of decision processes. As Glimcher and colleagues articulated several years ago:

“Ultimately, economics is a biological science. It is the study of how humans choose. That choice is inescapably a biological process. Truly understanding how and why humans make the choices that they do will undoubtedly require a neuroeconomic science.”

(Glimcher et al., 2005)

Furthermore, these critiques of ecological validity and utility of neural correlates may be addressed by future research. Studies that use existing in and out-groups, such as individuals from warring factions, could complement the experiments in this dissertation by providing information about the progression of ingroup bias that minimal groups cannot. Moreover, this area of research would also be well served by using repeated interaction paradigms. Allowing participants to learn from and react to others’ behavior – instead of one-shot interactions – would also afford researchers the chance to better understand the neural pathways of social cognition and biases in real time.

Concluding Remarks

Without social interaction, our understanding of human behavior is incomplete. Social preferences are what makes humans human. Despite decades of research from psychology, anthropology, biology and economics, how social preferences arise and vary across contexts remains an open question. This dissertation addresses this gap using a variety of economic games and neuroimaging techniques that allow for a tractable modeling of cooperation and competition. The findings suggest that while social preferences are linked to neural structure, they can also adapt to environmental factors as well as beliefs about interaction partners. We have seen that interacting with ingroup or outgroup members, taking decisions publicly or privately, and knowing we may interact with others again affect cooperative behavior. These results highlight the importance of understanding how prosociality can be affected and lay the foundations for policy makers to further those social environments that encourage prosocial behavior.