Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature
: _____ _____
Minwoo (Daniel) Lee Date

Genetic and Neural Basis of Cultural Norm Acquisition

By

Minwoo (Daniel) Lee
Doctor of Philosophy

Anthropology

James K. Rilling, Ph.D.
Advisor

Gregory Berns, Ph.D.
Committee Member

Melvin Konner, M.D., Ph.D.
Committee Member

Adriana Lori, Ph.D.
Committee Member

Larry Young, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D., MPH
Dean of the James T. Laney School of Graduate Studies

Date

Genetic and Neural Basis of Cultural Norm Acquisition

By

Minwoo (Daniel) Lee

M.S., Behavioral and Cognitive Neuroscience, Korea University, 2016
B.A., Psychology, Korea University, 2013

Advisor: James K. Rilling, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Anthropology
2022

Abstract

Genetic and Neural Basis of Cultural Norm Acquisition

By Minwoo (Daniel) Lee

Cultural norm acquisition is a process through which individuals learn normative beliefs and values prevalent in their cultural environment. A developing body of literature has suggested that a set of genes may modulate the way the brain processes normative social feedback from others, thereby contributing to individual variations in cultural norm acquisition. Yet, the specific intermediate mechanisms that support such "social sensitivity" remain elusive. The primary aim of this dissertation project was to explore the genetic and neural substrates of the cultural norm acquisition process, with a specific focus on genetic variation in the oxytocin receptor gene (*OXTR*). 195 healthy adult participants (Neuroimaging arm $N = 50$, Behavioral arm $N = 145$) performed three cognitive tasks in an imaging genetics experiment. The first task measured participants' ability to detect subtle emotional cues that convey evaluative social feedback (i.e., facial micro-expressions). The second task measured their ability to discriminate the authenticity of social feedback (i.e., genuine vs. posed smiles). The third task measured participants' susceptibility toward conformity pressure imposed on the domain of moral values and virtues (i.e., moral conformity). Participants' behavioral and functional magnetic resonance imaging (fMRI) data were analyzed with respect to a single nucleotide polymorphism in *OXTR* rs53576 and multi-locus genetic profile scores (MPS) that reflected the level of *OXTR* expression in the brain. We found that *OXTR* rs53576 G homozygotes detected facial micro-expressions better than the A allele carriers. This genetic modulation was associated with increased activations in the brain areas implicated in attentional control. Furthermore, G homozygotes were more likely to erroneously judge posed social cues as genuine, which was linked with decreased activations in the brain areas involved with mentalizing. Lastly, participants with higher MPS showed greater moral conformity. This effect was mediated by decreased activations in the brain area implicated in conflict processing. Despite a need for further replication, these findings illuminate specific neuro-cognitive pathways through which *OXTR* may facilitate or hinder the cultural norm acquisition process across individuals. It also suggests the potential utility of MPS as a means to characterize and explain various high-level social phenotypes in humans.

Genetic and Neural Basis of Cultural Norm Acquisition

By

Minwoo (Daniel) Lee

M.S., Behavioral and Cognitive Neuroscience, Korea University, 2016
B.A., Psychology, Korea University, 2013

Advisor: James K. Rilling, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Anthropology
2022

Acknowledgements

Completing my dissertation would have been an insurmountable goal, if it had not been for the support of family, friends, mentors, and many others whom I feel blessed to have met and interacted with over the past six years.

First and foremost, I would like to thank my advisor Dr. Jim Rilling for his support throughout my graduate training at Emory. His continuous encouragement, patience and intellectual guidance gave me confidence and courage to carry on whenever I struggled with academic and personal challenges.

I wish to express my gratitude to the four members of my dissertation committee who contributed to my project in many different ways. Dr. Mel Konner is one of the most kind-hearted and knowledgeable people I have ever met. His expertise in anthropology, human biology, and developmental psychology, as well as his ability to navigate multiple levels of analyses have inspired my work tremendously. Dr. Larry Young was the person I always turned to, whenever there are puzzling discoveries. With his vast knowledge in molecular genetics and neural mechanisms underlying social cognition, Dr. Young never hesitated to share insights and creative ideas to improve my research. Dr. Gregory Berns provided me with essential methodological training in fMRI data acquisition and analysis, which later became the backbone of my research. His feedback allowed me to perfect the experimental design and an overall conceptual framing of my dissertation project. Lastly, it was thanks to Dr. Adriana Lori's dedication and patience that I was able to bring a proper closure to my research despite challenges posed by the global pandemic.

I am indebted to the Department of Anthropology for letting me have a safe and secure ground to pursue my graduate training. Without the help of Lora McDonald, Jill Marshall, Kay Norgard, and Brian Banks, I would have been lost in the swamp of administrative procedures long before I was even beginning to formulate my research questions. I also thank Drs. Chikako Ozawa-de Silva and Bradd Shore for the warm, personal welcome and encouragement they gave me when I first joined the department.

I want to thank my colleagues and friends at Emory - Christina Rogers, Megan Beney Kilgore, Jordan Marin, Scott Schuner, and Luisa Rivera, just to name a few- for helping me with various aspects of my research and personal lives. I am particularly grateful to Lynnet Richey for the friendship we shared in and outside the lab: her brilliance, personality, and expertise in ovens were a source of great joy and inspiration that I leaned on during my toughest stretches. I also thank Jessie Doan and Winston Leung, who always helped me escape the lab and do something completely unrelated to cultural norm acquisition. Pyungwon Kang, Heejung Jung, Yeehyun Yoon, Seungsoo Kim, and Gabi Caceres also hold a dear place in my heart for their emotional support and comradeship across the globe.

Lastly, I would like to express my deepest gratitude to my beloved family. Their unwavering support was the single most important reason why I was able to finish this long and daunting journey, not to mention the fact that I would not have chosen to pursue this path in the first place without their encouragement and dedication.

Table of Contents

List of Tables and Figures

Tables

Figures

Supplementary Materials

Chapter 1

Introduction

Anthropologists have long studied the complex layers of human phenomena emerging at the juncture of the self and its cultural milieu. Culminating in the notion of embodiment, or the fundamental entanglement between social environment and the subjective experience of the body, a continuing attempt to depict how our perception and behaviors are "socially informed" characterizes a central tenet of anthropological inquiry (Bourdieu and Nice 1977, Csordas 1990, Campbell and Garcia 2009, Bourdieu 1977, Campbell, Ophir, and Phelps 2009). The current dissertation project seeks to address the issue of embodiment, with a specific focus on the patterns of moral becoming in humans.

Morality is broadly defined as the interlocking set of values, norms, and practices that prescribe how we ought to relate to one another (Haidt 2008). It has been considered as a hallmark of human sociality not only because it is rarely identified outside human societies (Ayala 2010), but also because it exerts a great influence on people's everyday language and thought process (Lambek 2010, Fassin 2014), as well as social interaction (Wojciszke, Bazinska, and Jaworski 1998, Goodwin, Piazza, and Rozin 2014) One notable aspect of morality with respect to embodiment is its bio-cultural origin: despite a surprising degree of cross-cultural, or even cross-specific overlap in behavioral domains that are imbued with normative significance (Shweder et al. 1997, Graham

et al. 2011, Sheskin and Santos 2012), specific moral values and rules endorsed by different groups of people vary greatly depending on their respective cultural backgrounds (Haidt and Joseph 2004).

These findings suggest that morality in humans reflect both evolutionarily preserved physiological processes and the diversifying influences of individuals' socio-cultural environment. Such a biocultural nature of human morality, however, inevitably raises an important question: how individuals across the globe, endowed with a finite set of biological and psychological dispositions that are arguably universal, grow to acquire a wide variety of culture-specific normative behaviors and moral values (Haidt and Joseph 2004). That is, what are the specific mechanisms that mediate the process through which inputs from the social environment are embodied and promote the acquisition of cultural norms and moral values within individuals?

The primary goal of the current proposal is to elucidate proximate mechanisms underlying the acquisition of moral values and cultural norms in humans. The involvement of learning and social influence vis-à-vis the internalization of norms has been widely recognized across disciplines (Bandura 1973, Konner 2010, Graham et al. 2011). However, our understanding of specific biological and cognitive mechanisms that enable such an intertwinement is still limited by the theoretical and methodological divide between various fields of research (Campbell and Garcia 2009).

In the following subsections of this chapter, I will review how the topic of *cultural norm acquisition* has been discussed in four major disciplines in the biological and social sciences: anthropology, psychology, neuroscience, and genomics. Each subsection will include a brief summary of relevant research within these fields, as well as their theoretical or methodological limitations. Then, I will

conclude this chapter with a description of specific research hypotheses, empirical predictions, and methodology of this dissertation project.

Anthropological perspectives on cultural norm acquisition

Cultural anthropology of morality: Broadly put, the study of morality in cultural anthropology has been centered around two philosophical paradigms, each of which is characterized by a different conceptualization of morality (Fassin 2014).

The first approach originates from the early theoretical work of Emile Durkheim. In "*The Determination of Moral Facts*," Durkheim viewed morality as a "system of rules of conduct" which "interest our sensibility to a certain extent and appear to us as desirable" (Durkheim 1906). A sizable body of ethnographical works that focused on providing a detailed descriptive account of rules and moral principles across human groups can be counted as examples (Read 1955). More recent cases of the Durkheimian definition of morality would also include various cross-cultural projects that sought to identify the common but differentially expressed themes or "domains" of moral values. For instance, Richard Shweder and colleagues (1997) surveyed past literature on "suffering," including his own works on Hindu India, and proposed an original taxonomy of three moral domains or codes (i.e., autonomy code, community code, and divinity code) that "may encompass all moral systems in the world." (Shweder et al. 1997)

The second approach is aligned more closely with Michel Foucault's definition of morality. In "*The Use of Pleasure*," Foucault discussed three dimensions of morality (Foucault 2012). The first

dimension, which borders with the Durkheimian notion of morality, refers to a "set of values and rules of action that are recommended to individuals through the intermediate of prescriptive agencies such as the family, educational institutions, churches." The second dimension is "the real behaviors of individuals in relation to the rules and values that are recommended to them." The third and final dimension refers to "the manner in which one ought to form oneself as an ethical subject acting in reference to the prescriptive elements that make up the code" (Foucault 2012). Foucault's discussion centered mainly on the third dimension, which captures the "ethical subjectivation," or the process through which an agent continuously strives to become an "ethical subject" with respect to societal rules and moral principles. The ethnographic literature based on this second approach (Mahmood 2011, Asad 1993) tends to focus less on providing a bird-eye view on the moral codes and rules shared within a specific community per se, but instead delves into the subjective states of individuals who "conduct themselves in accordance with their inquiry about what a moral life is." (Fassin 2014). This approach is well-depicted in an ethnographic work by Saba Mahmood (2005) on Muslim piety movements in Egypt, where she described how the women of the mosque movement cultivated an embodied practice of personal piety in response to the influx of modern western feminist theory on agency, freedom and gender that poses a distorted depiction of the Muslim world (Mahmood 2011).

The topic of cultural norm acquisition intersects with these two genealogies of moral anthropology. The first line of works concerns the sources of the normative influences, as they outline the systems of values, rules, and principles that give rise to moral behaviors and beliefs of individuals in specific social environments. The second line of research can unveil the phenomenological dimension of cultural norm acquisition, where individuals navigate the complex web of moral standards while trying to establish and negotiate their "moral selfhood" (Simon 2009). It also reveals the everyday

language that ordinary people use to describe in the process of internalizing or dissenting from the normative influences around them (Lambek 2010).

These approaches have offered insights into the diversity and particularities of moral lives found in different human communities and individuals. Yet, however, their focus on local idiosyncrasies and emphasis on phenomenological terms may limit our understandings of potential similarity of moral judgment and patterns of moral development across cultures or at the level of species. Notably, recognizing this limitation, there is currently a growing realization within the field that combining cultural anthropological insights of cultural norm acquisition with various methods and theories in other fields such as evolutionary sciences, psychology, and neuroscience, could lead us to fuller descriptions of how our "ethical life" arises from the intersections of biology, culture, and history (Keane 2015, Seligman and Brown 2009, Northoff et al. 2006, Chiao et al. 2010).

Adaptive value of social norms and norm enforcement Relatedly, biological anthropology and various subfields in evolutionary sciences have approached the topic of cultural norm acquisition with respect to its adaptive utility, especially in terms of how it spreads uniquely human forms of cooperation. For instance, evolutionary modeling literature has shown that the existence of stable group norms can facilitate the identification of in-group members and promote the selective exchange of cooperative interactions among them, which could increase the average fitness of implicated individuals (McElreath, Boyd, and Richerson 2003). The effects of social norms on maintaining in-group cooperation are known to be stronger when these norms are internalized by group members (Gintis 2003, Gavrilets and Richerson 2017), and actively enforced to punish deviant behaviors (Chudek and Henrich 2011, Boyd and Richerson 1992). Empirical findings from field- and laboratory experiments have revealed a converging picture with the results of these

simulation studies. Across human populations worldwide, the domains of morality tend to be centered around various themes related to in-group identification and cooperation (van Schaik et al. 2014, Curry, Mullins, and Whitehouse 2019). Furthermore, the sensitivity towards norm violation has been identified universally in both second- and third-party social interaction (Henrich et al. 2001, Marlowe et al. 2008, Blake et al. 2015, House et al. 2013, McAuliffe et al. 2017). When there is no means to enforce norms, the level of cooperation in repeated social interaction has been shown to decay rapidly (Fehr and Fischbacher 2004).

Cultural transmission and process of social learning Biological anthropologists have also studied the modes of cultural transmission and social learning processes in humans through which these cooperative norms spread between individuals. The mode of cultural transmission concerns who is transmitting and who is acquiring cultural information. It is often expressed using directional terms (e.g., vertical, oblique, and horizontal transmission) similar to genetic transmission between different generations of organisms (Cavalli-Sforza and Feldman 1981). Vertical transmission of culture involves children learning from directly their parents. Horizontal transmission refers to the exchange of cultural information among individuals of similar age group. Finally, oblique transmission occurs between individuals of distinct generations or age groups, including extended kin group or local population (Cavalli-Sforza and Feldman 1981). Different modes of cultural transmission are known to be used for different domains of cultural information. For instance, a recent meta-analysis on the patterns of social learning in modern hunter-gatherer societies suggests that vertical transmission plays a major role in the acquisition of non-subsistence skills (e.g., manufacturing), while oblique transmission prevails when it comes to the development of language (Garfield, Garfield, and Hewlett 2016). The internalization of moral values and various social norms has been shown to be largely mediated by vertical and oblique transmission (Garfield,

Garfield, and Hewlett 2016), with the specific sources of social influence changing with the age of the learner and the domain of moral values being taught (Lew-Levy et al. 2017). For instance, sharing behaviors in many small-scale societies tend to be promoted initially via vertical transmission during infancy, and later via oblique transmission during early and middle childhood (Boyette 2013, Lew-Levy et al. 2017).

These vertical and oblique cultural transmissions may involve different social learning processes or the specific way social information influences individual learning (Hoppitt and Laland 2008). Among a dozen forms of social learning that have been identified and studied across taxa (Laland 2004), teaching and imitation are two types of social learning processes that have been most consistently associated with the acquisition of social norms and moral values across human societies (Garfield, Garfield, and Hewlett 2016). Specifically, studies have found that the former often takes place in the form of direct verbal instructions or commands from adults that deliver positive or negative reinforcements (Bakeman et al. 1990, Ho et al. 2017). The latter usually involves copying the behaviors in everyday social life where normative behaviors are exchanged by various members of the community (Endicott and Endicott 2014, Bakeman et al. 1990).

Primatology and comparative psychology of morality and social norms Lastly, extensive research has been carried out to determine whether morality and social norms exist outside the human society, and if so, how they are transmitted and learned across individuals. In fact, some non-human species behave as though they have "norms," some of which seems moral to human eyes (De Waal 1991, Silk and House 2011). For instance, findings in primatology and comparative behavioral sciences suggest that various non-human primate species may be equipped with prosocial sentiments in the domains including interpersonal harm/care (de Waal and van Roosmalen 1979,

Burkart et al. 2007), trust (Engelmann and Herrmann 2016), generosity (Jaeggi and Van Schaik 2011), in-group bias (Mahajan et al. 2011), and fairness (Brosnan and De Waal 2003, 2014). However, it has been pointed out that these characteristically moral behaviors among non-human primates are predominantly observed during the second-party social interaction and not necessarily mediated by the enforcement or gradual acquisition of social norms (Buckholtz and Marois 2012, Silk and House 2011). For instance, high-ranking male chimpanzees are known to intervene when there is dispute among low-ranking conspecifics, a behavior called as "policing" (Goodall 1986, Von Rohr et al. 2012, Rudolf von Rohr, Burkart, and Van Schaik 2011). Yet, in many cases, policing in chimpanzees and other primates is linked with the assertion of social and sexual interest of interveners (Rudolf von Rohr, Burkart, and Van Schaik 2011), or could reflect their annoyance and frustration caused by the disturbance (Goodall 1986). Evidence for systematic norm enforcement in chimpanzees that involves the principled delivery of reward and punishment by disinterested third parties is again scarce, if not absent (Rudolf et al., 2011 (Rudolf von Rohr, Burkart, and Van Schaik 2011, Riedl et al. 2012).

Summary and limitation In all, the topic of cultural norm acquisition has received much attention in multiple subfields in anthropology. Cultural anthropologists provided ethnographic records of the patterning of local moral domains endorsed across human societies. They also provided detailed descriptions of the subjective experience of individuals who navigate the complex web of moral codes and rules to achieve what they view as "ethical life." Biological anthropologists and evolutionary scientists have contributed to our understanding of the 1) evolutionary scenarios that might give rise to cooperative social norms in humans, 2) how they spread, and 3) the adaptive significance of norm enforcement in comparison to non-human primates.

An important limitation of the anthropological literature on cultural norm acquisition is that it generally falls short of revealing the proximate mechanisms through which individuals internalize social norms within species. For the cultural anthropological approach, this was due to its strong emphasis on studying the qualitative dimension of the norm learning processes and to its relatively narrow focus on the experience shared within a particular community or individuals. The lack of mechanistic specificity in biological anthropology can be attributable to the notion of "phenotypic gambit" ((Fawcett, Marshall, and Higginson 2015), or the idea that the evolution of complex traits can be modeled without necessarily considering the specific mechanisms underlying them (Grafen 1991). Evolutionary modeling and comparative literature on human social learning and norm enforcement likewise made somewhat unrealistic assumptions that the capacity for social learning and norm enforcement is under tight genetic control and varies only meaningfully between species (Mesoudi et al. 2016). Individual variations existing within species, by contrast, have often been considered as noise or error. Notably, there is a growing call for incorporating more detailed mechanistic knowledge from neuroscience and psychology into the literature (Rittschof and Robinson 2014), as it will ultimately improve the accuracy of formal models and allow researchers to detect novel evolutionary dynamics that would not be observable under the models based on the phenotypic gambit (Mesoudi et al. 2016).

Psychological perspectives on cultural norm acquisition

One of the most important pioneering attempts in psychology to study cultural norm acquisition was made by developmental and social psychologists in the early- and mid-20th century, especially through the lens of children's moral development. It was widely recognized that children show

differential levels of moral sensitivity as they grow older, and researchers sought to understand what drives this change.

Cognitive-Developmental tradition and moral development One influential view, championed by Jean Piaget and Lawrence Kohlberg, posited that children go through invariant stages of moral development, with the later phases requiring more sophisticated cognitive mental faculty such as perspective taking (Piaget 1932, Kohlberg and Kramer 1969). As they emphasized the close entanglement of cognitive and moral development, those subscribing to this 'cognitive-developmental tradition' considered private reasoning done by individuals the key determinant of whether they can advance from lower to higher levels of moral development (Piaget 1932, Carpendale 2000). The contribution of social influence, by contrast, was thought to be peripheral. While Piaget and Kohlberg both recognized that children would be motivated to avoid sanction and seek approval from others (Piaget 1932, Kohlberg and Kramer 1969, Haidt 2008), such social reinforcement processes were deemed as something one needs to overcome during the early phase of development to reach the highest level of moral competence defined by the endorsement of the universal moral principles such as justice and fairness (Cowan et al. 1969).

As one of the first principled endeavors in the field that aimed to theorize how humans grow to embody norms and values over time, theories in the cognitive-developmental tradition became the "mainline" of the moral psychological research and had lasting impacts on the related literature (Haidt 2007). However, its central assertions that 1) moral development follows a single trajectory leading to the narrow sets of universal moral principles, and that 2) reasoning is the key driver of that process also beget multiple lines of critiques, each of which ultimately led to the deepening of our understanding of cultural norm acquisition in humans.

Social learning theory of moral development The first major opposition stemmed from the findings that people with different cultural or demographic backgrounds tended to endorse distinct norms and values (Bandura and McDonald 1963, Gilligan and Attanucci 1988). Such diversity could not be easily reconciled with the notion of morality drawn by the stage models, which assumed a single, fixed, and universal sequence of moral development. Some theorists, including Kolhberg himself, interpreted the variable patterning of moral values as an indication that some human groups are deficient in moral capacity (Kohlberg and Kramer 1969). Yet, many others had come to believe that the moral development process could be much more flexible and contextualized than what had previously been thought. Facing the discrepancy, social learning theorists proposed an alternative framework. Unlike the cognitive-developmental tradition that focused on individuals' cognitive architecture as a primary driving force underlying moral development in humans, social learning theorists emphasized the role of external forces in cultural norm acquisition. For instance, the early work of Albert Bandura showed that children would readily imitate specific types of social interaction (e.g., aggressive vs. non-aggressive action performed on the "Bobo" doll) performed by adults (Bandura, Ross, and Ross 1961). Later, Bandura also tested whether children's moral judgments were contingent on various forms of social reinforcement (Bandura and McDonald 1963). In an experimental setting originally invented by Piaget, children of varying ages read two stories. Each of these vignettes described either a well-intentioned behavior that incurred considerable material damage or an ill-intentioned behavior that resulted in minor consequences. Children were asked to determine which of the cases was "naughtier." In one condition, children saw an adult model whose answers were always in opposition to their responses. The experiment rewarded the model's behavior using verbal approval such as "Very good" and "That's good." The children were also rewarded if they adopted the

model's moral judgment instead of theirs. The second condition was identical to the first, except that the positive reinforcement was delivered only to the adult models. Only children received the reward for adopting the model's responses in the last condition. Piaget initially found that older children were more likely to choose the latter, weighing more on the intention behind the action (i.e., Subjective responsibility). Younger children under age 10, on the other hand, selected the former more often, showing increased sensitivity towards the outcome of the action (i.e., Objective responsibility). The social influence manipulation introduced by Bandura should not have had changed this pattern if Piaget's view that children follow the universal sequence of moral development were to be substantiated. However, the results showed that a considerable number of children shifted their views, regardless of their age (Bandura and McDonald 1963). Bandura concluded that "children's moral orientations can be altered and even reversed by the manipulation of response-reinforcement contingencies and by the provision of social models(Bandura and McDonald 1963).

Social influence: psychology of conformity and obedience to authority Besides the works of social learning theorists, researchers who studied conformity and obedience behaviors further revealed the general malleability in moral judgments and normative behaviors caused by various social factors, even among adults. For instance, a seminal study by Solomon Asch showed that people have a strong tendency to follow the opinions of the majority regardless of the accuracy of the social consensus (Asch 1956). Such conformity effect was initially demonstrated in the domain of perceptual decision making such as determining the length of lines, yet later replicated in experimental paradigms involving judgments in moral dilemmas and economic games (Kim et al. 2016, Kundu and Cummins 2013). In a similar vein, multiple social psychologists such as Stanley Milgram (Milgram 1974) and Philip Zimbardo (i.e., The Stanford Prison Experiment) (Zimbardo

et al. 1971) showed that people are disturbingly susceptible to the power of authority and social role expectations, even in the situation when these may induce harm to others (Cialdini and Goldstein 2004b). Lastly, longitudinal studies on cultural assimilation documented the gradual adoption of local behavioral norms among immigrants in the United States, which suggests that various social processes can alter adults' moral values and beliefs beyond temporary behavioral shifts often observed in the laboratory experiments (Miller et al. 2009).

This converging evidence in developmental and social psychology has gradually eroded the critical tenets of the Kolhbergian and Piagetian theories that a single, linear progression to moral universals can characterize the moral becoming of humans. Instead, they have reinstated the importance of social influence in moral development in humans, effectively accounting for the apparent within- and between-group variability in norms and values exhibited by both children and adults.

Yet another important body of research called for a critical reappraisal of the cognitive-developmental approach to human moral development. Inspired by evolutionary perspectives on the adaptive utility of emotions (Wilson 1975, Cosmides and Tooby 2000), and also by cognitive psychological literature on the "automaticity" found in social behaviors (Bargh and Chartrand 1999, Aarts, Dijksterhuis, and Custers 2003), doubt was cast on the alleged primacy of reasoning and deliberation in human moral judgments.

Dual-process model of morality The key theoretical support for this new trend came from the "dual-process" framework. Dural-process framework posited that human behaviors are explained by the crosstalk between two distinct yet mutually interacting systems: affective and cognitive processes (Evans 2004, Kahneman 2011) While this view was not inherently related to the studies

on human morality and cultural acquisition, its focus on the fast-acting, affective information processing led to the major revisions in the way psychologists viewed the structure of the human moral mind, which, in turn, provided insights into its evolutionary and developmental origin. Jonathan Haidt (2000) proposed the "social intuitionist model" based on the "moral dumbfounding" effect (Haidt, Bjorklund, and Murphy 2000), where people fail to provide rational justifications to their own moral judgments beyond stating the immediate gut reactions they experienced before making the decisions. The involvement of emotion and intuition in moral judgments was demonstrated further by a series of laboratory experiments utilizing so-called "trolley dilemmas." In the classic trolley dilemma, people are typically asked to indicate whether it would be morally acceptable to sacrifice one person for saving a larger number of people from a runaway trolley (Foot 1967). Philosophers have long been interested in whether and how people's responses to the question change across the different permutations of the original vignettes. For example, it was found that lay people were more inclined to make the sacrifice when they could stop the train by pressing a lever (i.e., trolley dilemmas). However, this majority response reversed when stopping the train required a more "personal" intervention such as pushing another innocent person off from the bridge (i.e., footbridge dilemma). A seminal study by Joshua Greene and his colleagues (2001) found that brain regions implicated in emotional processing such as the amygdala or ventromedial prefrontal cortex showed increased activations when participants responded to the footbridge-type vs. trolley-type dilemmas. Interestingly, in a series of follow-up behavioral studies, the authors found that characteristically utilitarian judgments (i.e., sacrificing one to save five) made in the footbridge-type dilemmas took significantly longer (Greene et al. 2001, Greene et al. 2004), and this difference became greater when participants did not have enough cognitive resources (Greene et al. 2008). These results suggested that people make use of their internal emotional feedback to guide their moral judgments. Specifically, the observed difference in the reaction time could be the

result of participants using cognitive control to override negative emotional arousal associated with incurring direct physical harm before they made sacrificial judgments for the footbridge-type dilemmas, which was absent for the classical trolley-type dilemmas. In the later studies, researchers indeed confirmed that the increased activations in the limbic cortex closely tracked people's subjective experience of negative affective states (Shenhav and Greene 2014).

Dual-process model of morality and reinforcement learning The dual-process models of moral judgments offered a new opportunity for researchers to study these "cognitive-slow vs. emotional-fast" aspects of human morality with increased mechanistic specificity. For example, Fiery Cushman suggested that these two processes can be re-expressed based on the notion of value, a concept that had widely been employed in economics and behavioral neuroscience to capture the motivating properties underlying instrumental actions (O'Doherty 2014). Cushman proposed that judgments in moral dilemmas involve two types of value computation processes: action and outcome values. The former and latter represent the level of appetitiveness or aversiveness of performing a specific action in a specific context (e.g., Pushing the lever), and the predicted outcome caused by the action (e.g., Sacrificing one person and saving the five lives), respectively.

Reframing the emotional and cognitive architecture of moral judgments in terms of values was a major conceptual breakthrough in psychology that had a direct implication for studying the cultural norm acquisition process. That is, how people come to endorse specific moral values and norms can be analyzed from the angle of a well-defined computational framework detailing how the action values and outcome values are learned and combined across individuals (Gęsiarz and Crockett 2015). For example, people's responses in the "footbridge dilemma" will be systematically different depending at least partially on the aggregated reinforcement history they received as

children for hitting or pushing others (e.g., negative social feedback such as punishment, scolding, verbal instruction about the negative consequences of violence, or frowned faces) (Crockett 2013, Gęsiarz and Crockett 2015).

A plethora of evidence now shows the generalizability of this approach to the development of normative behaviors and moral values across individuals and human populations. For example, David Rand (2014) proposed that sets of behaviors leading to advantageous outcomes in everyday life will give rise to generalized intuition, or "social heuristics," that automatically trigger those behaviors whenever one is facing a relevant situational cue (Rand et al. 2014). In computational terms, social heuristics in one population would reflect the collection of action-value representations shaped by repeated exposure to rewards or punishments based on the population-specific local norms and rules. Indeed, experimental evidence has shown that different human groups (e.g., the US vs. India) exhibit distinct prosocial heuristics in economic decision-making games (e.g., public good game), depending on the local prevalence of altruism, or how often altruistic behaviors would lead to advantageous outcomes for individuals (Capraro et al. 2017).

Summary and limitation In sum, psychologists have been at the forefront of studying the patterns and the mechanisms of cultural norm acquisition. The cognitive-developmental tradition and its assumptions regarding the universal stages of moral development and the centrality of reasoning subsequently sparked many lines of research weighing how various sources of social influences shape the diverse patterning of norms across different human groups and also the role of emotional processes in driving human morality. Together, these efforts gave rise to the new theoretical framework that incorporates the concept of values to analyze how specific types of moral values are internalized and affect normative behaviors and judgment. Especially, this last approach is

considered promising when it comes to analyzing the specific computational processes underlying cultural norm acquisition beyond the stage models or dichotomous distinction between emotion and cognition.

Notwithstanding this potential, however, the nascent psychological literature on cultural norm acquisition suffers similar shortcomings as the anthropological studies: the paucity of empirical evidence showing how these models regarding value computations are represented in the brain. For instance, questions such as "how social influence that shapes one's moral judgments can be connected to the established neural signatures of reinforcement learning, and moral cognition" have not yet been thoroughly studied and remains an important topic of future research.

Cultural norm acquisition in social neuroscience

Social neuroscience is a promising approach that could add a more detailed mechanistic framework to the existing lines of research in anthropology and psychology. Social neuroscience is a thriving field of study that focuses on the neural basis of social cognition and social behavior (Cacioppo et al. 2007). With the advent of advanced neuroimaging technologies such as functional magnetic resonance imaging (fMRI), the field has provided a unique window into the neural mechanisms of complex social behaviors in humans. Two areas of social neuroscience research hold special relevance to the topic of cultural norm acquisition: the neuroscience of morality and reinforcement learning.

Dual-process theory of explicit moral judgments It has been widely recognized among researchers that patients with brain damage often show deficits in social behaviors, including the inability to act in accordance with the moral standards of society (Damasio 1999). The most dramatic illustration of this is the case of Phineas Gage, who suffered a lesion in the medial orbitofrontal cortex due to a railroad accident, which subsequently took away his "balance between intellectual faculties and animal propensities." Gage soon developed a series of behaviors that were considered "out of his character," such as uttering "the grossest profanity" and showing "little deference for his fellows" (Twomey 2010). Many clinical cases showing the link between brain damage and social defects similar to Phineas Gage's appeared in the literature in the mid to late 20ths century and subsequently formed the foundations of modern moral neuroscience (Gazzaniga 2005)

Modern moral neuroscientists aim to characterize the brain mechanisms that support moral cognition and behaviors in humans. One major line of research within the field focuses on the neural correlates of moral decision-making, which were typically measured using experimental paradigms where participants make explicit judgments of right and wrong. Jorge Moll was the first to measure brain responses to a series of short verbal descriptions or graphical depictions of moral violations (Moll, Eslinger, and Oliveira-Souza 2001, Moll et al. 2002, Moll et al. 2005). This pioneering work was soon followed by another seminal study by Joshua Greene, who employed the trolley dilemmas to measure the neural correlates of judgments in moral dilemmas (i.e., "Is it morally permissible to sacrifice one life for saving five lives?") (Greene et al. 2001). The key findings from these studies suggested that human moral judgments have both affective and cognitive components (Greene and Haidt 2002, Moll, de Oliveira-Souza, and Eslinger 2003). Neuroscience of explicit moral judgments matured with numerous follow-up studies that focused on how activations within these brain regions vary across different human groups, including

clinical (Koenigs et al. 2007, Ciaramelli et al. 2007, Glenn, Raine, and Schug 2009, Moran et al. 2011, Yoder, Porges, and Decety 2015);, cross-cultural (Han, Glover, and Jeong 2014), and developmental samples (Harenski et al. 2012).

Neural mechanisms underlying prosocial behaviors Another important piece of literature that has deepened our knowledge of our moral brain comes from studies on human prosocial behaviors. Unlike the early works in moral neuroscience that focused largely on the judgments of right or wrong, these studies investigated the neural basis of a wider range of behaviors or cognitive functions that are considered moral and normative in human societies. Some researchers adopted various experimental paradigms from behavioral economics and evolutionary game theory that were originally designed to study the patterns of human cooperation, such as direct reciprocity, generalized altruism, and prosocial norm enforcement (Lee 2008). For example, Rilling et al (2002) used a sequential "prisoner's dilemma game (PDG)" to simulate repeated social interaction in humans and found that the brain regions implicated in reward-processing such as the caudate nucleus showed increased activations to reciprocated cooperation (Rilling et al. 2002). Similarly, Sanfey and colleagues (2006) used a version of "ultimatum game (UG)" to investigate the neural correlates of perceived (un)fairness in social interaction and their behavioral relevance. The authors found significant activations in the anterior cingulate cortex (ACC) and the anterior insula (AI) in response to unfair offers, which subsequently predicted rejection of the unfair offer (Sanfey et al. 2003). The activation within the ACC and AI were also found when in the context of altruistic punishment, where participants acted as disinterested third-party and delivered costly punishment to those who violated fairness norms in a dictator game (DG) (Strobel et al. 2011). In fact, results from recent activation likelihood estimation (ALE) meta-analyses further confirmed that the perception of norm-abiding behaviors (e.g., cooperation, fair offer in UTG and DG) and norm-

deviation (e.g., unreciprocated cooperation, unfair offer in UTG and DG) consistently incurred the activations within the striatum (e.g., striatum and caudate nucleus) and ACC/AI, respectively (Feng, Luo, and Krueger 2015, Zinchenko and Arsalidou 2018).

Moral cognition in relation to uniquely human social cognition Lastly, inseparable from the advancement of moral neuroscience was an attempt to identify the neural basis of uniquely human social cognition. The two most notable themes from this endeavor would be empathy and theory of mind (ToM). While neither empathy nor ToM is inherently linked with human moral judgments and prosocial behaviors (Decety and Cowell 2014, Decety 2021), a plethora of empirical and meta-analytic evidence has still revealed the close entanglement between empathy, ToM and human moral cognition (Bzdok et al. 2012, Eres, Louis, and Molenberghs 2018).

Empathy, generally defined as an ability to mirror other peoples' affective states (Eisenberg and Miller 1987), had been studied extensively even before the advent of social neuroscience. Psychologists studied it as a proximate mechanism to altruism and prosocial development (Eisenberg and Mussen 1989, Batson et al. 2016, Batson et al. 1981) Primatologists and comparative cognitive scientists sought to understand the phylogeny of empathy by studying whether non-human animals, including apes show signs of empathic responding to others' distress (Preston and De Waal 2002b, Preston and de Waal 2002a, Bartal, Decety, and Mason 2011). Adding to the existing literature, social neuroscientists provided more detailed views into the neural mechanisms that support empathy in humans and animals. Tania Singer, for example, demonstrated that the brain regions that process the direct (e.g., the ACC and AI), first-hand experience of pain also become activated during the observation of a loved one's pain. This provided neural evidence of the affective state-sharing mechanisms described in the earlier psychological literature (Singer

et al. 2004). Singer et al (2004) soon inspired a myriad of follow-up studies that look into the developmental trajectory of empathy (Decety 2015), its relation to psychopathology (Yoder et al. 2015), and, importantly, explicit moral judgments (Crockett et al. 2010, Decety and Cowell 2018) and various prosocial behaviors such as charitable donation, helping and caregiving ((Hein and Singer 2010, Feldman 2016, Rilling and Mascaro 2017).

Studies on theory of mind (ToM) in social neuroscience also share a similar root with the literature on empathy. The ability to "understand another's cognitive status and perspectives" in humans was initially studied among developmental psychologists and anthropologists (Underwood and Moore, 1982 (Underwood and Moore 1982, Eisenberg and Miller 1987, Luhrmann 2011). Social neuroscientists later adopted the experimental paradigm and stimuli that had been used by these earlier scholars (e.g., False belief task), which required participants to dissociate their own and others' beliefs based on information uniquely available to themselves (Gallagher et al. 2000) . The results of these studies indicated that the neural representation of other's intent, motivation, and belief is found in the multiple brain regions, including the medial prefrontal cortex (mPFC), superior temporal sulcus (STS), precuneus, temporal pole (TP), and right temporoparietal junction (rTPJ) (Gallagher and Frith 2003). Later studies focused on delineating the differential functional contributions of these individual brain regions to ToM performance (Saxe and Powell 2006, Schurz et al. 2014, Schaafsma et al. 2015), and how the activation within the "ToM network" is modulated by demographic factors including age (Gweon and Saxe, 2013 (Gweon and Saxe 2013), sex (Gao et al. 2019), and culture (Kobayashi, Glover, and Temple 2006) as well as social deficits such as autism (Silani et al. 2008, Lombardo et al. 2011). Researchers also investigated how ToM intersects with the brain regions involved in moral judgments and prosocial behaviors as the ability to understand other's intention and motivations were thought to be crucial for judging moral

responsibility (Young et al. 2007) or engaging in affiliative actions such as helping, caring, and behavioral coordination in repeated social interaction (Brüne and Brüne-Cohrs 2006)). For example, Liane Young and colleagues showed a consistent involvement of rTPJ in belief encoding during moral judgments, especially when people evaluated good-intentioned behaviors that caused great harm or ill-intentioned behaviors that resulted in small harm (Young and Saxe 2009). The authors also confirmed that the effects of belief on explicit moral judgment diminished when the activity of the rTPJ was disrupted with transcranial magnetic stimulation (Young, Camprodon, et al. 2010). James K. Rilling and colleagues found that the significant activations in the ToM network (e.g., dorsomedial prefrontal cortex, dmPFC; rostral anterior cingulate cortex, rACC) when participants had to determine whether to reject unfair offers in UG or to cooperate in PDG based on their partners' actions (Rilling et al. 2004). Intriguingly, some areas within the ToM network (e.g., STS) showed stronger activations in response to the actions of human partners vs. a computer partner, potentially indicating the unique recruitment of mentalizing during social interaction (Rilling et al. 2004).

Paradigm shift to domain-general account of moral cognition One of the key questions that motivated the early studies in moral neuroscience was whether it would be possible to find the "moral module" in the brain, a single, or network of brain regions dedicated to moral cognition and prosocial behaviors in humans (Young and Dungan 2012). This endeavor was in part influenced by the long-lasting philosophical notion of human uniqueness (Saxe 2006), and more recently by a proposal that humans are equipped with "universal moral grammar" that is similar to our seemingly inherent capacity for language acquisition (Mikhail 2007). However, studies on explicit moral judgments, prosocial behaviors, and social cognition have gradually revealed rather an opposite picture that human morality is represented "everywhere but nowhere" in the brain (Young

and Dungan 2012). In other words, it is now widely believed that our brain does not have a neural circuitry exclusive for moral cognition or prosociality. Rather, our moral competence is a result of the complex interaction between social and emotional information processing which could also participate in non-moral cognition (Young and Dungan 2012).

This 'domain-general' account of moral cognition was further accelerated by a contemporaneously emerging idea of the neural "common currency." This hypothesis suggested that the human brain operates on a single scale of value which allows us to compare and choose from multiple decision options that lead to different types of utility (Levy and Glimcher 2012). This proposal concerning the "root of all values" gained empirical support by multiple fMRI studies showing that both social (e.g., smile and social approval) and non-social (e.g., money and food) rewards are represented and integrated in the overlapping regions in the brain, most notably in the ventromedial prefrontal cortex (vmPFC), or medial orbitofrontal cortex (mOFC) (Levy and Glimcher 2012). In fact, the activation within the vmPFC had been reported routinely in moral neuroscience literature (Ciaramelli et al. 2007, Greene et al. 2001, Greene et al. 2004, Koenigs et al. 2007, Glenn, Raine, and Schug 2009, Harenski and Hamann 2006, Heekeren et al. 2003), although its specific functions in moral judgments and behavior were often debated (Young, Bechara, et al. 2010). Yet, with relevance to the neural common currency hypothesis, Shenhav and Greene (2010) demonstrated that the domain-general valuation mechanisms involving vmPFC and nucleus accumbens (NAcc) are recruited for encoding the probability and magnitude of harm during judgments in moral dilemmas (Shenhav and Greene 2010). Later, researchers further showed that vmPFC creates an integrative value signal that predicts moral judgments by combining both affective and cognitive appraisal of moral dilemmas represented in the brain regions such as amygdala and dmPFC, respectively (Shenhav and Greene 2014, Hutcherson et al. 2015)Hutcherson et al., 2015). Based

on these findings, a major trend in moral neuroscience has now moved from identifying the unique neural signature of morality or prosociality to exploring the connection between moral cognition and other domain-general information processing in the brain.

Such a paradigm shift in moral neuroscience had important implications in the study of cultural norm acquisition in social neuroscience. That is, most previous studies in moral neuroscience largely centered around the question of how people make moral judgments or perform normative actions as opposed to non-moral judgments or norm-deviant behaviors, rather than how they learn moral values or modify existing normative behaviors and beliefs. Therefore, while experimental findings and methods in moral neuroscience could provide us with a reference point for studying the neural representation of morality and norms, they offered limited insights into how learning and social influence interact with the neural circuitries involved in moral cognition. With the new approach that emphasizes the link between moral cognition and domain-general valuation mechanisms in the brain, however, researchers can address the question of cultural norm acquisition based on the theoretical and formal framework that concerns how our brain computes and modifies values.

Neuroscience of reinforcement learning Reinforcement learning (RL) is the most influential framework that could provide a window to the neural mechanisms underlying cultural norm acquisition. The RL framework has long been employed in cognitive science to characterize how behavioral changes occur through experience (Gęsiarz and Crockett 2015, Joiner et al. 2017). Its core foundation was derived from Markov decision process, which models how artificial agents should make decisions and learn from interactions with the environment to achieve certain goal states (Sutton and Barto 1998). For example, an agent moves through different environments or

states, which are defined by currently available actions and their corresponding outcomes. The agent computes a *reward function* to calculate the expected outcomes of certain states with respect to its goal- typically reward maximization. The agent also utilizes the reward function to develop a set of preferred actions known as *policy*. Finally, the agent may also have the model of the environment or *transition function* that defines how given actions in one stage lead to the next states ((Gęsiarz and Crockett 2015, Joiner et al. 2017). There exists a wide variety of RL models with different mathematical formulations and assumptions to characterize each of these components. Yet, in essence, they describe a chain of a decision where the agent forms a predictive model of the future outcomes associated with certain actions and constantly compares the predicted outcome with the actual experiences as it proceeds from one state to another.

The fundamental attribute of RL framework that drives this iterative modeling process is *prediction error* (PE), which corresponds to the discrepancy between the predicted vs. experienced reward outcomes (Schultz 2000). These PE signals are used to update the reward function and, eventually, the actions of an agent navigating its environment. The concept of PE and its role in determining the value of actions was initially used in machine learning and psychological literature. However, it was later incorporated into the neuroscientific literature after a series of animal studies confirmed the existence of the predictive coding mechanisms in the brain and its underlying computational properties (Rangel, Camerer, and Montague 2008).

It was well-known from conditioning experiments that an arbitrary stimulus with no intrinsic reward value (e.g., the sound of bells) will be perceived as rewarding after being associated repetitively in time with the appetitive stimulus (e.g., Food delivery) (Dickinson 1980). It was also shown that dopamine (DA) neurons in the midbrain (e.g., ventral tegmental area, VTA) mediate

the observed transfer of the appetitive value from a rewarding object to a predictive stimulus (Schultz, Dayan, and Montague 1997, Schultz 2016). Yet, the seminal study by Schultz and colleagues further showed that the specific response property of the DA neurons resembled that of the PE signals: the activity of DA neurons increased for the rewards greater than the predicted reward and decreased for the reward smaller than the predicted reward. For the fully anticipated rewards, DA neurons stopped responding (Schultz, Dayan, and Montague 1997, Schultz, Apicella, and Ljungberg 1993). The neural signature of PE in the midbrain was initially identified in Rhesus monkeys (Schultz, Dayan, and Montague 1997) but later also confirmed in other animals, including humans (Pan et al. 2005, Cohen et al. 2012).

The PE signals in the human brain were initially identified in the experimental paradigms in which participants must build an internal model of the task environment to maximize their reward. For example, in a widely used probabilistic learning task, participants are presented with several pairs of arbitrary symbols. Each symbol of these pairs is assigned with a high or low probability of reward, which is not known to participants. To achieve maximum rewards, participants should learn about the action-outcome contingencies of different pairs of symbols on a trial-and-error basis. In computational terms, action values for choosing high- vs. low-reward symbols should be calculated to make decisions in each trial and then updated based on the predicted outcome of the action (e.g., delivery of high reward). Participants' choices made for each trial are later fitted against a mathematical model that depicts an RL algorithm (e.g., temporal difference model, TD) to derive model parameters including PE (O'Doherty et al. 2003) . Based on this approach, early studies found that PE signals were most consistently represented in the striatum (Valentin and O'Doherty 2009), with evidence of possible functional dissociation between the caudate nucleus and nucleus accumbens (O'doherty et al. 2004). Later studies further revealed that a distributed

network of brain regions outside the striatum such as vmPFC, dACC, AI, and VTA also contribute to different aspects of PE (Garrison, Erdeniz, and Done 2013, Joiner et al. 2017, Fouragnan, Retzler, and Philiastides 2018).

Neuroscience of 'social' reinforcement learning Most research on RL and PE has been conducted in the context of individual learning or the learning process that does not involve any social components. Intriguingly, however, a growing body of evidence suggests that PE signals can also be triggered by a variety of social information. For example, it has been shown that PE signals can be induced vicariously in the vmPFC and VS when people observe others receive rewards for their actions during a reinforcement learning task. Markedly, this social information later led to more rapid learning performance in the identical task (Burke et al. 2010). Studies have also found that mismatches between one's belief about others' future actions and the actual actions performed by others also trigger error signals similar to the non-social PE in the nucleus accumbens and the anterior cingulate cortex (Zhu, Mathewson, and Hsu 2012, Joiner et al. 2017). Lastly, social feedback from others has been known to be a source of PE. For instance, facial expressions of positive or negative emotions, which would indicate social approval or disapproval, can elicit positive or negative PE signals in the overlapping brain regions that produce PE signals for non-social rewards or punishment (Lin, Adolphs, and Rangel 2011, Jones et al. 2011). It also appears that social feedback need not be visual or inherently evaluative to incur PE signal, as the mere gap between self and others opinions has been shown to effectively trigger error-related activations in the NAcc and dACC, which subsequently led to behavioral changes (Klucharev et al. 2009, Zaki, Schirmer, and Mitchell 2011, Izuma 2013, Levorsen et al. 2021).

The existence of such "social reinforcement learning" mechanisms that guide the gradual acquisition or modification of behaviors and beliefs via PE offers a promising avenue for studying the neural basis of cultural norm acquisition. As reviewed in previous sections, moral values and normative behaviors often arise from feedback from other individuals, which have been shown to recruit domain-general value computation mechanisms in the brain (Cushman, Kumar, and Railton 2017). Therefore, it would be possible to characterize the process through which people's normative behaviors and moral values emerge in the face of social feedback using the RL framework and PE.

However, although researchers have utilized the RL framework to study various facets of non-social and social cognition and behaviors such as neural responses to primary/secondary rewards (O'Doherty et al. 2003, Kim, Shimojo, and O'doherty 2010, Lin, Adolphs, and Rangel 2011), strategic behaviors (Seo et al. 2014, Lee, Seo, and Jung 2012), preference formation (Klucharev et al. 2009, Izuma 2013), and addiction (Mollick and Kober 2020), the possible application of the RL framework on studying cultural norm acquisition has not been adequately explored (Cushman, Kumar, and Railton 2017).

Summary and Limitation In all, social neuroscience has offered unique insights into the neural representation of moral judgments and normative behaviors, which psychological and anthropological endeavors have often lacked. Furthermore, social neuroscience also offers researchers a conceptual and mathematical framework that can illuminate the brain mechanisms subserving social learning. However, these two lines of research have not yet been brought together to study cultural norm acquisition or social learning of moral values and normative behaviors.

Cross-fertilization between these two subfields within social neuroscience would thus be necessary to effectively close this gap.

Cultural norm acquisition in social genomics

Social genomics investigates how specific social environments influence the expression and function of genes. Social regulation of gene expression has long been studied with animal models, especially in relation to the low-level physiological and morphological traits such as growth rates in bees, sex switching in cichlids, and body size, coloring, and immune responses in primates (Robinson, Fernald, and Clayton 2008, Tung et al. 2012, Cole et al. 2012, Powell et al. 2013).

Gene-environmental Interaction framework Social genomics research on human subjects has often focused on the genetic susceptibility towards a wide range of diseases and psychopathologies under specific social environments. For instance, Caspi and colleagues (2003) found a significant association between a length polymorphism (i.e., 5-HTTLPR) in the serotonin transporter gene (i.e., SLC64A) and depression, with the short allele (i.e., S allele) conferring a greater risk than the long allele (i.e., L allele). Intriguingly, the authors also found that the link between the S allele and the increased likelihood of showing depressive symptoms was only present when the carriers were exposed to early life stress (Caspi et al. 2003). Such environment-dependent association between genes and their phenotypic outcomes was also identified for other genes in addition to 5-HTTLPR, such as the gene encoding monoamine oxidase A (i.e., MAOA) (Caspi et al. 2002, Kim-Cohen et al. 2006), dopamine D4 receptor (DRD4) (Bakermans-Kranenburg and Van Ijzendoorn 2011, Sasaki 2013), and serotonin transporter protein (SERT) and receptor (5HTR1A) (LeClair,

Janusonis, and Kim 2014). These findings altogether solidified the gene-environment interaction framework (Ottman 1996), which centered on the idea that environmental conditions can moderate the psychological or cognitive outcomes of a specific genetic sequence or that genetic predispositions may change the strength of the association between the environment and an outcome.

Gene-culture interaction framework Of relevance to cultural norm acquisition, a nascent body of research in social genomics has attempted to extend the scope of the gene-environment interaction framework such that it can be more widely applicable to explain the development of various social behaviors shaped by one's cultural surroundings. That is, unlike most previous studies in the gene-environment interaction literature that focused on the variability in individuals' personal environment such as stressful home settings (Taylor et al. 2006), this new approach centers on "culture" as a constellation of beliefs, values, practices, and products that constitute shared meaning system (Geertz 1973) that shapes the psychology and behaviors of a specific population (Sasaki et al. 2016). Naturally, those who study this "gene-culture interaction" focus less on the notion of "susceptibility genes" that may lead to aberrant physical or mental conditions in a specific personal environment. Instead, they look for "plasticity/sensitivity genes" that aid the acquisition of various culture-specific phenotypes, such as the value of independence and autonomy in Western society, or that of interdependence and harmony in East Asian countries ((Kim et al. 2011, Sasaki 2013, Sasaki et al. 2016).

One of the first supporting evidence of the gene-culture interaction framework came from Kim et al (2011), where the authors showed that a single-nucleotide polymorphism (SNP) in the human oxytocin receptor gene (OXTR) is implicated in the development of culture-specific psychological

traits across different societies. For example, East Asians who are homozygous to the G allele of OXTR rs53576 were more likely to show increased emotional regulation (Kim et al. 2011, Kim et al. 2010) and less social support seeking in the face of stress. Remarkably, the same genotype was associated with the opposite patterns of psychosocial phenotype, such as increased emotional expression and more social support seeking in Americans (Kim et al. 2011, Kim et al. 2010). Drawing upon the well-documented cross-cultural differences in collectivistic vs. individualistic values in Eastern vs. Western societies, the authors suggested that the OXTR rs53576 was associated with the increased sensitivity towards normative social influences specific to each cultural cluster. Soon after the initial proposal, a separate group of researchers also identified a similar pattern of culture-specific genotype-phenotype association for the dopamine D4 receptor gene (*DRD4*), which has a variable number tandem repeat (*VNTR*) polymorphism. Specifically, those with 7 and 2 repeats, as opposed to 4 repeats, were shown to have a more interdependent and independent self-construal in East Asian countries and North American countries, respectively (Kitayama et al. 2014). Lastly, the short allele of the serotonin transporter polymorphism (*5-HTTLPR*) was significantly associated with the increased sensitivity towards to the disappearance of facial expressions in Japanese students but not in American students. This finding was consistent with the greater emphasis on social harmony among the former compared to the latter (Ishii et al. 2014), which might have rendered participants more sensitively respond to the potential signs of social rejection and ostracism.

Synthesizing the multiple lines of findings, Shinobu Kitayama and colleagues recently proposed the “norm sensitivity” hypothesis: a group of genes, rather than producing a fixed behavioral phenotype, could systematically enhance people’s ability to process evaluative social inputs from

others via altered neuromodulation. This could, in turn, facilitate the acquisition of norms and values widely shared and enforced in a given cultural environment (Kitayama et al. 2016).

Summary and Limitation Despite the empirical support, the social genomics approach to cultural norm acquisition based on the gene-culture interaction framework and norm-sensitivity still has limitations. One major issue is that most previous studies that identified an association between genes and culture-specific psychosocial traits have relied almost entirely on survey methods. Given that data collected using self-report questionnaires often fail to reflect respondents' actual behaviors (Stone et al. 1999), heavy reliance on survey methods raises the question of whether so-called "plasticity" or "sensitivity" genes have actual behavioral significance. This points to a more fundamental limitation to the existing literature on gene-culture interaction: the lack of a mechanistic model. Variation in genes such as *OXTR*, *DRD4*, and *5-HTTLPR* is likely related to intermediate phenotypes in the brain that produce diverging behavioral responses. Therefore, studying how these genes affect information processing in the brain is necessary to fully characterize how the association between the sensitivity/plasticity genes and culture-specific psychosocial traits emerges. For instance, the exact contribution of the sensitivity/plasticity genes to cultural norm acquisition would be better understood if the intermediate phenotypes of these genes in the brain, or "neuro-endophenotypes", are explored with reference to the neural mechanisms underlying social learning and moral cognition discussed in the previous section. Lastly, there has been no systematic attempt to search for social sensitivity/plasticity genes based on their effects on gene expression within the brain, although most candidate genes that have previously been identified so far are implicated in neuromodulation of which the specific downstream effects depend heavily on the patterns of receptor expression in the brain. Instead, researchers have often specified a list of genes mostly based on their reported behavioral effects.

Yet, this approach fails to identify specific mechanisms by which these genes can affect brain function.

In sum, defining the role of candidate genes such as OXTR, DRD4, and 5-HTTL in cultural norm acquisition requires further investigation, including 1) controlled experiments that involve both behavioral and neuroimaging data acquisition, 2) systematic investigation of different cognitive processes that these genes could modulate their higher-level phenotypes, and 3) identification of target sensitivity genes and their variations with respect to their influences on gene expression in the brain.

The current research: bringing existing lines of research together

The primary goal of this dissertation project is to explore specific biological and neuro-cognitive pathways based on which individuals learn moral values and norms from social feedback. As reviewed in the previous sections, various unbridged interdisciplinary gaps limit our understanding of cultural norm acquisition.

Notwithstanding its phenomenological descriptions and sophisticated evolutionary analyses of morality and social learning, anthropological literature has not given sufficient attention to the specific proximate mechanisms through which moral values are acquired and internalized within individuals. Psychologists have proposed multiple theories on the cognitive architecture of human morality and provided more specific mechanistic grounds for how our moral sensitivity can develop based on various sources of social influences. However, these models have not yet been

adequately linked with the literature on the neurobiology of value and reward, although these concepts have gained increasing prominence in psychological theories of moral learning and cultural norm acquisition. Social neuroscience taps into the level of analysis missing in anthropological and psychological literature. The rise of moral neuroscience and reinforcement learning theories has much to offer for understanding the neural basis of human moral cognition and how moral values become internalized within individuals. Yet, these two topics have been studied independently of one another, which kept the question of whether and how individuals learn moral values and normative behaviors via RL mechanisms unaddressed. Lastly, the gene-culture interaction framework in social genomics has pointed to the genetic basis of cultural norm acquisition, yet without specifying the intermediate biological pathways between the implicated genes and culture-specific behavioral or psychological phenotypes.

This dissertation project aims to overcome these limitations by employing 1) a methodology that can connect the genetic, neural, and behavioral levels of analysis and by 2) adopting an integrated analytic framework that specifies multiple neuro-cognitive processes involved in cultural norm acquisition within individuals.

Imaging genetics as the key methodological framework

Imaging genetics offers a unique window to intermediate mechanisms in the brain that link genetic polymorphisms with their behavioral or psychological phenotypes. As discussed earlier with regards to the "phenotypic gambit," a mere statistical association between high-level phenotypes and genes does not establish biological significance nor reveal a specific mechanistic underlying

the link, imaging genetics has been regarded as a valuable methodological strategy for "extending statistical evidence with biological data" (Bigos and Weinberger 2010).

By employing imaging genetics, this dissertation addresses the need for an integrative methodological framework that incorporates multiple levels of analysis. That is, it can shed light on the specific neurocognitive underpinnings of the cultural norm acquisition process, which are often not discussed in anthropological, psychological, and social genomics literature. For this dissertation project, I recruited 252 healthy adult volunteers. Among those who completed the study procedure (N=200), the full behavioral data were obtained from 200 individuals. The blood-oxytocin level-dependent (BOLD) fMRI data were collected from 50 individuals. Genetic data were collected from 192 individuals. Behavioral data and fMRI data were analyzed in conjunction with the genetic data, which focused on *OXTR*.

OXTR as the primary candidate gene

This dissertation project specifically focuses on the oxytocin receptor gene (*OXTR*), which regulates the signaling of the neuropeptide oxytocin (OT). Of course, it would be unrealistic to claim that a single gene would mediate a phenomenon as complex as cultural norm acquisition. Still, our focus on *OXTR* is based on three reasons. First, OT is known to regulate a wide range of social cognition and behaviors relevant to cultural norm acquisition. Second, despite the multiple genes that have been identified as "sensitivity" or "plasticity" genes in the gene-culture interaction literature, the genetic variations in *OXTR* (e.g., rs53576) have yielded the most consistent findings for a wide range of behavioral and psychological domains such as emotional regulation (Kim et al. 2011)), social support seeking (Kim et al. 2010) , empathy (Luo, Ma, et al. 2015), and psychiatric

conditions such as depression, psychological well-being (Sasaki et al. 2016), and loneliness (LeClair et al. 2016). Lastly, the social function of OT and its mechanisms of action have been extensively studied in various mammalian species, including humans, which could be translated into a more detailed mechanistic model. In the following subsections, I will summarize the neurophysiological basis of OT signaling effects on mammalian and human sociality. Then, I will present how they can be relevant for studying the cultural norm acquisition process.

OT secretion and mechanisms of actions OT is a nine-amino-acid neuropeptide that evolved from an ancient precursor in invertebrates at least 600 million years ago. Mammalian OT is synthesized in two distinct classes of neurons in the hypothalamus: magnocellular and parvocellular neurons. Magnocellular neurons are located in the paraventricular (PVN), supra-optic (SON), and accessory nuclei of the hypothalamus. They project to the posterior pituitary gland (i.e., neurohypophysis) and form neurohemal contacts with local fenestrated capillaries, where OT is secreted into the bloodstream.

Magnocellular neurons also form axon collaterals to various forebrain structures such as the prefrontal cortex, striatum, hippocampus, and amygdala, facilitating coordinated OT release in both the central and peripheral nervous systems (Valstad et al. 2017). Smaller than magnocellular neurons in size and number, parvocellular neurons do not directly project to the neurohypophysis. Instead, they connect to the brain stem, midbrain, and spinal cord, thereby regulating various autonomic functions (Jurek and Neumann 2018). Parvocellular OT neurons in the PVN also project to the ipsilateral SON and to the contralateral PVN, in which they are conjoined by magnocellular neurons to regulate the effect of OT on the pain processing and stress response (Eliava et al. 2016).

Magnocellular and parvocellular neurons contain large dense core vesicles (LDCVs) which contain and release OT upon exocytosis. The release of OT can occur at axon terminals or soma and dendrites of an OT neuron. Somatodendritic release of LDCVs lets OT gradually diffuse to the hypothalamic and other brain regions that are not directly innervated by OT neurons (Zheng et al. 2014). Yet, the effectiveness of such volume transmission is unclear, as the amount of OT released from dendrites may be insufficient to modulate neuronal activities outside 55-120 uM radius from the release sites (Chini, Verhage, and Grinevich 2017). Therefore, it is thought that axonal release of OT, which could have more rapid effects on the targeted brain regions outside the hypothalamus (Johnson and Young 2017), is likely responsible for the long-range neuromodulation caused by OT (Grinevich and Neumann 2021).

OXTR and its regulatory role in mammalian sociality Once released, OT must bind to *OXTR* to exert its downstream effects. The location and density of *OXTR* in the brain differ significantly across mammalian taxa, which is believed to be the basis of some species-specific social behaviors (Donaldson and Young 2008). One seminal study, for instance, showed that the mating preference between two closely related vole species (i.e., monogamous prairie voles vs. meadow and montane voles) was contingent on the *OXTR* expression in the nucleus accumbens (Insel and Shapiro 1992). Evidence showing the link between the patterns of *OXTR* distribution and cross-specific variations in social behaviors across primate species is also emerging (Mustoe, Taylor, and French 2018, Staes et al. 2014). All non-human primate species (e.g., macaque, titi monkey, chimpanzee, and marmoset) that have been investigated to date show *OXTR* expression in the nucleus basalis of Meynert (MBN), a major source of cholinergic input to the rest of the brain. However, dense *OXTR* binding in the NAcc has been observed in common marmosets, which may be linked with some facets of sociality unique to marmosets, such as social monogamy and cooperative breeding

(Freeman and Young 2016). Similarly, one recent study utilizing receptor autoradiography found that *OXTR* is absent in the reward-sensitive regions in chimpanzee brains, such as nucleus accumbens (NAcc) and ventral pallidum (VP) (Rogers et al. 2021). This is in stark contrast to the human brain, which shows a much more wide-spread *OXTR* expression beyond the striatum and basal ganglia (Quintana et al. 2019). While specific molecular functionality of *OXTR* in the reward-sensitive regions in the primate brains is not fully established due to various technical and ethical challenges, this result may reflect the neurochemical foundations of the observed differences in human vs. chimpanzee sociality, such as pair-bonding and parenting behaviors (Rogers et al. 2021, Freeman and Young 2016).

Notably, the pattern of central *OXTR* expression also differs in a within-species fashion, and allelic alterations in *OXTR* are known to play a key regulatory role in this variation. A striking demonstration of this comes from a recent study on male prairie voles (King et al. 2016), where the authors showed that a single nucleotide polymorphism (SNP) in *OXTR* (i.e., NT213739) was associated with region-specific *OXTR* mRNA expression in the brain. Specifically, NT213739 regulated the gene expression in the nucleus accumbens but not in the insula. Importantly, the level of mRNA expression in the nucleus accumbens showed a strong linear association with the local receptor binding density, which, in turn, predicted the OT-related social behaviors such as pair-bonding tendency (King et al. 2016).

Studies on primate species, including humans, also suggest that *OXTR* polymorphisms can modulate social cognition and behaviors via their regulatory effects on the central OT signaling (Staes et al. 2014). For example, many human fMRI studies have found the associations between *OXTR* SNPs and the activities in the brain areas expressing *OXTR*, including, but not restricted to,

the ventral striatum (Loth et al. 2014), caudate nucleus (Feng et al. 2015), amygdala (Marusak et al. 2015, Waller et al. 2016), and anterior cingulate cortex (Luo, Li, et al. 2015). Similarly, the effects of intra-nasally administered OT on social behaviors, and their underlying brain activations are known to be modulated by *OXTR* genotypes (Feng et al. 2015, Marsh, Henry, et al. 2012, Chen, Kumsta, et al. 2015, Watanabe et al. 2017). Lastly, there is a growing body of evidence showing a direct link between *OXTR* and OT signaling in the human brain, with allelic variations in *OXTR* SNPs influencing OT binding (Freeman et al. 2018) as well as the level of local mRNA expression (Reuter et al. 2017, Almeida et al. 2022).

Functions of OT and its underlying neural mechanisms Then, what functions does OT play once it binds to *OXTR* in the brain? OT has long been studied in relation to the biological functions it serves at the periphery, such as lactation and parturition in females (Churchland and Winkielman 2012). Yet, numerous studies have shown the far-reaching effects of central OT signaling, especially on complex social cognition and behaviors.

OT is perhaps most well-known for its effects on promoting social affiliation, such as parent-infant and conjugal bonding, cooperation, and behavioral coordination (MacDonald and MacDonald 2010, Shamay-Tsoory and Abu-Akel 2016). OT has a popular reputation as a "love hormone" and was once called a "moral molecule" in the science community (Zak 2013). Indeed, consistent with this "prosocial hypothesis" of OT, early pharmacological studies using intranasal administration of OT (INOT) found that those treated with OT showed increased interpersonal trust (Kosfeld et al. 2005) and altruism (Zak, Stanton, and Ahmadi 2007) in economic decision-making games. INOT also promotes conformity (Stallen, Smidts, and Sanfey 2013) and behavioral, as well as neural coordination among in-group members (Arueti et al. 2013, Mu, Guo, and Han 2016). Those who

investigated the behavioral and psychological correlates of various *OXTR* SNPs (e.g., rs53576) also found similar results with respect to empathy (Rodriguez (Rodrigues et al. 2009, Gong et al. 2017), mentalizing (Laursen et al. 2014), sensitive parenting (Feldman et al. 2012), and prosocial temperament (Tost et al. 2010).

The OT-induced upregulation of social affiliation may reflect OT's s modulatory inputs to the brain regions central to social cognition and behaviors. INOT treatment is known to enhance neural representations of positive social stimuli such as interpersonal touch (Scheele et al. 2014), smiling faces (Gamer, Zurowski, and Büchel 2010), and reciprocated cooperation in the economic decision-making game (Rilling et al. 2012). Notably, the effects of OT were often found in the brain regions implicated in reward and motivation, such as the VTA, Nacc, caudate nucleus, and mOFC/vmPFC (Grace et al, 2018 (Grace et al. 2018). These findings suggest that OT could promote affiliative behaviors by enhancing approach-related motivation ("Is affiliation important?") and the reward value of positive social interaction (e.g., "Is affiliation happy?") (Bartz (Bartz et al. 2011, Bartz 2016).

Yet, it should be noted that the impact of the effects of OT on social affiliation has been shown to be context-dependent and person-specific. For instance, many studies showed that INOT does not unconditionally promote affiliative social interaction when other individuals are unfamiliar (Declerck, Boone, and Kiyonari 2010) or from an outgroup (De Dreu et al. 2011). OT can also promote feelings of envy and gloating (Shamay-Tsoory et al. 2009), elicit defense-motivated competition (De Dreu and Kret 2016), protective response to aversive stimuli (Striepens et al. 2013), and active aggression towards outgroup (Zhang et al. 2019). Results from animal studies also

paralleled these findings, with central OT signaling predicting maternal aggression in rodents ((Ferris 2005, Bosch et al. 2005).

Facing this curious discrepancy, the "social salience hypothesis" of oxytocin (Shamay-Tsoory and Abu-Akel 2016) alternatively proposed that the role of OT in social cognition can be better characterized by its involvement in salience encoding in the brain. That is, OT could enhance the perceptual representation of socially relevant stimuli and increase attentional orientation towards the social cues (Walum and Young 2018). In a recent animal study, for example, Marlin et al (2015) tested the impact of the endogenous release of OT on rats' maternal behaviors (e.g., pup retrieval). While virgin female rats are typically insensitive to distress calls made by pups, this basic tendency was reversed when OT was optogenetically induced, causing the virgin female rats to engage in pup-directed care. Consistent with the social salience account of OT, the authors found that OT increased the signal/noise ratio of socially important stimuli by balancing the magnitude and timing of inhibitory and excitatory responses in the left auditory cortex. This pattern of cortical activity resembles that of experienced mothers (Marlin et al. 2015).

INOT treatment in humans has also been shown to improve detection and recognition of both positive and negative emotional expressions (Groppe et al. 2013), increase attentional orienting towards social cues regardless of valence (Domes, Sibold, et al. 2013), and selectively promote learning based on social-, but not non-social information (Hu et al. 2015). These effects are known to be mediated by activities in the brain areas that encode perceptual salience as well as attentional control such as the NAcc, superior colliculus, amygdala, and anterior cingulate cortex (Shamay-Tsoory and Abu-Akel 2016).

Functions of OT in multi-stage human decision making These diverging hypotheses are sometimes framed as mutually exclusive. However, it has been suggested that the effects of OT may differ depending on the specific information processing stages in the brain and that OT's role in enhancing social salience and social affiliation may concurrently contribute to social cognition and behaviors (Bartz et al. 2011). For example, increased OT signaling in the brain may help individuals to better attend to various social cues (e.g., smiling face) and amplify their reward values which could subsequently increase approach behaviors to obtain the reward. Here, OT's roles in promoting the salience of social cues and increasing the reward of the stimulus cues are not necessarily at odds with one another (Bartz et al. 2011). Rather, both explanations should be systematically combined in a unified framework to fully characterize the effects of OT on social cognition and behaviors in a specific context.

Accordingly, Piva and Chang (2018) recently proposed an analytic approach for studying the role of OT in complex social decision-making in humans. Similar to the schematics used by cognitive scientists to study attention and response selection mechanisms (Johnson and Proctor 2004), this framework divides the human social decision-making process into multiple, mechanistically separable stages in the brain () (Piva and Chang 2018). The first two stages (e.g., sensory input and sensory perception) correspond to the *perception of stimulus* in which an organism detects relevant social cues from its environment. These stages are followed by the *valuation and decision formation stage*, during which the organism assigns the value of the perceived social stimulus and incorporates this information to generate a set of outcome-maximizing choices in a given social context. The last stage is the *behavioral output* step, where the best action is executed.

While OT could influence any of these sequential processes to modulate the final behavioral responses, the specific function of OT may differ between the early and later stages of the decision-making. For example, OT's role in enhancing the salience of social stimulus may influence the perceptual stages more than the valuation and decision formation stages, while the OT-related upregulation of social affiliation and approach motivation may modulate the valuation and response formation stage more strongly. Also, the degree to which each of these processing stages influences the final behavioral outputs may also vary, depending on the specific social contexts ((Piva and Chang 2018).

Therefore, this framework offers researchers a useful tool to deconstruct a phenomenon of their interest into multiple sequential information processing stages. Researchers can also test if any of these stages is modulated by OT and if such OT-induced modulation produces behavioral outputs meaningfully different from baseline (e.g., placebo). Finally, based on these observations, researchers can infer what specific functions of OT, social salience, social affiliation, or both, contribute to producing the behavioral outputs in the given task environment.

OXTR **and cultural norm acquisition: a principled approach**

This analytic framework can also be applied to delineate the intermediate neuro-cognitive mechanisms that mediate genetic variations in *OXTR* and the internalization of moral values and norms. This will be especially important for addressing the lack of mechanistic models in the previous literature (e.g., social genomics).

OXTR, social salience and the perception of evaluative social cues One possible mechanism through which genetic variations in *OXTR* can influence cultural norm acquisition during the early stages of social decision-making is increasing the salience of evaluative feedback from others, such as faces with positive or negative emotional expressions (Shamay-Tsoory and Abu-Akel 2016). It is well-known that moral violations trigger specific emotional states such as anger and disgust (Hutcherson and Gross 2011)). Moreover, people often rely on these facial expressions to determine if their behaviors are socially appropriate (Cushman, Kumar, and Railton 2017). Thus, accurately detecting and decoding various evaluative social cues during the *perception* stage may form a ground for learning moral values and normative behaviors in one's social environment. As reviewed above, many INOT studies have shown that the drug treatment improves face and emotion processing (Leppanen et al. 2017, Shahrestani, Kemp, and Guastella 2013), as well as other psychological traits associated with sensitive social perception such as empathy and ToM (Domes, Heinrichs, Michel, et al. 2007, Hurlemann et al. 2010). Therefore, it is possible that genetic variations in *OXTR* can also modulate brain mechanisms implicated in these functions, thereby affecting the accurate perception of social cues.

OXTR, social affiliation, and valuation Besides its influence on the perception of social cues, genetic variations in *OXTR* can also be linked with cultural norm acquisition by affecting the later stages of social decision-making. Specifically, *OXTR* could activate approach-related motivations and affiliation goals among individuals (Bartz 2016) thereby affecting the *valuation* of social alignment and cohesion. It is also possible that *OXTR* SNPs further affect the *decision formation* to induce behavioral responses to achieve social alignment and cohesion when there is a mismatch between self and others. One way to determine the effects of *OXTR* SNPs on valuation and decision formation would be to measure the neural responses associated with social reinforcement learning.

Studies have shown that the error-monitoring mechanisms in the brain are activated in response to the perceived social misalignment, and the resultant PE signals lead to subsequent behavioral adjustments to close this gap (Shamay-Tsoory et al. 2019, Klucharev et al. 2009, Zaki, Schirmer, and Mitchell 2011, Levorsen et al. 2021). Therefore, it is possible that *OXTR* SNPs may influence the value of social alignment, which will modulate 1) the magnitude of PE signals in the brain regions such as NAcc and dACC in the face of social misalignment (Klucharev et al. 2009). The larger PE signals may promote 2) conformity behaviors to reduce the mismatch. While the relationship between *OXTR* and PE has not been directly tested in the previous literature, INOT treatment has been shown to induce behavioral conformity and behavioral synchrony (Stallen et al. 2012, Feldman et al. 2012). This finding suggests the similar involvement of *OXTR* in social reinforcement learning.

A strategy linking OXTR SNPs with OT signaling in the brain: A crucial issue that needs further clarification is how to model the effect of *OXTR* SNPs on central OT signaling. It has been repeatedly demonstrated that INOT passes the blood-brain-barrier via trigeminal and olfactory nerve fibers and increases the availability of OT molecules in the cerebrospinal fluid (CSF) (Quintana et al. 2021). This finding allows researchers to establish that various behavioral and psychological changes following INOT treatment at least partially reflect enhanced OT signaling in the brain. However, results from most behavioral- or pharmaco-genetics studies on human *OXTR* tend to be mechanistically elusive. It is because researchers often identify their candidate *OXTR* SNPs based on behavioral association studies without considering the actual regulatory effects of the target alleles on the receptor expression in the brain (Feldman et al. 2016).

One way to address this limitation is to infer the level of endogenous OT signaling based on one's *OXTR* genotypes. Specifically, since *OXTR* SNPs regulate the level of OT receptor expression across different parts of the brain, it is possible to create a cumulative index of *OXTR* expression for each individual by 1) identifying the high-expressing allele for the *OXTR* SNPs expressed in a target brain region of interest (ROI), and 2) calculating the total number of high-expressing allele summed across all the SNPs. The resulting output, or a multi-locus profile score (MPS), would correspond to the overall level of *OXTR* expression within a ROI, and thus the local OT signaling. This approach is similar to creating polygenic risk scores in clinical literature (Dima and Breen 2015). Yet, the MPS created in this way will allow researchers to interpret the effects of *OXTR* on their downstream phenotypes in a more biologically grounded way.

In this dissertation project, the effect of *OXTR* was modeled in two ways. The first approach focused on *OXTR* rs53576, one of the most extensively studied *OXTR* SNPs. Especially, the G allele of *OXTR* rs53576 has shown associations with sensitive social perception and affiliative behaviors (Li et al. 2015). Therefore, we first compared the data from G homozygotes (GG) vs. the A allele carriers (AA/AG). The second approach was based on the *OXTR* MPS score as described above. Specifically, I focused on seven *OXTR* SNPs (including rs53576) that regulate receptor expression in the brain regions implicated in reward processing and valuation such as the NAcc, caudate nucleus, and dACC and BA9. We note that the second approach was preferentially applied to analyze the data of Task 3, which were expected to recruit the brain areas in which the levels of the *OXTR* expression are known to be regulated by the seven target SNPs (See "Specific aim and Prediction"). The full list of these SNPs and specific procedures for calculating the MPS will be described further in the later chapters (i.e., Chapter 4).

Specific Aims and Predictions

Based on the role of OT in multi-stage decision-making in humans and its application to the cultural norm acquisition process, I conducted an imaging genetics experiment with three tasks. These tasks were designed to measure the association between *OXTR* and sensory input/perception and valuation/decision formation.

Task 1 (i.e., the facial emotion detection task) aimed to examine the impact of sensitivity alleles in *OXTR* on the detection of evaluative social cues. I hypothesized that the G allele of *OXTR* rs53576 (e.g., G homozygotes vs. the A allele carriers) would facilitate cultural norm acquisition by improving the initial detection of subtle social cues that are associated with positive and negative emotional reactions. At the neural level, we anticipated that this effect would be accompanied by increased activation in the core brain regions involved with face processing (i.e., posterior superior temporal sulcus, pSTS; Inferior frontal gyrus, IFG) and emotion perception (e.g., bilateral amygdala; NAcc, caudate nucleus, and ventromedial prefrontal cortex, vmPFC), as well as affective empathy (i.e., Anterior insula, AI; dorsal anterior cingulate cortex/medial frontal cortex, dACC/MFC), and cognitive empathy (i.e., right temporo-parietal junction, rTPJ; medial prefrontal cortex, mPFC, and precuneus, PC). The specific coordinates of the target ROIs and sizes were obtained from previous activation-likelihood estimate (ALE) meta-analyses on explicit and implicit evaluation of facial expressions (Dricu and Frühholz 2016), positive and negative facial emotion (Fusar-Poli et al. 2009), perceptual-affective empathy (Fan et al. 2011), cognitive-empathy/Theory of Mind (ToM) (Schurz et al. 2014) and reward processing (Liu et al. 2011).

Task 2 (i.e., the smile authenticity judgment task) aimed to examine the impact of sensitivity alleles in *OXTR* on discrimination of the authenticity of social cues. As in Task 1, I hypothesized that individuals homozygous for the G allele of rs53576 would show increased ability to determine the authenticity of facial cues compared to the A allele carriers. At the neural level, we test whether perceptual discrimination of posed vs. genuine smiles would be supported by the ROI listed above. Specifically, we focus on a subset of brain areas involved in face processing and ToM such as the STS, IFG, and mPFC, as the joint contribution of brain networks regulating attention, sensori-motor simulation and mentalizing are known to be important for determining the authenticity of emotional cues (Paracampo et al. 2017, McGettigan et al. 2015).

Task 3 (i.e., the moral conformity task) measured if high endogenous OT signaling would influence the degree to which people value social alignment and modify their moral preference when there is social misalignment. It was hypothesized that increased OT signaling in the brain (i.e., high *OXTR* MPS), would increase people's tendency to modify their behavior in response to social feedback. Drawing upon the influential neuro-cognitive model of social learning and conformity, we predicted that the behavioral conformity effect would be mediated by genetic modulation of the PE-related activations in brain regions such as the nucleus (NAcc) (e.g., increased activations in response to perceived social alignment), caudate nucleus and the pMFC/dACC (e.g., increased activations in response to perceived social misalignment), (Klucharev (Klucharev et al. 2009, Zaki, Schirmer, and Mitchell 2011, Izuma 2013).

What comes next?

In the following chapter, I will describe a more detailed theoretical background, specific hypotheses, and methodology specific to Task 1 (Chapter 2), Task 2 (Chapter 3), and Task 3 (Chapter 4), along with their corresponding results and discussion. In Chapter 5, I will summarize the main findings of this dissertation project and discuss the implication of this research in anthropology and other neighboring disciplines.

Chapter 2

A common oxytocin receptor gene (*OXTR*) polymorphism modulates neural responses to negative facial micro-expressions

Chapter Abstract

In this chapter, I investigated whether the *OXTR* variant rs53576 is associated with people's ability to correctly detect the valence of rapidly presented, dynamic facial expressions of emotions. In a novel facial emotion detection task, participants (Neuroimaging arm $N = 43$, Behavioral arm $N = 131$) watched a series of video clips showing dynamic facial expressions of happiness, disgust, anger, and neutral states. The positive and negative stimuli also varied in duration and the intensity of the emotional expressions (i.e., micro- vs. macro-expression). In each trial, participants determined the valence of the emotional expression. The BOLD fMRI signals recorded during the face perception and participants' overall behavioral task performance were analyzed with respect to their genotypes. We report four key findings:

1. The accurate perception of facial micro- and macro-expressions recruited the brain areas implicated in dynamic face processing, emotion perception, as well as affective and cognitive empathy.
2. Micro-expressions recruited activations in a much wider network of brain regions than what was found in previous literature, potentially due to the dynamicity of the stimuli.
3. G homozygotes showed increased ability to detect micro-expressions, although this effect was only found in the neuroimaging arm.
4. G homozygotes showed increased BOLD responses to negative micro-expressions in brain regions involved with attentional control: the supramarginal gyrus, STS, AI, and IFG.

These results are in line with the social salience hypothesis of OT and suggest that the genetic variations in OXTR may regulate the early perception of evaluative feedback in our daily social interaction.

Keywords: *OXTR*, micro-expressions, social salience, attentional control

Introduction

Oxytocin (OT) is a neurohypophysial hormone known to modulate social behavior and cognition across a wide range of mammals including humans. One prominent example from humans is the role of OT in facial emotion perception. Intranasal administration of OT (INOT) has been shown to improve recognition of basic emotions such as anger, fear, disgust, and happiness in both healthy and clinical populations (Leppanen et al. 2017, Shahrestani, Kemp, and Guastella 2013). INOT also promotes cognitive and emotional empathy (Bartz et al. 2011, Domes, Heinrichs, Michel, et al. 2007, Hurlemann et al. 2010) (Geng et al. 2018), which can lead to accurate perception of facial emotions (Penton-Voak et al., 2007; Olderbak and Wilhelm, 2017). Supporting these behavioral findings, studies using functional magnetic resonance imaging (fMRI) have shown that INOT modulates activity in face- and emotion-sensitive brain regions such as the amygdala (Domes, Heinrichs, Gläscher, et al. 2007), the anterior cingulate cortex (Luo et al. 2017, Wang et al. 2017), and the ventral striatum (Scheele et al. 2013), even when stimuli are presented subliminally (Luo et al. 2017, Kanat et al. 2015).

The facilitative role of OT in social perception depends on OT receptor expression in the brain, which is regulated by the oxytocin receptor gene (*OXTR*). Notably, naturally occurring genetic variation in *OXTR* is associated with region-specific expression of the OT receptor in the brain (King et al. 2016, Almeida et al. 2022, Reuter et al. 2017). Such alteration of receptor expression is associated with differential social phenotypes across species and within-species individuals. In a recent animal study, for example, a single-nucleotide polymorphism (SNP) in *OXTR* (i.e., NT213739) was shown to predict partner preference in male prairie voles, with the T allele

selectively associated with both increased partner preference and increased receptor expression in the nucleus accumbens, but not in the insula (King et al. 2016).

Taken together, these converging lines of evidence suggest that genetic variation in human *OXTR* and its modulatory effects on endogenous OT signaling in the brain may account for the individual difference in facial emotion perception. The human *OXTR* is located on chromosome 3p25, spans 17 kb, and includes three introns and four axons. Paralleling the animal literature, numerous behavioral and imaging genetics studies have shown that *OXTR* SNPs are associated with various facets of human sociality, including empathy (Gong et al. 2017, Wu, Li, and Su 2012), emotional response to social cues (Choi, Minote, and Watanuki 2017), parenting behaviors (Michalska et al. 2014), personality traits (Connelly et al. 2014, Creswell et al. 2015) and psychiatric disorders (LoParo and Waldman 2015).

Notwithstanding the mounting evidence showing the broad involvement of *OXTR* in human sociality, the possible link between *OXTR* SNPs and facial emotion perception, as well as its underlying neural mechanisms have not yet been thoroughly investigated. Notably, most previous studies that used functional magnetic resonance imaging (fMRI) to examine the effects of *OXTR* SNPs on face perception have used a passive viewing paradigm (Loth et al. 2014) or employed static face images portraying fully-expressed emotions (Kou et al. 2020). Yet, facial expressions in real life are always dynamic, and the specific contents of emotions are often actively repressed or concealed (Ekman and Friesen 1974, Shen et al. 2019). Therefore, the design features employed in previous studies may not adequately reveal the cognitive and neural mechanisms recruited for the perception of subtle dynamic facial expressions of emotion, and how this process could be modulated by *OXTR*.

The present study aims to investigate the role of genetic variation in *OXTR* on the perception of facial expressions of emotion, while addressing the limitations of previous studies. Building upon previous literature on the facilitative effects of INOT on face processing and empathy (Wu, Li, and Su 2012, Domes, Heinrichs, Michel, et al. 2007), we hypothesized that the levels of OT signaling in the brain, indexed by *OXTR* genotypes, would be associated with the ability to correctly detect dynamic facial expressions of emotion. To test this hypothesis, we conducted an imaging genetics study in which participants' genetic information was analyzed in relation to their behavioral and their blood-oxygen-dependent (BOLD) fMRI response to dynamic facial expressions of emotion.

We devised a novel, ecologically valid facial emotion detection task where participants viewed a series of subtle dynamic facial expressions of emotion, called micro-expressions (Shen et al. 2019, Ekman 2009). Micro-expressions refer to the brief facial expressions that reveal emotions a person attempts to conceal (Ekman 2009). They are distinguished from "macro-expressions" which portray stronger facial emotions that usually last for a longer duration (e.g., 500ms or longer) (Shen et al. 2019, Yan et al. 2013). The distinction between micro- vs. macro-expression was initially made to study deception and lie detection (Ekman 2009). Yet, it has gained increasing recognition across multiple disciplines as an ecologically valid framework to study social perception mediated by facial emotions (Shen et al. 2019). That is, it uniquely captures the fact that people may not necessarily "wear their heart on their sleeves" due to various social considerations such as cultural norms of emotional display (Matsumoto, Yoo, and Fontaine 2008), conflict aversion (Chiang et al., 2012), deception (Ekman and Friesen 1974), and impression management (Grandey et al. 2005).

In this study, we specifically focused on the effects of *OXTR* rs53576 on the perception of facial micro-expressions. Rs53576 is one of the most extensively studied *OXTR* SNPs that has been shown to regulate a wide range of social cognition and behaviors relevant for face and emotion processing. For example, the G allele of the SNP has been implicated in enhanced empathy (Wu, Li, and Su 2012), face recognition (Skuse et al. 2014), and physiological responses to positive or negative social cues (Smith et al. 2014, Auer et al. 2015, Michalska et al. 2014). Evidence also suggests that *OXTR* rs53576 exerts a modulatory influence on OT signaling in the brain (Marsh, Yu, et al. 2012, Luo, Ma, et al. 2015, Watanabe et al. 2017), potentially by regulating the OT receptor expression in the brain regions important for social cognition (Almeida et al. 2022). This makes the *OXTR* rs53576 an ideal candidate for studying genetic modulation of endogenous OT effects on facial emotion perception.

Building upon the general hypothesis that *OXTR* will be involved in the perception of subtle facial expression of emotion, we made a series of predictions that concerned 1) the validity of the task (**Prediction 1a, 1b, 2b, 2c**, and **2d**), and 2) the proposed association between allelic variants of rs53576 (i.e., GG vs. AA+AG) with the behavioral task performance as well as its neural correlates (**Predictions 1c, 1d**, **2e**).

Behaviorally, we predict that participants, regardless of their genotypes, will correctly perceive macro- and micro-expression above chance level (**Prediction 1a**). We also predicted that %Hit for the micro-expressions would be generally lower than that for the macro-expressions due to their rapid and subtle presentation (**Prediction 1b**). Next, we predicted that G homozygotes would show enhanced behavioral performance in the facial emotion detection task (**Prediction-1c**). Lastly, such positive association would be more pronounced for micro-expressions (**Prediction-1d**), which, by

definition, would require greater perceptual capacity, and thus be more contingent on the facilitative effects of the G allele.

In terms of neural activity, we predicted that the perception of micro-expression and macro-expression would be associated with enhanced activity in brain regions sensitive to changeable aspects of faces (e.g., posterior superior temporal sulcus, pSTS; Inferior frontal gyrus, IFG) (Dricu and Frühholz 2016) (**Prediction-2a**), and facial emotion processing (e.g., bilateral amygdala, bilateral nucleus accumbens, NAcc; right caudate nucleus, and ventromedial prefrontal cortex, vmPFC) (Fusar-Poli et al. 2009) (**Prediction-2b**). We also predicted that our task would incur activations in the brain areas involved with emotional (e.g., Anterior insula, AI; dorsal anterior cingulate cortex/medial frontal cortex, dACC/MFC) (Fan et al. 2011, Liu et al. 2011) and cognitive empathy (e.g., right temporo-parietal junction, rTPJ; medial prefrontal cortex, mPFC, and precuneus, PC) (Schurz et al. 2014) (**Prediction-2c**). Macro-expressions will more strongly activate these regions compared to micro-expressions (**Prediction-2d**). Finally, we predicted that G homozygotes will show increased activation in these brain regions, especially for micro-expressions (**Prediction-2e**). Our research hypotheses and analysis protocols were pre-registered and available at Open Science Framework: https://osf.io/rpcxv/.

Methods

Participants

A total of 195 healthy adult volunteers were recruited from Emory University and the surrounding community. We excluded those with a history of psychiatric or neurological illness, and those currently taking psychoactive drugs. All eligible participants were randomly assigned to either the neuroimaging (Total $N = 50$, Female $N = 29$) or behavioral arm (Total $N = 145$, Female $N = 90$). The sample size for each condition was determined based on *a priori* power analysis. Details of the power analysis and participant allocation strategy amid the COVID-19 pandemic are provided in **Supplementary materials S2-1**. The demographic characteristics of the final study samples are summarized in **Table 2-1**.

Materials and Procedures

Pre-experiment online survey

Once enrolled, participants visited an online study portal (i.e., Research Electronic Data Capture, REDCap: https://www.project-redcap.org) to complete written informed consent and the pre-experiment questionnaires. All materials and study procedures were approved by Emory University Institutional Review Board (IRB00112525).

***Demographic survey*:** Participants reported their age, sex, and ethnicity. We also collected data on political self-identification and religiosity, which were not used for analysis in this study.

***Psychological questionnaires*:** We measured a broad range of personality traits that have been shown to correlate with various forms of social sensitivity. These variables include trait empathy

(i.e., IRI) (Davis 1983), need for cognition (Need for Cognition Scale, NfC) (Cacioppo and Petty 1982), self-monitoring (i.e., Social Monitoring Scale, SM) (Lennox and Wolfe 1984), and impression management (i.e., Fear of Negative Evaluation, FN) (Leary 1983).

As described in the pre-registered protocol, demographic and psychological variables were used for exploratory analyses, and to test for any baseline differences between genotype groups. Descriptive statistics for the psychological characteristics across the two experimental arms are provided in **Supplementary materials S2-2**.

Saliva sample collection

Participants in the behavioral and neuroimaging condition visited Laboratory for Darwinian Neuroscience and Facility for Education and Research in Neuroscience (FERN), respectively. Upon arrival, participants provided their saliva sample using Oragene DNA self-collection kits (OGR-600, DNA Genotek Inc., Ontario, Canada). Participants were asked to refrain from eating and drinking 30 minutes prior to their visit to ensure the quality of the saliva samples.

Main Task

Following the saliva sample collection, participants performed a novel facial emotion detection task implemented in Psychtoolbox 3 on MATLAB (The MathWorks, Natick, 2015). Participants in the neuroimaging arm performed the task inside an MRI scanner located at FERN. Those in the

behavioral arm completed the identical task in a testing room located at Lab for Darwinian Neuroscience.

Facial emotion detection task: A schematic representation of the facial emotion detection task is presented in **Figure 2-1**. Each trial started with a fixation cross that lasted for a jittered interval (1000-5000ms). Participants were then presented with a video clip (3000ms) showing a face with a dynamic expression of emotions or neutral state. Emotional faces displayed one of three affective states: happiness, disgust, or anger (See "Experimental Stimuli" below). On each trial, participants were asked to determine the valence of the facial emotions (i.e., Positive, Neutral, and Negative), and to rate the perceived intensity of the emotional expression on a 5-point likert scale (1=Very subtle, 5=Very strong). The intensity judgment was omitted if participants selected "Neutral" during the initial valence judgment. The chosen option was highlighted in red for 500ms, and the trial ended with a fixation point. Participants completed a total of 48 trials. Participants also performed two unrelated tasks in addition to the facial emotion detection task in a counter-balanced order. The description of these tasks is provided in full at: https://osf.io/rpcxv. At the end of the experiment, participants received $40 (Behavioral condition) or $50 (Neuroimaging condition) as compensation.

Experimental stimuli: All experimental stimuli were created with the CAS(ME)2 database (Qu et al., 2016). The database includes video recordings of 250 spontaneous dynamic macro-expression and 53 micro-expressions obtained from 22 Chinese adults (Qu et al., 2016). Following previous literature, the cut-off point of 500ms was used to define Macro- vs. Micro-expression (Yan et al. 2013). All facial expressions included in the CAS(ME)^2 database were naturally induced by having participants watch a series of emotion-laden videos, and later annotated using the Facial

Action Coding System (FACS) (Ekman, 1997). In addition, face models themselves also provided subjective reports of affective states experienced during the emotion-inducing procedures. To ensure the validity of the emotional expressions, we first selected a set of experimental stimuli (*N* = 70) from the original CAS(ME)^2 videos for which the FACS-based annotation and the face models' subjective reports matched one another. The selected videos specifically depicted three basic emotions that convey evaluative social feedback (i.e., Anger, Disgust, and Happiness) (Hutcherson and Gross 2011). We ran a separate pilot study with an independent group of participants (*N* = 28). As in the main task, respondents were asked to determine the valence (i.e., Positive vs. Negative vs. Neutral) of the sample emotional expressions. The stimuli with the %Hit below chance level (i.e., 33%) were excluded. Based on the pilot study, a total of 48 videos (i.e., Micro-expression *N* = 18 Macro-expression *N* = 18, and Neutral expression *N*=12) from four female and four male face models were selected as the final stimuli. Specific characteristics of the stimuli for each expression category are summarized in **Supplementary materials S2-3**.

Data Collection and Analysis

Genetic data acquisition: Participants' DNA were extracted from saliva samples. The rs53576 genotypes was determined by Axiom™ Precision Medicine Research Array (Affymetrix) and TaqMan SNP Genotyping Assays using a ViiA7 Real Time PCR System for genotype resolution (Applied Biosystems, Foster City, CA). For quality control in SNP genotyping, each 384 well genotyping plate contained multiple duplicate wells and positive and negative controls. 106 Ancestry-Informative markers were used to account for potential population stratification. These markers discriminated European, African, East Asian, and Native American origins. We used a

structure software (Pritchard, Stephens, and Donnelly 2000) to estimate proportions of chromosomal ancestry based on K (the number of source populations). Principal components analysis (PCA) was calculated to account for population stratification. The first two principal components from this analysis were used in the analyses as covariates to control for population stratification.

***Neuroimaging data acquisition*:** All neuroimaging data were acquired using a 3-Tesla Siemens MAGNETOM Prisma MRI scanner. T1-weighted anatomical images were obtained using a 3D magnetization-prepared rapid gradient-echo (MPRAGE) sequence with a Generalized auto-calibrating partial parallel acquisition (GRAPPA) factor of 3. The T1 scan protocol, optimized for 3 Tesla, used the following imaging parameters: the repetition time (TR) = 1900ms, inversion time (TI) = 900ms and echo time (TE) = 2.27ms, a flip angle of 9˚, a volume of view of 256×256×176 mm^3, a matrix of 256×256×176, and isotropic spatial resolution of 1.0×1.0×1.0 mm^3. FMRI data were acquired using an Echo-Planar Imaging (EPI) sequence for blood-oxygen-level-dependent (BOLD) fMRI. EPI images were collected in an interleaved fashion with the following parameters: TR = 1,200ms, TE = 30ms, matrix = 74*74, Field of View = 220mm, isotropic in-plane resolution = 3.0 mm, slice thickness = 3.0 mm, 54 axial slices with no gap in between and no phase oversampling.

Behavioral data analysis

All raw data were processed and analyzed with MATLAB R2015b and SPSS version 28 (Armonk, NY: IBM Corp). The analyses described below applied to the behavioral data obtained from either the behavioral arm or neuroimaging arm.

Testing the association among *OXTR* genotype, personality traits, and task performance: A series of independent sample *t*-tests were performed to explore any baseline difference in personality traits (i.e., Empathy, FNE, NfC, and SM) across G homozygotes (i.e., GG) and the A allele carriers (i.e., AA/AG).

Comparing the average task performance against chance level (Prediction 1a): We computed the overall proportion of trials where participants correctly identified the valence of emotional expressions (i.e., Global %Hit). To examine if participants' performance exceeded chance level, we used a series of one-sample *t*-tests comparing Global %Hit as well as the average hit rates calculated for each expression category (i.e., micro-, macro-, and neutral expressions) with chance level (i.e., 1/three choice options = 0.33).

Testing the effect of *OXTR* on the average task performance (Prediction 1b-1c): As described in the pre-registered protocol, the relationship between *OXTR* SNPs and detection accuracy (**Prediction 1-c, and 1-d**) was tested in a 2 (*OXTR* Genotype: GG vs. AA/AG) × 2 (Expression Type: Macro- vs. Micro) × 2 (Valence: Positive vs. Negative) repeated measure analysis of variance (RMANOVA) on %Hit for each expression category. This model was also used to examine if participants showed higher task performance for the macro-expressions compared to micro-expressions (**Prediction-1b**). We did not directly analyze the task performance for the neutral expressions due to the ceiling effect, with the average %Hit higher than 90.

Building upon this baseline model, we extended our analysis with additional statistical terms modelling the main effects of demographic variables that have previously been shown to influence face perception such as age and sex (Olderbak et al. 2019), as well as ethnicity (Meissner and Brigham 2001). The effect of Genotype × Sex interaction was also included to explore possible sex-dependent effects of OT (Rilling et al. 2014). Participants' ethnicity was modelled using on the first two principal components yielded from genotyping process. A significant main effect of the *OXTR* genotype, and an *OXTR* genotype × Expression Type interaction in these RMANOVA or RMANCOVA models would support **Prediction 1-c** and **1-d**, respectively.

For the RMNOVA and RMACOVA models, Bonferroni correction was applied to *post-hoc* pairwise comparisons for any significant main effect or interaction effect (Two-tailed $\alpha = .05$). Greenhouse-Geisser correction was applied to adjust degrees of freedom and *p*-values in case of violation of sphericity. We note that the results of data exploration, including the analyses on personality traits and mood were not corrected for multiple comparisons.

Neuroimaging data analysis

Neuroimaging data analyses were performed with the Oxford Center for Functional Magnetic Resonance Imaging of the Brain's software library (FSL v6.0.3, http://www.fmrib.ox.ac.uk/fsl/).

Preprocessing: Our preprocessing pipeline included 1) motion correction using MCFLIRT (Jenkinson et al. 2002), 2) skull-stripping using FSL's Brain Extraction Tool (BET), 3) slice timing

correction, 4) high-pass temporal filtering with a filter width of 100s, 5) spatial smoothing using a Gaussian kernel of full width at half maximum (FWHM) of 6mm, 6) spatial registration of fMRI images to high-resolution T1 images (i.e., Boundary-Based-Registration), and 7) spatial normalization to the standard Montreal Neurological Institute (MNI) 2 mm brain using FLIRT (Greve and Fischl 2009) with affine transformation (i.e., 12 degrees of freedom). EPI distortion correction based on field map images was not performed, although it was part of the pre-registered protocol, as the algorithm did not consistently yield optimal registration across participants.

1st level analysis: We convolved each trial with a double-gamma hemodynamic response function (HRF) in FSL. The main univariate GLM included seven explanatory variables (EVs) and their temporal derivatives corresponding to the following events: Presentation epochs for the positive and negative micro-expressions (i.e., $Micro_{negative}$, $Micro_{positive}$), macro-expressions (i.e., $Macro_{negative,}$ $Macro_{positive}$), neutral expressions (i.e., Neu). We also modelled the valence judgment epoch (i.e., Valence), and intensity judgment epoch (i.e., Mag) as events of no interest. Six additional EVs were added to account for head motion. The activations associated with the fixation point was not modelled and served as an implicit baseline.

2nd level analysis: For group-level analyses, we ran a series of mixed-effect analyses (i.e., FLAME1) using fMRI Experts Analysis Tool (FEAT) in FSL. As our main hypotheses centered around the effect of Expression Type, we computed the following contrasts and their reverse contrasts: 1) [$Micro_{negative+positive}$+$Macro_{negative+positive}$] > Neutral, 2) $Macro_{negative+positive}$ > $Micro_{negative+positive}$, 3) $Macro_{negative}$ > $Micro_{negative,}$ 4) $Macro_{positive}$ > $Micro_{positive}$. The first two eigenvectors from the principal components analysis were mean centered and included as covariates of no interest to capture the ethnic variation of the study sample. The contrast estimates

obtained from the second level GLM were thresholded as below to test **Prediction 2-a** through **Prediction 2-e**.

Search volume and thresholding: Both whole-brain analyses and exploratory ROI analyses were used to test our predictions. **Predictions 2-a, 2-b**, and **2-c** were examined with one sample *t*-tests comparing the contrast estimates calculated from the main GLM with zero. The effects of *OXTR* (**Prediction 2-d** and **2-e**) were tested by splitting the study sample based on participants' genotype (i.e., rs53576 GG vs. AA/AG) and comparing the genotype groups with two-sample *t*-tests.

For the whole-brain analyses, the *Z*-statistic (i.e., Gaussianized *t*) images were thresholded using a cluster-defining threshold of $Z > 3.1$(i.e., voxel-wise one-tailed $p < .001$) and a family-wise error (FWE)-corrected cluster-level significance threshold of $p < .05$ (one-tailed).

For the exploratory ROI analyses, we examined a total of 11 brain regions implicated in dynamic face processing and emotion perception (i.e., posterior superior temporal sulcus, pSTS; Inferior frontal gyrus, IFG, bilateral amygdala; NAcc, caudate nucleus, and ventromedial prefrontal cortex, vmPFC), as well as affective empathy (i.e., Anterior insula, AI; dorsal anterior cingulate cortex/medial frontal cortex, dACC/MFC), and cognitive empathy (i.e., right temporo-parietal junction, rTPJ; medial prefrontal cortex, mPFC, and precuneus, PC). The specific coordinates of the target ROIs and sizes were obtained from previous activation-likelihood estimate (ALE) meta-analyses on explicit and implicit evaluation of facial expressions (Dricu and Frühholz 2016), positive and negative facial emotion (Fusar-Poli et al. 2009), perceptual-affective empathy (Fan et

al. 2011), cognitive-empathy/Theory of Mind (ToM) (Schurz et al. 2014) and reward processing (Liu et al. 2011).

For each ROI, voxel-wise, one-sample *t*-tests were used to determine if the group-level contrast estimates were significantly different from zero. The Z-statistic images were thresholded at $p < 0.05$ (one-tailed), corrected for multiple comparisons across all ROI voxels (i.e., Small Volume Correction, SVC) (Poldrack 2007). Only activations including 5 or more voxels were reported and considered for follow-up analyses. We used *Featquery* to extract and visualize the BOLD responses (e.g., %signal change) for any significant fMRI results.

Results

Behavioral arm

Sample characteristics: Data from eight participants were not included in the analysis due to technical errors or failure to understand the task instruction. Data from six participants were further excluded due to error in genotyping and missing ancestry data. The final study sample size was $N = 131$. There was no significant difference in personality traits (All *P*s > .075), demographic characteristics (All *Ps* > .754), and sex ratio (p=.551) between the two genotype groups.

The average task performance of participants (Prediction 1a): The average %Global Hit ($M = 74.67$, SD = 9.05) was significantly above chance (i.e., 33%), $t_{(130)} = 52.69$, $p < .001$, Cohen's $d = 9.05$. The average hit rate for Macro- ($M = 83.03$, $SD = 11.63$), Micro- ($M = 55.22$, $SD = 15.45$)

and Neutral (M = 91.28, SD = 10.4) expressions also exceeded the chance level performance (All Ps <.001).

The association between *OXTR* and behavioral task performance (Prediction 1b, 1c): The 2 (*OXTR* genotype) x 2 (Expression Type) x 2 (Valence) RMANOVA on the average hit rates revealed a significant main effect of Expression Type, $F_{(1, 128)} = 232.23$, $p < .001$, $\eta_p^2 = .645$, and Valence, $F_{(1, 128)} = 129.57$, $p < .001$, $\eta_p^2 = .503$. Post-hoc pairwise comparisons for the main effect of Expression Type and Valence revealed that the average %Hit was higher for macro-expressions vs. micro-expressions ($p < .001$), and for positive expressions vs. negative expressions ($p < .001$), respectively **Figure 2-2.** There was also a significant interaction between Expression Type and Valence, $F_{(1, 128)} = 241.84$, $p < .001$, $\eta_p^2 = .654$. This interaction was driven by a significant difference in %Hit for negative macro-expression vs. micro-expressions ($p < .001$). Participants performed equally well for positively valenced macro- and micro-expressions ($p = .602$). An RMANCOVA including the demographic variables revealed a significant interaction between sex and valence, $F_{(1, 128)} = 4.55$, $p = .035$, $\eta_p^2 = .038$, with female participants showing higher %Hit for negative macro- and micro-expression (M = 66.06, SD = 12.7) compared to male participants (M = 57.6, SD = 17.8) (p =.013). We found no evidence of significant genetic modulation of the task performance in either of the models (All $Ps > .089$). The *OXTR* genotype did not affect participants' intensity judgments (All Ps >.177) or reaction time data either (All Ps >.219).

Neuroimaging arm

Sample characteristics: Data from seven participants were excluded from the analyses due to technical errors during the task, or genotyping failure. The final sample size was $N = 43$. There was no significant difference in personality traits (All *P*s > .205), demographic characteristics (All *Ps* > .089), or sex ratio (p=.531) between the two genotype groups.

Behavioral Results

The average task performance of participants (Prediction 1a): The average %Global Hit (M = 63.84, SD = 11.16) was significantly above chance (i.e., 33%), $t_{(44)}$=18.541, p<.001, Cohen's d=2.76. The average hit rate for Macro- (M = 72.1, SD = 16.13), Micro- (M = 40.7, SD = 14.6) and Neutral (M= 86.1, SD = 14.2) expressions also exceeded the chance level performance (All *P*s <.001).

The association between *OXTR* and behavioral task performance (Prediction 1b, 1c): The 2 (*OXTR* genotype) x 2 (Expression Type) x 2 (Valence) RMANOVA on the average hit rates revealed a significant main effect of Expression Type, $F_{(1, 41)} = 182.97$, $p < .001$, η_p^2=.817, and Valence, $F_{(1, 41)} = 82.79$, $p < .001$, $\eta_p^2 = .669$, as well as a significant interaction between these two factors, $F_{(1, 41)} = 38.91$, $p < .001$, $\eta_p^2 = .487$. Post-hoc pairwise comparisons showed that the average %Hit was higher for macro-expressions vs. micro-expressions ($p < .001$), and, also for positive expressions vs. negative expressions ($p < .001$), as in the results of the behavioral arm. the differential %Hit for macro- vs. micro-expressions was more pronounced for negative expressions (p<.004) than for positive expressions (p<.001) **(Table 2-2)** With respect to the predicted effect of

OXTR, we did not find significant effect of the *OXTR* genotype (All *Ps* > .07). The *OXTR* genotype did not affect participants' intensity judgments (All *P*s >.602) or RTs (All *P*s >.079).

Notably, the inclusion of demographic variables (i.e., sex, age, and ethnicity) revealed a significant interaction between *OXTR* genotype and Expression Type, $F_{(1, 36)} = 4.691$, $p = .037$, $\eta_p^2 = .115$. Our Bonferroni-corrected, post-hoc analyses showed that this result was driven by the G homozygotes who performed better than A allele carriers for the micro-expressions (Mean difference: 13.43, p=.018), but not for the macro-expressions (Mean difference: 3.02, p=.565) **(Table 2-2).** We also found a significant main effect of age, $F_{(1, 36)} = 4.987$, $p = .032$, $\eta_p^2 = .122$, with older individuals showing a decreased Global %Hit **(Supplementary materials S2-4).** Participants' sex and ethnicity did not significantly modulate the task performance (All *Ps* >.155).

Overall, although **Prediction 1a** and **1b** were supported, our baseline RMANOVA model did not reveal conclusive evidence supporting **Prediction 1-c** or **1-d**. Yet, after adjusting for the demographic variables, the predicted main effect of *OXTR* genotype as well as the interaction between the *OXTR* genotype × Expression Type became significant, which held up even when the personality traits were further controlled for.

Neuroimaging Results

Testing the Main Predictions

Whole-brain analysis

The effect of the Expression Type (Prediction 2a-2c): The contrast between macro- and micro-expressions vs. neutral expressions (i.e., [Macro+Micro > Neutral]) yielded widespread activation in brain areas involved in dynamic face processing (**Prediction 2a**), emotion perception (**Prediction 2b**) as well as affective and cognitive empathy (**Prediction 2c**). Specifically, significant clusters were found in the bilateral inferior frontal gyrus (IFG), bilateral occipital-temporal cortex (LOC), the posterior superior temporal sulcus (pSTS), the left superamarginal gyrus (SMG), middle temporal gyrus (MTG), temporal pole (TP), lateral orbitofrontal cortex (lOFC), right anterior insula (AI), left supramarginal gyrus, and medial frontal cortex/dorsal anterior cingulate cortex (MFC/dACC). The reverse contrast revealed one significant cluster in the frontal pole (**Figure 2-3, Table 2-3**). The direct contrast [$Macro_{negative+positive}$ > $Micro_{negative+positive}$] did not yield significant clusters in either direction. When the comparison between macro-expression vs. micro-expressions was made more specifically within the negatively valenced stimuli (i.e., $Macro_{negative}$ > $Micro_{negative}$), a subset of brain areas identified from the contrast [Macro+Micro > Neutral] emerged again. That is, negative macro-expressions incurred stronger activations in the IFG, AI, dACC, LOC, and TP. In contrast, the reverse contrast revealed only one cluster in the primary visual cortex **(Table 2-3).** The contrast [$Macro_{positive}$ > $Micro_{positive}$] yielded stronger activations in the superior parietal lobule, precuneus, and middle frontal gyrus. The reverse contrast did not show any significant activations **(Table 2-3)**.

The interaction between *OXTR* and Expression Type (Prediction 2e): We found a significant genetic modulation in left supramarginal gyrus (SMG), with the A allele carriers showing greater contrast estimates for [$Macro_{negative}$ > $Micro_{negative}$]. This interaction was driven by G homozygotes who showed relatively higher activations within the region for negative micro-expressions,

compared to the A allele carriers **(Figure 2-4, Table 2-3).** The effect of *OXTR* was not found in other contrasts of interest.

ROI analysis

The effect of the Expression Type (Prediction 2a-2c): The results of the exploratory ROI analysis overlapped mostly with the findings from the whole-brain analyses. The contrast [Macro + Micro > Neutral] revealed significant voxels within the ROIs implicated in dynamic face processing (i.e., STS and IFG), facial emotion (i.e., amygdala), emotional empathy (i.e., the AI, and MFC/dACC), and cognitive empathy (i.e., rTPJ). The direct comparison between macro- vs. micro-expressions showed that the former incurred stronger activations in all ROIs identified for the contrast [Macro + Micro > Neutral], except for the rTPJ, which did not show differential activations across the expression categories. The reverse contrast (i.e., Micro > Macro) did not yield any significant voxels, indicating that no ROI was uniquely involved in the perception of the facial micro-expressions. The contrast between [$Macro_{negative}$ > $Micro_{negative}$] and [$Macro_{positive}$ > $Micro_{positive}$] revealed that both positive and negative macro-expressions incurred the greater activations in the dACC and mPFC compared to positive and negative micro-expressions. Again, no ROIs showed increased activations for micro-expressions regardless of valence. The full results of the ROI analyses concerning the effect of Expression Type are reported in **Supplementary materials S2-6**.

***OXTR* rs53576 and the perception of micro vs. macro-expressions (Prediction 2e):** Consistent with the results of the whole-brain analysis, we found significant genetic modulation for the contrast [$Macro_{negative}$ > $Micro_{negative}$] in multiple ROIs including the right STS, bilateral AI, and

IFG. These interaction effects were driven by G homozygotes showing equally strong activations for the negatively valenced macro-expressions and micro-expressions. The A allele carriers, in comparison, consistently exhibited lower activations for the negative valenced micro-expressions relative to macro-expressions (**Figure 2-5a-c**).

Discussion

A plethora of evidence shows that exogenous oxytocin administration enhances face and emotion perception in humans. However, it has remained unclear if such facilitatory effects of oxytocin (OT) would also be linked with individual differences in the endogenous OT signaling, which is regulated by allelic variations in a single oxytocin receptor gene (*OXTR*). This study investigated whether and how a single nucleotide polymorphism (SNP) in *OXTR* modulates the perception of dynamic facial micro-expressions.

Our first behavioral prediction **(Prediction 1a)** was supported. Overall, participants in the neuroimaging and behavioral arm successfully discriminated the valence of macro- and micro-expressions above chance level. Also, consistent with **Prediction 1b**, the average %Hit for the macro-expressions was significantly higher than that for the micro-expression, which is consistent with the previous findings that people typically find micro-expressions more challenging to recognize (Ekman 2009). We also found a negative relationship between participants' age and task performance among participants in the neuroimaging arm. A similar pattern emerged in data obtained from the behavioral arm, although the effect did not reach statistical significance. This

result replicated a previous finding on the background demographic variable that affects the perception of micro-expression (Hurley et al. 2014).

Notably, the %Hit for the positive micro-expressions (i.e., happiness) was higher than %Hit for the negative micro-expressions (i.e., disgust and anger). In fact, evidence indicates that facial expressions of happiness tend to be recognized much faster (Leppänen and Hietanen 2004) and more accurately (Esteves and Öhman 1993) as opposed to negatively valenced emotions such as anger, sadness, and disgust. The exact sources of such a processing advantage for positive facial expressions are not fully understood (Leppänen and Hietanen 2004). Still, this overlap suggests that our experimental setups successfully recruited the face processing mechanisms in a way consistent with what has been reported in the literature, despite the novel stimuli and the task structure used in the study.

Previous studies based on exogenous OT treatment (INOT) have strongly suggested the role of OT in face and emotion perception (Shahrestani, Kemp, and Guastella 2013). Given that the G allele of *OXTR* rs53576 has been repeatedly shown to enhance the efficacy of INOT (Watanabe et al. 2017) and facilitate the effects of OT on social affiliation and social salience at the level of brain activation and behavior (Marsh, Yu, et al. 2012, Michalska et al. 2014, Luo, Ma, et al. 2015, Watanabe et al. 2017), we hypothesized that the G allele carriers would show increased ability to perceive facial macro- and micro-expressions.

We did not find conclusive evidence that supports the predicted relationship between the *OXTR* genotype and behavioral task performance. In terms of overall task performance (i.e., Global %Hit), G homozygotes were not significantly different from the A allele carriers (**Prediction 1c**). Yet,

they showed higher average %Hit for micro-expression **(Prediction 1d).** Consistent with the previous study showing that OT may not necessarily lead to a valence-specific enhancement in visual perception (Domes, Heinrichs, Gläscher, et al. 2007, Guastella, Mitchell, and Dadds 2008), we did not find significant effects of the stimuli valence regardless of the expression categories. Overall, our results are in line with previous findings that the G allele of the rs53576 may have facilitatory effects on social cognition, including face perception and emotion recognition (Skuse et al. 2014, Lucht et al. 2013, Stanković et al. 2019). This study could extend the existing body of research by showing that the conducive effects of the rs53576 G allele could also be found in the perception of dynamic facial micro-expressions, which are thought to closely resemble real-life social communication. However, it is important to point out that the link between the *OXTR* genotype and %Hit rate was not replicated in the larger behavioral sample. Therefore, our results concerning the behavioral effects of the *OXTR* rs53576 should not be overinterpreted until further replication is made.

Consistent with the **Prediction 2a through 2c**, the average effects of the macro- and micro-expressions (i.e., [Macro+Micro] > Neutral) were represented in the brain regions previously implicated in 1) dynamic face processing (e.g., STS), 2) emotion perception (e.g., amygdala), and 3) affective and cognitive empathy (e.g., dACC, AI, rTPJ, and mPFC). Largely overlapping regions of activations were found when Macro- and Micro- expressions were compared separately with Neutral expressions **(Supplementary materials S2-5).** Such broad recruitment of functionally separable brain areas has been commonly found in experiments using various face perception tasks (Dricu and Frühholz 2016). Notably, however, it runs counter to the findings from some previous studies where the perception of facial micro-expression incurred the stronger BOLD signals or ERP components (e.g., N170) in relatively restricted brain regions in the parietal lobe (e.g., Inferior

parietal lobule) and fusiform gyrus, respectively (Zhang et al. 2020, Zhang et al. 2018). We suspect that this discrepancy may reflect the heterogeneity in the experimental stimuli that have been employed in the related literature to model the micro- and macro-expressions. Previous neuroimaging studies on micro-expression perception have relied on "synthesized micro-expressions," which were created by embedding a static, fully-expressed emotional face between a continuous visual stream of neutral faces (Shen et al., 2012 (Shen, Wu, and Fu 2012, Zhang et al. 2014, Zhang et al. 2020). The presentation duration of the emotional faces was typically below 100ms (e.g., 60ms) (Zhang et al. 2020). By contrast, participants in this study viewed naturally induced, dynamic emotional expressions, with the average lengths of the micro- and macro-expressions being approximately 350ms and 1400ms, respectively. It is possible that the relatively *longer* presentation of emotions and the richness of the visual information conveyed in the *dynamic* facial expressions may have contributed to the widespread activations observed in this study.

Here, one important conceptual question may be raised as to the relative validity of the synthesized- vs. naturally induced emotional expressions for studying the neurocognitive mechanisms subserving the perception of micro-expressions in humans. It is beyond the purpose of this study to pit one approach against another. Yet, evidence suggests that the limit of the facial skeletal muscle contraction time is around 100ms (Ito, Murano, and Gomi 2004). Accordingly, it has been proposed that researchers should avoid using facial micro-expression stimuli shorter than this duration for achieving maximum ecological validity (Yan et al. 2013). Therefore, while the human brain can detect emotional faces presented subliminally (e.g., 13ms) (Whalen et al. 2004), using the synthesized micro-expressions may fall short of revealing the neural activations that reflect the full temporal dynamics and visual features of micro-expressions in real life.

One related issue is whether there are any neural signatures uniquely associated with the perception of the micro-expressions vs. macro-expressions. In this study, direct contrasts between the macro- and micro-expressions (i.e., Macro vs. Micro) revealed no suprathreshold activations preferentially associated with the micro-expressions. This result persisted when a more focused comparison was made between either negatively or positively valenced macro- and micro-expressions. Our data suggest common neural mechanisms underlying the perception of both types of facial expressions. This is in line with the existing evidence the subliminal vs. supraliminal presentation of *static* emotional expressions recruited a largely overlapping network of brain regions, including the STS, IPS, IFG, FFA, and LO (Prochnow et al. 2013). It should be noted that fMRI is not best suited for analyzing the early brain activations associated with the rapidly occurring visual events (Kable 2011). While the macro- and micro-expressions may incur similar spatial patterns of BOLD responses, the specific temporal fluctuations of the neural activities within the implicated brain areas may still be different. For instance, many electrophysiological studies using electroencephalogram (EEG) have isolated the early event-related potential (ERP) components (e.g., N170) that are associated with the perception of facial expressions with short vs. longer durations (Shen, Wu, and Fu 2012). Studies using multi-modal imaging techniques (e.g., EEG-MRI) would be a promising approach to fully uncover the temporal and spatial patterns of the neural activations that support the accurate perception of macro- vs. micro-expressions (Müller-Bardorff et al. 2018).

At the whole-brain level, the *OXTR* genotype significantly modulated activation in the left supramarginal gyrus (SMG) in response to the negatively valenced macro-expressions vs. micro-expressions. This effect was driven by G homozygotes, who showed greater activation for the negative micro-expressions compared to the A allele carriers. Our finding is not strictly consistent

with **Prediction 2d,** as the SMG is not typically considered part of the neural systems directly involved with dynamic face processing, emotion perception, or various facets of empathy. Rather, it points to the possibility that the genetic variations in the *OXTR* may contribute to the perception of facial micro-expressions by modulating more basic cognitive processes that subserve these high-level functions.

Specifically, the supramarginal gyrus has been widely implicated in both voluntary and stimulus-driven shifts of visual attention (Corbetta, Patel, and Shulman 2008). Previous evidence has also shown the role of the supramarginal gyrus in orienting attention to affective information during early visual processing, most notably positive and negative facial expressions (Narumoto et al. 2001). Considering these findings, the observed intergroup differences in the SMG may reflect the increased allocation of attentional resources to subtle emotional cues among G homozygotes, relative to the A allele carriers.

Our exploratory ROI analyses also revealed the identical patterns of genetic modulations in multiple brain areas, including the superior temporal sulcus (STS) and bilateral anterior insula (AI). In the inferior frontal gyrus (IFG), the A allele carriers showed elevated average activations regardless of the expression type than did G homozygotes. Yet, G homozygotes showed the equally strong levels of IFG activations in responses to both macro- and micro negative expressions, while the A allele carriers showed significantly smaller activations for the negative micro-expressions.

The AI and IFG have frequently been identified in experiments measuring affective empathy and mirroring (Singer et al. 2004, Budell, Jackson, and Rainville 2010), which can aid the accurate identification of emotions. Given that these brain regions were identified in the contrasts involving

negative emotions in this study, it is possible that increased activation within these areas may reflect participants' empathic responses to subtle emotional cues. Yet, an exploratory correlation analysis showed that the activation in neither of the areas showed associations with the self-report measure of empathy (e.g., empathic concern) (All Ps >.735). Also, G homozygotes and the A allele carriers in this study did not significantly differ in trait empathy. These results suggest that the BOLD responses in these ROIs may reflect more general cognitive processes that subserve affective empathy, such as salience encoding and attentional reorientation.

In fact, the AI and IFG have been widely implicated in attentional processes in the human brain. The AI, along with dACC, is a critical node in the salience network that monitors external events and guides adaptive behaviors (Menon and Uddin, 2010; Seeley et al., 2007). Unlike the dACC, which is more directly involved with conflict processing and response selection, the AI is thought to be implicated in the neural encoding of perceptually salient sensory cues (Menon and Uddin 2010). Similarly, the IFG comprises a ventral frontoparietal attention network that reorients visual and auditory attention to salient and behaviorally relevant stimuli in the environment, especially when such targets are unexpected (Corbetta, Patel, and Shulman 2008). Notably, OT-induced increase in IFG and AI among healthy and clinical subjects have been reported in the face processing tasks that did not involve affective mirroring or emotional empathy (Gordon et al. 2013, Domes, Heinrichs, et al. 2013). These findings add to the possibility that the increased activations in the AI and IFG in response to negative micro-expression may reflect the enhanced salience encoding and attentional functions among G homozygotes compared to the A allele carriers.

Our interpretation of the AI and IFG activations is further supported by the fact that the G homozygotes showed greater average activations in the right STS in response to negative micro-

expressions compared to the A allele carriers. As part of the core neural system for face processing, the role of the STS in facial expressions and gaze processing has been extensively studied and documented (Haxby, Hoffman, and Gobbini 2000). Of relevance to the finding in this study, selective visual attention to facial emotion has been shown to modulate the BOLD responses within the STS more strongly than the more general deployment of attention towards faces per se (Narumoto et al. 2001). In the same vein, repeated transcranial magnetic stimulation (TMS) delivered either on the left or right STS has been shown to undermine emotion recognition (Sliwinska and Pitcher 2018). Such recruitment of the STS during the perception of facial expression is known to occur as early as 150-200ms after the presentation of relevant stimuli (Sato et al. 2008). These findings suggest that the stronger activation in the STS among G homozygotes in response to facial micro-expressions may represent the more effective allocation of visual attention to subtle, rapidly presented emotional cues.

In all, our imaging genetics analyses suggest that *OXTR* rs53576 G homozygotes showed relatively stronger activations for the negative micro-expressions compared to the A allele carriers in the multiple brain areas implicated in salience encoding and attentional control. Our findings are consistent with the social salience hypothesis of OT, which suggests that OT enhances visual attention to socially relevant stimuli, thereby allowing an organism to better execute adaptive behavioral responses in a given environment ((Shamay-Tsoory and Abu-Akel 2016, Quintana and Guastella 2020, Ma et al. 2016). Although tentative, our behavioral findings showed that G homozygotes were more proficient at detecting facial micro-expressions. While there was no statistically significant interaction between the *OXTR* genotype, Expression Type and Valence, our exploratory analysis revealed that the genetic modulation of the %Hit for the micro-expressions was more pronounced for the negative micro-expressions **(Supplementary materials S2-7).**

Therefore, it is possible that the allelic variations in *OXTR* rs53576 may regulate the perception of facial micro-expressions that convey subtle, yet motivationally important emotional cues primarily through the neural encoding of salience and attentional control in the brain.

Chapter Summary and Conclusion

Accurate decoding of facial expressions of emotion is pivotal to human social communication. We investigated whether a common polymorphism in the *OXTR* rs53576 can modulate our ability to perceive rapidly presented, dynamic emotional cues, namely, facial micro-expressions. Despite the novel experimental paradigm and stimuli used in this study, we generally replicated the patterns of behavioral responses to the macro- and micro-expressions observed in previous research. We also broadened the existing literature by showing that the accurate perception of macro- and micro-expressions may be mediated by neural activations in a much broader, functionally dissociable network of brain regions than what has previously been identified. Most importantly, our imaging genetics analyses suggest a potential contribution of *OXTR* rs53576 to the more efficient recruitment of attentional mechanisms to process emotionally salient social cues. While this result adds to the known role of *OXTR* SNPs in the perception and recognition of static faces and emotional expressions, future studies involving larger samples would be necessary to firmly establish the behavioral relevance of the observed patterns of neuromodulation associated with the *OXTR* genotypes.

Figures and Tables

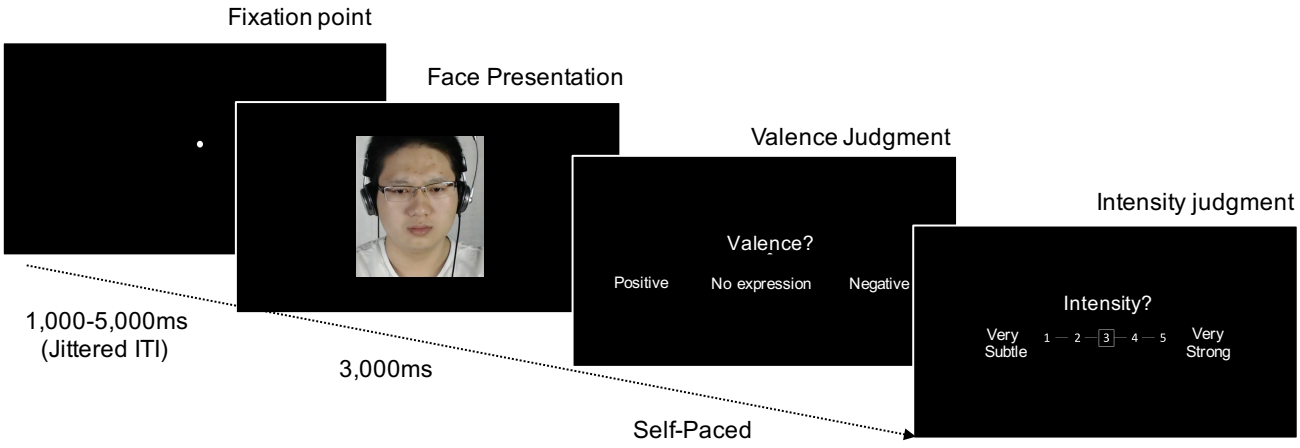


Figure 2-1. The trial structure of the face emotion detection task.

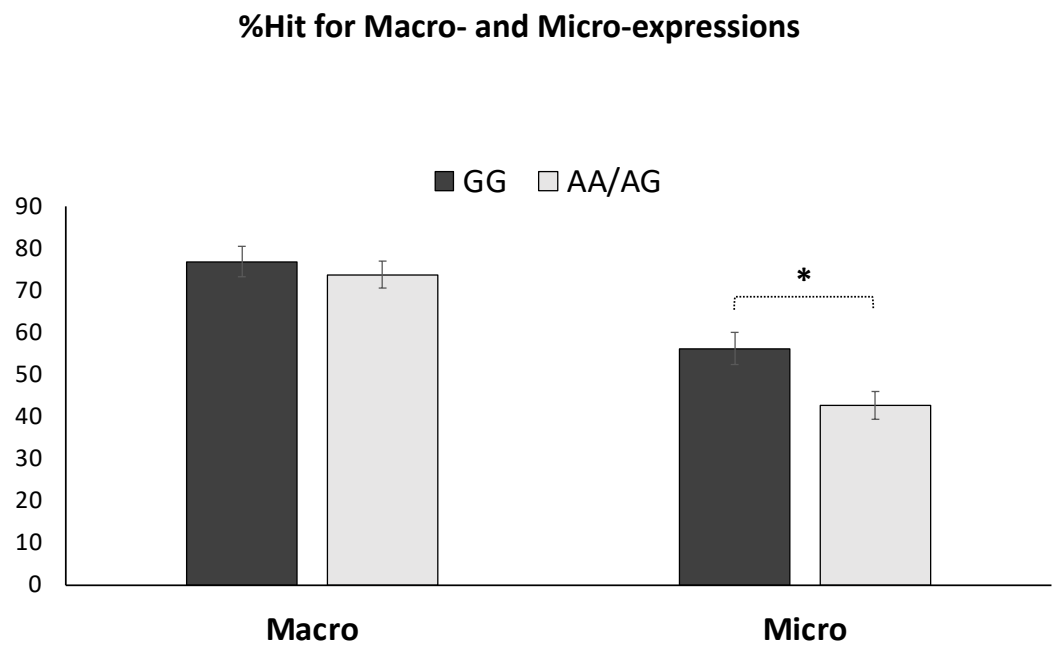


Figure 2-2. The average %Hit for the macro vs. micro-expressions by the *OXTR* rs53576 genotype. On average, G homozygotes showed significantly higher task performance for the micro-expressions, controlling for the effect of demographic variables such as age, sex, and ethnicity *p <.01). The error bars denote 95% confidence interval (CI).

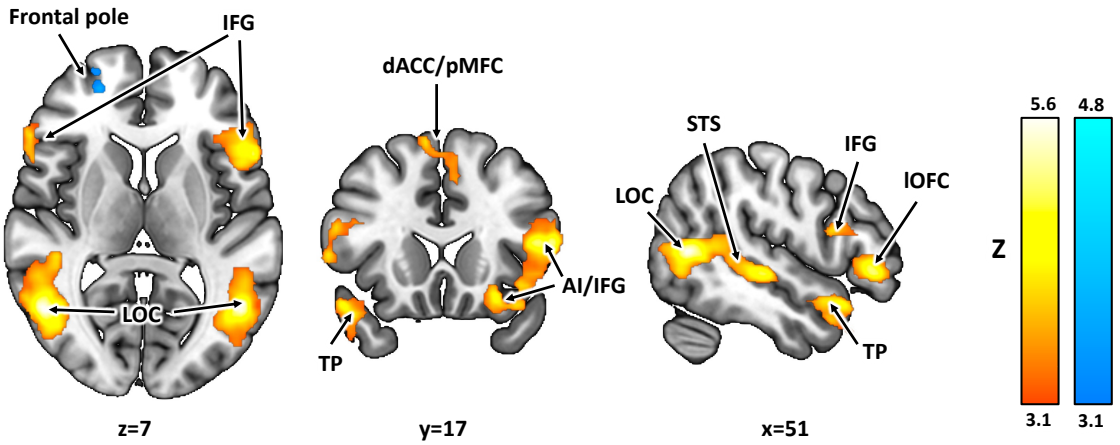


Figure 2-3. Brain activation for the macro- and micro expression vs. Neutral expression. Z-statistic images for the contrast [Macro + Micro] > Neutral (orange) and [Macro + Micro] < Neutral (orange) (blue) were obtained and corrected at a cluster-defining threshold of Z > 3.1 (Voxel-wise one-tailed p <.001), and a FWE-corrected cluster-level significance level of p < .05. (IFG, Inferior frontal gyrus; STS, superior temporal sulcus; LO, lateral occipital cortex; MFC/dACC, middle frontal cortex/dorsal anterior cingulate cortex; lOFC, lateral orbitofrontal cortex; dACC/pMFC, dorsal anterior cingulate cortex/posterior medial frontal cortex; TP, temporal pole)

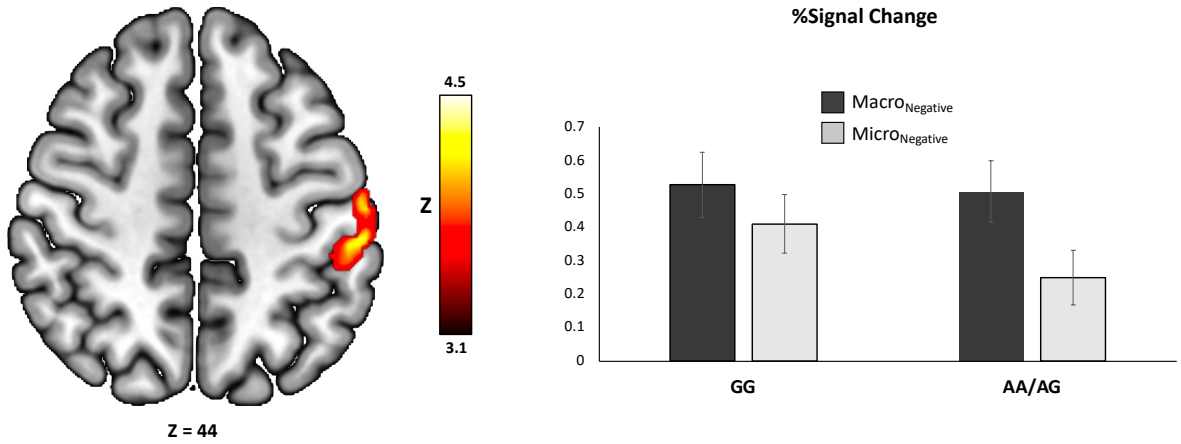


Figure 2-4. The *OXTR* genotype modulated the BOLD responses within the SMG in response to negative macro-vs. micro-expressions. *OXTR* rs53576 G homozyogtes showed relatively enhanced activation within the SMG in response to negative micro-expressions compared to the A allele carriers. The %signal change was extracted from the peak voxel [x=-58,y= -z=26] within the SMG for a visualization purpose.

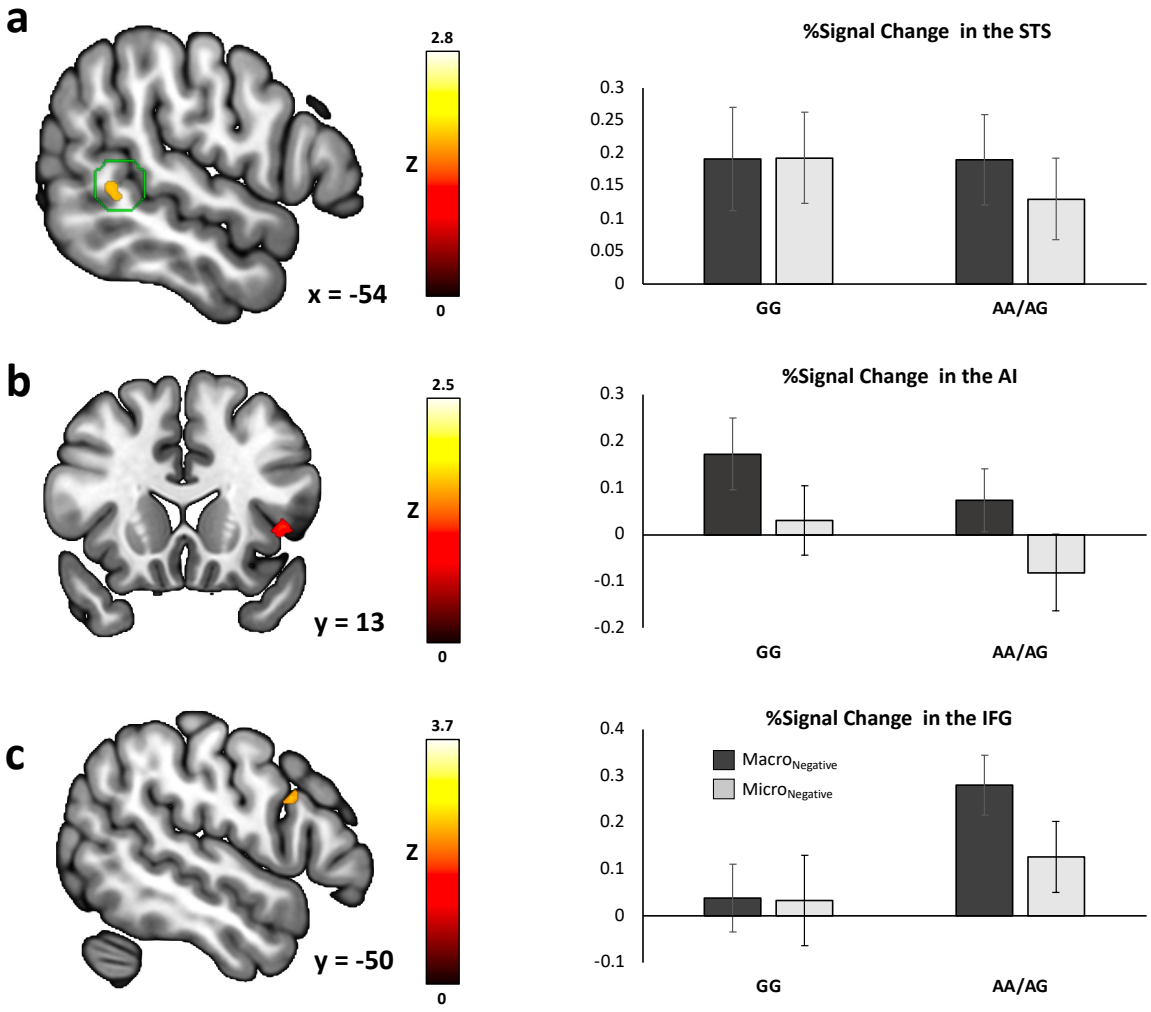


Figure 2-5. Genetic modulation of activation within the STS (a), AI (b), and IFG (c) for the contrast between the negative micro-expressions vs. macro-expressions. Exploratory ROI analyses showed that G homozygotes showed equally strong activations in these brain regions for negative the macro- and micro expressions, compared to the A allele carriers. Significant voxels (yellow) were overlayed with a functional ROI mask (green). All results were thresholded with small-volume corrected $p < .05$.

Table 2-1. Participants demographics and genotype composition

Demographics	**Neuroimaging Condition**		**Behavioral Condition**	
	N	*%*	*N*	*%*
***OXTR* Genotype**				
GG	19	42	42	30
AA/AG	24	48	99	70
Gender				
Female	26	60	90	37
Male	17	40	54	63
Ethnicity				
Asian	17	42	60	42
African American	6	13	27	19
Caucasian	18	40	33	24
Hispanic	2	5	17	12
Others	-	-	4	3

Table 2-2. Participants' average behavioral task performance.

Demographics	**Neuroimaging Condition**		**Behavioral Condition**	
	M	*SD*	*M*	*SD*
Macro				
Negative	67.6	23.2	83.4	16.2
Positive	81.4	17.1	82.3	12.3
Micro				
Negative	36.8	16.3	42.5	19.6
Positive	69.7	20.3	80.5	18.3

Table 2-3. Whole-brain activations for the effects of the Expression Type and the *OXTR* genotype.

Brain region	**Cluster size (voxels)**	**Max Z value**	**MNI coordinate**		
			X	**Y**	**Z**
Macro+Micro > Neutral					
Left IFG	2460	5.77	-56	20	12
Right LO*	1853	6.23	50	-62	8
Left LO**	1072	6.11	-50	-62	8
Right IFG	907	5.86	54	30	-2
Right TP	702	5.43	54	6	-18
dACC/pMFC	605	4.48	-6	12	54
Left SMG	181	4.6	-54	-42	28
Macro+Micro < Neutral					
Right frontal pole	1583	4.9	27	82	55
$Macro_{negative} > Micro_{negative}$					
Left IFG	1178	4.81	-54	28	4
dACC/pMFC	267	4.23	-8	16	46
Right TP	243	4.22	48	16	-22
Right LO*	187	4.48	48	-60	0
$Macro_{negative} < Micro_{negative}$					
Primary visual cortex	209	4.33	-22	-98	10
$Macro_{positive} > Micro_{positive}$					
Superior parietal lobe	1153	4.95	50	-66	36

Left cerebellum	307	4.17	-36	-70	-40
Middle frontal gyrus	229	4.02	28	20	58
Precuneus	200	4.31	0	-72	46
$Macro_{positive} < Micro_{positive}$					
No activation					
AA/AG ($Macro_{negative} > Micro_{negative}$) > GG ($Macro_{negative} > Micro_{negative}$)					
Left SMG	366	4.88	-58	-26	48

*Cluster encompassing the right pSTS

**Cluster encompassing the left pSTS, TPJ and left LO

Chapter 2 Supplementary materials/Appendix

S2-1. Sample size determination and participant allocation strategy

A priori-power analysis

We used the G*Power to conduct a priori power analysis. The reference effect sizes were taken from previous studies that investigated neural (GG vs. AA, Cohen's d =.81) and behavioral effects (GG vs. AA+AG, Cohen's d = .49) of OXTR rs53576 on social cognition involving face and emotion perception (Luo et al., 2015; Rodrigues et al., 2009). With the type-I error rate set to α = .05, the power analysis showed that a total of N=50 (e.g., 25 GG and 25 AA+AG) are required to provide 80% power for detecting a significant main effect of genotype on the neural response associated with socio-emotional processing. For the behavioral task, the power analysis yielded a required sample size of N=144 (GG=53, AA+AG = 91). In sum, by recruiting 200 participants, the current project is expected to have sufficient statistical power for detecting true effects at both neural and behavioral levels.

Participant allocation strategy

Before the COVID-19 pandemic, (~March 2020) participants were pseudo-randomly assigned to the neuroimaging or behavioral arm based on their genotype (i.e., OXTR rs53576, G/A). That is, genotyping was performed prior to the group assignment. As it was more difficult to recruit participants for the neuroimaging arm due to the additional screening criteria, those homozygotes for the A or G allele were prioritized to be included in the neuroimaging arm of the study whenever possible. However, the experimenter was blind to the specific genotypes of participants.

To facilitate data collection amid a COVID-19 pandemic, the recruitment protocol was modified such that participants were randomly assigned to either the neuroimaging or behavioral arm regardless of their genotypes. Genotyping was performed after neuroimaging/behavioral data collection. The experimenter remained blind to the specific genotype of participants.

S2-2. Personality traits of participants

S2-2-1. Behavioral arm

Table S2-1. Personality Traits across the two OXTR genotype group

Characteristics	GG		AA/AG		*t*(129)	*p*
	M	*SD*	*M*	*SD*		
Empathy	68.9	9.96	72.45	10.2	-1.79	.075
FNE*	38.2	10.7	40.4	11.6	-1.01	.316
NfC**	3.71	.59	3.57	.65	1.140	.256
SM***	13.3	3.72	13.38	3.10	.005	.996

*Fear of negative evaluation; **Need for Cognition; ***Social Monitoring

S2-2-2. Neuroimaging arm

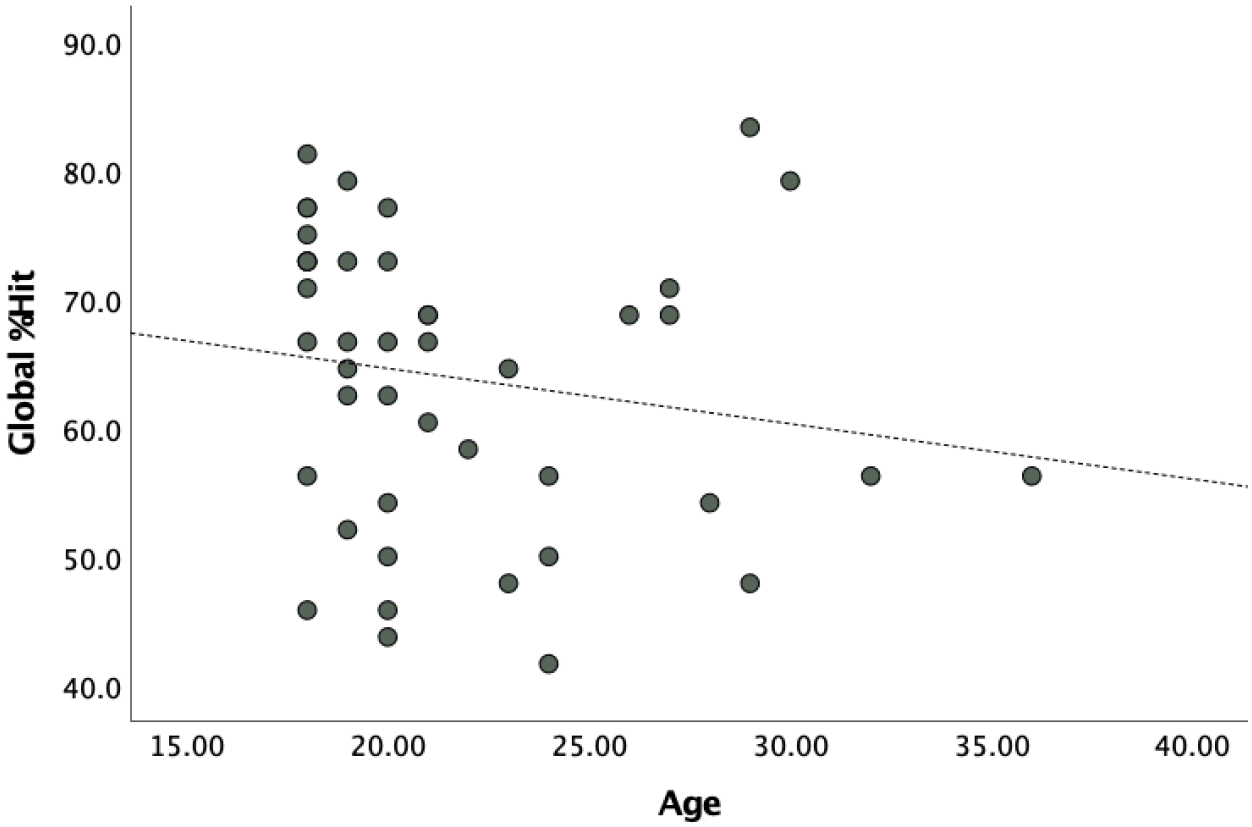
Table S2-2. Personality Traits across the two OXTR genotype group

Characteristics	GG		AA/AG		*t*(41)	*p*
	M	*SD*	*M*	*SD*		
Empathy	71.6	11.56	66.6	13.3	1.28	.205
FNE	35.6	6.65	36.6	10.5	-.392	.697
NfC	3.58	.67	3.6	.74	-.097	.923
SM	12.7	4.5	13.4	4.61	-.485	.631

S2-3. Stimuli Characteristics

StimulNumber	Category	Emotion	Gender	Onset	Apex	Offset	Length	CASME2 ModelNumber
1	Macro	Anger	M	105.866667	106.3	106.666667	0.8	37
2	Macro	Anger	M	63.4333333	63.86666667	65	1.56666667	24
3	Macro	Anger	F	18.5666667	19.06666667	20.2666667	1.7	15
4	Macro	Anger	M	38.4	38.73333333	39.2333333	0.83333333	27
5	Macro	Anger	F	18.7333333	19.56666667	20.3	1.56666667	32
6	Macro	Anger	F	18.4666667	18.96666667	20.5666667	2.1	16
7	Macro	Disgust	F	26.0333333	26.33333333	26.7333333	0.7	15
8	Macro	Disgust	M	15.6666667	15.93333333	16.3333333	0.66666666	25
9	Macro	Disgust	M	45.2	45.46666667	46	0.8	24
10	Macro	Disgust	F	59.7333333	59.96666667	61.2	1.46666667	16
11	Macro	Disgust	M	29.6666667	29.9	30.3666667	0.7	27
12	Macro	Disgust	F	12.2	12.53333333	13.1	0.9	22
13	Macro	Happiness	F	37.8333333	38.26666667	40.1666667	2.33333334	16
14	Macro	Happiness	M	28.0333333	29.06666667	30.4333333	2.4	25
15	Macro	Happiness	M	67.3333333	68.13333333	69.1333333	1.8	37
16	Macro	Happiness	F	81.9	82.53333333	83.6666667	1.76666667	15
17	Macro	Happiness	F	28.6	29.1	30.3	1.7	32
18	Macro	Happiness	M	36.0333333	36.73333333	37.7666667	1.73333334	27
19	Micro	Anger	F	71.8333333	72.1	0	0.26666667	15
20	Micro	Anger	M	72.7	72.9	73.1	0.4	24
21	Micro	Anger	F	131.066667	131.4333333	0	0.36666667	16
22	Micro	Anger	M	42.0666667	42.3	0	0.23333333	24
23	Micro	Anger	M	8.93333333	9.066666667	9.26666667	0.33333333	32
24	Micro	Anger	M	65.8	65.96666667	66.2	0.4	27
25	Micro	Disgust	M	12.4666667	12.56666667	12.9333333	0.46666667	27
26	Micro	Disgust	M	33.6	33.76666667	33.9	0.3	25
27	Micro	Disgust	F	11.1	11.53333333	0	0.43333333	16
28	Micro	Disgust	F	17	17.5	0	0.5	22
29	Micro	Disgust	F	34.3666667	34.56666667	34.8	0.43333333	15
30	Micro	Disgust	M	29.1333333	29.33333333	29.6	0.46666667	27
31	Micro	Happiness	M	28.2	28.7	0	0.5	37
32	Micro	Happiness	M	58.3666667	58.73333333	0	0.36666666	37
33	Micro	Happiness	M	61.2	61.4	61.6	0.4	24
34	Micro	Happiness	F	4.6	4.733333333	4.93333333	0.33333333	15
35	Micro	Happiness	F	76.1666667	76.36666667	76.5	0.33333333	16
36	Micro	Happiness	F	8.6	8.8	9.06666667	0.46666667	32
37	Neutral	Neutral	F				-	22
38	Neutral	Neutral	F				-	22
39	Neutral	Neutral	M				-	37
40	Neutral	Neutral	F				-	32
41	Neutral	Neutral	F				-	32
42	Neutral	Neutral	M				-	27
43	Neutral	Neutral	M				-	24
44	Neutral	Neutral	M				-	37
45	Neutral	Neutral	F				-	16
46	Neutral	Neutral	M				-	25
47	Neutral	Neutral	F				-	15
48	Neutral	Neutral	M				-	25

S2-4. Zero order correlation between age and task performance



S2-5. Whole-brain activations for the effects of the Macro > Neutral and Micro > Neutral

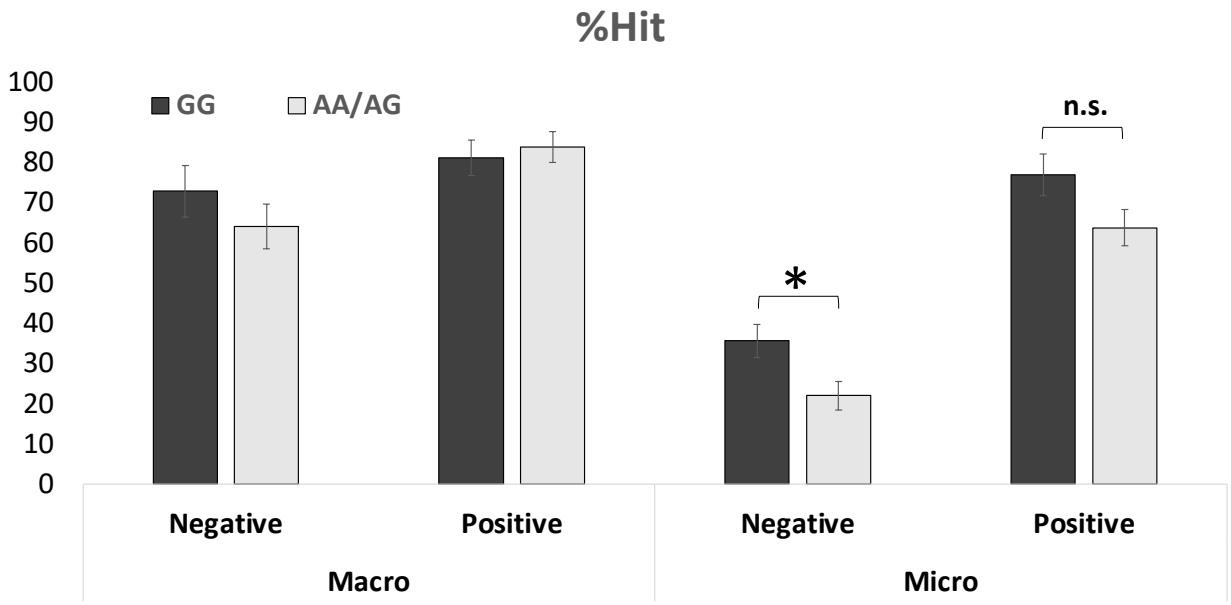
Brain region	Cluster size (voxels)	Max Z value	MNI coordinate		
			X	Y	Z
Macro > Neutral					
Left IFG/AI	4008	6.14	-40	24	10
Right TP/IFG	2214	5.65	48	20	-20
Left LO/STS	1946	5.88	52	-60	6
Right LO/STS	1231	5.94	-50	-60	8
Left dACC	1170	6.75	-4	12	54
Left SMG	886	5.48	-48	-34	42
Left cerebellum	236	4.56	-20	-66	-48
Right cerebellum	209	4.9	22	-66	-46
Macro < Neutral					
Right Frontal pole	883	5.0	34	56	24
Micro > Neutral					
Right LO/STS	1064	5.17	50	-62	8
Left LO/STS	533	5.08	-50	-60	8
Left IFG/AI	343	4.93	-48	18	12
Right IFG/AI/TP	340	4.98	54	32	0
Micro < Neutral					
Right frontal pole	1179	4.47	42	36	38

S2-6. The results of ROI analyses

Main Contrast /Brain region	Cluster size (voxels)	Max Z value	Peak Voxel MNI coordinate		
			X	Y	Z
Macro+Micro > Neutral					
Amygdala (R)	7	3.36	18	-6	-12
dACC (L)	105	4.00	-6	22	44
IFG (L)	302	5.54	-56	20	14
IFG (R)	120	4.48	54	16	18
AI (R)	96	3.28	46	26	-2
AI (L)	33	3.89	-44	22	0
STS (R)	84	4.17	56	-44	8
TPJ (R)	172	5.65	52	-60	10
Macro+Micro < Neutral					
Precuneus (R)	79	3.13	2	-56	30
$Macro_{All}$ > $Micro_{All}$					
Caudate nucleus (R)	32	2.54	18	4	16
Amygdala (R)	81	4.38	20	-4	-14
dACC (L)	130	2.86	-8	22	34
IFG (L)	21	3.24	-54	22	28
AI (R)	46	2.72	44	28	-6
AI (L)	12	2.82	-44	18	-4
mPFC (L)	141	3.12	-2	56	24

STS (R)	45	2.88	56	-40	4
$Macro_{All} < Micro_{All}$					
No activation					
$Macro_{negative} > Micro_{negative}$	58	3.33	8	40	-6
dACC (L)	116	3.23	-8	24	34
IFG (L)	158	4.3	-56	22	16
AI (R)	40	3.33	42	30	-2
AI (L)	31	3.7	-44	18	-4
mPFC (L)	17	2.42	-4	54	26
STS (R)	5	2.49	56	-40	6
$Macro_{negative} < Micro_{negative}$					
No activation					
$Macro_{positive} > Micro_{positive}$					
Caudate nucleus (R)	14	2.4	18	8	16
Amygdala (R)	39	3.61	20	-4	-16
dACC (Bilateral)	21	2.68	0	30	42
mPFC (R)	132	3.03	2	58	20
Precuneus (R)	115	3.32	10	-50	38
STS (R)	34	2.65	62	-40	2
TPJ (R)	5	2.91	52	-60	26
$Macro_{positive} < Micro_{positive}$					
No activation					

S2-7. The effect of the *OXTR* genotype on the perception of positive and negative micro-expressions.



In the 2 (*OXTR* Genotype: GG vs. AA/AG) × 2 (Expression Type: Macro- vs. Micro) × 2 (Valence: Positive vs. Negative) repeated measure analysis of variance (RMANOVA) on %Hit for each expression category, the effect of the *OXTR* genotype on the perception of micro-expressions tended to be pronounced for the negative expressions (p = .03) than positive expressions (p=.09), although the three-way interaction did not reach statistical significance (p=.155).

Chapter 3

The neural basis of smile authenticity judgments and the possible modulatory role of the oxytocin receptor gene (*OXTR*)

Chapter Abstract

In this chapter, I sought to answer two questions: 1) what neuro-cognitive mechanisms are subserving the correct identification of genuine vs. posed smiles, and 2) how genetic variation in *OXTR* modulates these mechanisms. Participants (Neuroimaging arm $N = 43$, Behavioral arm $N = 131$) viewed a series of recordings that depicted one of three dynamic facial expressions: a genuine smile, posed smile, or no smile. Participants discerned the authenticity of the smiles. The BOLD fMRI signals recorded during the face perception and participants' overall behavioral task performance were analyzed with respect to their genotypes. The main findings of this experiment can be summarized as follows:

1. Overall, smiling faces recruited brain regions involved with dynamic face perception, emotion processing, sensorimotor simulation, cognitive and emotional empathy.
2. Brain regions implicated in sensorimotor simulation (e.g., the putamen, secondary somatosensory cortex), mentalizing (e.g., dmPFC), and reward processing (e.g., mOFC) contribute to the correct identification of genuine vs. posed smiles.
3. Individual differences in overall perceptual accuracy (i.e., d-prime) and decision bias (e.g., response criteria) were represented in the IFG and dACC, each of which is known for mirroring mechanisms and conflict processing.
4. G homozygotes in the neuroimaging arm tended to judge posed smiles as genuine erroneously, and such liberal decision bias was associated with reduced activations in the mPFC and rTPJ. This result, however, needs further replication.

Overall, data suggest that the joint contribution of sensorimotor simulation and metalizing is key to the accurate smile authenticity judgment. Increased social affiliation due to OT may interfere with this process by making people employ a "gullible interaction style."

Keywords: smile authenticity, sensorimotor simulation, mentalizing, *OXTR,* decision bias

Introduction

Smiles are versatile tools for social communication in humans. Often perceived as an honest display of pleasure, smiling faces effectively capture visual attention in both infants and adults (Hayward et al. 2018), influence perceivers' mood (Neumann and Strack 2000), and act as a reward that reinforces behaviors (Lin, Adolphs, and Rangel 2012). People also use smiles to infer affiliative motives of others in social interaction, showing a tendency to trust and to be attracted to those who display smiles (Martin et al. 2017).

Not all smiles, however, are made equal. While emotional expressions in humans are often involuntarily triggered (Ekman and Friesen 2003), people can deliberately pose smiles regardless of their true feelings or motives in order to reap the positive social outcomes that the smiles may facilitate. For instance, those with high trait psychopathy are known to put up fake smiles frequently and skillfully to build a positive public self-image and deceive others (Porter, Ten Brinke, and Wallace 2012, Ten Brinke et al. 2017). Similarly, people in service industries or educational institutions are often required to pose smiles to enhance consumer experiences (Grandey et al. 2005) and perceived efficacy of teaching in the classroom (King 2016).

Since human smiles are multi-purpose signals that can serve either affiliative or manipulative goals (Martin et al. 2017), the ability to discriminate among them may be adaptive and promote successful social navigation. Interestingly, despite our species-wide proficiency at reading facial expressions, considerable variation exists in the capacity for accurate smile authenticity judgments (McGettigan et al. 2015). For instance, psychologists have found that people scoring high in trait

empathy are better at detecting the "Duchenne markers (Ekman, Davidson, and Friesen 1990)," or the muscles around the eye (i.e., orbicularis oculi) uniquely engaged during the genuine smiles. Neuroscientists have also started to ask how genuine vs. posed emotional expressions are represented in the brain, and what neurocognitive mechanisms subserve our ability to discern different categories of smiles. Paracampo and colleagues (2017) found that transcranial magnetic stimulation (TMS) of brain regions involved in face processing (e.g., inferior frontal gyrus, IFG; ventral somatosensory cortex, SI) subsequently interfered with smile authenticity judgments (Paracampo et al. 2017). Similarly, McGetiggan and colleagues (2015) used functional magnetic resonance imaging (fMRI) to examine brain activations associated with auditory perception of natural and fake laughter. The results indicated that increased activations within the mentalizing network (e.g., the medial prefrontal cortex, mPFC), and sensorimotor areas (e.g., secondary somatosensory cortex) tracks individual participants' task performance (McGettigan et al. 2015).

While offering invaluable insights into the brain mechanisms underlying smile authenticity judgments, these pioneering studies had several methodological limitations. First, the TMS technique is best suited for establishing the causal involvement of a relatively small number of brain areas (Kable 2011). Therefore, it is not able to identify a broader pattern of brain activations that could concurrently emerge during smile authenticity judgments. Also, TMS cannot adequately target the subcortical structures, which are known to play an essential role in smile processing (O'Doherty et al. 2003). Second, the perception of static- vs. dynamic facial expressions of emotions is known to incur dissociable patterns of brain activations (Johnston et al. 2013, Kilts et al. 2003). However, multiple studies have examined the neural correlates of genuine vs. posed smiles using static images of emotions (McLellan et al. 2010, Mega, Gigerenzer, and Volz 2015) or non-visual stimuli (McGettigan et al. 2015). This begs the question of how facial dynamicity

influences the smile authenticity judgments and their neural substrates less addressed. Lastly, emerging evidence suggests that both structural and functional characteristics of the brain areas central to human social cognition and behaviors can be regulated by a set of genes and their allelic variations (Bigos and Weinberger 2010). This raises the possibility that smile authenticity judgments may also be contingent on genetic factors that regulate the psychological or neuro-cognitive mechanisms underlying smile authenticity judgments.

The goal of the current study is to extend the existing evidence on smile authenticity judgments by addressing these shortcomings. We conducted *an fMRI experiment* in which participants viewed a series of *dynamic* facial expressions depicting genuine vs posed smiles. Building upon the aforementioned findings, we hypothesized that the perception and discrimination of genuine vs. posed smiles would be linked with neural mechanisms that subserve face and emotion processing, empathy, mentalizing, and sensorimotor simulation(Paracampo et al. 2017, McGettigan et al. 2015). Furthermore, we explored the possible genetic underpinnings of the neural mechanisms that subserve smile authenticity judgments, as human social cognition and behaviors are influenced by a wide array of genetic variants linked with the structure and functional architectures of the brain (Bigos and Weinberger 2010). We specifically focus on the role of the human oxytocin receptor gene (*OXTR*). Genetic variations in *OXTR* are known to regulate the pattern of oxytocin receptor expression in the brain, thereby modulating the function of the neuropeptide oxytocin (OT) (King et al. 2016). OT is known to regulate a wide range of mammalian social behaviors by enhancing the salience and reward value of socially relevant cues (Shamay-Tsoory and Abu-Akel 2016). Related to smile authenticity judgments in humans, intranasally administered oxytocin (INOT) has been shown to promote gaze to faces (Domes, Heinrichs, Gläscher, et al. 2007, Tollenaar et al. 2013), improve facial emotion recognition (Shahrestani, Kemp, and Guastella 2013), and modulate

the neural representation of positive and negative social cues (Domes, Heinrichs, Gläscher, et al. 2007, Gamer, Zurowski, and Büchel 2010, Groppe et al. 2013). INOT can also upregulate psychological or physiological processes that may support the smile authenticity judgment, such as mentalizing ((Domes, Heinrichs, Michel, et al. 2007), empathy (Hurlemann et al. 2010), and automatic motor simulation of facial or bodily movements (Pavarini et al. 2019, Korb et al. 2016) in both clinical and healthy populations (Shinotoh 2020, Keech, Crowe, and Hocking 2018).

Evidence from INOT studies raises the possibility that naturally occurring allelic variations in *OXTR* may be systematically linked with smile authenticity judgments in humans by regulating endogenous OT signaling in the brain and its various downstream phenotypic effects. Consistent with this possibility, many behavioral and imaging genetics studies have found associations between single nucleotide polymorphisms (SNPs) in *OXTR* and various facets of human sociality, including face perception (Skuse et al. 2014, Melchers et al. 2013, Burkhouse et al. 2016), as well as empathy (Lucht et al. 2013, Gong et al. 2017). However, a direct investigation into the link between *OXTR* and smile authenticity judgment has not yet been made. In this study, we examined whether *OXTR* SNPs are associated with the ability to discern genuine from posed smiles and how this genetic modulation is represented in the brain activation.

We specifically examined 1) differences in the neural response to posed vs genuine smiles, 2) whether OXTR SNPs are associated with the ability to discern genuine from posed smiles and 3) how this genetic modulation is represented in the brain. We conducted an imaging genetics experiment where participants determined the authenticity of dynamic facial expressions of smiles. Participants' behavioral responses and the blood-oxygen dependent (BOLD) fMRI signal measured during the task were analyzed in conjunction with their genetic data. Our imaging genetics analyses

specifically centered on the G allele of rs53576, which has widely been implicated in sensitive social cognition and behaviors (Li et al. 2015, Gong et al. 2017, Chander et al. 2021). While the effects of a single gene on higher-level phenotypes tend to be small and thus prone to false positives (Bogdan et al. 2017), our focus on rs53576 was motivated further by multiple independent lines of research showing that the G allele of rs53576 modulates the effects of INOT on various brain regions important for social cognition (Marsh, Yu, et al. 2012, Luo, Ma, et al. 2015, Watanabe et al. 2017).

Based on the general hypothesis that *OXTR* rs53576 will be associated with the accuracy of smile authenticity judgment, we made the specific predictions as below:

Behavioral Predictions: First, we predicted that participants, on average, will successfully discriminate genuine from posed smiles conveyed in dynamic facial expressions (**Prediction 1a**). Second, we predicted that G homozygotes of *OXTR* rs53576, with the facilitative role of the G allele in social cognition and behavior, will show higher discrimination accuracy (**Prediction-1b**) compared to the A allele carriers.

fMRI Predictions: We predicted that the perception of genuine and posed smiles, compared to the control stimuli without smiles, would be associated with increased activations in the brain regions involved in the processing of dynamic facial expressions and positive emotion perception (i.e., superior temporal sulcus, STS; inferior frontal gyrus, IFG; nucleus accumbens, NAcc; medial orbitofrontal cortex, mOFC; caudate nucleus, and amygdala) (Dricu and Frühholz 2016, Fusar-Poli et al. 2009, Liu et al. 2011) (**Prediction 2a**), as well as affective empathy (i.e., anterior insula, AI, medial frontal cortex/dorsal anterior cingulate cortex, dACC) and cognitive empathy (i.e.,

mentalizing) (i.e., right temporoparietal junction, rTPJ; medial prefrontal cortex, mPFC; precuneus) (**Prediction 2b**) (Fan et al. 2011, Schurz et al. 2014). We also predicted that genuine smiles would be associated with stronger activations in subsets of these brain regions compared with posed smiles **(Prediction 2c)**. Lastly, we predicted that the participants homozygous to the G allele of *OXTR* rs53576 would be linked with stronger activation in these target brain regions compared with A allele carriers **(i.e., Prediction 2d).**

Methods

Participants

We recruited 193 healthy adults from Emory University and the surrounding community. Volunteers who had a history of psychiatric or neurological illness, as well as those who were currently on psychoactive drugs were excluded. All eligible participants were then assigned into either the neuroimaging (N = 50, Female N = 29) or behavioral arm (N = 145, Female N = 89) based on *a priori* power analysis. The results of the power analysis and the participant allocation methods are described in **Supplementary materials (S3-1).** The demographics of the final study samples for the neuroimaging and behavioral arms are summarized in **Table 3-1.**

Materials and Procedures

All study materials and experimental procedures were approved by Emory University Institutional Review Board (IRB00112525) and pre-registered at https://osf.io/d3x85.

Pre-experiment online survey

Participants ptovided written informed consent and completed pre-experiment questionnaires via an online study portal (i.e., Research Electronic Data Capture, REDCap: https://www.project-redcap.org).

***Demographic survey*:** Participants indicated their age, sex, ethnicity, political self-identification (1=Very conservative, 5=Very liberal) and religiosity (1=Very religious, 5=Not at all religious). Data on political self-identification and religiosity were not used for the current study.

***Psychological questionnaires*:** Participants completed a set of psychological questionnaires designed to measure personality traits that could be associated with face and emotion perception. These variables included affective/cognitive empathy (i.e., Interpersonal Reactivity Index, IRI) (Davis 1983), need for cognition (Need for Cognition Scale, NfC) (Cacioppo and Petty 1982), self-monitoring (i.e., Social Monitoring Scale, SM) (Lennox and Wolfe 1984), and impression management (i.e., Fear of Negative Evaluation, FN) (Leary 1983).

Saliva sample collection

Participants who finished the online survey were subsequently invited to the study sites located on the Emory University campus. Participants who were assigned to the behavioral arm visited Laboratory for Darwinian Neuroscience. Upon arrival, participants provided their saliva samples using Oragene DNA self-collection kits (OGR-600, DNA Genotek Inc, Ontario, Canada). Those

in the neuroimaging arm went through the same procedure at Facility for Education and Research in Neuroscience (FERN).

Main Task

Following the saliva collection, participants performed a novel smile authenticity judgment task. Participants in the neuroimaging arm performed the task inside an MRI scanner located at FERN. Participants in the behavioral condition performed the identical task inside a testing room in the Laboratory for Darwinian Neuroscience. The task was implemented in Psychtoolbox 3 on MATLAB (The MathWorks, Natick, 2015b).

Smile authenticity judgment task: Participants watched a series of short video clips (3000-6000ms) depicting one of three types of dynamic facial expressions: a genuine smile, a posed smile, or a neutral expression. After viewing each video clip, participants indicated if the smile was genuine, posed, or absent using a keypad. The chosen option was highlighted in red for 500ms. Each trial (N = 60; Genuine smile N = 20, Posed smile N = 20, Neutral expression N = 20) was separated by a jittered inter-trial interval (ITI) with a fixation point (1000-5000ms) (**Figure 3-1**). Participants also performed two unrelated cognitive tasks before and after the authenticity judgment task in a counterbalanced order. The results of these extra tasks are not discussed in the current manuscript. At the end of the experiment, participants were debriefed and received either $40 (Behavioral condition) or $50 (Neuroimaging condition) as compensation.

Experimental stimuli: The smile videos were created with the UvA-NEMO smile database (Dibeklioğlu et al., 2015). The UvA-NEMO database consists of 1,240 high resolution (1920×1080 pixels), high-frame (50 fps), illumination-controlled recordings of genuine and posed smiles obtained from 400 individuals (i.e., face models). Genuine smiles were spontaneously induced using video segments, and posed smiles were elicited by asking the face models to smile as realistically as possible. The authenticity of the smiles was cross-checked by two trained annotators based on action units and facial dynamics (Dibeklioğlu, Salah, and Gevers 2012).

To select the experimental stimuli, we conducted a separate pilot study involving an independent group of participants ($N = 25$) who rated the face models in terms of their facial attractiveness, and smile intensity. Based on the results of the pilot study, we identified face models whose genuine and posed smile were rated equal on perceived smile intensity and overall attractiveness. Then, we chose 20 face models of varying ages (i.e., 8 to 60), sexes (Female $N = 10$), and ethnicity. From each of the 20 face models selected through this process, one control stimulus (i.e., No expression) was created by extracting a section of the genuine- or posed smile videos displaying a neutral expression. The average length of the videos used for the smile categories were matched (Genuine $M = 3.4$s $SD = .88$), Posed: $M = 3.2$s, $SD = .89$ No smile: $M = 3.2$s, $SD = .40$). All videos were edited with Adobe Premiere Pro CC (2014 release). Detailed results of the pilot study are available in **Supplementary Materials S3-2.**

Data Analysis

Neuroimaging data acquisition: Neuroimaging data were acquired with a 3-Tesla Siemens MAGNETOM Prisma MRI scanner. T1-weighted high resolution anatomical images (i.e., T1 scans) were acquired using a 3D magnetization-prepared rapid gradient-echo (MPRAGE) sequence with a Generalized auto-calibrating partial parallel acquisition (GRAPPA) factor of 3. The T1 scan protocol used the following imaging parameters: the repetition time (TR) = 1900, inversion time (TI) = 900 and echo time (TE) = 2.27ms, a flip angle of 9˚, a field of view of 256×256 mm^3, a matrix of 256×256, and isotropic spatial resolution of 1.0×1.0×1.0 mm^3. fMRI data were acquired using an Echo-Planar Imaging (EPI) sequence for blood-oxygen-level-dependent (BOLD) fMRI. EPI images were collected in an interleaved fashion with the following parameters: TR = 1200ms, TE = 30ms, matrix = 74x74, Field of View = 220mm, isotropic in-plane resolution = 3.0 mm, slice thickness = 3.0 mm, 54 axial slices with no gap in between and no phase oversampling.

Genetic data acquisition: Participants' DNA was extracted from saliva samples. Individual participants' genotype was extracted and analyzed with Axiom™ Precision Medicine Research Array (Affymetrix) and TaqMan SNP Genotyping Assays with a ViiA7 Real Time PCR System for genotype resolution (Applied Biosystems, Foster City, CA). For quality control in SNP genotyping, each 384 well genotyping plate contained multiple duplicate wells and positive and negative controls. 106 Ancestry-Informative markers were used to account for potential population stratification. These markers discriminated European, African, East Asian, and Native American origins. We used a structure software (Pritchard, Stephens, and Donnelly 2000) to estimate proportions of chromosomal ancestry based on K (the number of source populations). Principal components analysis (PCA) was calculated account for population stratification. The first two principal components from this analysis were used in the analyses as covariates to control for population stratification.

Behavioral data analysis

All behavioral data were processed and analyzed with MATLAB R2020a and SPSS version 28 (Armonk, NY: IBM Corp).

We calculated the proportion of the correct trials for each smile type (i.e., Genuine, posed, and no smiles), and averaged them to create a measure of overall behavioral task performance (i.e., Global %Hit). Following previous studies that employed a signal-detection framework to analyze the perception of the smile authenticity (Paracampo et al. 2017, McLellan et al. 2010, McLellan et al. 2012), we also computed the parameters for decision bias (i.e., response criterion C) and *d*-prime (i.e., *d*') based on the hit rate (i.e., the number of trials that participants correctly identified genuine smiles*100/20; HR) and false alarm rate (i.e., the number of trials where participants misclassified posed smiles as genuine smiles*100/the total number of posed smile trials, FA). The neutral expressions were not considered as they contained no true signal or noise (i.e., genuine or posed smiles).

Prediction 1a: To test whether participants' performance was significantly above chance, we used one-sample *t*-tests comparing 1) Global %Hit with 0.33 (i.e., 1/three choice options) and 2) the group-averaged *d*' values with zero.

Prediction 1b: The association between the *OXTR* genotype and the measures of discrimination accuracy was tested with a univariate GLM framework. The model included the *OXTR* genotype, participants' sex, and age as primary predictors, and the signal detection parameters (i.e., *d*' and C)

as dependent variables. Complementing this approach, we also used a repeated measure analysis of covariance (RMANCOVA) model to test the effects of the smile category (i.e., Genuine vs. Posed vs. No smile), *OXTR* genotype (i.e., GG vs. AA+AG), participants' sex (i.e., Female vs. Male), and age on the proportions of the correct responses for each smile category.

All statistical tests for behavioral data analyses were performed with the type-I error rate of $\alpha = .05$ (two-tailed). Bonferroni-corrected *post hoc* pairwise comparisons were made for significant main effects or interactions. Greenhous-Geisser corrected degrees of freedom and *p*-values were reported for the RMANOVA models if the sphericity assumption was violated. The results of the exploratory analyses involving the personality traits were not adjusted for multiple comparisons.

Exploratory analysis

We examined whether there existed any significant baseline differences in personality traits and demographic variables between the two *OXTR* genotype groups. In case such variables were identified, we performed exploratory analyses with those variables included as covariates.

Neuroimaging data analysis

Neuroimaging data analyses were performed using the Oxford Center for Functional Magnetic Resonance Imaging of the Brain's software library (FSL v6.0, http://www.fmrib.ox.ac.uk/fsl/).

Preprocessing: Our preprocessing pipeline included 1) motion correction using the MCFLIRT (Jenkinson, Bannister, Brady, & Smith, 2002), 2) skull-stripping using FSL's Brain Extraction Tool

(BET), 3) slice timing correction, 4) high-pass temporal filtering with a filter width of 100 seconds, 5) spatial smoothing using a Gaussian kernel of full width at half maximum (FWHM) of 6 mm, 6) spatial registration of fMRI images to high-resolution T1 images (i.e., Boundary-Based-Registration), and 7) spatial normalization to the standard Montreal Neurological Institute (MNI) 2 mm brain (i.e., Affine transformation using 12 degrees of freedom) using FNIRT (Greve & Fischl, 2009). Data from four participants were excluded due to technical errors during the scan, or missing ancestry data. We further excluded data from two participants who misunderstood the task instructions. The final sample size for the neuroimaging was $N = 44$.

1st level analysis: Functional images were analyzed with a univariate general linear model (GLM) approach. At the single-subject level, the raw data from each trial were convolved with a double-gamma hemodynamic response function (HRF) in FSL. To identify neural activations specifically associated with the correct identification of genuine and posed smiles, we included explanatory variables (i.e., EVs) that can separately model the brain responses associated with the correct vs. incorrect smile authenticity judgments for both smile categories. Specifically, the model included the following EVs and their temporal derivatives: Gen_{Hit}, Pos_{Hit}, $Gen_{Incorrect}$, $Pos_{Incorrect}$, and No expression (NE). We also modelled the decision phase during which participants made authenticity judgments as an event of no interest (i.e., Decision). Please note that the distinction between correct vs. incorrect trials was not made for NE trial because participants' performance for the control stimuli showed a ceiling effect with the mean %Correct greater than 90%. In addition to the task-related EVs, six motion regressors and first two principal components that captured the of the study sample were included to account for head motion and population structure, respectively.

2nd level analysis: A series of mixed-effect analyses (i.e. FLAME1) were carried out with the fMRI Experts Analysis Tool (FEAT) in FSL. To test our predictions regarding the neural activations associated with the 1) perception of genuine and posed smiles, and 2) correct smile authenticity judgments, we specified the following contrasts and their reverse contrasts: 1) $Smile_{All}$ > NE, and 2) $Gen_{Hit} > Pos_{Hit}$.

Both whole brain analyses and ROI analyses were used to test our predictions**. Prediction 2-a, 2-b,** and **2-c** were examined with one sample *t*-tests comparing the contrast estimates yielded from the main GLM with zero. **Prediction 2-d** was examined by dividing the study sample into two groups based on participants' genotype (i.e., rs53576 GG vs. AA/AG) and performing a two-sample t-test.

For the whole-brain analyses, the Z-statistic images were thresholded initially with a cluster-defining threshold of Z > 3.1(voxel-wise one-tailed $p < .001$). We then applied a familywise error (FWE)-corrected cluster significance level of one-tailed $p < .05$ to any supra-threshold activations.

ROI analyses centered on the brain regions previously implicated in dynamic face processing and positive emotion evaluation (i.e., posterior superior temporal sulcus, pSTS; Inferior frontal gyrus, IFG, bilateral amygdala; NAcc, right caudate nucleus, and medial orbitofrontal cortex, mOFC), as well as empathy (i.e., affective empathy: anterior insula, AI, dorsal anterior cingulate cortex/medial frontal cortex, dACC/MFC; cognitive empathy: right temporo-parietal junction, rTPJ; medial prefrontal cortex, mPFC, and precuneus, PC). The coordinates and size of these ROIs were determined based on the activation likelihood estimation (ALE) meta-analyses showing the brain regions recruited during explicit evaluation of facial expression (Dricu and Fruhholz, 2016),

positive and negative facial emotion (i.e., amygdala; Fusar Poli et al., 2009), perceptual-affective empathy (i.e., ACC and AI) (Fan et al. 2011), ToM/Mentalizing (i.e., pSTS, rTPJ, mPFC, PC) (Schurz et al. 2014, Dricu and Frühholz 2016), and reward processing (i.e., Nacc, Caudate, and mOFC) (Liu et al. 2011).

For each ROI defined, voxel-wise, one-sample *t*-tests were used to test if the activation values obtained from the contrasts [$Smile_{All}$ > NE], and [Gen_{Hit} > Pos_{Hit}] were significantly different from zero. The Z statistic images were thresholded at $p < 0.05$ (one-tailed) corrected for multiple comparisons across all ROI voxels within each ROI based on Gaussian Random Field Theory (i.e., Small Volume Correction, SVC) (Poldrack 2007).

Results

Behavioral Arm

Sample characteristics: The demographic and personality characteristics of G homozygotes and the A allele carriers are summarized in **Table 3-1.** There was no significant between-group difference in personality traits (All *P*s >.058), demographic variables (All Ps >.944), and sex ratio (p=.703). Four participants were excluded dye ti genotyping failure or technical errors that resulted in incomplete data. The final sample size in the behavioral arm was N = 141.

Prediction 1a: Participants overall performance in the smile authenticity judgment task

The average Global %Hit (M = 75.57, SD = 8.18) was significantly above chance ($t_{(138)}$ = 108.874, p<.001, Cohen's d = 9.23). The proportion of correct trials for each stimulus category (i.e., Genuine, Posed, and No expressions) also exceeded the chance level (All Ps <.001). The average d-prime (M = .953, SD = .61) was significantly larger than zero ($t_{(138)}$=18.476, p<.001, Cohen's d=1.567) indicating the discrimination accuracy above-chance.

Prediction 1b and 1c: Association between *OXTR* and the smile authenticity judgment

The univariate GLM analysis on the discrimination accuracy (i.e., d') revealed no evidence of a significant main effect or interaction effect involving the *OXTR* Genotype (All Ps >.525). Similarly, participants' sex and ethnicity had no effects on the average d' (All Ps > .094). We found the significant main effect of age (B -.016, p=.03), indicating that older participants tended to show slightly decreased sensitivity (**Supplementary material S3-2**). We found no significant effects of *OXTR* genotype (All Ps >.289), age (p=.123), sex (All Ps>.122), or ethnicity (All Ps>.603) on decision bias.

Neuroimaging Arm

Behavioral Results

Sample characteristics: The demographic and personality characteristics of G homozygotes and the A allele carriers are summarized in **Table 3-2**. No significant intergroup differences in these background variables were identified.

Prediction 1a: Participants overall performance in the smile authenticity judgment task

Participants' Global %Hit (M = 74.81, SD = 7.63) was well above chance ($t_{(44)}$ = 36.757, p<.001, Cohen's d = 5.47). The proportion of correct trials for each stimulus category (i.e., Genuine, Posed, and No expressions) also exceeded the chance level (All Ps <.001). The group-averaged d-prime (M = .882, SD = .55) was significantly larger than zero ($t_{(44)}$=10.58, p<.001, Cohen's d=1.578) indicating discrimination accuracy above chance.

Prediction 1b and 1c: Association between *OXTR* and the smile authenticity judgment

The results of the univariate GLM analysis on the discrimination accuracy (i.e., d') showed no evidence of a significant main effect or interaction effect involving the *OXTR* Genotype (All Ps >.131). However, G homozygotes (M=-.19 SD=.28) employed a significantly more liberal response criteria (i.e., C) than did the A allele carriers (M=-.09 SD=.30), $F_{(1, 40)}$ = 4.73, p=.024, η_p^2=.121. That is, G homozygotes were more likely to erroneously judge posed smiles as genuine smiles. This group difference in decision bias was more significant (p=.018*)* after baseline differences in personality traits across the *OXTR* genotype groups were controlled for **(Figure 3-2a).**

Given the significant intergroup differences in decision criteria, a follow-up analysis performed on the average %Hit rates for each smile category to determine whether the effect was driven by either hit rate or false alarm rate. A consistent pattern of results emerged in the 2 (*OXTR* Genotype) x 2 (Sex) x 3 (Smile category) RMANCOVA. While there was no evidence of the main effect of *OXTR* genotype (p=.966) (Prediction 1b), we found a significant interaction between *OXTR* genotype and

Smile category, $F_{(1.699, 64.554)} = 4.775$, p=.016, η_p^2=.112 (Prediction 1c). Consistent with the results of signal-detection analysis, a post-hoc analysis revealed that G homozygotes were more likely to judge posed smiles as genuine ($F_{(1, 38)} = 6.273$, p= .017, η_p^2=.142), compared to the A allele carriers (i.e., High false alarm rate) **(Figure 3-2b)** . No other effects involving the *OXTR* genotype turned out significant (All *Ps* = .127).

In sum, our behavioral data in the neuroimaging arm provided mixed support for the predicted effects of *OXTR*. While the *OXTR* genotypes did not generally enhance or diminish participants' overall discrimination accuracy (i.e., *d*'), we found evidence relevant to **Prediction 1c**: G homozygotes showed a liberal decision bias, which led them to more often mistakenly judge posed smiles as genuine.

Neuroimaging Results

Prediction 2a and 2b: Neural correlates of perceiving genuine and posed smiles

Our whole-brain analysis revealed the patterns of brain activations broadly consistent with **Prediction 2a and 2b**. Specifically, correctly identified genuine and posed smiles, compared to neutral expressions (i.e., $Smiles_{All}$ > NE), were associated with greater activations in brain regions involved in dynamic face processing (e.g., bilateral STS, MFG, IFG, and lateral occipital cortex), emotion perception (e.g., amygdala, thalamus, dorsal and ventral striatum), as well as affective and cognitive empathy (e.g., pMFC/dACC, AI, and mPFC)**.** In comparison, the reverse contrast (i.e., NE > $Smiles_{All}$) yielded significant activations in the precuneus, bilateral middle frontal gyrus,

posterior cingulate cortex (PCC), bilateral superior occipital cortex, bilateral somatosensory cortex, and rostral anterior cingulate cortex (rACC) **(Figure 3-3, Table 3-4).**

ROI analyses confirmed suprathreshold activations in all predicted brain regions. The direction of these activations was consistent with the results of the whole-brain analysis, with the contrast [$Smiles_{All}$ > NE] revealing significant voxels in all ROIs, except for the precuneus and mOFC, which showed greater activations for the reverse contrast **(Supplementary materials S4)**.

Prediction 2c: Neural mechanisms supporting the correct identification of genuine vs. posed smiles

At the whole-brain level, the correctly identified genuine smiles, compared to the correctly identified posed smiles (i.e., Gen_{Hit} > Pos_{Hit}), were associated with stronger activations in the brain areas previously implicated in cognitive empathy or mentalizing, such as the mPFC, posterior cingulate cortex (PCC), and precuneus (PC). Lastly, the secondary somatosensory cortex and putamen, which are involved in sensorimotor simulation and facial mimicry (Ross and Atkinson, 2020), showed greater activations for genuine vs. posed smiles. No clusters turned out significant for the reverse contrasts (i.e., Pos_{Hit} > Gen_{Hit}) **(Figure 3-4, Table 3-4).**

ROI analyses comparing [Gen_{Hit} > Pos_{Hit}] reavealed the significant activations in the brain areas involved in ToM (i.e., mPFC, precunues, and rTPJ), and emotion processing (i.e., NAcc, caudate, mOFC, and amygdala). No significant activations were identified in the ROIs implicated in dynamic face perception (i.e., IFG, right STS) and affective empathy (i.e., MFC/dACC, and AI). The reverse contrast revealed no significant activations. **(Supplementary material S4).**

To gain further insights into the brain areas that were specifically associated with the individual differences in task performance, we explored whether activations within any of these brain areas were linearly tracking the sensitivity and decision bias parameters calculated from the signal detection analysis. To this end, subject-specific *d'* and C values were mean-centered and included in the GLM as continuous covariates. The results showed that the level of activations in the right IFG for the contrast [$Smiles_{All}$ > NE] were positively associated with individual participants' *d*' values (r=.560, p<.001) **(Figure 3-5a**). In comparison, variations in the decision bias were represented in the dACC and its neighboring mPFC regions, with larger contrast estimates from [Gen_{Hit} > Pos_{Hit}] predicting more conservative decision bias (r=.502, p<.001) **(Figure 3-5b)**.

Prediction 2d: *OXTR* and the neural correlates of smile authenticity judgment

At the whole brain level, no significant difference between G homozygotes and A carriers was found for the main contrasts of interests (i.e., [$Smiles_{All}$ > NE], [Gen_{Hit} > Pos_{Hit}]). However, ROI analyses revealed significant genetic modulation in the mPFC and rTPJ. Contrary to our prediction that G homozygotes would show increased activations in the brain areas implicated in smile authenticity judgments, the A allele carriers showed greater average contrast estimates for [Gen_{Hit} > Pos_{Hit}] than did G homozygotes in the mPFC **(Figure 3-6a).** The A allele carriers also exhibited stronger activation in the rTPJ for the contrast [$Smiles_{All}$ > NE] compared to G homozygotes **(Figure 3-6b)**.

Behavioral and personality correlates of the *OXTR* effects

Given the significant genetic modulation within the rTPJ and mPFC, we examined whether the activations in these ROIs correlated with participants' task performance, especially the significant intergroup difference in decision bias (i.e., C) found in the behavioral data analysis. Our exploratory correlation analyses found that activations within the mPFC for the contrast [Gen_{Hit} > Pos_{Hit}] were significantly associated with individual participants' decision bias (r = .379, p = .011), with the larger contrast estimates linked with more conservative responses **(Figure 3-7a)**. Between the two signal detection parameters that are theorized to influence decision bias (i.e., Hit rate and False alarm rate), the contrast estimates within the mPFC showed a significant negative correlation with the false alarm rate (r = -.329, p = .029). Activations in the rTPJ were not associated with decision bias.

We also tested the associations between a personality trait variable and the contrast estimates (i.e., [Gen_{Hit} > Pos_{Hit}]) within the mPFC and rTPJ to explore the possible psychological processes that may be associated with the activations within these ROIs. We focused on the perspective-taking subcomponent of IRI, as both ROIs are widely implicated in mentalizing (Schurz et al., 2014, Dricu and Fruholz, 2016; Saxe et al., 2003; Gallagher and Frith, 2003). The results showed that the activations within the mPFC for [Gen_{Hit} > Pos_{Hit}] positively correlated with perspective taking (r = .456, p = .002) **(Figure 3-7b),** suggesting that mentalizing could be a one cognitive pathway through which the genetic variations in *OXTR* modulate participants' decision bias in authenticity judgments.

Discussion

Brain activations associated with the perception of genuine and posed smiles

Face perception in humans is not only subserved by the "core neural system" that encodes low-level facial features, but also by the "extended system" that integrates non-perceptual information such as emotion or intention into the perceptual analysis (Haxby, Hoffman, and Gobbini 2000, Duchaine and Yovel 2015). Consistent with this model, as well as our predictions (i.e., **Prediction 2a, 2b**), the contrast between all types of smiles and control stimuli (i.e., [$Smile_{All}$ > NE]) yielded activations that encompassed both the core- and extended neural system for face perception. First, we found the multiple brain areas in the ventral- (e.g., the lateral occipital cortex, occipital fusiform gyrus, and anterior temporal cortex) and dorsal face processing pathway (e.g., the posterior and anterior superior temporal gyrus, and inferior frontal gyrus) (Duchaine and Yovel 2015). The involvement of these regions during the smile authenticity judgment task is not surprising**,** as decoding facial emotions depend on the perceptual analysis of both invariant facial features (Beaudry et al. 2014) and movements in eyes and mouth (Bassili 1979, Atkinson and Adolphs 2005), which are primarily carried out in the ventral and dorsal stream, respectively. Second, we found significant activations in the subcortical structures involved with approach motivations, arousal, and reward processing such as the bilateral amygdala, caudate nucleus, ventral striatum, and midbrain (e.g., VTA). Activations in these regions are likely to reflect participants' motivational and emotional reactions to smiling faces (Strathearn et al. 2008, Bhanji and Delgado 2014), which can be automatically triggered by highly salient social stimuli such as smiling or angry faces, irrespectively of explicit task goals (Lebreton et al. 2009, Adolphs and Spezio 2006). Lastly, genuine and posed smiles elicited greater BOLD signals in the cortical areas that have previously been implicated in affective (e.g., AI, IFG, and SMA) and cognitive empathy (e.g., rTPJ, and dmPFC) (Fan et al. 2011). Evidence suggests that understanding others' emotional states

require both affective mirroring based on perceptual cues (e.g., facial expressions), as well as top-down efforts to form the mental representation of others' inner states and beliefs (De Waal and Preston 2017). The joint recruitment of affective and cognitive empathy has also been found in studies where participants needed to understand ambiguous conversation between individuals (Mathersul, McDonald, and Rushby 2013), comics that can be interpreted in multiple ways (Völlm et al. 2006), and face and body stimuli depicting mixed emotions (Amting et al. 2009). Similarly, since genuine and posed smiles in this study were both positively valenced yet could be linked with the opposite intents (e.g., affiliation vs, deception), it is possible that participants relied on both affective and cognitive empathy to successfully discern the true social signals.

Brain activations associated with the correct identification of genuine vs. posed smiles

To identify brain activations more specifically associated with the discrimination of genuine vs. posed smiles, we compared the BOLD responses to the genuine vs. posed smiles that were correctly identified by participants (i.e., $Gen_{Hit} > Pos_{Hit}$). Consistent with **Prediction 2c**, we found activation in brain regions that partially overlapped with the results of the previous contrast. These areas included bilateral putamen, left secondary somatosensory cortex, precuneus (PC), posterior cingulate cortex (PCC), and medial orbitofrontal cortex (mOFC), with the significant clusters extending to the rostral ACC and mPFC.

Activations in the putamen and the secondary somatosensory cortex (S2) are often found in experimental tasks involving face processing or emotion recognition. One of the well-known mechanisms through which these two structures support social perception is facial mimicry, or involuntary activation of muscles upon viewing facial expression (Likowski et al. 2012). Many

human neuroimaging studies have reported the activation in the putamen during facial mimicry of basic emotions such as happiness, fear, and disgust (Rymarczyk et al. 2019, Iwase et al. 2002). Likewise, the S2, an important node of the extended mirror neuron network in the human brain (Pineda 2008, Keysers 2009), has been shown to contribute to emotion recognition through a somatic simulation of perceived facial expression (Pitcher et al. 2008, Hussey and Safford 2009). Facial mimicry of emotions is believed to be crucial for our ability to experience and comprehend emotions (Wood et al. 2016). The most dramatic examples of the sensorimotor grounding of our emotional competence comes from clinical literature. For example, neuronal degeneration or lesions in the basal ganglia structures including the putamen is known to hamper patients' capacity for emotional expression (Calder et al. 2000) and emotion perception (Prenger and MacDonald 2018). Repetitive transcranial magnetic stimulation (rTMS) applied on the somatosensory cortices is also known to disrupt recognition of facial expressions or emotional prosody (Van Rijn et al. 2005). Facial mimicry is pivotal to smile authenticity judgments. Perception of genuine smiles activates the Duchenne markers including zygomaticus major (ZM) and orbicularis oculi (Korb et al. 2014). When such a muscle movement is physically inhibited, people's ability to distinguish genuine smiles from posed smiles declines (Rychlowska et al. 2014). These findings strongly suggest that facial mimicry triggered by genuine smiles might have played an important role in the accurate facial authenticity judgments in this study.

Activations in the PC, PCC, and mPFC have been mostly interpreted with respect to cognitive empathy, or theory of mind (ToM). Increased recruitment of the ToM network is thought to aid emotion recognition and face processing, which often requires explicit mental state inference and processing of socially relevant information such as eyes, faces, and body (Mitchell and Phillips 2015). In the specific context of smile authenticity judgments, one study has shown the similar

recruitment of mPFC and precuneus when participants made distinctions between "Real vs. Posed" laughter (McGettigan et al. 2015). Our results add to this finding and suggest a general role of ToM in determining the authenticity of emotional expressions, irrespectively of specific modalities.

Two cautionary notes regarding the interpretation of our findings are worth mentioning. First, as with many other brain areas, the PC, PCC, and mPFC are involved with many functions in social and non-social cognition and the recruitment of these regions are commonly found outside the context of ToM or emotion perception. Thus, in principle, the increased activations in these cortical midline structures may not be attributable to mentalizing. However, our interpretation is supported by exploratory analysis that found positive linear associations between participants' self-report ratings on perspective taking (PT) and the contrast estimates extracted from dmPFC. Therefore, it is plausible that our findings at least partially reflect ToM-related cognitive processes. Second, it is also worth pointing out that ToM is a multidimensional construct. In other words, different brain regions within the ToM network can play non-overlapping roles in social cognition and behaviors depending on specific experimental contexts (Schurz et al. 2014, Saxe and Powell 2006). Our data do not allow us to parse out functionally dissociable contributions of PC, PCC, and dmPFC to mentalizing, which should be addressed in future studies.

Activations in the mOFC may reflect the experience of reward associated with the perception of genuine vs. posed smiles. Smiling faces serve as positive reinforcers almost universally in human social interaction (Martin et al. 2017, Godoy et al. 2005). Similar to non-social rewards such as money or juice, smiling faces have been shown to independently incur or amplify the reward-related activations in the mesolimbic dopaminergic pathway including ventral striatum or medial orbitofrontal cortex (Lin, Adolphs, and Rangel 2012, O'Doherty et al. 2003). The reward value of

smile has been shown to be encoded in the brain even after a brief exposure (17ms), and continuously shape affiliative behaviors afterwards (Chen, Whalen, et al. 2015). Notably, meta-analytic evidence suggests that genuine smiles are typically perceived as more rewarding than posed smiles (Gunnery et al., 2016). Remarkably, it has also been found that people were willing to sacrifice the opportunity to receive monetary rewards to view genuine smiles, but not posed smiles (Shore and Heerey 2011). These findings suggest that increased activations in the mOFC is likely to reflect participants' experience of reward and value in response to genuine vs. posed smile.

Here, an intriguing question would be whether the activations within the mOFC were due to participants' *subjective* experience of rewards associated with the perception of authentic smiles, or to *objective* characteristics of the experimental stimuli besides smile authenticity, such as smile intensity. This point is relevant especially given that genuine smiles often involve stronger muscle movements and are perceived to be stronger than posed smiles (Gunnery and Ruben 2016). Yet, this is unlikely to have had a major influence on our data for several reasons. First, we matched the average intensity and duration of the smile videos across the two smile categories through a pilot study. Second, an overlapping cluster of activation in the mOFC was found in our exploratory analyses comparing the BOLD responses associated with 1) the genuine smiles categorized as genuine vs. posed (i.e., $Gen_{Hit} > Gen_{Incorrect}$), and 2) the posed smiles erroneously categorized as genuine vs. posed (i.e., $Pos_{Incorrect} > Pos_{Hit}$), although the effect size of the latter contrast was weaker **(Supplementary Materials S3-5).**

Neural indices of individual difference in smile authenticity judgments

As we confirmed multiple brain areas involved with the perception and correct identification of genuine vs. posed smile, we also asked how and where participants' overall perceptual accuracy is represented in the brain. Our analyses based on a signal detection framework provided insights into this question.

While participants in the behavioral or neuroimaging arm successfully discriminated the authenticity of smiles **(Prediction 1a),** significant individual differences were found for both perceptual sensitivity (i.e., *d'*) and response bias (i.e., C). We found that activations in the right IFG (i.e., pars opercularis) during the presentation of genuine and posed smiles, as opposed to the control faces, were positively correlated with the *d*'. The IFG is considered a critical node in the putative human mirror neuron system (Molenberghs, Cunnington, and Mattingley 2009, Hecht et al. 2013). It is recruited during both action observation and execution (Caspers et al. 2010, Keysers 2009), which may form a sensorimotor grounding for action understanding, imitation, and empathy (Rizzolatti and Craighero 2004, Shamay-Tsoory et al. 2009, Shamay-Tsoory 2011). Virtual lesion of the IFG via transcranial magnetic stimulation (TMS) or transcranial direct current stimulation (tDCS) has been shown to impair interpersonal motor synchrony (Enticott et al., 2012), emotion recognition (Keuken et al. 2011), empathy for pain (Li et al. 2021), and most relevantly, smile authenticity judgments (Paracampo et al. 2017). These findings strongly suggest that variation in the smile authenticity judgment task performance in this study was contingent on the neural mechanism that allows individuals to mirror and translate other's emotional expressions into their own internal affective states.

Individual difference in the response criteria (i.e., C) was represented largely in the anterior dACC, with the cluster extending towards the dorsal part of the mPFC. Specifically, those who adopted

conservative response criteria showed greater activations in the dACC for the genuine smiles relative to the posed smiles. One possible interpretation of this result is that the activations within the dACC in this study reflected the degree of response conflict that participants experienced while determining the authenticity of the smiles. The dACC has been widely implicated in conflict monitoring and cognitive control (Botvinick, Cohen, and Carter 2004) in the experiment settings where people must choose from multiple, mutually incompatible, yet equally permissible response options (Ebitz and Hayden 2016). For instance, the dACC has been shown to increase activation when participants had to select between sensory or semantic stimuli with opposing valences (Wittfoth et al. 2009, Nohlen, van Harreveld, and Cunningham 2019), conflicting social categories (Stolier and Freeman 2017), and facial expressions depicting ambiguous affective states (Ito et al. 2017). Similar to these studies, participants in our experiment also had to categorize a series of smile stimuli into one of two competing perceptual categories. This design feature might have been a major source of response conflict, especially for those with conservative decision criteria who tend to search for stronger visual evidence of signal (i.e., genuine smile) before making a decision (Zehetleitner and Mueller 2010).

The dACC is also involved with a wide variety of cognitive and emotional processes beyond response conflict, such as error/gap monitoring (Critchley et al. 2005) or pain processing (Eisenberger and Lieberman 2004). Yet, participants were not provided with any performance feedback which would have elicit error-related activities in the dACC or its surrounding medial prefrontal cortex (e.g., prediction error or error-related negativity; (Joiner et al. 2017, Charles et al. 2013). In addition, our task did not induce either direct or vicarious pain among participants. Therefore, while it is often challenging to determine specific cognitive or psychological properties

underlying dACC activation, these alternative functions are not likely to have contributed substantially to the results of this study.

The association between OXTR rs53576 and smile authenticity judgments

Lastly, we explored the relationships between the allelic variation in *OXTR* SNP rs53576 and the behavioral and neural correlates of smile authenticity judgments. Based on the previous studies that implicated the G allele of rs53576 in sensitive social cognition, we predicted that G homozygotes would show better behavioral performances, and enhanced activations in the brain regions that subserve correct identification of genuine vs. posed smiles.

Overall, we did not find conclusive evidence that the *OXTR* genotype modulated the behavioral task performance in the smile authenticity judgment task **(Prediction 1b)** or the brain responses to genuine vs. posed smiles **(Prediction 2d)** in the predicted directions.

At the behavioral level, G homozygotes in the neuroimaging arm endorsed more *liberal* response criteria and showed higher false alarm rates than did the A allele carriers. That is, G homozygotes were more likely to erroneously judge posed faces to be genuine. This effect cannot be explained by baseline intergroup differences in personality traits or demographic characteristics, as two genotype groups were comparable in terms of those variables. One possible explanation for this unexpected finding is that the G homozygotes may have had increased approach and affiliative motivations, which could introduce positive biases in social perception and behaviors. Previous evidence suggests that people administered with INOT tended to show increased risk tolerance and decreased sensitivity towards negative social outcomes. For instance, INOT has been linked with

reduced aversion towards negatively conditioned social cues (Petrovic et al. 2008) and prolonged interpersonal trust following defection (Baumgartner et al. 2008). These findings indicate that OT alters people's social perception and behaviors such that they could engage more readily in affiliative interactions with others (Bartz 2016). Of relevance to the face authenticity judgment, Pfundmair et al (2017) and colleagues found that INOT treatment interfered with correct discrimination between lies and truth statements in both males and females. Moreover, this effect was driven by the increased false alarm, or incorrectly judging lies to be truth statements (Pfundmair, Erk, and Reinelt 2017). Similarly, our results that G homozygotes exhibited more liberal decision bias may also reflect the "gullible interaction style" associated with endogenous OT signaling (Pfundmair, Erk, and Reinelt 2017).

At the neural level, we found significant genetic modulations within two of our main ROIs: rTPJ and mPFC. The significant activations in the rTPJ emerged when the effect of all smiles was compared with that of the control stimuli (i.e., AA[$Smile_{All}$ - NE]- GG[$Smile_{All}$ - NE]). Genetic modulation within the mPFC emerged when the comparison was made between the trials where participants correctly discerned genuine vs. posed smiles (i.e., AA[G_{Hit} - P_{Hit}] – GG[G_{Hit} - P_{Hit}]). Intriguingly, the A allele carriers showed greater average contrast estimates than did G homozygotes in both ROIs. Given that the rTPJ and mPFC are parts of the ToM network that play critical role in accurate mental state inference and the allocation of attentional resources to salient social targets (Frith and Frith 2006, Young, Dodell-Feder, and Saxe 2010, Krall et al. 2015), the relatively subdued activations within these ROIs among G homozygotes may be linked with the liberal decision bias exhibited by G homozygotes. Consistent with this interpretation, our exploratory analysis indeed revealed that mPFC activations that showed maximum genetic

modulation positively correlated with both perspective taking and participants' decision bias, with lower contrast estimates associated with low perspective taking and more liberal response criteria.

Our behavioral and imaging genetics analyses point to the possibility that genetic variation in the *OXTR* may modulate smile authenticity judgments via liberal decision bias, which was represented in the activations in the mentalizing network. Here, it is crucial to note that the association between *OXTR* genotype and response criteria only turned out significant in the neuroimaging arm. Therefore, our results should be interpreted with caution and considered tentative until further replication is made.

Chapter Summary and Conclusion

Smiles are versatile tools that can serve either affiliative or deceptive goals in human social communication. This study aimed to investigate the neural mechanisms underlying our ability to discern genuine smiles from posed smiles. Employing an imaging genetic approach, we also explored the possible link between a single nucleotide polymorphism (SNP) in the oxytocin receptor gene (*OXTR*) rs53576 and smile authenticity judgments. Our analyses showed that multiple brain areas that have previously been implicated in dynamic face processing, reward encoding, motor mimicry, affective and cognitive empathy were recruited during the perception and correct identification of genuine vs. posed smiles. Especially, activations within the inferior frontal gyrus (IFG) and dorsal anterior cingulate cortex (dACC) linearly tracked the perceptual sensitivity (i.e., *d*' prime) and response criterion (i.e., C) of individual participants, respectively. Albeit preliminary, we also found that individuals homozygous for the G allele of *OXTR* rs53576

tended to categorize posed smiles as genuine, which was represented in the decreased activations within the medial prefrontal cortex (mPFC) and right temporoparietal junction (rTPJ).

Although the neural correlates of human smiles have been extensively studied in various experimental contexts, only few attempts have been made to date to examine the neurocognitive mechanisms underlying smile authenticity judgments. To our knowledge, the present study is the first to address this gap using an fMRI task and dynamic facial expressions of smiles as experimental stimuli. We also extended the scope of existing literature by investigating the possible role of *OXTR* in smile authenticity judgments. Although our findings could illuminate the multiple brain mechanisms recruited during the smile authenticity judgments, few outstanding questions and methodological limitations merit a mention. First, it is not clear whether our behavioral and neuroimaging results can be generalized into the authenticity judgments for other basic emotions such as anger and sadness (Krall et al. 2015). Some preliminary evidence exists as to the emotion-specific effects on the authenticity judgments (McLellan et al. 2012, Mega, Gigerenzer, and Volz 2015), but these data were obtained based on small sample size (e.g., $N = 6$) (McLellan et al. 2012) or still images (Mega, Gigerenzer, and Volz 2015). These methodological shortcomings limit the generalizability and ecological validity of the findings, which leaves room open for future studies. Second, while our data strongly suggest that motor mimicry and mirroring mechanisms are important for smile authenticity judgment, the current study lacked independent measures to directly test the actual recruitment of those processes during the task. This could be addressed with multi-modal neuroimaging, as in some previous studies that combined EMG with fMRI task paradigm, to investigate the temporal and functional coupling between facial muscle movements and BOLD signals (Likowski et al. 2012, Rymarczyk et al. 2019). This approach, while not uncommon in emotion research, has not yet been applied specifically to study smile authenticity

judgments and is thus worth pursuing. Finally, an additional investigation would be necessary to understand the link between *OXTR* and smile authenticity judgments. While our results were explicable with the known function of OT and social behaviors (e.g., social affiliation and behavioral approach), the observed genetic modulation was at best modest and not consistent across the neuroimaging and behavioral arms. Future studies should involve 1) larger sample sizes and combine 2) multiple *OXTR* SNPs to examine the robustness of our findings.

Figures and Tables

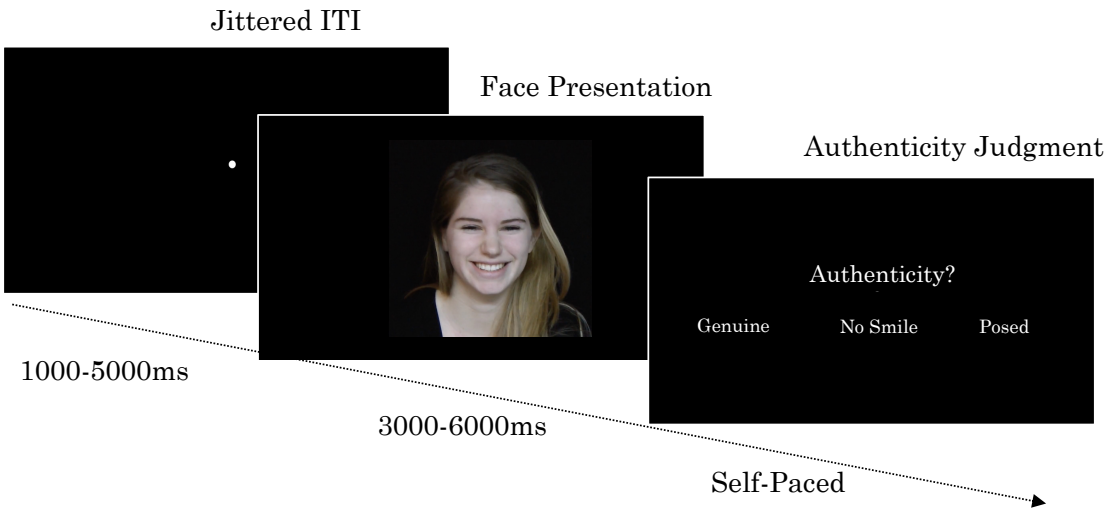


Figure 3-1. The schematic representation of the smile authenticity judgment task.

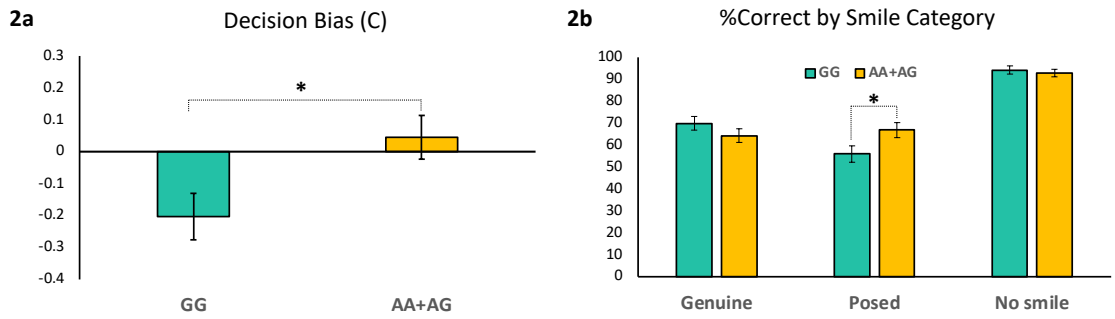


Figure 3-2. The intergroup difference between decision bias (i.e., response criterion, C) (2a), **and the average %Correct for each smile category (2b)**. The means are adjusted for participants' Age and NfC. Error bars indicate standard errors (SEM). *p<.05 (Bonferroni-corrected).

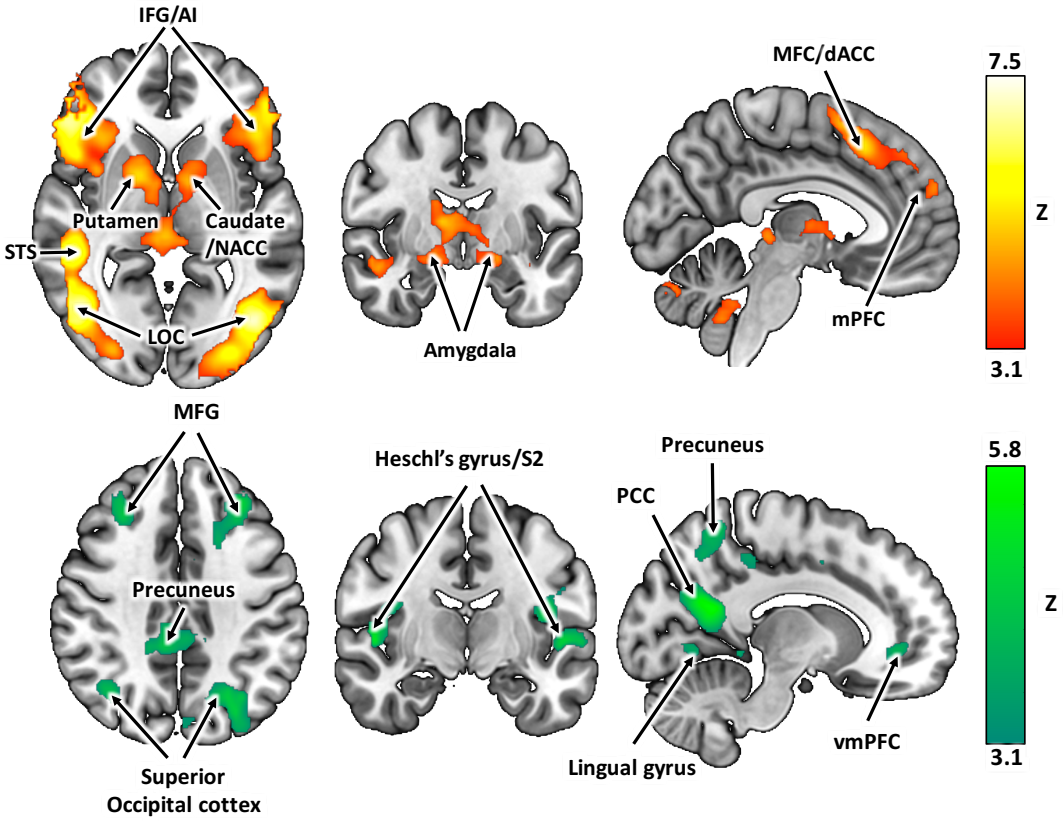


Figure 3-3. Brain activations associated with correct identification of genuine or posed smiles vs. neutral expressions. The top images (Red-Yellow) and bottom (Blue-Green) represent the contrast [Gen_{Hit} + Pos_{Hit} > Neu] and its reverse, respectively. All output images were shown at the cluster-forming threshold of Z > 3.1 (voxel-level p<.001), and cluster-level FWE corrected p<.05 (Inferior frontal gyrus, IFG; Superior temporal sulcus, STS; Nucleus accumbens, NAcc; medial frontal cortex/dorsal anterior cingulate cortex, MFC/dACC; Secondary somatosensory cortex, S2; medial prefrontal cortex, mPFC; middle frontal gyrus, MFG; posterior cingulate cortex, PCC; medial orbitofrontal cortex, mOFC).

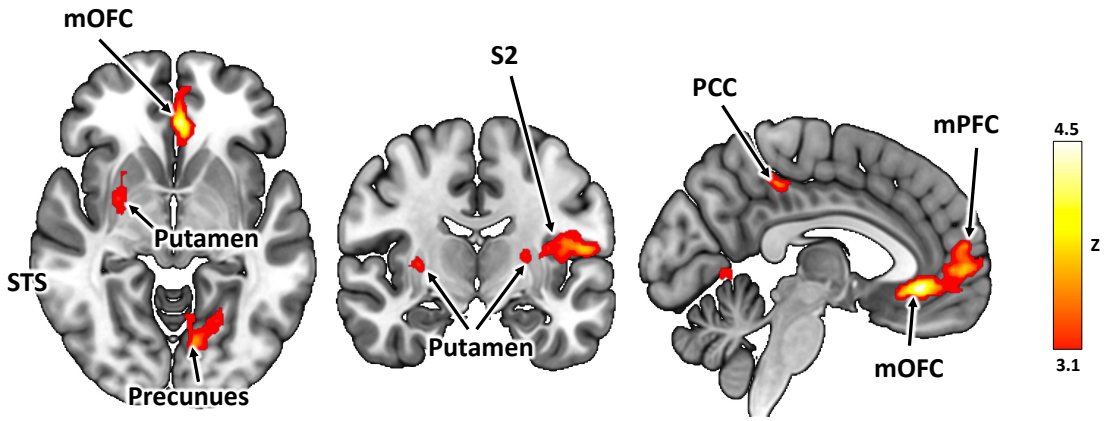


Figure 3-4. Brain activations associated with correct identification of genuine vs. posed smiles (i.e., Gen_{Hit} > Pos_{Hit}). The reverse contrasts reveal no active clusters. All output images were shown at the cluster-forming threshold of Z > 3.1 (voxel-level p<.001), and cluster-level FWE corrected p<.05 (medial prefrontal cortex, mPFC; lateral occipital cortex, LOC; posterior cingulate cortex, PCC).

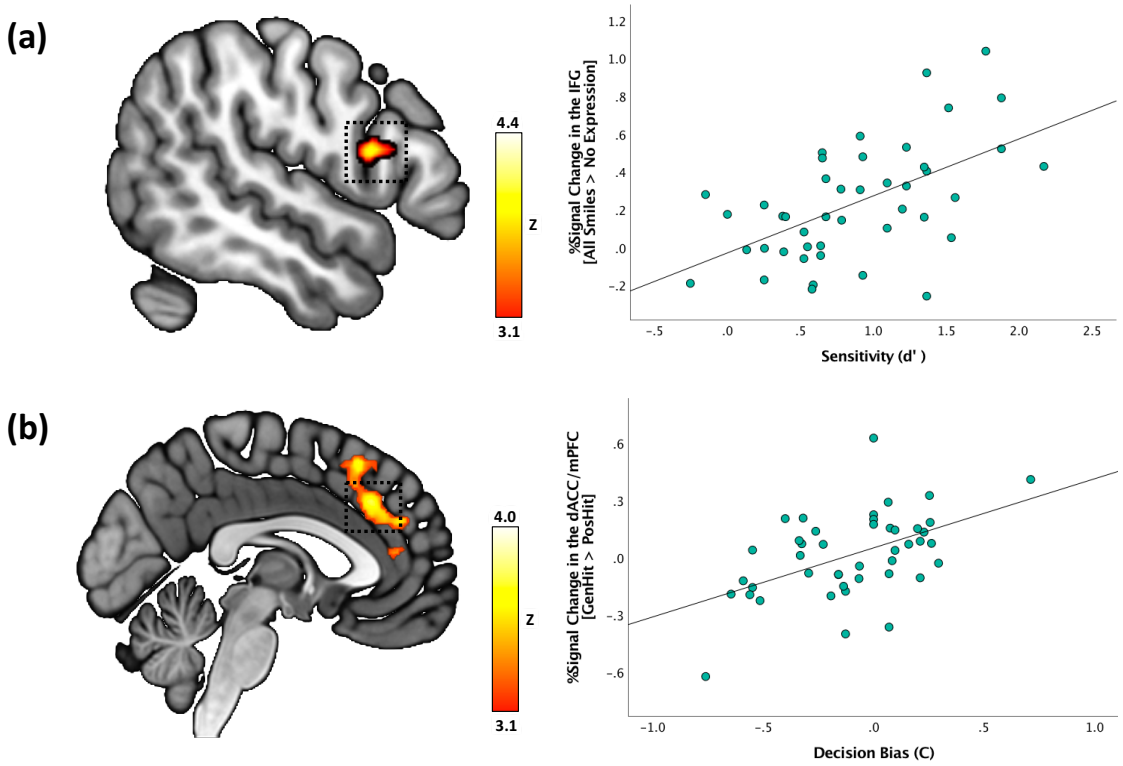


Figure 3-5. Associations between signal detection parameters and activations in the IFG and dACC/mPFC. BOLD signals within the dACC/mPFC significantly correlated with individual differences in sensitivity **(a)** and decision bias **(b)**, respectively. The results are thresholded with whole-brain, cluster level (Z>3.1), FWE-corrected $p<.05$. Activation values were extracted from the peak voxel coordinates within each ROI for plotting (dorsal anterior cingulate cortex, dACC; medial prefrontal cortex, mPFC).

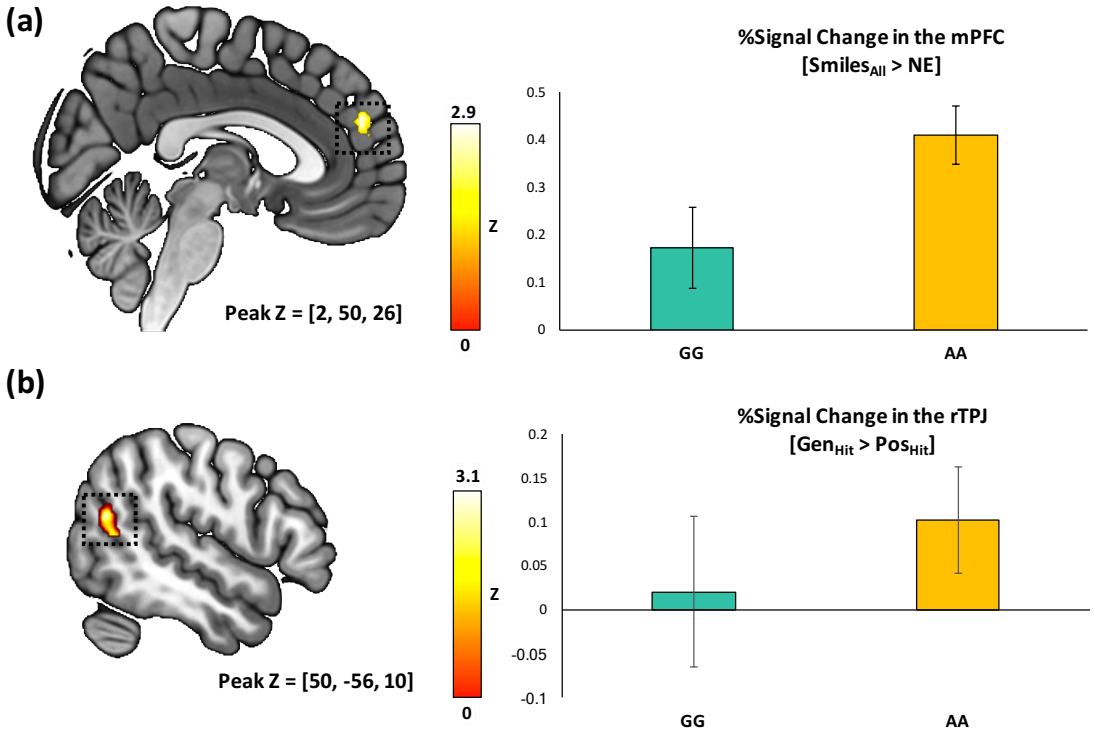


Figure 3-6. The results of ROI analyses showing significant genetic modulations of the activations in the TPJ and mPFC. The *OXTR* genotype significantly modulated neural activation within the mPFC **(a)** and rTPJ **(b)**. The results are thresholded with small volume correction at voxel-wise, FWE-corrected $p<.05$. Activation values were extracted from the peak voxel coordinates within each ROI for plotting (medial prefrontal cortex, mPFC; right temporoparietal junction, rTPJ).

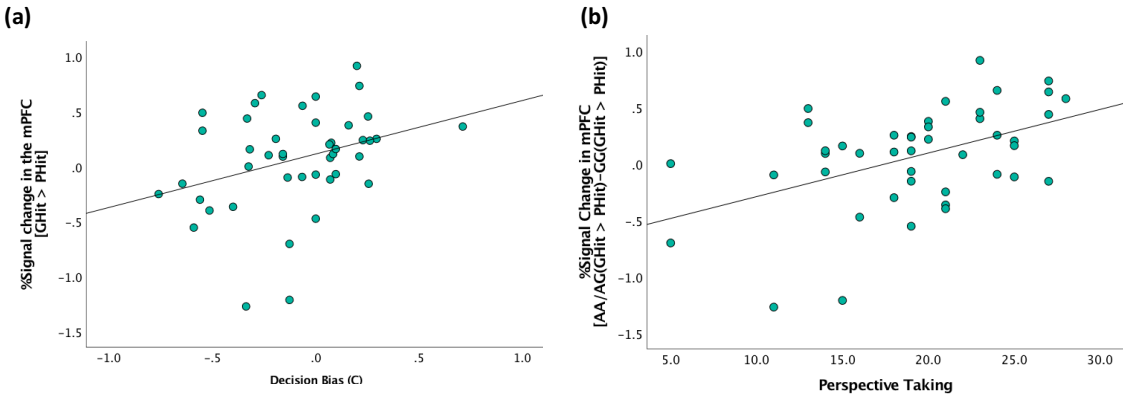


Figure 3-7. Brain-behavior correlations found in the mPFC. Conservative response criteria were associated with the greater differences in BOLD signals within the mPFC for the authenticity judgments for genuine vs. posed smiles **(a)**, which was also associated with perspective taking **(b)**.

Table 3-1. Demographics and genotype composition of the study sample

Demographics	**Neuroimaging Condition**		**Behavioral Condition**	
	N	*%*	*N*	*%*
***OXTR* Genotype**				
GG	19	42	42	29.8
AA/AG	25	57	99	70
Gender				
Female	26	60	89	37
Male	18	40	52	63
Ethnicity				
Asian	20	44	60	42
African-American	5	11	27	19
Caucasian	17	40	35	24
Hispanic	2	5	17	12
Others	-	-	2	3

Table 3-2. Results of independent sample *t*-tests on personality and demographic traits between the *OXTR* genotype groups in the behavioral arm.

Variables	**GG (*N*=42)**		**AA+AG (*N*=99)**		$t_{(139)}$	*p*
	M	SD	M	SD		
Age	23.6	6.7	23.7	7.7	-.071	.944
Empathic concern	19.3	3.6	19.8	3.4	-.770	.443
Perspective taking	18.1	4.7	19.7	4.2	-1.909	.058
NfC	3.6	.5	3.6	.7	.913	.363
FNE	37.1	11.7	40.2	11.3	-1.439	.152
SM	13.3	3.7	13.4	3.1	-.355	.723

Table 3-3. Results of independent sample *t*-tests on personality and demographic traits between the *OXTR* genotype groups in the neuroimaging arm.

Variables	**GG (*N*=19)**		**AA+AG (*N*=25)**		$t_{(42)}$	*p*
	M	SD	M	SD		
Age	22.2	4.6	20.7	3.9	1.191	.240
Empathic concern	18.6	3.8	19.6	3.7	-.956	.344
Perspective taking	17.9	5.9	20.0	5.2	-1.403	.168
NfC	3.9	.4	3.5	.7	1.924	.061
FNE	38.0	11.2	35.6	9.5	.813	.420
SM	13.7	4.2	13.7	4.7	.417	.679

Table 3-4. Summary of the whole-brain activations associated with the successful identification of genuine, posed, and neutral smiles.

Main Contrast/ Brain region	**Cluster size (voxels)**	**Max Z value**	**Peak Voxel MNI coordinate**		
			X	**Y**	**Z**
$Smiles_{All}$ > No Expression (NE)					
Right LOC*	18,482	7.57	50	-64	6
Left IFG**	3,865	6.39	-42	36	0
Thalamus/Striatum***	3,750	5.91	12	-12	-10
MFC/dACC****	2,468	6.37	2	8	64
Cerebellum	282	4.93	2	-52	-34
$Smiles_{All}$ < No Expression (NE)					
Left Precuneus/PCC	6,311	5.84	-14	-58	20
Left Heschl's gyrus*****	1,130	4.82	-50	-8	0
Left MFG	995	5.54	-30	32	36
Right MFG	817	4.95	28	12	56
Left Lingual gyrus	761	5.49	-30	-42	-8
Right Heschl's gyrus	604	3.72	50	-8	2
Right Lingual gyrus	413	4.76	30	-38	-12
Left mOFC/rACC	372	4.27	-16	34	-10

Right superior occipital cortex	340	4.79	38	-72	34
Gen_{Hit} > $Posed_{Hit}$					
mPFC	674	4.31	-10	58	0
Left Lingual gyrus	659	4.45	-14	-72	-8
Right Precuneus	493	4.08	12	-56	10
Right Putamen	419	4.02	30	-12	4
Left STS/Central operculum	361	4.00	-54	-40	6
Left LOC	339	4.50	-40	-74	8
PCC	178	3.75	-4	-30	42
Left primary somatosensory cortex	174	3.91	-50	-20	44

*Cluster extending to the right STS and middle temporal gyrus; **Cluster extending to the left AI and ventrolateral prefrontal cortex; ***Cluster extending to the bilateral amygdala; ****Cluster extending to the mPFC; *****Cluster extending to the central operculum

Chapter 3 Supplementary materials

S3-1. Sample size determination and participant allocation strategy

A priori-power analysis

We used the G*Power to conduct a priori power analysis. The reference effect sizes were taken from previous studies that investigated neural (GG vs. AA, Cohen's d =.81) and behavioral effects (GG vs. AA+AG, Cohen's d = .49) of OXTR rs53576 on social cognition involving face and emotion perception (Rodrigues et al. 2009, Luo, Li, et al. 2015). With the type-I error rate set to α = .05, the power analysis showed that a total of $N = 50$ (e.g., 25 GG and 25 AA+AG) are required to provide 80% power for detecting a significant main effect of genotype on the neural response associated with socio-emotional processing. For the behavioral task, the power analysis yielded a required sample size of $N = 144$ (GG=53, AA+AG = 91). In sum, by recruiting 200 participants, the current project is expected to have sufficient statistical power for detecting true effects at both neural and behavioral levels.

Participant allocation strategy

Before the COVID-19 pandemic, (~March 2020) participants were pseudo-randomly assigned to the neuroimaging or behavioral arm based on their genotype (i.e., OXTR rs53576, G/A). That is, genotyping was performed prior to the group assignment. As it was more difficult to recruit participants for the neuroimaging arm due to the additional screening criteria, those homozygotes for the A or G allele were prioritized to be included in the neuroimaging arm of the study

whenever possible. However, the experimenter was blind to the specific genotypes of participants.

To facilitate data collection amid a COVID-19 pandemic, the recruitment protocol was modified such that participants were randomly assigned to either the neuroimaging or behavioral arm regardless of their genotypes. Genotyping was performed after neuroimaging/behavioral data collection. The experimenter remained blind to the specific genotype of participants.

S3-2. Pilot Experiment for Stimuli Selection

All stimuli were drawn from the UvA-NEMO database. We first identified the face models for which both spontaneous and posed smile stimuli were available. Only the spontaneous and posed smiles that were equivalent in terms of the gaze directions, head and shoulder movements, and the exposure of teeth following smile onset were considered. A total of 158 smiles from 79 actors were selected from this initial screen process.

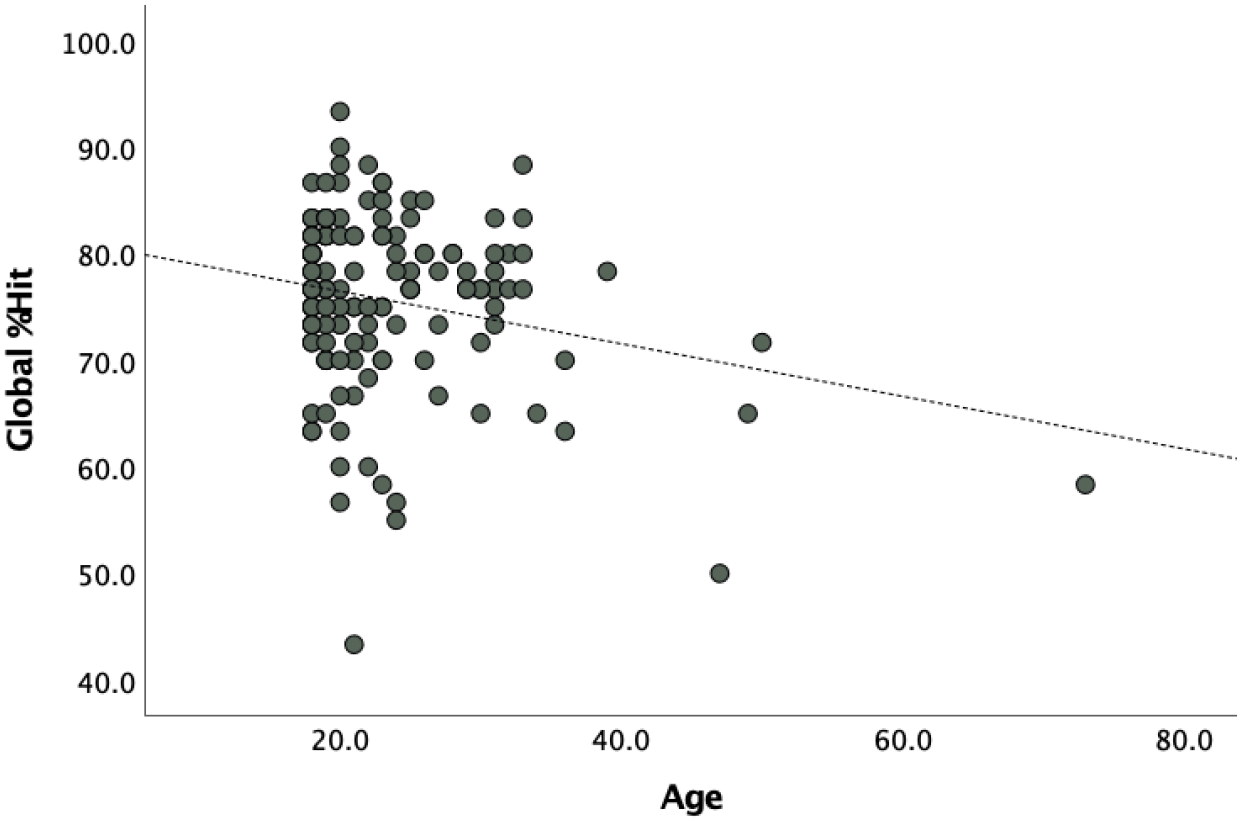
Then, a separate group of 25 adults were recruited from Amazon Mechanical Turk (https://www.mturk.com). The participants in the pilot study were directed to a separate website (https://www.surveymonkey.com/r/VV8SRCX) where they 1) determined the smile authenticity of the stimuli, and 2) evaluate them in terms of the attractiveness of the face models and the intensity of facial muscle recruitment. This procedure was to ensure that smile authenticity judgments were not biased by extraneous factors. Participants who completed the rating process received 12 US dollar as compensation.

The smiles videos with the average hit rate below chance (i.e., 50%) were excluded. For the remaining stimuli, we selected 40 stimuli from 20 face models with varying ages (Ranged between 14-65) and ethnicities. One genuine smile and posed smile were included for each face model to minimize the perceptual differences between the two stimulus categories (Genuine $N = 20$, Posed $N = 20$). The average lengths, %Hit rate, attractiveness, and intensity ratings for the genuine and posed smiles were matched, as shown below.

Table S3-2-1. Results of pilot ratings for genuine and posed smile stimuli

Characteristics	Genuine		Posed		*t*(19)	*p*
	M	*SD*	*M*	*SD*		
Lengths	3.40	.88	3.2	.89	1.00	.330
%Hit	66.3	12.7	59.8	15.7	1.44	.166
Intensity	2.78	.63	2.92	.57	-1.23	.235
Attractiveness	55.2	9.7	53.6	11.4	1.51	.147

S3-3. Linear association between participants' age and task performance in the behavioral condition

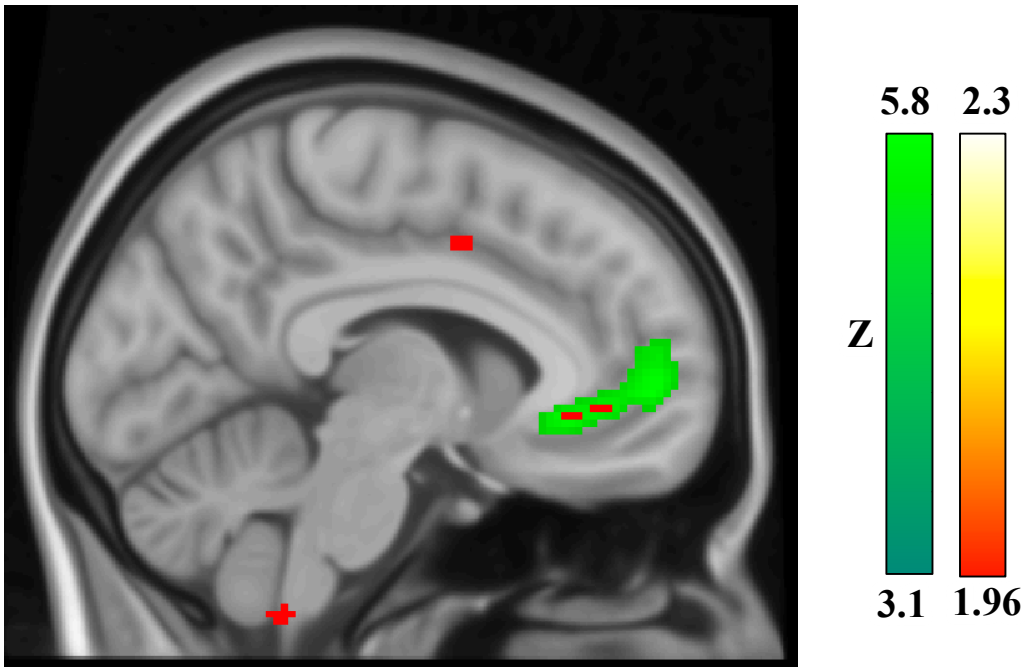

100.0
90.0
80.0
70.0
60.0
50.0
40.0
Global %Hit
20.0
40.0
60.0
80.0
Age

S3-4. The results of ROI analysis on the average activations for the contrast [$Smiles_{All}$ vs. No Expression] and [Gen_{Hit} vs. Pos_{Hit}]

Main Contrast /Brain region	Cluster size (voxels)	Max Z value	Peak Voxel MNI coordinate		
			X	Y	Z
$Smiles_{All}$ > No Expression					
NAcc (L)	150	5.31	-10	-2	-2
NAcc (R)	69	4.63	12	8	0
Amygdala (R)	246	4.69	12	-12	-12
Amygdala (L)	175	4.77	-18	-6	-12
TPJ (R)	357	7.03	52	-58	10
mPFC (R)	140	5.88	4	58	26
STS (R)	122	6.41	56	-40	6
IFG (R)	461	6.86	50	20	24
IFG (L)	276	6.25	-52	20	14
AI (R)	171	6.24	42	30	-6
AI (L)	33	5.07	-44	22	0
dACC	199	5.31	-6	22	44
$Smiles_{All}$ < No Expression					
mOFC	58	3.33	8	40	-6
Precuneus	30	3.62	8	-60	26

Gen_{Hit} > Pos_{Hit}					
NAcc (L)	19	3.7	-2	12	-4
Caudate nucleus					
mOFC (R)	270	4.63	-4	36	-8
Amygdala (R)	27	3.45	-24	0	-10
Amygdala (L)	9	3.56	24	9	-10
Gen_{Hit} < Pos_{Hit}					
No activations					

S3-5. mOFC activations associated with the participants' subject perception of smile authenticity



Activations associated with [Gen_{Hit} > $Gen_{Incorrect}$] (Green), and 2) the posed smiles erroneously categorized as genuine vs. posed [$Pos_{Incorrect}$ > Pos_{Hit}] (Red). Significant activations were found in the overlapping region in the mOFC.

Chapter 4

Enhanced endogenous oxytocin signaling modulates neural responses to social misalignment and promotes conformity in humans: A multi-locus genetic profile approach

Chapter Abstract

In this chapter, I examined the neurocognitive mechanisms through which OT may promote social conformity. I used a multi-locus genetic profile score (MPS) defined from the allelic variant in 7 *OXTR* SNPs, including rs53576, that are linked with higher OT receptor expression in the human brain. Participants (Neuroimaging arm $N = 45$, Behavioral arm $N = 144$) played a novel card-sorting task where they evaluated the relative importance of two moral values presented in pairs. Conformity pressure was imposed in a form of opposing majority opinions. Whether and how much participants shifted their initial value preference following the majority opinions were analyzed with respect to the *OXTR* MPS. We found that:

1. Overall, participants showed behavioral conformity.
2. Both the NAcc and pMFC/dACC increased activations to social misalignment vs. social alignment, although these were not likely to be reinforcement learning signals.
3. Those with higher *OXTR* MPS, thus enhanced endogenous OT signaling, showed decreased pMFC/dACC activations in response to perceived social misalignment.
4. The dampening of the pMFC/dACC activation predicted stronger conformity.

These findings suggest that 1) the brain areas comprising the gap/error detection mechanisms contribute to the moral conformity via mechanisms distinct from prediction errors, and 2) enhanced OT signaling in the brain may increase conformity by amplifying the subjective value of social affiliation, which reduces the internal conflict associated with overriding personal moral preference. Lastly, this study points to the potential value of using MPS as a measure of region-specific OT signaling in the human brain.

Keywords: social conformity, moral values, *OXTR, multilocus genetic profile score*, pMFC/dACC, response conflict

Introduction

Oxytocin (OT) is a nonapeptide synthesized in the hypothalamus and released into the circulation via neurohypopheseal system. Most early studies on OT were centered on its roles in lactation and parturition in females (Churchland and Winkielman 2012). Yet, decades of research have revealed that OT signaling in the brain modulates a wide range of complex social behaviors and cognition beyond reproductive functions (Johnson and Young 2017). Especially, OT has been referred to as a "binding" or "herding" hormone (MacDonald and MacDonald 2010, Stallen and Sanfey 2015, Xu, Becker, and Kendrick 2019) due to its facilitative effects on social bonding and cohesion in mammalian species (Carter 2014).

Conformity, or modifying one's behaviors or beliefs to align with others (Cialdini and Goldstein 2004a), is one of the key mechanisms through which OT promotes social cohesion in humans(Shamay-Tsoory et al. 2019, De Dreu and Kret 2016). For instance, intranasal administration of OT (INOT) has shown to increase behavioral conformity in aesthetic preference judgment (Stallen and Sanfey 2015), facial attractiveness judgment (Huang, Kendrick, and Yu 2014), episodic memory retrieval (Edelson et al. 2015) and even dishonest actions (Aydogan et al. 2017). The effects of OT on conformity tend to be stronger when the social influence is imposed by in-group members (Stallen et al. 2012) in a competitive context (Aydogan et al. 2017). These findings suggest that OT may be an essential building block of our evolved group psychology that helps us navigate a complex social environment, especially through the rapid acquisition of knowledge, values and norms endorsed and performed by others in one's environment(De Dreu and Kret 2016, Xu, Becker, and Kendrick 2019).

The nascent literature on the link between OT and conformity, and its possible adaptive significance in human evolution, raises the question as to the proximate mechanisms that mediate such association: how does OT signaling in the brain affect behavioral conformity? Yet, our understanding of the specific neuro-cognitive processes underlying the OT-induced conformity effects is limited.

One possibility is that OT modulates the neural representations of the perceived social alignment and misalignment in a way that promotes conformity. Recent studies using functional magnetic resonance imaging (fMRI) have shown that the susceptibility towards social influence is closely linked with reward processing and error/gap detection mechanisms in the brain (Shamay-Tsoory et al. 2019, Wu, Luo, and Feng 2016). Specifically, the nucleus accumbens (NAcc), one of the key nodes in the mesolimbic dopaminergic pathway, showed decreased activations when there was a mismatch between the behaviors of self and others (Klucharev et al., 2009; Zaki et al., 2011; Izuma and Adolphs, 2013; Lin et al., 2018; Levorsen et al., 2021). By contrast, the same self-other gap elicited the increased activations in the posterior medial frontal cortex (pMFC) including the dorsal anterior cingulate cortex (dACC), which are involved in conflict monitoring (Klucharev et al., 2008; Izuma and Adolphs, 2013; Lin et al., 2018; Wei et al., 2018; Levorsen et al., 2021).

The exact psychological and computational properties of the activations within the NAcc and pMFC/dACC are debated (Izuma, 2013; Levorsen et al., 2021). Still, a widely-accepted view holds that they may reflect individuals' subjective experience of reward and conflict associated with the self-other alignment or the lack thereof, while also serving as reinforcement learning signals (e.g., prediction error, PE) (Klucherev et al., 2008(Shamay-Tsoory et al. 2019)). Consistent with this

interpretation, activations within the NAcc and pMFC/dACC have been shown to track the magnitude of self-other gap (Izuma et al., 2013) and the degree of subsequent behavioral conformity (Lin et al., 2018). A recent study has also found the spatial overlap between the PE signals measured in a reinforcement learning task (i.e., probabilistic learning task; Levorsen et al., 2021) and the neural representations of social misalignment within the NAcc and pMFC/dACC.

Then, how could OT interface with the reward processing and gap/error detection mechanisms in the context of social conformity? According to the social salience hypothesis (Shamay-Tsoory and Abu-Akel 2016), one of the main functions of OT is to enhance the salience of social stimuli, regardless of their specific valence (Shamay-Tsoory and Abu-Akel 2016). For instance, INOT treatment has been shown to modulate the neural responses to both positive and negative social cues in the human brain, especially within the brain regions implicated in visual attention such as the amygdala (Domes et al., 2007; Gamer et al., 2010), dACC (Scheele et al., 2014; Gorka et al., 2015; Li et al., 2015), and ventral tegmental area (VTA) (Scheele et al., 2013; Groppe et al., 2013). OT is also known to upregulate affiliative motivations and approach behaviors in humans (Bartz, 2016; Piva and Chang, 2018). Supporting this view, INOT treatment was shown to increase activations in the reward-sensitive areas in the brain including the caudate nucleus and ventral striatum in response to appetitive stimuli or reciprocated cooperation (Scheele et al., 2013; Gregory et al., 2015; Rilling et al., 2012).

The role of OT in enhancing social salience and affiliative motivations, when considered together with the neural underpinnings of conformity, points to the possibility that OT strengthens behavioral conformity by making the self-other alignment more salient and rewarding. This effect would be represented in the increased activations in the NAcc when there is social alignment

(Bahnji and Delgado, 2014; Lin et al., 2018; Nook and Zaki, 2015). Alternatively, but not mutually exclusively, OT may also enhance the neural encoding of error and conflict in the pMFC/dACC when discrepancy is found between self and others (Klucharev et al., 2009; Lin et al., 2018). Such alteration in the NAcc and pMFC/dACC activations could function as PE signals and subsequently lead to stronger conformity.

This study aims to investigate the neural mechanisms that mediate OT and behavioral conformity by testing this possibility. We specifically focused on the genetic variations in the oxytocin receptor gene (*OXTR*), which are known to influence various social phenotypes by regulating the OT receptor the expression in the brain (King et al., 2016; Reuter et al., 2020; Almeida et al., 2022).

In the context of social conformity, a single nucleotide polymorphism (SNP) *OXTR* rs53576 holds special relevance. Both empirical and meta-analytic evidence suggests that the G allele of rs53576 is associated with a wide array of social cognition and behaviors relevant for social conformity (Li et al., 2015). For example, individuals homozygous for the G allele have been shown to endorse culture-specific normative behaviors and psychological traits more strongly than the A allele carriers across different societies, potentially due to its role in facilitating the sensitivity towards evaluative social feedback (Kim et al., 2011; Kim et al., 2014; Kitayama et al., 2017). The G allele carriers also show heightened sensitivity towards social rewards (Feng et al., 2015; Damiano et al., 2014) and sensitive to social cues (Choi et al., 2017). Lastly, allelic variations in *OXTR* rs53576 are associated with the OT receptor mRNA expression in the brain regions directly implicated in social conformity and reinforcement learning signals (i.e., PE) such as the NAcc (Losdale et al., 2013) and dACC (Almeida et al., 2022). These findings altogether suggest that *OXTR* rs53576 is a

promising genetic marker that may modulate behavioral conformity and its underlying neural mechanisms.

Concerns have been raised regarding the exclusive focus on rs53576 in the process of modeling the link between endogenous OT signaling and human sociality. This is because the effects of a single SNP on complex traits are known to be very small (Bakermans-Kranenburg et al., 2014). One way to avoid this issue is to build a composite index with multiple genes or SNPs that are theorized to have similar functions. Polygenic risk cores or allele-dosage scores are examples of this approach (Torkamani et al., 2018). While fruitful (Belsky and Israel., 2014; Hernandez et al., 2017; Davis et al 2019), this method also has limitation in that it relies solely on the behavioral or psychological phenotypes of the target gene(s) to construct the index. This precludes researchers from interpreting the specific function of the implicated genes with reference to the actual physiological influences in the brains, such as receptor expression.

To address these caveats, we devised a novel index of endogenous OT signaling (i.e., *OXTR* multi-locus profile score, OXTR MPS) based on the multiple OXTR SNPs that have overlapping receptor expression profiles in the brain with that of OXTR rs53576. Specifically, we used the expression quantitative trait loci (eQTL) data from the GTEx (http://gtexportal.org) to translate individual participants' genotypes for seven *OXTR* SNPs, including *OXTR* rs53576, into the overall level of *OXTR* expression in the brain (See "Methods"). Compared to the conventional MPS created from distal phenotypes of the implicated genes, the MPS used in this study should offer a more direct window into the mechanisms through which endogenous OT signaling modulates its neural or behavioral phenotypes.

To test the link between *OXTR* MPS and conformity, we conducted an imaging genetics experiment that encompasses genetic, behavioral, and neural levels analyses (Falk et al., 2012). In a novel social conformity task, participants were presented with a series of word pair denoting various moral values and virtues widely recognized in the United States. Participants rated the relative importance of these words as guiding principles in their lives (i.e., value preference ratings) and then learned how most other participants responded to the same word pairs. Later in the experiment, participants repeated the task a second time. The conformity effect was defined as the changes in the value preference ratings that occurred between the first and second sessions of the task. The effects of participant-specific *OXTR* MPS values were analyzed with the behavioral data and blood oxygen level-dependent (BOLD) fMRI signals.

Based on the hypothesized link between *OXTR* and conformity, we made two sets of predictions. First, we predicted that 1) our novel experimental paradigm would induce behavioral conformity as in many previous studies that examined the conformity effect in non-moral domains (**Prediction 1a**), and that 2) the presentation of social feedback that is either consistent or inconsistent with the ratings of self will incur activations in the NAcc and pMFC/dACC (**Prediction 2a)**.

Second, we predicted that individuals with enhanced OT functions in the brain (i.e., higher *OXTR* MPS), due to their heightened sensitivity towards social reward and rejection, would show a stronger tendency to change their behaviors than the A allele carriers (i.e., AA/AG) to align with others **(Prediction 1b).** Lastly, the greater behavioral conformity among G homozygotes would be subserved by increased activations in the NAcc and pMFC/dACC in response to social feedback consistent and inconsistent with participants' value preference ratings, respectively **(Prediction 2b).**

Methods

Participants

A total of 194 (Female $N = 119$) adults over age 18 were recruited from Emory University and its surrounding community (Age $M = 22.9$, $SD = 6.8$). Volunteers were screened for past or current psychological or neurological illness. Those who are currently taking psychoactive drugs were further excluded. All eligible participants were assigned into either the neuroimaging ($N = 50$) or behavioral arm ($N = 144$) based on *a priori* power analysis. The two arms were implemented for assessing the robustness of the findings. The results of the power analysis are presented in **Supplementary Information (S4-1).** The demographics of the final study samples are summarized in **Table 4-1.**

Materials and Procedures

All experimental procedures and study materials were approved by Emory University Institutional Review Board (IRB00112525). This study was pre-registered at https://osf.io/d3x85.

Pre-experiment online survey

Once enrolled, participants accessed the online study portal (i.e., REDCap, https://www.project-redcap.org) to complete written informed consent, demographic survey, and a set of psychological

questionnaires. They also responded to an additional survey (i.e., cultural value survey; See below) which were used to create a participant-specific experimental stimulus for the main task.

***Demographic survey*:** We collected data on participants' age, gender, ethnicity, political self-identification (1=Very conservative, 5=Very liberal) and religiosity (1=Very religious, 5=Not at all religious). Descriptive statistics for the demographic variables were provided in **Supplementary materials (S4-2).** The results of analyses involving demographic variables were reported in the "Results" section, whenever appropriated.

***Psychological questionnaires*:** We measured a broad array of personality traits that are known to correlate with sensitive social behaviors. These variables included empathy (i.e., Interpersonal Reactivity Index, IRI; Davis, 1983), social anxiety (i.e., The Liebowitz Social Anxiety Scale, LSAS; Liebowitz, 1987), need for cognition (Need for Cognition Scale, NfC; Cacciopo and Petty, 1982), self-monitoring (i.e., Social Monitoring Scale, SM; Lenox and Wolfe, 1984), and impression management (i.e., Fear of Negative Evaluation, FNE; Leary, 1983). We specifically focused on NfC, SM and FNE as they have previously been shown to predict behavioral conformity (Haugtvedt et al., 1992; Scher et al., 2007; Kim et al., 2021).

***Cultural value survey*:** The cultural value survey consisted of English nouns and adjectives (N = 108) depicting moral values (e.g., fairness, loyalty, care etc.) and virtues (e.g., reliability, logicality, bravery etc.) widely recognized in the United States. The stimuli (i.e., virtue words) were adopted from psychological and anthropological literature on morality and character strengths (Keseber and Keseber, 2012; Schwartz, 2012; Graham et al., 2013). Through a separate pilot study (Participant *N*=30), we selected a broad array of values and virtues perceived as positive, familiar, and

unambiguous. Participants in the main experiment also rated the virtue words on a 5-point Likert scale in terms of the personal importance ("How important do you consider this value/virtue/trait as a guiding principle of your everyday behavior?"), familiarity ("Are you familiar with the meanings of this value/virtue/character trait?") and moral relevance ("How much do you consider this value/virtue/character trait to be relevant when you make moral evaluations on yourself or others?). Participants submitted all survey responses prior to their visit to the lab for the experiment.

Saliva sample collection

Participants in the behavioral- and neuroimaging condition visited Laboratory for Darwinian Neuroscience and Facility for Education and Research in Neuroscience (FERN) at Emory Atlanta campus, respectively. Upon arrival, participants provided their saliva samples using Oragene DNA self-collection kits (OGR-600, DNA Genotek Inc, Ontario, Canada) and proceeded to the main experiment.

Main Task

All participants performed a novel card choice task designed to induce conformity pressure by majority feedback. Similar to the experimental paradigm used in previous studies (Klucherev et al., 2008; Izuma et al., 2010; Zaki et al., 2011), the card choice task was repeated twice. Participants in the neuroimaging arm performed the first session inside an MRI scanner and the second session outside the scanner. Those assigned to the behavioral arm performed all two sessions in a test room without MRI scan. The schematic representation of the card choice task is shown in **Figure 4-1**.

Session 1 (S1)*:* The purpose of the first session was to measure participants' baseline response pattern unaffected by social influence. Each trial started with a fixation point appearing at the center of the screen. After a jittered interval (1-5s), participants were presented with a pair of cards with their back sides facing up. The cards flipped after 2s, revealing one virtue word printed on each card. Participants were asked to indicate the relative importance of the two virtue words using a 6-point scale. Unlike in many previous experimental paradigms in the conformity literature (Klucherev et al., 2008; Izuma et al., 2010; Zaki et al., 2011), we presented the virtue words in pairs. This design feature was used to increase the variability in participants' ratings, as all virtue words were positively valenced and could thus lead to a negatively skewed response pattern. The word pairs presented in each trial were matched for personal importance-, familiarity-, and moral relevance ratings submitted by each participant prior to the task.

The task had two trial types. In the "social feedback" trials, participants' ratings were followed by a prompt (i.e., "majority response") and visual feedback (i.e., a red box) ostensibly showing a majority response collected from previous participants. In approximately 25% of trials, the majority response matched the rating given by participants. In 50% of trials, the majority response deviated by either ±2 or ±3 points from participants' ratings. An adaptive algorithm was used to keep the overall ratio of positive and negative feedback, as well as their average magnitude close to equal. In the "control feedback" trials, which were approximately 25% of the trials, participants viewed a different prompt (i.e., "No Data) that the majority response was not available due to incomplete data. All feedback lasted for 3000ms and was replaced by a fixation point. Participants completed a total of 54 trials.

Filler Tasks: Between the first and second session of the task, participants completed two unrelated cognitive tasks (e.g., Facial emotion perception and authenticity judgment task) which lasted for approximately 30 minutes. The order of the filler cognitive tasks was counterbalanced across participants. The results of the filler tasks are not discussed in the present manuscript.

***Session 2 (S2)*:** After the filler tasks, participants performed a modified version of the card choice task a second time. Participants had not been informed of the second session in advance to prevent demand characteristics. As participants repeated the task, they 1) rated the same virtue word pairs a second time, and 2) recalled the majority responses for each word pair presented during S1. Each of these two ratings was used to determine 1) whether any changes in participants' ratings between S1 and S2 were driven by social feedback, and 2) how accurately participants recalled the majority opinions, respectively. For the word pairs initially presented during the control feedback trials (i.e., "No Data"), participants were instructed to make their best guesses on what the majority responses would have been, had the data been collected from the same majorities. Upon completion of the task, participants filled out a post-experiment questionnaire.

Post-experiment questionnaire

Previous studies have shown that the degree of social conformity can be influenced by the specific impressions of referent groups (Izuma et al., 2010). Hence, participants were asked to rate "other participants" who ostensibly provided the majority feedback during S1 on overall likability (1=Not at all, 5=Very much), morality (1=Not at all moral, 5=Very much moral), and similarity to self (1=Not at all similar, 5=Very much similar) using a 5-point likert scale. Participants also indicated whether the majority feedback was consistent with their own pre-existing beliefs or knowledge

about the cultural values and virtues endorsed in the United States (i.e., Belief update: 1=Much more dissimilar than I had thought, 5=Much more similar than I had thought). Participants who complete the experimental procedures were debriefed and received compensation (Behavioral arm: $40; Neuroimaging arm: $50).

Data Acquisition and Analysis

Neuroimaging data acquisition: For structural scans, T1-weighted images were acquired using a 3D magnetization-prepared rapid gradient-echo (MPRAGE) sequence with a Generalized auto-calibrating partial parallel acquisition (GRAPPA) factor of 3. The T1 scan protocol, optimized for 3 Tesla, used the following imaging parameters: a repetition time/inversion time/echo time (TR/TI/TE) of 1900/900/2.27ms, a flip angle of 9˚, a volume of view of 256×256×176 mm^3, a matrix of 256×256×176, and isotropic spatial resolution of 1.0×1.0×1.0 mm^3. Functional images were acquired using an Echo-Planar Imaging (EPI) sequence for blood-oxygen-level-dependent (BOLD) fMRI. EPI images were collected in an interleaved fashion with the following parameters: TR=1200ms, TE=30ms, matrix=74*74, Field of View=220mm, isotropic in-plane resolution=3.0 mm, slice thickness=3.0 mm, 54 axial slices with no gap in between and no phase oversampling.

Genetic data acquisition Participants' DNA data were extracted from saliva samples. Each participant's genotype was determined by Axiom™ Precision Medicine Research Array (Affymetrix) and TaqMan SNP Genotyping Assays using a ViiA7 Real Time PCR System for genotype resolution (Applied Biosystems, Foster City, CA). For quality control in SNP genotyping, each 384 well genotyping plate contained multiple duplicate wells and positive and negative

controls. 106 Ancestry-Informative markers were used to account for potential population stratification. These markers discriminated European, African, East Asian, and Native American origins. We used a structure software (Pritchard, 2000) to estimate proportions of chromosomal ancestry based on K (the number of source populations). Principal components analysis (PCA) was calculated account for population stratification. The first two principal components from this analysis were used in the analyses as covariates to control for population stratification.

Genetic data analysis based on *OXTR* multi-locus profile score For each participant, we computed an OXTR multi-locus profile score (MPS) to approximate the level of *OXTR* expression in the brain. First, we used the GTEx database (https://gtexportal.org) to find *OXTR* SNPs that show similar expression profiles in the brain as *OXTR* rs53576. *OXTR* rs53576 is directly associated with the receptor expression in 6 brain areas (i.e., NAcc, caudate nucleus, putamen, the frontal cortex including BA9, and hippocampus). Therefore, only the SNPs linked with the receptor expression in at least five overlapping brain regions were considered and selected among imputed SNPs. This yielded a total of 6 SNPs **(Table 4-2).** Then, based on the expression quantitative trait loci (eQTL) data available in GTEx, the allele associated with higher receptor expression in the striatum and other brain regions were identified for each SNP (i.e., High-expressing allele). Next, for each participant, we calculated the number of the high-expressing allele for each *OXTR* SNP (i.e., 0, 1, or 2). The final MPS values were calculated by summing the allele counts across 7 SNPs, including rs53576 that regulate receptor expression in these brain areas (i.e., 0-14).

Behavioral data analysis

Behavioral data obtained from either the behavioral arm or neuroimaging arm were processed and analyzed with MATLAB R2015b (The MathWorks, Natick, 2015) and SPSS version 28 (Armonk, NY: IBM Corp).

Testing main predictions

Prediction 1a: We performed a one-sample *t*-test to compare the proportion of social feedback trials in which participants shifted their ratings towards the majority opinions with zero (i.e., %Conformity). The same analysis was also repeated for the proportion of trials where participants did not show any behavioral shift (i.e., %Resist), or moved farther away from the majority opinion (I.e., %Anti-conformity).

Next, to examine the specific directions and magnitude of the decision shifts following different types of social feedback, we calculated the average changes in the value preference ratings between S1 and S2 for each participant (i.e., decision shifts) across three feedback conditions: consistent social feedback, inconsistent social feedback with either positive or negative deviation, as well as no-feedback. To control for the overall decision shift that may take place between S1 and S2 independently of social feedback, the average decision shift values calculated for the social feedback trials were centered within each participant using the average decision shift occurred for the "No-Feedback" trials. With the resulting indices of the normalized decision shift, we defined a linear mixed model (LMM) with Feedback Type (i.e., Consistent social feedback vs. positive inconsistent social feedback vs. negative inconsistent social feedback) and sex (i.e., male vs. female) included as fixed factors and subjects as a random factor. The main effect of Feedback Type, Sex, and their interaction were modelled. Participants' age and the first two principal components that

captured the ethnic variation of the study sample (i.e., ethnicity) were also included as continuous covariates to explore the possible main effect of the demographic variables. The unstructured covariance matrix in SPSS was used to model the relationship between different levels of repeated measures (i.e., Feedback Type). Restricted maximum-likelihood estimation was applied with 100 iterations to reduce bias in random effect variance estimation.

Prediction 1b: We tested whether the *OXTR* MPS was associated with the behavioral conformity. First, the effects of *OXTR* MPS on the overall degree of conformity (%Conformity) were tested in a univariate GLM, with the *OXTR* MPS, sex, age, and ethnicity included as predictors. We also applied the same GLM to the average %Resist and %Anti-conformity. Lastly, we repeated the aforementioned LMM analysis on the average normalized decision shift ratings, with the *OXTR* MPS included as a continuous covariate.

Along with the main analyses, we also examined the possible effect of "regression towards the mean (RTM)," which can confound the behavioral shifts observed in social conformity experiments (Yu and Chan, 2015). We confirmed that the effect of self-other gap on decision shifts was significant after controlling for participants' initial ratings (Izuma and Adolphs; Nook and Zaki, 2015; Yu and Chan, 2015) **(Supplementary materials S4-3).**

All statistical tests were performed with the type-one error rate of α=.05 (two-tailed). Bonferroni correction was applied for any significant main effects or interaction effects. Cohen's *d* is reported wherever appropriate as a measure of the effect size.

Exploratory analyses

As indicated in the pre-registered protocol, we also performed a series of exploratory analysis testing 1) the correlation between behavioral conformity, personality traits, and social impression ratings and 2) the effects of *OXTR* MPS on recall accuracy. All relevant procedures and results of the exploratory analyses are reported in **Supplementary materials S4-4 and S4-5**, and the "Results" section whenever appropriate. All statistical tests for the exploratory analyses were performed with the type-one error rate of α=.05 (two-tailed). The results of the exploratory analyses were not corrected for multiple comparisons.

Neuroimaging data analysis

Neuroimaging data analysis were carried out with the Oxford Center for Functional Magnetic Resonance Imaging of the Brain's software library (FSL v6.0, http://www.fmrib.ox.ac.uk/fsl/).

Preprocessing: Preprocessing of images included 1) motion correction using the MCFLIRT (Jenkinson, Bannister, Brady, & Smith, 2002), 2) skull-stripping using the Brain Extraction Tool (BET), 3) slice timing correction, 4) high-pass temporal filtering with a filter width of 100 seconds, 5) spatial smoothing using a Gaussian kernel of full width at half maximum (FWHM) of 8mm, 6) registration of fMRI images to high-resolution T1 images (i.e., Boundary-Based-Registration), and 7) spatial normalization to the standard Montreal Neurological Institute (MNI) 2 mm brain (i.e., Affine transformation using 12 degrees of freedom) using FLIRT (Greve & Fischl, 2009).

1st level analysis: FMRI data were analyzed with a univariate general linear model (GLM) approach. For the first-level analyses, data from each trial were convolved with a double-gamma

hemodynamic response function (HRF) in FSL. Following the pre-registered protocol, our main GLM was defined to identify the neural signatures underlying the perception of gap/error (i.e., GLM1) (Klucharev et al., 2009; Wake et al., 2019).

The model included explanatory variables (EVs) corresponding to the following events: presentation of a word pair (i.e., Word), decision prompt (i.e., Prompt), value preference judgments (i.e., Decision), consistent social feedback (i.e., Feedback-C), inconsistent social feedback (i.e., Feedback-IC), and non-social control feedback (i.e., NF).

In addition, one exploratory GLM was derived from the main GLMs to test whether the activations in the NAcc and pMFC/dACC exhibited the properties of PE and if *OXTR* MPS modulated such PE signals. In the context of social conformity, a PE signal should be tracking the distance between self and other (Izuma and Adolphs, 2013; Bhanji and Delgado, 2014). Therefore, the inconsistent social feedback EV in the exploratory GLM was weighted by the trial-by-trial gap between participants' value preference ratings and majority feedback during S1 (i.e., Feedback-IC_{PM}). The original Feedback-IC EV with a constant height was also included to model the average effect of the inconsistent social feedback. Feedback-IC_{PM} was orthogonalized with respect to the Feedback-IC to capture the unique BOLD responses proportionally varying with the magnitude of self-other gap (Mumford et al., 2011).

2nd level analysis: A series of mixed-effect analyses (i.e., FLAME1) were carried out with the FSL's fMRI Experts Analysis Tool (FEAT). To assess the brain responses associated with the *perception* of inconsistent social feedback, the following contrasts were defined for GLM1: [Consistent > NF], [Inconsistent > NF], and [Consistent > Inconsistent].

The first two PCA eigenvectors from the genetic relationship matrix (GRM; http://gump.qimr.edu.au/gcta; Yang et al., 2010) were mean centered and included in all models as covariates to control for variance due to population structure.

Search Volumes and Thresholding: We used both ROI analyses and whole-brain analyses to test our hypotheses. **Hypothesis 2-a** and **2-b** were examined with one sample *t*-tests that compared the average contrast estimates yielded from the first level GLM to zero. Hypotheses 2-c was also tested by modeling subject-specific *OXTR* MPS scores as a mean-centered, continuous covariate in the second level GLM. The slope was compared with zero by one-sample *t*-test.

For the main ROI analyses, we focused on two brain areas that are known to 1) express *OXTR* and 2) influence social conformity: pMFC/dACC and bilateral NAcc (Wu, Luo, and Feng 2016). The MNI coordinates and the sizes for our spherical ROIs were drawn from a recent coordinate-based activation likelihood estimation (ALE) meta-analysis on the neural signature of social conformity (Wu, Luo, and Feng 2016). The pMFC/dACC (BA32, x = 41, y = 6, z= 34) and bilateral NAcc ROIs (L: x = -6, y = 16, z = -4; R: x = 10, y = 16,z = -2) encompassed the anterior midcingulate cortex (BA8), and the head of the caudate nucleus, respectively (Wu, Luo, and Feng 2016). For each ROI, voxel-wise, one-sample *t*-tests were used to determine if the contrast estimates obtained from the main GLM are significantly different from zero. The resultant Z statistic images were thresholded at $p<0.05$ corrected for multiple comparisons across all voxels within each ROI based on Gaussian Random Field Theory (i.e., a small volume correction, SVC).

To explore the brain activations outside these primary ROIs, we performed whole-brain analyses. All Z-statistic images were initially thresholded using a cluster-forming threshold of Z > 3.1(i.e., voxel-wise 1-tailed $p<.001$), and then with a family wise error (FWE)-corrected cluster significance threshold of $p<.05$.

Results

Behavioral Arm

Sample characteristics

Data from six participants were excluded due to task errors or genotyping failure. The average *OXTR* MPS did not show significant association with age, religiosity, political self-identification, sex, or personality traits (All *P*s > .136).

Behavioral Results

Prediction 1a: On average, participants completed 25 inconsistent social feedback trials throughout the task. The one-sample *t*-test on %Conformity (M = 30.6, SD = 12.5) against zero turned out significant, $t_{(138)}=28.972$, $p < .001$, indicating that participants exhibited behavioral conformity significantly above zero. The average %Resist (M = 55.1, SD = 14.6) and %Anti-conformity (M=14.2, SD=8.2) were also significantly above zero (All *Ps* <.001).

There was no significant difference in the average value preference ratings measured across S1 (*M*=3.47, SD=.20) and S2 (*M*=3.49, *SD*=.23) (*p*=.09). Yet, the LMM analysis performed on the normalized decision shifts between S1 and S2 yielded a significant main effect of Feedback Type (*F*=61.79, *p*<.001). Post-hoc tests revealed that the average decision shifts following the positive, negative inconsistent social feedback and consistent social feedback differed across all levels (All *Ps*<.001) **(Figure. 2a)**. Participants' sex did not have a significant main effect or interaction (All *P*s >.208). The effect of social feedback remained significant after controlling for participants' age, ethnicity (i.e., two principal component eigenvectors), which showed no association with the patterns of decision shift (All *P*s > .403).

Prediction 1b: The univariate GLM performed on the %Conformity, %Resist, and %Anti-conformity, revealed no significant effects involving *OXTR* MPS (All *P*s>.147). Participants' age, sex, and ethnicity were not associated with these indices (All *Ps* >.138). Similarly, the LMM analysis with the *OXTR* genotype as a between-subjects fixed factor found no evidence of significant effect of *OXTR* genotype (All Ps >.094)**.**

In sum, although significant behavioral conformity was present among participants in the behavioral arm, the *OXTR* genotype did not significantly influence the observed patterns of decision shifts.

Neuroimaging Arm

Sample characteristics

Data from five participants were excluded from analysis due to task error and incomplete ancestry data. Average *OXTR* MPS did not correlate with participants' age, sex, personality traits, religiosity, and political self-identification (All *Ps*>.085).

Behavioral Results

Prediction 1a: On average, participants completed 26 inconsistent social feedback trials. The one-sample *t*-test comparing %Conformity (*M*=28.7, *SD*=13.2) against zero turned out significant, $t_{(44)}$=14.597, $p < .001$, showing that participants exhibited behavioral conformity on approximately 30% of the inconsistent social feedback trials. The average %Resist (*M*=55.43, *SD*=15.5) and %Anti-conformity (*M*=15.8, *SD*=7.9) also exceeded zero (All *Ps* <.001).

As in the behavioral arm, the average value preference ratings did not significantly change across S1 and S2 (*p*=.846). The LMM analysis performed on the normalized decision shifts between S1 and S2 also revealed a significant main effect of Feedback Type (*F*=43.906, $p < .001$). Our *post-hoc* pairwise comparisons turned out significant across all levels (All *Ps* < .05), indicating that participants shifted their decisions consistently with the types of social feedback they received during S1 **(Figure 4-2b).** Participants' sex did not significantly modulate the patterns of decision shifts. The effect of social feedback persisted ($p < .001$) after controlling for participants' age and ethnicity, which were not significantly associated with the behavioral conformity (All *P*s > .053).

Prediction 1b: The univariate GLM performed on the %Conformity, %Resist, and %Anti-conformity, revealed no significant effects involving *OXTR* MPS (All *P*s>.388). Likewise, the LMM analysis did not reveal any significant effects of the *OXTR* MPS on the magnitude of decision shift, (All Ps >.536). In sum, we did not find evidence of a significant genetic modulation of participants' task performance, when the analysis was performed without the imaging genetics data (See "Neuroimaging Results" Prediction 2b).

Neuroimaging Results

ROI analysis

Prediction 2a: Consistent and inconsistent social feedback and activations in the gap-detection mechanism

NAcc We found that the inconsistent social trials (i.e., Feedback-IC) was associated with the stronger average activations in the bilateral NAcc/caudate nucleus compared to non-social control feedback (i.e., NF). No activations were found for the contrast between the consistent social feedback trials (i.e., Feedback-C) and NF. The direct comparison between Feedback-IC and Feedback-C revealed that the former was associated with the stronger activations in the right caudate nucleus **(Figure 4-3a).** The NF trials were associated with stronger activations in the septum/subgenual region compared to either Feedback-IC or Feedback-C **(Supplementary materials S4-6**).

pMFC/dACC Both Feedback-IC and Feedback-C elicited stronger activations in the pMFC/dACC compared to NS. The direct comparisons between Feedback-IC and Feedback-C revealed the increased activations for the Feedback-IC in dACC, with the peak voxel found in the anterior midcingulate cortex **(Figure 4-3a)**. No significant activations were found with respect to the NF trials.

An exploratory parametric modulation analysis revealed that activations in either the NAcc or pMFC/dACC were not linearly tracking the magnitude of self-other gap in value preference ratings. The average effect of identified in the main GLM (i.e., Feedback-IC > Feedback-C) remained significant in both ROIs **(Supplementary Materials S4-7).**

To gain insights into the function of the BOLD responses observed within the NAcc and pMFC for the contrast [Feedback-IC > Feedback IC], we correlated the peak activations within these ROIs with behavioral measures of conformity. The results showed that the activations in the contrast estimates within the pMFC negatively predicted conformity behaviors (r = -293, p = .005) **(Figure 4-3b)**. No comparable activations were found in the NAcc (p=.345).

In summary, **Prediction 2a** was partially supported, as the perception of social feedback generally incurred increased activations in the brain areas implicated in the NAcc and pMFC/dACC. Intriguingly, however, the level of the BOLD responses within *both* ROIs were stronger when participants perceived social misalignment as opposed to social alignment, which is at odds with many previous findings (Klucherev et al., 2009). We also confirmed that the trial-by-trial fluctuations of the BOLD responses within the key ROIs were not associated with the gap between the value preference ratings of self vs. others.

Prediction 2b: Genetic modulation of the activations in the NAcc and pMFC/dACC during social feedback processing

NAcc Our analyses yielded no evidence that *OXTR* MPS significantly modulated the NAcc activations.

pMFC/dACC: There was a significant negative association between *OXTR* MPS and the contrast estimates from [Incon-Non > NF] **(Figure 4-4a)**. This relationship reflected the fact that those with the higher *OXTR* MPS showed the decreased activations in the pMFC/dACC in response to inconsistent social feedback (r = -.457 , p = .002).

The behavioral relevance of the activations within the pMFC/dACC were explored by extracting the contrast estimates from the voxels showing a peak genetic modulation within each ROI and correlating those values with individual participants' task performance. The results showed that the genetic modulation within the pMFC/dACC was linked with the degree of non-conformity. Specifically, the decreased activations within the ROI in response to inconsistent social feedback predicted greater conformity (i.e., %Conformity) (r = -.428, p = .003) **(Figure 4-4b).** Given the cross-correlations between *OXTR* MPS, the activations within the pMFC/dACC and conformity, we conducted a follow-up analysis to explore whether the *OXTR* MPS OT could lead to behavioral conformity by modulating the pMFC/dACC activity in response to inconsistent social feedback. The mediation model was performed with PROCESS MACRO in SPSS (Heyes et al., 2017). Indeed, the activations within the pMFC/dACC during the inconsistent feedback trials significantly mediated the relationship between *OXTR* MPS and %Conformity, with higher *OXTR* MPS leading

to the increased conformity by decreasing pMFC/dACC activations in response to inconsistent social feedback (Bootstrapped 95% CI[0.009, 0.2712]) **(Figure 4-4c).**

In a nutshell, **Prediction 2b** was partially supported in that *OXTR* did modulate the activations in the NAcc and pMFC/dACC, yet with an unanticipated direction. In other words, participants with higher *OXTR* MPS showed *decreased* BOLD signals in the NAcc and pMFC/dACC when receiving social misalignment. Specifically, the subdued activations in the pMFC/dACC were associated with greater conformity, revealing an indirect pathway linking higher *OXTR* MPS with the increased behavioral conformity.

Whole-brain Analysis

Our whole-brain analyses revealed a wider network of brain regions known to be associated with the feedback processing, which suggested continuity between previous findings and ours.

The main GLM showed that both consistent social feedback and inconsistent social feedback recruited largely overlapping brain areas including the pMFC/dACC, dorsomedial prefrontal cortex (dmPFC), precuneus, and AI, areas previously implicated in error/gap detection (Klucherev et al., 2009; Liu et al., 2013). Significant activations were also found in the brain areas implicated in action observation and execution, such as the bilateral inferior frontal gyrus (IFG) extending towards the ventrolateral orbitofrontal cortex (vlPFC), bilateral superior parietal lobule (SP), encompassing the precuneus. Notably, the significant activations in the dorsal caudate nucleus and thalamus emerged only with the [Feedback-IC > NF]. The direct contrasts between [Feedback-IC > Feedback-C] showed that Feedback-IC generally incurred greater activations in the

abovementioned regions including the dorsal caudate nucleus. No significant activations were found in [Feedback-IC < Feedback-C] **(Supplementary material S4-8).**

An exploratory parametric analysis again revealed no significant cluster associated with the trial-by-trial discrepancy between the value preference ratings of self and others. In all of the whole-brain analyses, we did not find evidence that *OXTR* significantly modulated the neural correlates of conformity and non-conformity at the whole-brain level.

Discussion

The neuropeptide oxytocin (OT) is known to upregulate social cohesion by promoting conformity, which allows individuals to adopt behaviors and beliefs prevalent in their social environments. Yet, specific neural mechanisms that subserve the "herding" effect of the OT are yet to be understood. We aimed to address this gap by investigating how single nucleotide polymorphisms (SNPs) in the human oxytocin receptor gene (*OXTR*), which regulates endogenous OT signaling in the brain, could account for the individual variations in conformity. This study introduced two novel design features. First, the conformity effect was measured in the domain of values and virtues that have high moral relevance. Second, to better capture the individual variations in endogenous OT signaling in a more biologically grounded way, we defined a multi-locus profile score (MPS) with seven *OXTR* SNPs, including *OXTR* rs53576, that share similar expression profiles in the brain.

Consistent with **Prediction 1a,** participants changed their average value preference ratings in accordance with majority opinions. The effects of social feedback remained significant after controlling for the demographic and personality variables, as well as RTM (Yu and Chan, 2015). This study extends the existing literature that investigated the conformity effects in non-moral domains, such as aesthetical preference (Klucherav et al., 2009; Izuma and Adolphs, 2010; Zaki et al., 2011; Nook and Zaki, 2015), memory (Edelson et al., 2011), or perceptual judgments (Asch, 1954; Berns et al., 2005;(Stallen et al. 2012), 2013). Our findings also suggest the generalizability of the moral conformity effect beyond the judgments in hypothetical moral dilemmas (Kundus and Cummins, 2013; Wei et al., 2017, 2019) or economic decision games (e.g., trust game and dictator game; Wei et al., 2017, 2019), which were recently shown to be susceptible to majority opinions.

Broadly consistent with **Prediction 2a**, we found that the perception of social feedback modulated the BOLD responses in the NAcc and pMFC/dACC (Shamay-Tsoory et al., 2016), two key nodes in the gap/error-monitoring system in the brain. Intriguingly, the specific patterns of responses within these ROIs did not entirely replicate the established neural signatures of social conformity.

Previous studies have found that the activations in the NAcc and pMFC/dACC proportionally decreased and increased with the degree of social misalignment, respectively (Klucherev et al., 2009; Izuma et al., 2013). In this study, however, inconsistent social feedback incurred greater average BOLD responses in *both* NAcc and pMFC/dACC masks. Our exploratory parametric modulation analyses also showed that voxels within these ROIs were not linearly tracking the magnitude of the gap, which suggests that these signals lack an important property of PEs. Overall, despite the recruitment of the key brain structures within the gap/error-monitoring mechanisms,

the psychological and computational properties underlying these activations were not readily explicable with the prevailing neurocognitive model of social conformity.

Striatal activations in social conformity experiments have often been interpreted as the experience of rewards associated with the agreement between self and others. Our data do not fit this reward-centered account of the striatal activity, as the contrast estimates in the ROI were larger for social misalignment compared to social alignment and non-social feedback. The direct comparisons between social alignment and non-social feedback did not yield any significant activations in the NAcc. These results suggest that the NAcc/caudate in this study may be encoding an element of social feedback specific to inconsistent social feedback, such as salience.

The human striatum is comprised of multiple, functionally heterogeneous populations of neurons (Prensa et al., 2003; Lauer and Heinsen, 1996), and evidence suggests that both dorsal and ventral striatum can encode socially-relevant information beyond the experience or anticipation of reward. Of relevance to our findings, studies have shown that the human NAcc and caudate nucleus can encode highly salient, yet non-rewarding event (Zink et al, 2003; Zink et al., 2004; Becerra et al., 2001; Jensen et al., 2007; Levita et al., 2009; Cooker and Knutson, 2008). Multiple activation likelihood estimation meta-analyses have also suggested that the NAcc and dorsal caudate nucleus suggested that the NAcc and dorsal caudate nucleus can show valence-independent activations to surprising social cues (Liu et al., 2012; Fouragnan et al., 2017). Inconsistent social feedback is motivationally significant and potentially aversive as they may lead to punishment and social rejection (Allen, 1965). Therefore, it is plausible that the activations in the caudate nucleus in this study may also reflect the salience of social misalignment.

It should be noted that some participants might still find the inconsistent social feedback rewarding, especially if they valued uniqueness. Since independence tends to be considered desirable in western societies (Singelis, 1994), perceiving the social misalignment may have contributed to the striatal activations via incurring the experience of rewards. However, we confirmed that the degree to which participants cared about the value of independence or autonomy, which was determined by individual participants' ratings in the cultural value survey (i.e., Personal importance rating), was not significantly associated with the contrast estimates in the caudate nucleus for inconsistent social feedback (All *Ps* = .457).

While unanticipated, the non-social control feedback trials induced stronger activations in the septal/subgenual regions than did the consistent and inconsistent social feedback trials. The lateral septum (SP) has previously been implicated in the control of mood (Thomas, 1988), motivation (Olds and Miller, 1954), movement (Sagvolden, 1976), and spatial cognition (Brioni, 1990). The lateral septum in rodents and primates has recently been implicated in the regulation of frustration induced by the omission of feedback (Henke, 1977; David, 2004) and the neural encoding of uncertainty (Monosov et al., 2015; Ledbetter et al., 2016). Similarly, human neuroimaging studies have found that the activations within the lateral septum guide social learning under volatility (Biele et al., 2011; Diaconescu et al., 2017; Iglesias et al., 2013). These findings point to the possibility that the relative increase in septal activation during the control feedback trials may reflect the neural encoding of the omission of social feedback, which increases uncertainty.

The voxels within the pMFC/dACC showed increased BOLD signal across 1) all types of social feedback compared to control feedback, especially during 2) the inconsistent social feedback trials than consistent social feedback trials. The exploratory whole-brain analysis comparing Feedback-

IC vs. Feedback-C contrasts also identified extra-striatal activations in the SPL, IFG, AI, and vlPFC. Our observation is a direct replication of previous findings on the involvement of pMFC/dACC during the inconsistent social feedback trials (Klucherev et al., 2009; Zaki et al., 2011; Zaki and Nook, 2014, Izuma et al., 2013). The results from the whole-brain analysis are also consistent with a recent proposal that the "action observation-execution system" (e.g., SPL, IFG, and vlPFC) play a central role in generating adaptive behavioral responses once social misalignment is detected in the brain regions such as pMFC/dACC (Shamay-Tsoory et al. 2019).

The pMFC/dACC is considered critical for translating external stimuli to autonomic signals, emotions, and context-specific adaptive behaviors (Vogts, 2016; Corbetta). Of relevance to this study, the pMFC/dACC has been widely implicated in conflict processing (Eisenberger, 2015). This area is also known to produce negatively-valenced PE following an unexpected, aversive outcome of behavior (Fouragnan et al., 2018). Therefore, the association between the pMFC/dACC activation and inconsistent social feedback observed in this study may be led by the experience of negative affect and conflict experienced during the inconsistent social feedback trials.

And yet, the specific sources of this *conflict* and its psychological nature merit further discussion. The peak activation in the pMFC/dACC for the [Feedback-IC > Feedback-C] negatively predicted the degrees of conformity. This is, in fact, opposite to the previous findings that greater pMFC/dACC activations in response to inconsistent social feedback led to increased conformity. That is, the BOLD responses within pMFC/dACC were *not* likely to reflect the negative affect or error-related signals induced by *social conflict* or the self-other gap *per se*. Rather, it may be driven by participants' *internal conflict* associated with overriding their own value priorities to achieve

social cohesion. Consistent with this interpretation, the increased activations within the pMFC/dACC in response to Feedback-IC vs. Feedback-C negatively predicted conformity.

In stark contrast to most previous neuroimaging studies on conformity, our experimental stimuli had strong thematic connections to moral values. Moral values are a constellation of beliefs and behavioral heuristics that are often highly internalized (Rand et al., 2012; Cushman et al., 2017), maintained by deontic rules (Berns et al., 2014; Pincus et al., 2014), and comprise core self-concept (Hornsey at al., 2003; Han et al., 2017). As a result, morally-laden beliefs and behaviors are known to be more impervious to social influences compared to mere customs and preferences (Turiel, 1983; Chituk and Sinnott-Armstrong, 2020). Thus, it is possible that participants in this study had to reconcile two competing motivations, namely, achieving social cohesion vs. maintaining a coherent moral identity and internalized social heuristics. Such heightened internal conflict, in turn, might have driven the activations in the pMFC/dACC in ways different from previous studies where participants were not exposed to the morally-laden choice options of personal importance.

Lastly, we found that *OXTR* MPS significantly modulated the activations in the pMFC. Specifically, participants with higher *OXTR* MPS showed decreased activations in pMFC/dACC in response to inconsistent social feedback. While this is consistent with **Prediction 2b**, we did not find any significant interaction between the *OXTR* MPS and the contrasts involving Feedback-IC_{PM}. This result suggests that the observed genetic modulation were computationally distinct from reinforcement learning signals.

Then, what would be the cognitive or psychological correlates of the *OXTR*-induced dampening of the MFC/dACC activations? Given that the BOLD responses within pMFC/dACC were likely to

reflect participants' internal response conflict, one plausible interpretation of our result would be that those with higher *OXTR* MPS experienced less conflict when presented with inconsistent social feedback.

OT is widely implicated in behavioral approach and social affiliation in both animals and humans (Bartz et al., 2016). It is well-established that OT is critical for the onset of parenting behaviors in various mammalian species (Pedersen et al., 1982, Kendrik et al., 1987; Williams et al.,1994). Increased OT signaling in the brain also facilitates conjugal bonding and social approach (King et al., 2016; Williams et al., 2020). Similarly, in humans, those administered with INOT were shown to exhibit prosocial and approach behaviors to a stranger (Kosfeld et al., 2005; Zak et al., 2005, Zak et al., 2007; Cohen et al., 2018) to strangers, even when such affiliative gestures were not honored (Baumgartner et al., 2008). These converging lines of research show that one of the key mechanisms through which OT regulates human sociality is to upregulate approach motivation and amplify the subjective values of social affiliation (Bartz et al., 2011; Bartz, 2016).

According to this view, it is possible that the higher *OXTR,* thus increased endogenous OT signaling, might have helped individuals "re-balance the scale" during the decision-making process by augmenting the value of social alignment (Bartz et al., 2016). Such added weights to the affiliative goals may subsequently reduce participants' internal conflict associated with overriding their personal value priorities.

What follows from this interpretation is that increased OT signaling may promote behavioral conformity via downregulation of the conflict-related activations in the pMFC/dACC. Our mediation model combining genetic, behavioral, and neuroimaging data indeed suggested this

possibility: higher *OXTR* MPS reduced the pMFC/dACC responses to social misalignment, which, in turn, increased conformity **(Prediction 1b).**

It is important to note that data analysis with behavioral task performance and *OXTR* MPS alone did not yield any significant effect of *OXTR* MPS on participants' behavioral performances in either the neuroimaging or behavioral arm. Our results from the exploratory mediation analysis thus need to be interpreted with caution. Still, the lack of direct associations in the mediation model does not preclude the existence of indirect paths which could potentially reveal more accurate patterns of associations among the implicated variables (Agler and De Boeck, 2017). Further replication would be necessary to validate the relationship between *OXTR*, the recruitment of pMFC/dACC, and moral conformity identified in this study.

Chapter Summary and Conclusion

Our data have two important implications for the study of social conformity and OT in social neuroscience. First, while consistent and inconsistent social feedback incurred differential neural activations in the gap/error detection mechanisms in the brain, the specific patterns of responses within these ROIs and their underlying cognitive processes were different from what has previously been found in the related literature. Most notably, our data suggest that the activations in the NAcc and pMFC/dACC were not the PE signals but other cognitive mechanisms such as salience encoding or response conflict. These processes, although often concurrently activated during feedback processing, are separable from reinforcement learning signals which tend to show specific computational properties (O'Doherty, 2014).

Our findings resonate with the ongoing discussions on the psychological and computational properties of the neural signature of conformity (Levenson et al., 2019; Izuma et al., 2013). We speculate that the specific cognitive functions represented in NAcc and pMFC may be critically contingent on the task environments and the behavioral domains that are subjected to normative social influence. In other words, the activations within the NAcc and pMFC/dACC should not be blindly interpreted as PE or reward/conflict-related signals associated with social alignment/misalignment. Likewise, the generalizability of our findings where *OXTR* MPS promoted conformity by reducing the activations in the pMFC/dACC should also be tested in different experimental paradigms.

Second, we utilized the multi-locus profile score (MPS), a novel index that allows us to translate individual participants' genotypes into the level of *OXTR* expression in the brain. This approach is similar to using polygenic risk scores or risk-allele dosages in behavioral and imaging genetics (Dima and Breene, 2015). Yet, as *OXTR* MPS in this study is indicative of the level of *OXTR* expression in specific brain regions instead of distal psychological or behavioral correlates, it could potentially offer 1) sensitive measures of detecting the effect of multiple genetic variants, and 2) more direct mechanistic insights into how OT signaling in the brain regulates its downstream social phenotypes. We used the brain receptor expression of *OXTR* rs53576 as our reference to select additional *OXTR* SNPs. Our choice was based on previous findings that *OXTR* rs53576 is widely implicated in social cognitive functions relevant for social conformity. Rs53576 is also known to be either directly or indirectly associated with the *OXTR* expression in the brain areas critical for social reward (e.g., striatum) and gap detection (e.g., dACC), among others (Almeida et al., 2022). It should be noted, however, that endogenous OT signaling in the brain is known to regulate social

cognition and behaviors in a region-specific way (King et al., 2016). Future studies should also investigate if *OXTR* SNPs that are differentially expressed across multiple brain regions have differential phenotypic effects.

In all, this study is the first to examine the specific neurocognitive mechanisms underlying OT-induced behavioral conformity. We presented putative evidence that those with higher *OXTR* expressions in the brain showed the greater decrease in the pMFC/dACC activations in the face of the opposing majority, which, in turn, positively predicted conformity in the domain of moral values and virtues. Given the role of pMFC/dACC in conflict processing, it is possible that the enhanced OT signaling in the brain reduced the level of internal conflict associated with making choices against their value priorities to achieve social cohesion. Future studies should employ different task environments and a larger sample size to further our understanding of the link between *OXTR* and conformity.

Figure and Tables

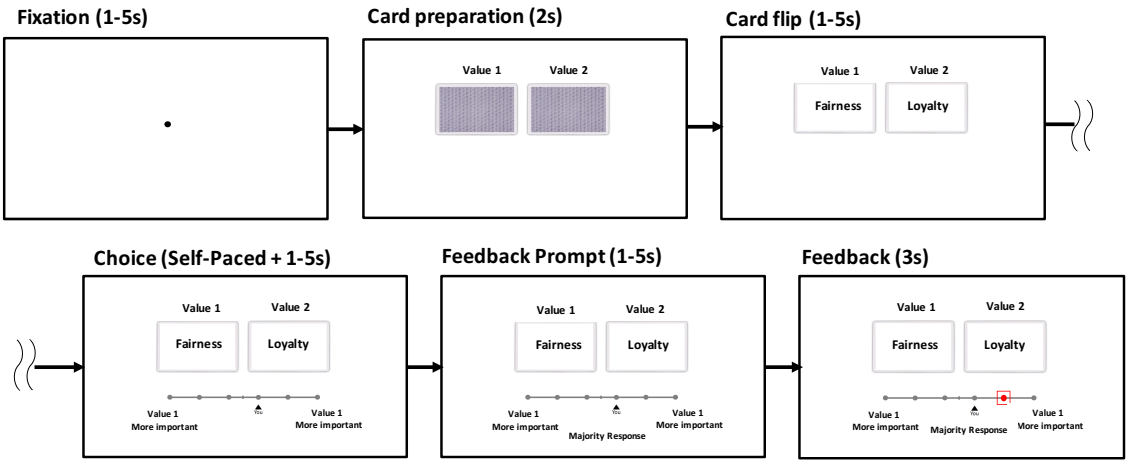


Figure 4-1. Schematic representation of the card sorting task. In each trial, participants indicated the relative importance between two values/virtues presented in pairs. In social feedback trials, participants' choice was followed by a red box ostensibly depicting a majority response made by previous participants. In non-social control feedback trials, participants were presented with a red prompt (i.e., "No data") that information about the majority response was not available. The BOLD fMRI signals during the "Feedback" phase was analyzed as the primary epoch of interest.

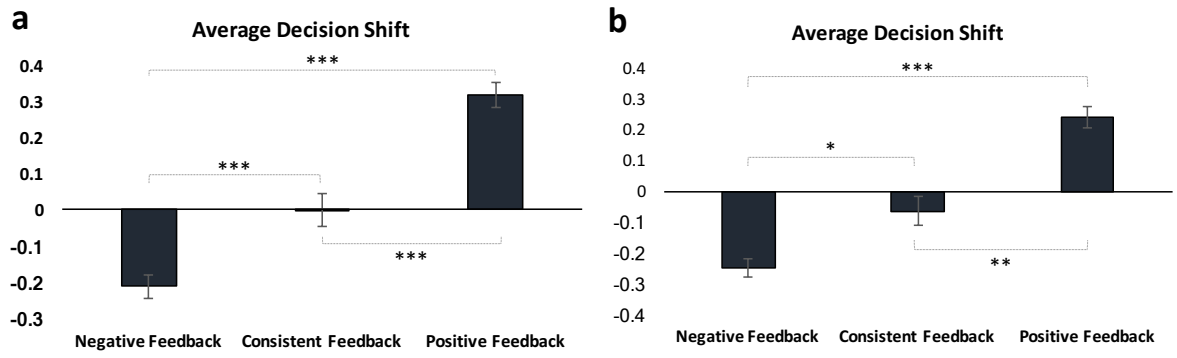


Figure 4-2. The average normalized decision shift across the different social feedback conditions. Participants in the behavioral arm (a) and neuroimaging arm (b) shifted their value preference ratings between S1 and S2 and the direction of the decision shift was consistent with the types of social feedback they received during S1. The means were adjusted for participants' age and two eigenvectors representing the ethnic variations of the study sample. (Error bars denote standard error means; *p<.05, **p<.01, ***p<.001.)

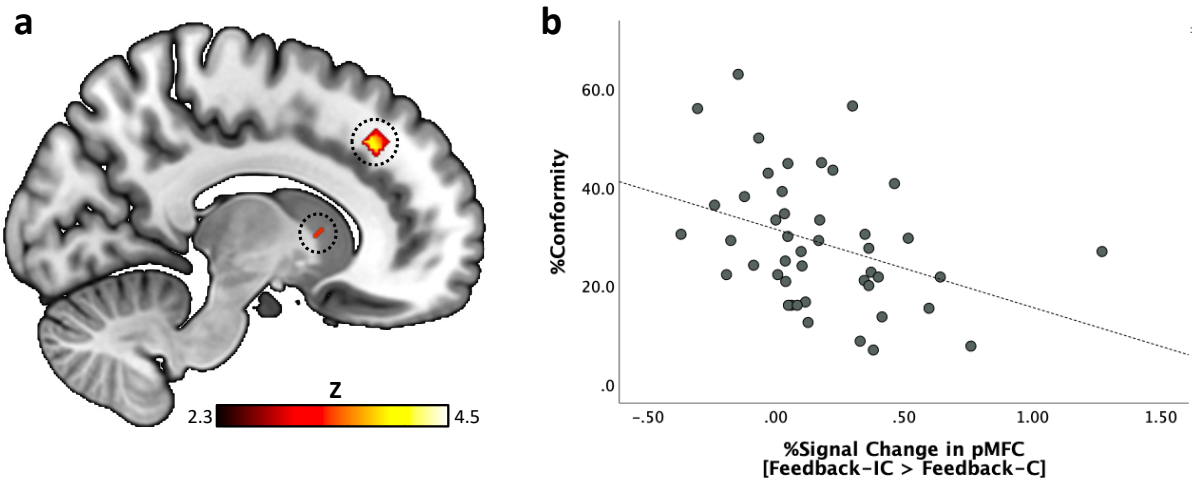


Figure 4-3. Significant voxels identified from the contrast [Feedback-IC > Feedback C]. Voxels in the right caudate nucleus (peak voxel MNI coordinates: x=12, y=14, z=6) and left pMFC/dACC (peak voxel MNI coordinates: x=-4, y=28, z=28) were significantly more active in response to inconsistent social feedback vs. consistent social feedback (a). The %Signal change within the pMFC negatively predicted subsequent conformity behaviors (b).

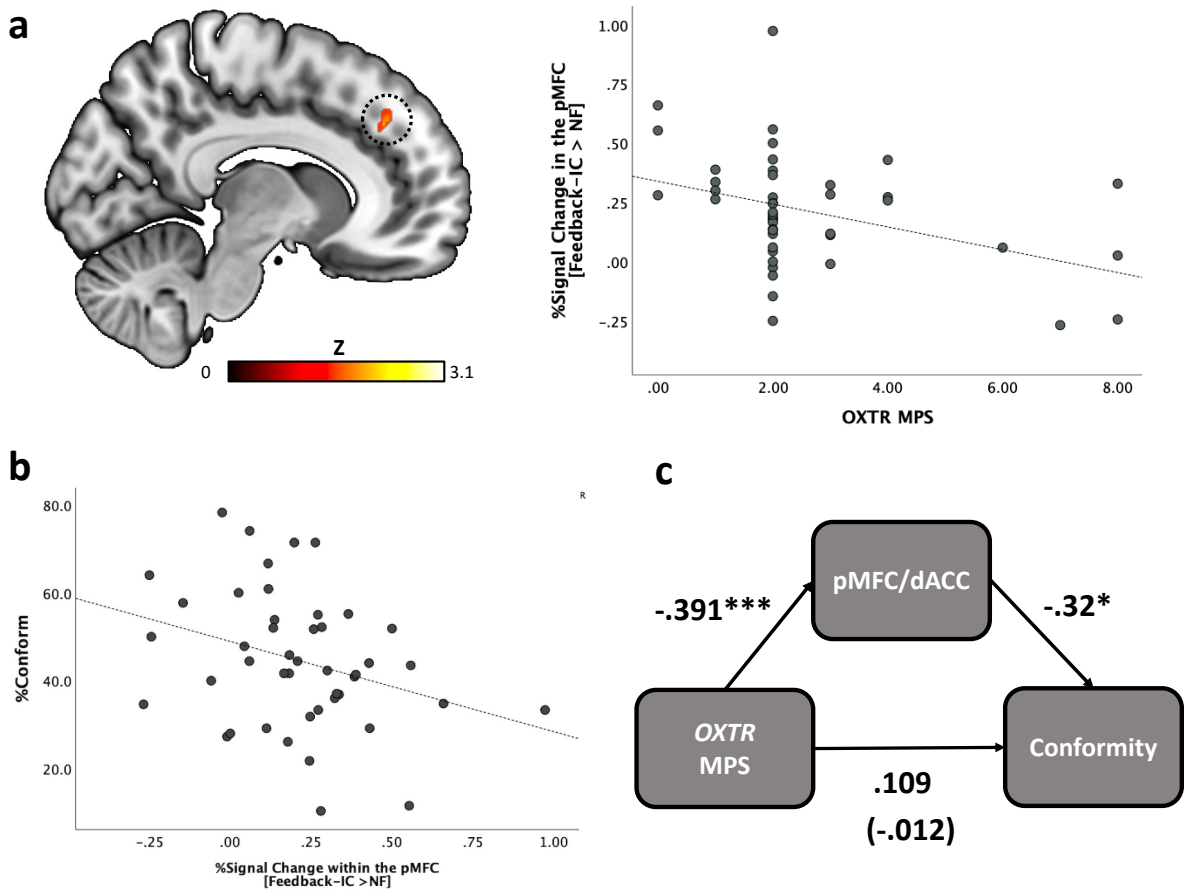


Figure 4-4. The genetic modulation within the pMFC lead to increased behavioral conformity. Voxels in the bilateral ventral caudate nucleus were significantly more active in response to consistent and inconsistent social feedback (peak voxel MNI coordinates: x=12, y=14, z=6). (Red). Voxels in the septal/subgenual region showed increased activations for control feedback condition (peak voxel at: x=2, y=16, z=-34) (Blue).

Table 4-1. Study sample demographics and genotype composition

Demographics	**Neuroimaging arm**		**Behavioral arm**	
	N	*%*	*N*	*%*
Gender				
Female	17	38	54	37
Male	28	60	90	63
Ethnicity				
Asian	20	47	60	42
Black	6	13	17	19
Caucasian	16	36	27	24
Hispanic	2	4	35	12
Others	-	-	4	3

Table 4-2. The list of the seven *OXTR* SNPs used for constructing MPS.

SNP ID	Type	Position	Allele (Frequency)*	
			HEA	LEA
rs73027838	Intron variant	Chr 3:8790184	C (0.1)	G (0.9)
rs9844525	Intron variant	Chr 3:8792759	A (0.7)	G (0.3)
rs3901927	Intron variant	Chr 3:8793793	G (0.4)	A (0.6)
rs77238791	Intron variant	Chr 3:8794567	A (0.1)	G (0.9)
rs56253322	Intron variant	Chr 3:8766599	A (0.1)	G (0.9)
rs73029733	Intron variant	Chr 3:8808030	A (0.1)	G (0.9)
rs53576	Intron variant	Chr 3:8762685	A (0.4)	G (0.6)

*HEA: high-expressing allele; LEA: low-expressing allele

Chapter 4 Supplementary materials

S4-1. Sample size determination and participant allocation strategy

A priori-power analysis

We used the G*Power to conduct a priori power analysis. The reference effect sizes were taken from previous studies that investigated neural (GG vs. AA, Cohen's d =.81) and behavioral effects (GG vs. AA+AG, Cohen's d = .49) of OXTR rs53576 on social cognition involving face and emotion perception (Rodrigues et al. 2009, Luo, Li, et al. 2015). With the type-I error rate set to $\alpha = .05$, the power analysis showed that a total of $N = 50$ (e.g., 25 GG and 25 AA+AG) are required to provide 80% power for detecting a significant main effect of genotype on the neural response associated with socio-emotional processing. For the behavioral task, the power analysis yielded a required sample size of $N = 144$ (GG=53, AA+AG = 91). In sum, by recruiting 200 participants, the current project is expected to have sufficient statistical power for detecting true effects at both neural and behavioral levels.

S4-2. Descriptive statistics for demographic variables

Variable	Behavioral		Neuroimaging	
	M	*SD*	*M*	*SD*
Age	23.67	7.42	21.6	1.88
Political orientation	3.91	0.91	4.42	3.25
Religiosity	2.27	1.24	2.91	2.99

S4-3. Test of regression of means (RTM)

Previous studies suggested that the degree of decision shifts observed in conformity experiments may be confounded by "regression towards the mean (RTM)" where extreme data points tend to become less extreme in a subsequent observation (Izuma and Adolphs 2013, Yu and Chen 2015, Huang, Kendrick, and Yu 2014). In a social conformity experiment that uses an adoptive algorithm to create social misalignment, the specific types of social feedback generated by the algorithm are often constrained by participants' initial ratings. That is, social feedback with an upward (downward) deviation from participants' self-ratings can be presented only when participants' ratings are sufficiently low (high) (Schnuerch, Schnuerch, and Gibbons 2015). Since extreme values are also more likely to be followed by the less-extreme mean in a subsequent measurement, the RTM effect may be convoluted with the feedback-driven decision shifts. Several proposals have been made to assess and correct for the RTM effects.

Following previous suggestions (Yu and Chen 2015), we defined a separate LMM that includes both participants' initial ratings and self-other gap in each trial as continuous predictors for the changes in decision shifts that occurred between S1 and S2. Unstructured covariance matrix and restricted maximum-likelihood estimation was used to better capture the random effect variance. Subjects and trial numbers were included as random factors. The analysis was performed on the combined dataset to test whether our experimental setup and task design were generally susceptible to RTM. Across the behavioral and neuroimaging arm of the experiment, the gap between self and others' ratings during S1 significantly predicted the decision shifts during S2 ($\beta_{\text{Gap}} = .024$, $p < .001$) after controlling for the effect of participants' initial ratings ($\beta_{\text{Initial Rating}} = .105$, $p < .001$).

S4-4. Exploratory analyses on the associations among personality traits, impression rating, and conformity

S4-4-1. Neuroimaging arm

Variable	1	2	3	4	5	6	7
1. Social monitoring (SM)	-						
2. Need for cognition (NfC)	-.133	-					
3. SM	.344*	.283	-				
4. Likability	-.240	-.021	-.02	-			
5. Perceived prosociality	-.038	-.253	.039	.376*	-		
6. Similarity to self	-.210	.054	.030	.553**	.389*	-	
7. %Conformity	.131	.107	-.081	.258	.081	.07	-

*p <.05; **p<.01

S4-4-2. Behavioral arm

Variable	1	2	3	4	5	6	7
1. Social monitoring (SM)	-						
2. Need for cognition (NfC)	-.254**	-					
3. SM	.191*	.081	-				
4. Likability	-.144	-.073	-.094	-			
5. Perceived prosociality	-.09	-.014	-.009	.376**	-		
6. Similarity to self	.05	.-.155	-.109	.539**	.369**	-	
7. %Conformity	-.072	.039	.027	-.065	.569	..938	-

*p <.05; **p<.01

S4-5. Exploratory analyses on the association between *OXTR* and memory accuracy

OT is known to enhance the social memory in both animals in humans (Skuse et al. 2014, Bielsky and Young 2004). The facilitatory effects of OT on social memory raises the possibility that individuals with enhanced OT signaling in the brain may recall the majority opinion more accurately during S2. To test this possibility, we correlated individual participants' *OXTR* MPS with the average distance between the actual majority feedback presented during S1, and recalled majority feedback recorded during S2. The significant relationship between these variables were found among participants in either behavioral or neuroimaging arm (All *Ps* >.121)

S4-6. Activations in the Septal/Subgenual Area

Brain region / Contrast	Cluster size (voxels)	Max Z value	Peak activation MNI coordinate		
			X	Y	Z
NF > Feedback-C					
Septum/Subgenual cortex	31	3.2	0	16	-4
NF > Feedback-IC	1428	5.78	16	-70	60
Septum/Subgenual cortex	166	4.36	0	16	-4

S4-7. The results of the ROI analysis in the exploratory parametric modulation model.

Brain region / Contrast	Cluster size (voxels)	Max Z value	Peak activation MNI coordinate		
			X	Y	Z
Feedback-IC > Feedback-C					
Caudate nucleus (R)	5	2.92	14	19	4
pMFC/dACC (L)	388	4.46	-4	28	38
Feedback-C > Feedback-IC					
No activation					
Feedback-IC > NF					
Caudate nucleus (R)	87	4.03	12	12	6
Caudate nucleus (L)	16	3.19	-12	12	2
pMFC/dACC (L)	389	6.71	6	34	41
NF > Feedback-IC					
Septal/subgenual cortex*	167	4.34	0	16	-4
Feedback-C > NF					
pMFC/dACC (R)	292	4.35	6	34	46
NF > Feedback-C					
Septal/subgenual cortex*	30	3.19	0	16	-4
Feedback-IC$_{PM}$					
No activation					

*only within the NAcc ROI.

S4-8. The results of the exploratory whole-brain analysis

Brain region / Contrast	Cluster size (voxels)	Max Z value	Peak activation MNI coordinate		
			X	Y	Z
Feedback-IC > Feedback-C					
dACC/pMFC	4436	5.71	-2	20	58
IFG/vlPFC (L)	3273	5.08	-48	33	8
Superior parietal lobe/PC (R)	2581	6.75	16	-70	60
Primary visual cortex (R)	2299	4.66	4	-90	8
dorsal caudate nucleus (R)	1456	4.72	14	6	14
IFG/vlPFC (R)	824	5.01	36	30	-8
MFG (R)	692	4.37	48	28	32
LO (L)	4.96	4.56	34	-78	34
Feedback-C > Feedback-IC					
No activation					
Feedback-IC > NF					
Superior parietal lobe/PC (L)	23233	6.9	-28	-54	44
AI (R)	16784	7.14	32	18	-14
AI (L)	7298	6.95	-32	29	-14
Caudate nucleus (R)	1713	5.21	10	12	12
Cerebellum (L)	731	5.47	-4	-52	-42
MTG (R)	533	4.79	62	-32	-8
PCC (L)	470	5.77	-2	-18	30

Brain region / Contrast	Cluster size (voxels)	Max Z value	Peak activation MNI coordinate		
			X	Y	Z
NF > Feedback-IC					
PCC (L)	1424	5.99	-12	-32	44
Angular gyrus (R)	746	4.98	56	-30	26
Posterior insula (R)	716	5.27	42	-4	-6
SMG (R)	663	5.08	-64	-28	26
Posterior insula (L)	517	4.63	-44	-4	0
Parahippocampal gyrus (L)	346	5.05	-30	-36	-14
MFG (L)	345	5.05	-30	34	38
Feedback-C > NF					
Superior parietal lobe/PC (R)	3524	5.61	42	-46	4
IFG/vlPFC (R)	2655	5.11	42	48	-10
Cerebellum	2025	4.7	-32	-68	-46
dmPFC	1846	4.49	6	50	44
Superior parietal lobe (L)	1391	4.37	-34	-54	46
MFG (R)	874	5.05	52	32	36
AI (L)	571	4.73	-28	16	-14
MTG (R)	450	4.11	-62	-32	34
OFC (L)	444	3.81	-10	60	-10
Primary visual cortex (R)	372	3.81	30	-96	-8

Brain region / Contrast	Cluster size (voxels)	Max Z value	Peak activation MNI coordinate		
			X	Y	Z
NF > Feedback-C					
Cuneus (R)	465	4.18	4	-88	32
Precuneus (L)	433	4.1	-6	-54	8

Chapter 5

Conclusion

Cultural norm acquisition is a process through which individuals embody values and norms prevalent in their respective social environments. This dissertation project aimed to explore specific genetic and neuro-cognitive mechanisms that mediate cultural norm acquisition.

I focused on the role of genetic variants in the oxytocin receptor gene (*OXTR*), as it has been widely implicated in various facets of mammalian and human sociality. Drawing upon recent theoretical models that highlighted 1) the role of *OXTR* in regulating social sensitivity (Kitayama et al., 2017; Sasaki et al., 2016) and 2) the need for principled analysis as to how oxytocin signaling modulates multi-stage decision making processes in humans (Piva and Chang, 2018), I conducted three imaging genetics experiments that investigated the following topics: how genetic variation in *OXTR* rs53576 modulates the intermediate neural mechanisms involved with 1) the detection of positive and negative facial micro-expressions, 2) the accurate identification of the authenticity of smiles, and 3) how *OXTR* MPS reactivity towards conformity pressure imposed on the domain of cultural and moral values.

In this final chapter, I will first summarize the key findings of this dissertation project. I will then discuss the implications of these results. Lastly, I will conclude the chapter with possible ways to extend the scope of this research in future projects.

Summary of Main Findings

***OXTR* and the neural basis of facial micro-expressions processing** Facial expressions of emotion are a primary medium of human social communication. While previous research has shown that INOT treatment and certain allelic variations in *OXTR* single nucleotide polymorphisms (SNPs) can promote the recognition of facial emotions (Lopatina, 2018), the generalizability of the related works has been limited by the fact that most of these studies employed *static* pictures of *fully-expressed* facial emotions, which are of relatively low ecological validity. Addressing these limitations, I explored the regulatory role of *OXTR* on peoples' ability to detect *dynamic facial micro-expressions*.

The specific behavioral/fMRI predictions of the experiment 1 is provided below **(Table 5-1)**. The results of the study are also summarized in **Table 5-2,** with the important findings numbered and highlighted in red.

Table 5-1. The main predictions of Experiment 1.

Prediction by data type	Description
Behavioral prediction	• **1-a:** Participants will perform above chance level. • **1-b**: Participants will show better performance for macro-exxressions. • **1-c**: *OXTR* rs5376 G homozygotes will show higher average task performance than the A allele carriers. • **1-d**: G homozygotes will perform better than the A allele carriers especially for micro-expressions.
fMRI prediction	• **2-a**: Perception of macro/micro-expressions will recruit brain regions sensitive to dynamic aspects of face **(e.g**., STS, IFG) • **2-b:** Perception of macro/micro-expressions will recruit brain regions sensitive to dynamic aspects of face and emotion perception (e.g., amygdala, Nacc, vmPFC). • **2-c:** Perception of macro/micro-expressions will recruit brain regions implicated in affective (AI, dACC) and cognitive empathy (e.g., rTPJ, mPFC, PC) • **2-d:** Activations in these regions will be stronger for facial macro-expressions compared to micro-expressions. • **2-e**: G homozygotes will show increased activations during the perception of the experimental stimuli, especially for the facial micro-expressions.

Table 5-2. Summary of the main findings of Experiment 1.

Data type	Summary of the main findings	Relevance to previous literature	Relevance to the main predictions
Behavioral/ fMRI findings	• Participants i) *successfully detected the valence of dynamic facial micro- and macro-expressions*,	i. Novel finding	i. Supported prediction 1a
	• Task performance was higher for ii) *macro-expressions,* and iii) *positively-valenced expressions.*	ii. Replication iii. Replication /Extension*	ii. Supported prediction 1b iii. Not relevant
	• Perception of facial micro- and macro-expressions, especially the latter, recruited **iv)** *brain regions implicated in dynamic face processing, emotion perception, and affective cognitive empathy*	iv. Replication /Extension**	iv. Supported prediction 2a-2d
	• Perception of micro-expressions recruited **v)** *a wider network of brain regions than what was previously reported.*	v. Novel finding	v. Supported prediction 1a
Findings related to *OXTR*	• G homozygotes in the neuroimaging arm showed **vi)** *increased task performance for facial micro-expressions*.	vi. Novel finding	vi. Supported prediction 1d, but not 1c.
	• Among G homozygotes, **vii)** *the brain regions implicated in attentional control (e.g., SMG, AI, IFG, and STS) showed increased activations in response to negative micro-expressions.*	vii. Novel finding	vii. Supported prediction 2b
Limitations/ Implication	• Finding vi) was not replicated in the behavioral arm. • Overall, genetic variations in *OXTR* may enhance the detection of subtle evaluative social feedback by regulating visual attention to relevant social cues. This is consistent with social salience hypothesis of OT. • This may facilitate cultural norm acquisition mediated by facial expression in everyday social interaction		

*, **Replication of well-established findings, yet using the novel experimental stimuli

***OXTR* and the neural basis of the smile authenticity judgments** It is widely known that people can fake facial expressions of emotions to manipulate others' social behaviors. Therefore, it becomes crucial for us to not only detect social cues from our environments but also to correctly determine the *authenticity* of such social cues. In a novel experiment where participants viewed a set of high-quality video recordings of dynamic expressions of genuine vs. posed smiles, we examined whether allelic variation in *OXTR* rs53576 was associated with individual differences in the ability to determine smile authenticity.

The specific behavioral/fMRI predictions of the experiment 2 are provided below **(Table 5-3)**, which will be followed by the summary of the main findings of the study **(Table 5-4).**

Table 5-3. The main predictions of Experiment 2.

Prediction by data type	Description
Behavioral prediction	• **1-a:** Participants will perform above chance level. • **1-b**: G homozygotes will show higher task performance than the A allele carriers.
fMRI prediction	• **2-a**: Perception of macro/micro-expressions will recruit brain regions sensitive to dynamic aspects of face **(e.g**., STS, IFG) and facial emotion (e.g., amygdala, Nacc, vmPFC). • **2-b:** Perception genuine and posed smiles will recruit brain regions implicated in affective (AI, dACC) and cognitive empathy (e.g., rTPJ, mPFC, PC) • **2-c:** Activations in these regions will be stronger for genuine smiles compared to poses smiles • **2-d**: G homozygotes will show increased activations in these areas when perceiving genuine vs. posed smiles.

Table 5-4. Summary of the main findings of Experiment 2.

Data type	Summary of the main findings	Relevance to previous literature	Relevance to the main predictions
Behavioral/ fMRI findings	• Participants i) *successfully discriminated genuine smiles from posed smiles.*	i. Replication /Extension*	i. Supported prediction 1a
	• Perception of smiling faces recruited **ii)** *brain regions implicated in dynamic face processing (e.g., STS), emotion perception (e.g., Amygdala), reward processing (e.g., NAcc, caudate nucleus, mOFC), and affective cognitive empathy (e.g., dACC, AI, mPFC).*	ii. Novel finding	ii. Supported prediction 2a-b.
	• Brain areas implicated in **iii)** *sensorimotor simulation (e.g., putamen, S2), reward (e.g., mOFC) and theory of mind (e.g., dmPFC) showed stronger activations for correctly identified genuine vs. posed smiles.*	iii. Replication /Extension**	iii. Partially supported prediction 2c.
	• The average activation in the **iv)** *IFG during the perception of genuine and posed smiles, as opposed neutral face linearly tracked participants' overall perceptual accuracy (d prime).*	iv. Novel finding	iv. Supported prediction 2b
	• Increased activations in the **v)** *dACC/dmPFC during the perception of genuine vs. posed smiles were associated with conservative decision bias.*	v. Novel finding	v. Not directly relevant.
Findings related to *OXTR*	• G homozygotes in the neuroimaging arm tended to **vi)** *erroneously judge posed smiles as genuine smiles (e.g., higher false alarm rate).*	vi. Novel finding	vi. Did not support prediction 1b.
	• Among G homozygotes, **vii)** *the activation within the dmPFC and rTPJ was reduced during the perception of genuine vs. posed smile.*	vii. Novel finding	vii. Did not support prediction 2d.
Limitations/ Implication	• Finding 6) was not replicated in the behavioral arm. • Overall, genetic variations in *OXTR* may enhance the detection of subtle evaluative social feedback by regulating visual attention to relevant social cues. This is consistent with social salience hypothesis of OT. • This may facilitate cultural norm acquisition mediated by facial expression in everyday social interaction		

*, ** Replication of previous findings using different stimuli modality

***OXTR* and the neural basis of moral conformity** Besides evaluative social cues conveyed via facial expressions, direct majority feedback is another important source of social influence that can guide the cultural norm acquisition process. Previous studies have implicated OT in social conformity, where people align their behaviors and beliefs with those of others to either gain correct knowledge of their social environment or to obtain approval (Cialdini and Goldstein, 2004). We explored whether the allelic variations in *OXTR* SNPs linked with enhanced OT receptor expressions (i.e., *OXTR* multi-locus genetic profile score) in the brain would regulate people's susceptibility towards conformity pressure, as well as its underlying neural mechanisms. To test and establish a more specific link between *OXTR* and the endorsement of normative behaviors, we investigated whether people change their personal priorities in moral values and virtues when confronted by an opposing majority.

The specific behavioral/fMRI predictions of experiment 3 are provided below **(Table 5-5)**, followed by the summary of the main findings of the study **(Table 5-6).**

Table 5-5. The main predictions of Experiment 2.

Prediction by data type	Description
Behavioral prediction	• **1-a:** Participants will show behavioral conformity in the domain of moral values and virtues • **1-b**: Individuals with higher *OXTR* MPS values will be more likely to conform following social misalignment.
fMRI prediction	• **2-a**: Perceived social misalignment and alignment with others will incur activations in the NAcc and pMFC/dACC, two key nodes in the gap/error detection mechanisms in the brain. • **2-b:** Participants with higher *OXTR* MPS will show either 1) higher activations in the NAcc in response to consistent social feedback, or 2) higher activations in the pMFC/dACC in response to inconsistent social feedback.

Table 5-6. Summary of the main findings of Experiment 3.

Data type	Summary of the main findings	Relevance to previous literature	Relevance to the main predictions
Behavioral/ fMRI findings	• Participants showed **i)** *behavioral conformity, shifting their value preference following majority feedback.*	i. Replication /Extension*	i. Supported prediction 1a
	• Inconsistent social feedback was associated with the increased activations in **ii)** *both the NAcc and pMFC/dACC.*	ii. Novel finding	ii. Partially supported prediction 2a
	• Activations in the pMFC/dACC during inconsistent social feedback trials **iii)** *negatively predicted conformity.*	iii. Novel finding	iii. Partially supported prediction 2a.
	• Neither **iv)** *the Nacc nor pMFC/dACC were encoding the magnitude of self-other gap in value preference ratings*.	iv. Novel finding	iv. Did not support prediction 2a.
Findings related to *OXTR*	• Participants with higher *OXTR* MPS showed **v)** *decreased activations in the pMFC/dACC in response to social misalignment, which in turn, positively predicted conformity.*	v. Novel finding	vi. Did not support prediction 2b.
	• The reduced pMFC/dACC activation in response to inconsistent social feedback **vi)** *mediated the link between higher OXTR MPS and increased conformity.*	vi. Novel finding	vii. Supported prediction 1b.
Limitations/ Implication	• Social alignment and misalignment in the domain of morality modulated the activations in the NAcc and pMFC/dACC. • Yet, the prevailing neuro-cognitive model of social conformity, which centers on reinforcement learning and prediction errors, cannot adequately explain our data. • The specific cognitive functions represented in NAcc and pMFC may be context-specific • The pMFC activations in this study may reflect internal conflict associated with overriding one's value preference. • Enhanced OT signaling in the brain may reduce the internal conflict by augmenting the value of social affiliation (over independence or non-conformity), which, in turn, promotes conformity. • *OXTR* may enhance cultural norm acquisition by modulating the value of affiliative goals. • Using *OXTR* MPS to capture the individual variations in endogenous OT signaling holds promise.		

*Replicated the social conformity effect in the domain of moral values and virtu

Significance

The topic of cultural norm acquisition has long been studied in multiple disciplines. However, it has rarely been examined in a way that brings genetic, behavioral, and neural data together, although the need for an integrative approach to human sociality has been recognized by many authors (Hofmann et al. 2014, Han and Ma 2015, Kitayama et al. 2016). This dissertation project investigated whether and how the genetic variation in *OXTR* modulates various intermediate neurocognitive mechanisms that may subserve the social learning of local norms and values. It thus has a general methodological implication for ongoing endeavors to connect different levels of analyses to achieve a more integrative understanding of the nature of complex human sociality.

Implications for anthropologists

How human biology is intertwined with social influences, and how human variations emerge at the confluence of these two inheritance systems have long been at the heart of the anthropological inquiry. Findings of this research project point to the genetic and neurocognitive mechanisms through which the diverse patterning of norms and moral values may emerge within individuals over time. Importantly, these mechanisms produce plastic responses to external feedback rather than fixed phenotypes. Therefore, the findings of this research overcome the alleged dichotomy between biology and social environment while highlighting the truly biocultural nature of human sociality.

Mechanism-based approaches to human variation have often been underappreciated in anthropology (Fawcett, Marshall, and Higginson 2015). The idea of the phenotypic gambit, for

example, holds that the evolution of complex traits can be modeled as if a single gene controlled them since natural selection will favor traits linked with high fitness, regardless of the specific mechanisms underlying them (Grafen 1991). While this is often an inevitable choice for modeling convenience, the phenotypic gambit may sometimes fail due to 1) ignoring mechanistic constraints that limit an organism's ability to behave optimally, 2) an over-simplifying assumption that there will be no individual differences in the ways organisms in a given species respond to the environment. Yet, findings from this research project show that people do not show the same neural or behavioral responses towards the identical social inputs (e.g., majority feedback or negative facial expressions), which could lead to differential phenotypic traits (e.g., differing degrees of conformity behaviors). Previous evidence from simulation studies shows that individual variation in such social learning capacity may contribute to the emergence of dramatically different evolutionary stable strategies (ESS) at the level of the population (Mesoudi et al. 2016, Fawcett, Marshall, and Higginson 2015). In all, findings of this study that highlight individual variation in the cultural norm acquisition process can inform anthropological endeavors to correctly characterize the evolution of various social phenotypes in humans.

Lastly, this research also has a more general implication for the study of social learning. Considerable cross-specific variation exists in social learning capacity across mammalian taxa, including primates (van Schaik and Burkart 2011). Many anthropological studies have focused on structural characteristics of the brain, such as the size (Reader and Laland 2002) and white matter connectivity (Hecht et al. 2013) to elucidate the proximate mechanisms underlying the uniquely human ability for social learning. This endeavor is important, as social learning is thought to be pivotal to cumulative culture and cultural evolution (Dean et al. 2012, Henrich 2015). Findings in this research point to the possibility that the neurochemical organization of

the brain may also play a role in shaping individuals' ability and motivation to detect behaviorally relevant social signals and change their behaviors accordingly. It has been shown that 1) the pattern of central *OXTR* expression varies considerably across species (Freeman and Young 2016) and that 2) different primate species may exhibit distinguishable neurochemical profiles (Raghanti et al. 2018). While this research project only concerned variation among human subjects, brain-wide *OXTR* expression profiles may also account for the cross-specific variation in social learning in general.

Implications for the psychologists and neuroscientists

Most psychological and neuroscientific endeavors have focused on 1) characterizing the cognitive architecture of human moral capacity and 2) identifying brain areas that support explicit moral judgments or prosocial behaviors. The question of "moral learning", or how individuals assign or change values to morally relevant behaviors and beliefs in the course of development, has received attention only recently (Cushman, Kumar, and Railton 2017). Findings from this research project provide preliminary evidence that the brain mechanisms involved in error-detection and reinforcement learning such as NAcc and dACC/pMFC may be involved in moral learning process, although the specific patterns of activations within these areas differed from the established neural signature of social feedback processing (Wu, Luo, and Feng 2016). In addition, our findings also suggest the possible role of *OXTR* in regulating this process. Altogether, the results of this study could potentially support an emerging view in psychology and social neuroscience that multiple domain-general valuation mechanisms play a critical role in the formation of moral values and beliefs. This research project also highlights the need for studying various genetic components that may alter the early level of information

processing (e.g., detection and discrimination of positively and negatively valenced social cues) that feeds into the value computation processes in the brain (Falk, Way, and Jasinska 2012).

This research project also has implications for the study of oxytocin (OT) and human sociality. Despite a considerable body of evidence supporting the facilitatory effects of OT on human social cognition and behaviors, concerns have been raised as to the 1) inhomogeneity in research methodology that could lead to diverging findings (McCullough, Churchland, and Mendez 2013), 2) relatively small sample size that inflates risk for false-positive findings (Walum, Waldman, and Young 2016), 3) and overreliance of a small subset of *OXTR* SNPs to model the effect of endogenous OT signaling (Feldman et al. 2016, Bakermans-Kranenburg and van IJzendoorn 2014). In two out of three experiments in this dissertation project, the significant effect of *OXTR* rs53576 was not reliably replicated across the behavioral and neuroimaging arms. The lack of robust effects of the *OXTR* genotype in this study suggests that replications should be made with a larger sample size and the novel research methodology that allows researchers to model the effect of individual SNPs more sensitively. *OXTR* MPS used in the third experiment can be one way to address this methodological limitation. The use of cumulative polygenic risk score or allele-dosage scores have yielded fruitful results in behavioral and imaging genetics (Bigos and Weinberger 2010). *OXTR* MPS constructed based on the actual expression levels may complement the existing approaches in the related disciplines, by offering a novel method to represent the 1) effect of multiple genetic factors 2) with an anatomically-grounded way. This would enable researchers to infer more specific intermediate mechanisms through which certain genetic polymorphisms may influence their downstream phenotypes.

Implications for the study of gene-culture interaction and norm sensitivity hypothesis in social genomics

As noted in the first chapter, the existing gene-culture interaction literature has relied heavily on the associations between certain genetic variations and high-level psychosocial traits defined in terms of survey responses (Kim et al. 2011, Sasaki 2013). The lack of specific mechanistic models, however, prevents researchers from drawing a firm conclusion that the implicated genes are indeed causally linked with the development of culture-specific phenotypes in a biologically meaningful way. The norm sensitivity hypothesis (Kitayama et al. 2016) predicted that a set of so-called "sensitivity genes" would modulate the reinforcement learning mechanisms in the brain such that some individuals can process the social feedback to update their behaviors more effectively than others. Findings from this dissertation project partially support this idea. Specifically, *OXTR* rs53576, and other *OXTR* SNPs that share a similar expression profile in the brain with *OXTR* rs53576, may contribute to the increased endorsement of cultural norms in two ways: 1) by modulating individuals' ability to detect subtle feedback cues associated with negative moral evaluation, such as disgust and anger; 2) augmenting the value of social affiliation, which promote social learning of values by means of moral conformity. However, our findings also suggest that the *OXTR* allele associated with the increased detection of the negative micro-expressions may simultaneously hamper the carriers' ability to discern genuine vs. posed emotional expressions. These mixed findings, while requiring further replications, suggest that genetic modulation of social sensitivity may not necessarily result in more efficient cultural norm acquisition depending on the specific types of normative social influence being imposed. Overall, our data indicate the need for more thorough characterization of the mechanisms mediating the cultural norm acquisition process.

Possible Improvements and Ideas for Future studies

Limitations of univariate analyses of BOLD fMRI signals

In this research project, the effect of *OXTR* rs53576 on BOLD signal has been modeled based on a univariate approach. However, evidence indicates that OT may not necessarily increase or decrease the activation level of individual brain regions. Instead, it may regulate social cognitive functions and behaviors by altering the overall connectivity among multiple brain regions (Johnson et al. 2017, Rilling et al. 2018). Therefore, it would be worthwhile to use a multi-variate approach (e.g., network analysis, Fornito et al., 2016), targeting brain regions that are known to 1) express *OXTR*, and 2) participate in cognitive functions relevant for the tasks used in this study such as face processing or reinforcement learning.

Beyond the genetic variations in the *OXTR*: the multiple routes to social sensitivity

While this study focused on the genetic variations in the *OXTR*, it would be wrong to assert that endogenous OT signaling is the sole determinant of individual variations in social sensitivity and thus more effective cultural norm acquisition process. For example, recent studies have suggested that there are at least three candidate genes that may operate similarly with *OXTR* in regulating social sensitivity and behavioral plasticity (i.e., *5HTR1A, SERT*, and *DRD4*) (Sasaki et al. 2016). LeClair et al (2014) empirically showed that a polygenic index created based on the allelic variations in these genes correlated with the degree of culture-specific psycho-behavioral traits among East Asian- and North American populations, just as the G allele of *OXTR* rs53576 did in the earlier studies (LeClair, Janusonis, and Kim 2014). The "multi-gene"

approach may allow us to better capture the genetic basis of cultural norm acquisition explored in this study.

Relatedly, the social effects of OT themselves are known to be dependent on other neuromodulators. Most notable examples of these include serotonin and dopamine (MacDonald and MacDonald 2010). Previous animal studies have shown that DA neurons in the midbrain (e.g., ventral tegmentum area) and serotonin signaling in the NAcc critically mediate the prosocial effects of OT (Hung et al. 2017, Dölen et al. 2013). These findings strongly suggest that research in social sensitivity will benefit from broadening the scope of investigation such that information about other genes that may regulate endogenous OT signaling in the brain could also be considered together with the *OXTR* genotypes. Most promising targets would include the dopaminergic, muscarinic, vasopressin and opioid genes that are known to co-express with OXTR in various brain regions (Quintana et al. 2019).

It would be worthwhile to also investigate how epigenetic mechanisms such as DNA methylation regulate the effects of *OXTR* SNPs on their social and non-social phenotypes (Andari and Rilling 2021). An emerging body of evidence shows that *OXTR* DNA methylation is linked with the etiology of autism spectrum disorder at the level of behavior and the brain (Puglia, Connelly, and Morris 2018, Andari et al. 2020). Yet, the specific environmental conditions that may influence the degree of *OXTR* DNA methylation (e.g., childhood maltreatment), and how different levels of methylation found in a specific CpG site(s) modulate the endogenous OT signaling and its downstream phenotypes either independently of or in tandem with the allelic variations in *OXTR* are largely not known. Future research should incorporate both methylation data as well as the *OXTR* genotypes such as *OXTR* MPS to study

how these variables influence various social phenotypes previously known to be influenced by OT, including social sensitivity.

Relatedly, according to the GTEx eQTL database, the G allele of *OXTR* rs53576 is associated with low OT receptor expression in the brain regions such as the striatum, hippocampus. A similar negative association between the G allele and *OXTR* mRNA expression has recently been identified in the human ACC (Almeida, 2022). This is puzzling as increased *OXTR* expression in the reward-sensitive brain regions has often been associated with enhanced sociality in animal studies (Donaldson and Young, 2008), and 2) the G allele has been linked with sensitive social cognition and behaviors in humans (Li et al. 2015). In this study, likewise, we found preliminary evidence that *rs*53576 could aid the perception of subtle facial expressions of emotion (e.g., micro-expressions).

These paradoxical findings imply that *OXTR* expression may not be equated blindly with enhanced social cognition. It is likely that the complex interplay between the *OXTR* expression, methylation, and the availability of endogenous OT would all contribute to the overall functionality of the OT system. For instance, rather paradoxically, a high *OXTR* level may reflect the compensatory upregulation of the receptor caused by the lack of receptor ligand availability in the system (Reuter et al. 2017). It is also possible that lower receptor density may increase the chance of receptor saturation, which could increase the effectiveness of the OT signaling. Lastly, the link between *OXTR* expression and enhanced sociality may vary across different brain regions or the domains of sociality measured in different experimental contexts. A further investigation would be necessary to determine the relationships between *OXTR* genotypes, methylation levels, and OT functionality.

Study samples with different demographic characteristics

The generalizability of research findings is critically constrained by the demographic characteristics represented in the study sample. Two demographic variables hold particular importance when it comes to the topic of cultural norm acquisition and *OXTR:* age and cultural background of participants.

Age is a critical factor that needs further consideration, as there exist "sensitive periods" during which people's brain, behaviors, and psychology are rapidly shaped by experience (Knudsen 2004). Middle childhood (Age 6-9) and adolescence (Age 10-13) are believed to be the two important sensitive periods in humans. Evidence suggests that people who are in or transitioning into these phases may respond differently to social and environmental influences, potentially due to the unique patterns of gene expression, bodily and neural development, hormonal shifts, and psychological maturation (DelGiudice 2018, Konner 2010). It is well-known, for example, that adolescents typically show heightened neural responses to social cues that imply approval or rejection from peers compared to younger or older individuals (Fuhrmann, Knoll, and Blakemore 2015). Intriguingly, central *OXTR* expression in animals and humans is known to be dynamic and changes throughout the lifespan (Vaidyanathan and Hammock 2017, Rokicki et al. 2022). The existence of the sensitive periods in humans, when considered together with the age-dependent changes in endogenous OT signaling in the brain, strongly suggests that the relationship between *OXTR* SNPs and social cognition tested and observed in this study may vary depending on the specific life stages of participants. For instance, it may be possible that the salience-enhancing property of the *OXTR* rs53576 G allele can be more pronounced among those in middle childhood or adolescence.

In fact, middle childhood seems to be a particularly important period during which individuals endorse and internalize behavioral norms and values specific to their respective communities (DelGiudice 2018). In a series of cross-cultural studies, Bailey House and colleagues convincingly demonstrated the developmental trajectories of prosocial behaviors across different human populations, with middle childhood (Age 6-9) being a critical watershed from which society-specific responses in economic decision games begin to emerge (House et al. 2013). Note that the present study involved healthy adult participants with a mean age of 24.5. In all, future studies should explore whether genetic variation in *OXTR* (or allelic alterations in other candidate sensitivity genes) exert age-specific effects on the behavioral or neural responses to normative social feedback conveyed in forms of positive or negative facial expressions or conformity pressure.

Besides participants' age, another important topic that should be addressed in follow-up studies on *OXTR* and social learning of norms and values is the cultural backgrounds of participants. Although ethnically diverse, the majority of the participants in this study were U.S. citizens or international students who have finished higher education in the United States. It is thus possible that their prior knowledge of social practices, norms, and values in the U.S. might have reduced the salience of normative social feedback implemented in this research (e.g., conformity pressure). In this regard, studying populations with limited cultural experiences in the U.S., such as immigrants or new international students, would offer us a better opportunity to examine and detect the link between *OXTR* and the cultural norm acquisition process taking place in adulthood. Supporting this possibility, anthropological and psychological literature on the acculturation process have reported evidence of heightened social sensitivity among immigrants (Sheung et al., 2011; Hynie, 1996; Christmas et al., 2014), especially when it comes to the domain of values (Al Wekhian, 2016; Rosenthal et al., 1989, Phalet et al., 2018).

In all, studying various populations with differential demographic traits would be important to gain further insights into the ways *OXTR* shapes social adaptation via cultural norm acquisition.

Bibliography

Aarts, Henk, Ap Dijksterhuis, and Ruud Custers. 2003. "Automatic normative behavior in environments: The moderating role of conformity in activating situational norms." *Social cognition* 21 (6):447-464.

Adolphs, Ralph, and Michael Spezio. 2006. "Role of the amygdala in processing visual social stimuli." *Progress in brain research* 156:363-378.

Almeida, Daniel, Laura M Fiori, Gary G Chen, Zahia Aouabed, Pierre-Eric Lutz, Tie-Yuan Zhang, Naguib Mechawar, Michael J Meaney, and Gustavo Turecki. 2022. "Oxytocin receptor expression and epigenetic regulation in the anterior cingulate cortex of individuals with a history of severe childhood abuse." *Psychoneuroendocrinology* 136:105600.

Amting, Jayna M, Jodi E Miller, Melody Chow, and Derek GV Mitchell. 2009. "Getting mixed messages: The impact of conflicting social signals on the brain's target emotional response." *Neuroimage* 47 (4):1950-1959.

Andari, Elissar, Shota Nishitani, Gopinath Kaundinya, Gabriella A Caceres, Michael J Morrier, Opal Ousley, Alicia K Smith, Joseph F Cubells, and Larry J Young. 2020. "Epigenetic modification of the oxytocin receptor gene: implications for autism symptom severity and brain functional connectivity." *Neuropsychopharmacology* 45 (7):1150-1158.

Andari, Elissar, and James K Rilling. 2021. "Genetic and epigenetic modulation of the oxytocin receptor and implications for autism." *Neuropsychopharmacology* 46 (1):241.

Arueti, Maayan, Nufar Perach-Barzilay, Michael M Tsoory, Barry Berger, Nir Getter, and Simone G Shamay-Tsoory. 2013. "When two become one: the role of oxytocin in interpersonal coordination and cooperation." *Journal of cognitive neuroscience* 25 (9):1418-1427.

Asad, Talal. 1993. *Genealogies of religion: Discipline and reasons of power in Christianity and Islam*: JHU Press.

Asch, Solomon E. 1956. "Studies of independence and conformity: I. A minority of one against a unanimous majority." *Psychological monographs: General and applied* 70 (9):1.

Atkinson, Anthony P, and Ralph Adolphs. 2005. "Visual emotion perception: Mechanisms and processes."

Auer, Brandon J, Jennifer Byrd-Craven, DeMond M Grant, and Douglas A Granger. 2015. "Common oxytocin receptor gene variant interacts with rejection sensitivity to influence cortisol reactivity during negative evaluation." *Hormones and behavior* 75:64-69.

Ayala, Francisco J. 2010. "The difference of being human: Morality." *Proceedings of the National Academy of Sciences* 107 (Supplement 2):9015-9022.

Aydogan, Gökhan, Andrea Jobst, Kimberlee D'Ardenne, Norbert Müller, and Martin G Kocher. 2017. "The detrimental effects of oxytocin-induced conformity on dishonesty in competition." *Psychological science* 28 (6):751-759.

Bakeman, Roger, Lauren B Adamson, Melvin Konner, and Ronald G Barr. 1990. "! Kung infancy: The social context of object exploration." *Child Development* 61 (3):794-809.

Bakermans-Kranenburg, Marian J, and Marinus H Van Ijzendoorn. 2011. "Differential susceptibility to rearing environment depending on dopamine-related genes: New evidence and a meta-analysis." *Development and psychopathology* 23 (1):39-52.

Bakermans-Kranenburg, Marian J, and Marinus H van IJzendoorn. 2014. "A sociability gene? Meta-analysis of oxytocin receptor genotype effects in humans." *Psychiatric genetics* 24 (2):45-51.

Bandura, Albert. 1973. *Aggression: A social learning analysis*: prentice-hall.

Bandura, Albert, and Frederick J McDonald. 1963. "Influence of social reinforcement and the

behavior of models in shaping children's moral judgment." *The Journal of Abnormal and Social Psychology* 67 (3):274.

Bandura, Albert, Dorothea Ross, and Sheila A Ross. 1961. "Transmission of aggression through imitation of aggressive models." *The Journal of Abnormal and Social Psychology* 63 (3):575.

Bargh, John A, and Tanya L Chartrand. 1999. "The unbearable automaticity of being." *American psychologist* 54 (7):462.

Bartal, Inbal Ben-Ami, Jean Decety, and Peggy Mason. 2011. "Empathy and pro-social behavior in rats." *Science* 334 (6061):1427-1430.

Bartz, Jennifer A. 2016. "Oxytocin and the pharmacological dissection of affiliation." *Current Directions in Psychological Science* 25 (2):104-110.

Bartz, Jennifer A, Jamil Zaki, Niall Bolger, and Kevin N Ochsner. 2011. "Social effects of oxytocin in humans: context and person matter." *Trends in cognitive sciences* 15 (7):301-309.

Bassili, John N. 1979. "Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face." *Journal of personality and social psychology* 37 (11):2049.

Batson, C Daniel, Nadia Ahmad, David A Lishner, and J Tsang. 2016. "Empathy and altruism." *Oxford handbook of hypo-egoic phenomena: Theory and research on the quiet ego*:161-174.

Batson, C Daniel, Bruce D Duncan, Paula Ackerman, Terese Buckley, and Kimberly Birch. 1981. "Is empathic emotion a source of altruistic motivation?" *Journal of personality and Social Psychology* 40 (2).

Baumgartner, Thomas, Markus Heinrichs, Aline Vonlanthen, Urs Fischbacher, and Ernst Fehr. 2008. "Oxytocin shapes the neural circuitry of trust and trust adaptation in humans." *Neuron* 58 (4):639-650.

Beaudry, Olivia, Annie Roy-Charland, Melanie Perron, Isabelle Cormier, and Roxane Tapp. 2014. "Featural processing in recognition of emotional facial expressions." *Cognition & emotion* 28 (3):416-432.

Bhanji, Jamil P, and Mauricio R Delgado. 2014. "The social brain and reward: social information processing in the human striatum." *Wiley Interdisciplinary Reviews: Cognitive Science* 5 (1):61-73.

Bielsky, Isadora F, and Larry J Young. 2004. "Oxytocin, vasopressin, and social recognition in mammals." *Peptides* 25 (9):1565-1574.

Bigos, Kristin L, and Daniel R Weinberger. 2010. "Imaging genetics—days of future past." *Neuroimage* 53 (3):804-809.

Blake, PR, K McAuliffe, J Corbit, TC Callaghan, O Barry, A Bowie, L Kleutsch, KL Kramer, E Ross, and H Vongsachang. 2015. "The ontogeny of fairness in seven societies." *Nature* 528 (7581):258.

Bogdan, Ryan, Betty Jo Salmeron, Caitlin E Carey, Arpana Agrawal, Vince D Calhoun, Hugh Garavan, Ahmad R Hariri, Andreas Heinz, Matthew N Hill, and Andrew Holmes. 2017. "Imaging genetics and genomics in psychiatry: a critical review of progress and potential." *Biological psychiatry* 82 (3):165-175.

Bosch, Oliver J, Simone L Meddle, Daniela I Beiderbeck, Alison J Douglas, and Inga D Neumann. 2005. "Brain oxytocin correlates with maternal aggression: link to anxiety." *Journal of Neuroscience* 25 (29):6807-6815.

Botvinick, Matthew M, Jonathan D Cohen, and Cameron S Carter. 2004. "Conflict monitoring and anterior cingulate cortex: an update." *Trends in cognitive sciences* 8 (12):539-546.

Bourdieu, P., Nice, R. . 1977. *Outline of a Theory of Practice* Vol. 16: Cambridge: Cambridge university press.

Bourdieu, Pierre, and Richard Nice. 1977. *Outline of a Theory of Practice*. Vol. 16: Cambridge university press Cambridge.

Boyd, Robert, and Peter J Richerson. 1992. "Punishment allows the evolution of cooperation (or anything else) in sizable groups." *Ethology and sociobiology* 13 (3):171-195.

Boyette, Adam Howell. 2013. *Social learning during middle childhood among Aka foragers and Ngandu farmers of the Central African Republic*: Washington State University.

Brosnan, Sarah F, and Frans BM De Waal. 2003. "Monkeys reject unequal pay." *Nature* 425 (6955):297.

Brosnan, Sarah F, and Frans BM de Waal. 2014. "Evolution of responses to (un) fairness." *Science* 346 (6207):1251776.

Brüne, Martin, and Ute Brüne-Cohrs. 2006. "Theory of mind—evolution, ontogeny, brain mechanisms and psychopathology." *Neuroscience & Biobehavioral Reviews* 30 (4):437-455.

Buckholtz, Joshua W, and René Marois. 2012. "The roots of modern justice: cognitive and neural foundations of social norms and their enforcement." *Nature neuroscience* 15 (5):655.

Budell, Lesley, Phillip Jackson, and Pierre Rainville. 2010. "Brain responses to facial expressions of pain: emotional or motor mirroring?" *Neuroimage* 53 (1):355-363.

Burkart, Judith M, Ernst Fehr, Charles Efferson, and Carel P van Schaik. 2007. "Other-regarding preferences in a non-human primate: Common marmosets provision food altruistically." *Proceedings of the National Academy of Sciences* 104 (50):19762-19766.

Burke, Christopher J, Philippe N Tobler, Michelle Baddeley, and Wolfram Schultz. 2010. "Neural mechanisms of observational learning." *Proceedings of the National Academy of Sciences* 107 (32):14431-14436.

Burkhouse, Katie L, Mary L Woody, Max Owens, John E McGeary, Valerie S Knopik, and Brandon E Gibb. 2016. "Sensitivity in detecting facial displays of emotion: Impact of maternal depression and oxytocin receptor genotype." *Cognition and Emotion* 30 (2):275-287.

Bzdok, Danilo, Leonhard Schilbach, Kai Vogeley, Karla Schneider, Angela R Laird, Robert Langner, and Simon B Eickhoff. 2012. "Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy." *Brain Structure and Function* 217 (4):783-796.

Cacioppo, John T, David G Amaral, Jack J Blanchard, Judy L Cameron, C Sue Carter, David Crews, Susan Fiske, Todd Heatherton, Marcia K Johnson, and Michael J Kozak. 2007. "Social neuroscience: Progress and implications for mental health." *Perspectives on Psychological Science* 2 (2):99-123.

Cacioppo, John T, and Richard E Petty. 1982. "The need for cognition." *Journal of personality and social psychology* 42 (1):116.

Calder, Andrew J, Jill Keane, Facundo Manes, Nagui Antoun, and Andrew W Young. 2000. "Impaired recognition and experience of disgust following brain injury." *Nature neuroscience* 3 (11):1077-1078.

Campbell, Benjamin C, and Justin R Garcia. 2009. "Neuroanthropology: evolution and emotional embodiment." *Frontiers in Evolutionary Neuroscience* 1:4.

Campbell, Polly, Alexander G Ophir, and Steven M Phelps. 2009. "Central vasopressin and oxytocin receptor distributions in two species of singing mice." *Journal of Comparative Neurology* 516 (4):321-333.

Capraro, Valerio, Brice Corgnet, Antonio M Espín, and Roberto Hernán-González. 2017. "Deliberation favours social efficiency by making people disregard their relative shares: evidence from USA and India." *Royal Society open science* 4 (2):160605.

Carpendale, Jeremy IM. 2000. "Kohlberg and Piaget on stages and moral reasoning." *Developmental Review* 20 (2):181-205.

Carter, C Sue. 2014. "Oxytocin pathways and the evolution of human behavior." *Annual review of psychology* 65:17-39.

Caspers, Svenja, Karl Zilles, Angela R Laird, and Simon B Eickhoff. 2010. "ALE meta-analysis of action observation and imitation in the human brain." *Neuroimage* 50 (3):1148-1167.

Caspi, Avshalom, Joseph McClay, Terrie E Moffitt, Jonathan Mill, Judy Martin, Ian W Craig, Alan Taylor, and Richie Poulton. 2002. "Role of genotype in the cycle of violence in maltreated children." *science* 297 (5582):851-854.

Caspi, Avshalom, Karen Sugden, Terrie E Moffitt, Alan Taylor, Ian W Craig, HonaLee Harrington, Joseph McClay, Jonathan Mill, Judy Martin, and Antony Braithwaite. 2003. "Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene." *Science* 301 (5631):386-389.

Cavalli-Sforza, Luigi Luca, and Marcus W Feldman. 1981. *Cultural transmission and evolution: A quantitative approach*: Princeton University Press.

Chander, Russell J, Karen A Mather, Rhiagh Cleary, Sarah A Grainger, Anbupalam Thalamuthu, Katya Numbers, Nicole A Kochan, Nicola J Armstrong, Henry Brodaty, and Julie D Henry. 2021. "The influence of rs53576 polymorphism in the oxytocin receptor (OXTR) gene on empathy in healthy adults by subtype and ethnicity: a systematic review and meta-analysis." *Reviews in the Neurosciences*.

Charles, Lucie, Filip Van Opstal, Sébastien Marti, and Stanislas Dehaene. 2013. "Distinct brain mechanisms for conscious versus subliminal error detection." *Neuroimage* 73:80-94.

Chen, Frances S, Robert Kumsta, Fabian Dvorak, Gregor Domes, OS Yim, Richard P Ebstein, and Markus Heinrichs. 2015. "Genetic modulation of oxytocin sensitivity: a pharmacogenetic approach." *Translational psychiatry* 5 (10):e664.

Chen, Pin-Hao A, Paul J Whalen, Jonathan B Freeman, James M Taylor, and Todd F Heatherton. 2015. "Brain reward activity to masked in-group smiling faces predicts friendship development." *Social psychological and personality science* 6 (4):415-421.

Chiao, Joan Y, Ahmad R Hariri, Tokiko Harada, Yoko Mano, Norihiro Sadato, Todd B Parrish, and Tetsuya Iidaka. 2010. "Theory and methods in cultural neuroscience." *Social cognitive and affective neuroscience* 5 (2-3):356-361.

Chini, Bice, Matthijs Verhage, and Valery Grinevich. 2017. "The action radius of oxytocin release in the mammalian CNS: from single vesicles to behavior." *Trends in pharmacological sciences* 38 (11):982-991.

Choi, Damee, Natsumi Minote, and Shigeki Watanuki. 2017. "Associations between the oxytocin receptor gene (OXTR) rs53576 polymorphism and emotional processing of social and nonsocial cues: an event-related potential (ERP) study." *Journal of physiological anthropology* 36 (1):1-10.

Chudek, Maciej, and Joseph Henrich. 2011. "Culture–gene coevolution, norm-psychology and the emergence of human prosociality." *Trends in cognitive sciences* 15 (5):218-226.

Churchland, Patricia S, and Piotr Winkielman. 2012. "Modulating social behavior with oxytocin: how does it work? What does it mean?" *Hormones and behavior* 61 (3):392-399.

Cialdini, Robert B, and Noah J Goldstein. 2004a. "Social influence: Compliance and conformity." *Annu. Rev. Psychol.* 55:591-621.

Cialdini, Robert B, and Noah J Goldstein. 2004b. "Social influence: Compliance and conformity." *Annual review of psychology* 55 (1):591-621.

Ciaramelli, Elisa, Michela Muccioli, Elisabetta Làdavas, and Giuseppe di Pellegrino. 2007. "Selective deficit in personal moral judgment following damage to ventromedial

prefrontal cortex." *Social cognitive and affective neuroscience* 2 (2):84-92.
Cohen, Jeremiah Y, Sebastian Haesler, Linh Vong, Bradford B Lowell, and Naoshige Uchida. 2012. "Neuron-type-specific signals for reward and punishment in the ventral tegmental area." *nature* 482 (7383):85-88.
Cole, Steven W, Gabriella Conti, Jesusa MG Arevalo, Angela M Ruggiero, James J Heckman, and Stephen J Suomi. 2012. "Transcriptional modulation of the developing immune system by early life social adversity." *Proceedings of the National Academy of Sciences* 109 (50):20578-20583.
Connelly, Jessica J, Jean Golding, Steven P Gregory, Susan M Ring, John M Davis, George Davey Smith, James C Harris, C Sue Carter, and Marcus Pembrey. 2014. "Personality, behavior and environmental features associated with OXTR genetic variants in British mothers." *PloS one* 9 (3):e90465.
Corbetta, Maurizio, Gaurav Patel, and Gordon L Shulman. 2008. "The reorienting system of the human brain: from environment to theory of mind." *Neuron* 58 (3):306-324.
Cosmides, Leda, and John Tooby. 2000. "Evolutionary psychology and the emotions." *Handbook of emotions* 2 (2):91-115.
Cowan, Philip A, Jonas Longer, Judith Heavenrich, and Marjorie Nathanson. 1969. "Social learning and Piaget's cognitive theory of moral development." *Journal of Personality and Social Psychology* 11 (3):261.
Creswell, Kasey G, Aidan GC Wright, Wendy M Troxel, Robert E Ferrell, Janine D Flory, and Stephen B Manuck. 2015. "OXTR polymorphism predicts social relationships through its effects on social temperament." *Social cognitive and affective neuroscience* 10 (6):869-876.
Critchley, Hugo D, Joey Tang, Daniel Glaser, Brian Butterworth, and Raymond J Dolan. 2005. "Anterior cingulate activity during error and autonomic response." *Neuroimage* 27 (4):885-895.
Crockett, Molly J. 2013. "Models of morality." *Trends in cognitive sciences* 17 (8):363-366.
Crockett, Molly J, Luke Clark, Marc D Hauser, and Trevor W Robbins. 2010. "Serotonin selectively influences moral judgment and behavior through effects on harm aversion." *Proceedings of the National Academy of Sciences* 107 (40):17433-17438.
Csordas, Thomas J. 1990. "Embodiment as a Paradigm for Anthropology." *Ethos* 18 (1):5-47.
Curry, Oliver Scott, Daniel Austin Mullins, and Harvey Whitehouse. 2019. "Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies." *Current Anthropology* 60 (1):47-69.
Cushman, Fiery, Victor Kumar, and Peter Railton. 2017. "Moral learning: Current and future directions." *cognition*.
Damasio, Antonio R. 1999. "How the brain creates the mind." *Scientific American* 281 (6):112-117.
Davis, Mark H. 1983. "Measuring individual differences in empathy: evidence for a multidimensional approach." *Journal of personality and social psychology* 44 (1):113.
De Dreu, Carsten KW, Lindred L Greer, Gerben A Van Kleef, Shaul Shalvi, and Michel JJ Handgraaf. 2011. "Oxytocin promotes human ethnocentrism." *Proceedings of the National Academy of Sciences* 108 (4):1262-1266.
De Dreu, Carsten KW, and Mariska E Kret. 2016. "Oxytocin conditions intergroup relations through upregulated in-group empathy, cooperation, conformity, and defense." *Biological psychiatry* 79 (3):165-173.
De Waal, Frans BM. 1991. "The chimpanzee's sense of social regularity and its relation to the human sense of justice." *American Behavioral Scientist* 34 (3):335-349.
De Waal, Frans BM, and Stephanie D Preston. 2017. "Mammalian empathy: behavioural manifestations and neural basis." *Nature Reviews Neuroscience* 18 (8):498-509.

de Waal, Frans, and Angeline van Roosmalen. 1979. "Reconciliation and consolation among chimpanzees." *Behavioral Ecology and Sociobiology* 5 (1):55-66.

Dean, Lewis G, Rachel L Kendal, Steven J Schapiro, Bernard Thierry, and Kevin N Laland. 2012. "Identification of the social and cognitive processes underlying human cumulative culture." *Science* 335 (6072):1114-1118.

Decety, Jean. 2015. "The neural pathways, development and functions of empathy." *Current Opinion in Behavioral Sciences* 3:1-6.

Decety, Jean. 2021. "Why empathy is not a reliable source of information in moral decision making." *Current Directions in Psychological Science* 30 (5):425-430.

Decety, Jean, and Jason M Cowell. 2014. "The complex relation between morality and empathy." *Trends in cognitive sciences* 18 (7):337-339.

Decety, Jean, and Jason M Cowell. 2018. "Interpersonal harm aversion as a necessary foundation for morality: A developmental neuroscience perspective." *Development and psychopathology* 30 (1):153-164.

Declerck, Carolyn H, Christophe Boone, and Toko Kiyonari. 2010. "Oxytocin and cooperation under conditions of uncertainty: the modulating role of incentives and social information." *Hormones and behavior* 57 (3):368-374.

DelGiudice, Marco. 2018. "Middle childhood: An evolutionary-developmental synthesis." In *Handbook of Life Course Health Development*, 95-107. Springer.

Dibeklioğlu, Hamdi, Albert Ali Salah, and Theo Gevers. 2012. "Are you really smiling at me? spontaneous versus posed enjoyment smiles." European Conference on Computer Vision.

Dickinson, Anthony. 1980. *Contemporary animal learning theory*: Cambridge University Press.

Dima, Danai, and Gerome Breen. 2015. "Polygenic risk scores in imaging genetics: usefulness and applications." *Journal of Psychopharmacology* 29 (8):867-871.

Dölen, Gül, Ayeh Darvishzadeh, Kee Wui Huang, and Robert C Malenka. 2013. "Social reward requires coordinated activity of nucleus accumbens oxytocin and serotonin." *Nature* 501 (7466):179-184.

Domes, Gregor, Markus Heinrichs, Jan Gläscher, Christian Büchel, Dieter F Braus, and Sabine C Herpertz. 2007. "Oxytocin attenuates amygdala responses to emotional faces regardless of valence." *Biological psychiatry* 62 (10):1187-1190.

Domes, Gregor, Markus Heinrichs, Ekkehardt Kumbier, Annette Grossmann, Karlheinz Hauenstein, and Sabine C Herpertz. 2013. "Effects of intranasal oxytocin on the neural basis of face processing in autism spectrum disorder." *Biological psychiatry* 74 (3):164-171.

Domes, Gregor, Markus Heinrichs, Andre Michel, Christoph Berger, and Sabine C Herpertz. 2007. "Oxytocin improves “mind-reading” in humans." *Biological psychiatry* 61 (6):731-733.

Domes, Gregor, M Sibold, L Schulze, Alexander Lischke, Sabine C Herpertz, and Markus Heinrichs. 2013. "Intranasal oxytocin increases covert attention to positive social cues." *Psychological medicine* 43 (8):1747-1753.

Donaldson, Zoe R, and Larry J Young. 2008. "Oxytocin, vasopressin, and the neurogenetics of sociality." *Science* 322 (5903):900-904.

Dricu, Mihai, and Sascha Frühholz. 2016. "Perceiving emotional expressions in others: activation likelihood estimation meta-analyses of explicit evaluation, passive perception and incidental perception of emotions." *Neuroscience & Biobehavioral Reviews* 71:810-828.

Duchaine, Brad, and Galit Yovel. 2015. "A revised neural framework for face processing." *Annual review of vision science* 1:393-416.

Durkheim, Émile. 1906. "1974." The Determination of Moral Facts." *Sociology and*

Philosophy. London: Routledge.
Ebitz, R Becket, and Benjamin Yost Hayden. 2016. "Dorsal anterior cingulate: a Rorschach test for cognitive neuroscience." *Nature neuroscience* 19 (10):1278-1279.
Edelson, Micah G, Maya Shemesh, Abraham Weizman, Shahak Yariv, Tali Sharot, and Yadin Dudai. 2015. "Opposing effects of oxytocin on overt compliance and lasting changes to memory." *Neuropsychopharmacology* 40 (4):966-973.
Eisenberg, Nancy, and Paul A Miller. 1987. "The relation of empathy to prosocial and related behaviors." *Psychological bulletin* 101 (1):91.
Eisenberg, Nancy, and Paul Henry Mussen. 1989. *The roots of prosocial behavior in children*: Cambridge University Press.
Eisenberger, Naomi I, and Matthew D Lieberman. 2004. "Why rejection hurts: a common neural alarm system for physical and social pain." *Trends in cognitive sciences* 8 (7):294-300.
Ekman, Paul. 2009. "Lie catching and microexpressions." *The philosophy of deception* 1 (2):5.
Ekman, Paul, Richard J Davidson, and Wallace V Friesen. 1990. "The Duchenne smile: Emotional expression and brain physiology: II." *Journal of personality and social psychology* 58 (2):342.
Ekman, Paul, and Wallace V Friesen. 1974. "Detecting deception from the body or face." *Journal of personality and Social Psychology* 29 (3):288.
Ekman, Paul, and Wallace V Friesen. 2003. *Unmasking the face: A guide to recognizing emotions from facial clues*. Vol. 10: Ishk.
Eliava, Marina, Meggane Melchior, H Sophie Knobloch-Bollmann, Jérôme Wahis, Miriam da Silva Gouveia, Yan Tang, Alexandru Cristian Ciobanu, Rodrigo Triana Del Rio, Lena C Roth, and Ferdinand Althammer. 2016. "A new population of parvocellular oxytocin neurons controlling magnocellular neuron activity and inflammatory pain processing." *Neuron* 89 (6):1291-1304.
Endicott, Karen L, and Kirk M Endicott. 2014. "Batek Childrearing and Morality." *Ancestral landscapes in human evolution: Culture, childrearing and social wellbeing*:108.
Engelmann, Jan M, and Esther Herrmann. 2016. "Chimpanzees trust their friends." *Current Biology* 26 (2):252-256.
Eres, Robert, Winnifred R Louis, and Pascal Molenberghs. 2018. "Common and distinct neural networks involved in fMRI studies investigating morality: an ALE meta-analysis." *Social neuroscience* 13 (4):384-398.
Esteves, Francisco, and Arne Öhman. 1993. "Masking the face: Recognition of emotional facial expressions as a function of the parameters of backward masking." *Scandinavian journal of psychology* 34 (1):1-18.
Evans, Jonathan St BT. 2004. "History of the dual process theory of reasoning." In *Psychology of reasoning*, 251-276. Psychology Press.
Falk, Emily, Baldwin Way, and Agnes Jasinska. 2012. "An imaging genetics approach to understanding social influence." *Frontiers in human neuroscience* 6:168.
Fan, Yan, Niall W Duncan, Moritz de Greck, and Georg Northoff. 2011. "Is there a core neural network in empathy? An fMRI based quantitative meta-analysis." *Neuroscience & Biobehavioral Reviews* 35 (3):903-911.
Fassin, Didier. 2014. *A companion to moral anthropology*: John Wiley & Sons.
Fawcett, Tim W, James AR Marshall, and Andrew D Higginson. 2015. The evolution of mechanisms underlying behaviour. Oxford University Press.
Fehr, Ernst, and Urs Fischbacher. 2004. "Social norms and human cooperation." *Trends in cognitive sciences* 8 (4):185-190.
Feldman, Ruth. 2016. "The neurobiology of mammalian parenting and the biosocial context of human caregiving." *Hormones and behavior* 77:3-17.

Feldman, Ruth, Mikhail Monakhov, Maayan Pratt, and Richard P Ebstein. 2016. "Oxytocin pathway genes: evolutionary ancient system impacting on human affiliation, sociality, and psychopathology." *Biological psychiatry* 79 (3):174-184.

Feldman, Ruth, Orna Zagoory-Sharon, Omri Weisman, Inna Schneiderman, Ilanit Gordon, Rina Maoz, Idan Shalev, and Richard P Ebstein. 2012. "Sensitive parenting is associated with plasma oxytocin and polymorphisms in the OXTR and CD38 genes." *Biological psychiatry* 72 (3):175-181.

Feng, Chunliang, Patrick D Hackett, Ashley C DeMarco, Xu Chen, Sabrina Stair, Ebrahim Haroon, Beate Ditzen, Giuseppe Pagnoni, and James K Rilling. 2015. "Oxytocin and vasopressin effects on the neural response to social cooperation are modulated by sex in humans." *Brain imaging and behavior* 9 (4):754-764.

Feng, Chunliang, Yue-Jia Luo, and Frank Krueger. 2015. "Neural signatures of fairness-related normative decision making in the ultimatum game: A coordinate-based meta-analysis." *Human brain mapping* 36 (2):591-602.

Ferris, Craig F. 2005. "Vasopressin/oxytocin and aggression." Novartis Foundation Symposium.

Foot, Philippa. 1967. "The problem of abortion and the doctrine of the double effect." *Oxford review* 5.

Foucault, Michel. 2012. *The history of sexuality, vol. 2: The use of pleasure*: Vintage.

Fouragnan, Elsa, Chris Retzler, and Marios G Philiastides. 2018. "Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta-analysis." *Human brain mapping* 39 (7):2887-2906.

Freeman, Sara M, Michelle C Palumbo, Rebecca H Lawrence, Aaron L Smith, Mark M Goodman, and Karen L Bales. 2018. "Effect of age and autism spectrum disorder on oxytocin receptor density in the human basal forebrain and midbrain." *Translational psychiatry* 8 (1):1-11.

Freeman, Sara M, and Larry J Young. 2016. "Comparative perspectives on oxytocin and vasopressin receptor research in rodents and primates: Translational implications." *Journal of neuroendocrinology* 28 (4).

Frith, Chris D, and Uta Frith. 2006. "The neural basis of mentalizing." *Neuron* 50 (4):531-534.

Fuhrmann, Delia, Lisa J Knoll, and Sarah-Jayne Blakemore. 2015. "Adolescence as a sensitive period of brain development." *Trends in cognitive sciences* 19 (10):558-566.

Fusar-Poli, Paolo, Anna Placentino, Francesco Carletti, Paola Landi, Paul Allen, Simon Surguladze, Francesco Benedetti, Marta Abbamonte, Roberto Gasparotti, and Francesco Barale. 2009. "Functional atlas of emotional faces processing: a voxel-based meta-analysis of 105 functional magnetic resonance imaging studies." *Journal of Psychiatry and Neuroscience* 34 (6):418-432.

Gallagher, Helen L, and Christopher D Frith. 2003. "Functional imaging of 'theory of mind'." *Trends in cognitive sciences* 7 (2):77-83.

Gallagher, Helen L, Francesca Happé, Nicola Brunswick, Paul C Fletcher, Uta Frith, and Chris D Frith. 2000. "Reading the mind in cartoons and stories: an fMRI study of 'theory of mind'in verbal and nonverbal tasks." *Neuropsychologia* 38 (1):11-21.

Gamer, Matthias, Bartosz Zurowski, and Christian Büchel. 2010. "Different amygdala subregions mediate valence-related and attentional effects of oxytocin in humans." *Proceedings of the National Academy of Sciences* 107 (20):9400-9405.

Gao, Yidian, Jack C Rogers, Ruth Pauli, Roberta Clanton, Rosalind Baker, Philippa Birch, Lisandra Ferreira, Abigail Brown, Christine M Freitag, and Graeme Fairchild. 2019. "Neural correlates of theory of mind in typically-developing youth: influence of sex, age and callous-unemotional traits." *Scientific reports* 9 (1):1-12.

Garfield, Zachary H, Melissa J Garfield, and Barry S Hewlett. 2016. "A cross-cultural analysis

of hunter-gatherer social learning." In *Social learning and innovation in contemporary hunter-gatherers*, 19-34. Springer.

Garrison, Jane, Burak Erdeniz, and John Done. 2013. "Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies." *Neuroscience & Biobehavioral Reviews* 37 (7):1297-1310.

Gavrilets, Sergey, and Peter J Richerson. 2017. "Collective action and the evolution of social norm internalization." *Proceedings of the National Academy of Sciences* 114 (23):6068-6073.

Gazzaniga, Michael S. 2005. *The ethical brain*: Dana press.

Geertz, Clifford. 1973. *The interpretation of cultures*. Vol. 5043: Basic books.

Geng, Yayuan, Weihua Zhao, Feng Zhou, Xiaole Ma, Shuxia Yao, Rene Hurlemann, Benjamin Becker, and Keith M Kendrick. 2018. "Oxytocin enhancement of emotional empathy: generalization across cultures and effects on amygdala activity." *Frontiers in neuroscience* 12:512.

Gęsiarz, Filip, and Molly J Crockett. 2015. "Goal-directed, habitual and Pavlovian prosocial behavior." *Frontiers in behavioral neuroscience* 9:135.

Gilligan, Carol, and Jane Attanucci. 1988. "Two moral orientations: Gender differences and similarities." *Merrill-Palmer Quarterly (1982-)*:223-237.

Gintis, Herbert. 2003. "The hitchhiker's guide to altruism: Gene-culture coevolution, and the internalization of norms." *Journal of theoretical biology* 220 (4):407-418.

Glenn, Andrea L, Adrian Raine, and Robert A Schug. 2009. "The neural correlates of moral decision-making in psychopathy." *Molecular psychiatry* 14 (1):5-6.

Godoy, Ricardo, Victoria Reyes-García, Tomás Huanca, Susan Tanner, William R Leonard, Thomas McDade, and Vincent Vadez. 2005. "Do smiles have a face value? Panel evidence from Amazonian Indians." *Journal of Economic Psychology* 26 (4):469-490.

Gong, Pingyuan, Huiyong Fan, Jinting Liu, Xing Yang, Kejin Zhang, and Xiaolin Zhou. 2017. "Revisiting the impact of OXTR rs53576 on empathy: A population-based study and a meta-analysis." *Psychoneuroendocrinology* 80:131-136.

Goodall, Jane. 1986. "The chimpanzees of Gombe: Patterns of behavior." *Cambridge Mass*.

Goodwin, Geoffrey P, Jared Piazza, and Paul Rozin. 2014. "Moral character predominates in person perception and evaluation." *Journal of personality and social psychology* 106 (1):148.

Gordon, Ilanit, Brent C Vander Wyk, Randi H Bennett, Cara Cordeaux, Molly V Lucas, Jeffrey A Eilbott, Orna Zagoory-Sharon, James F Leckman, Ruth Feldman, and Kevin A Pelphrey. 2013. "Oxytocin enhances brain function in children with autism." *Proceedings of the National Academy of Sciences* 110 (52):20953-20958.

Grace, Sally A, Susan L Rossell, Markus Heinrichs, Catarina Kordsachia, and Izelle Labuschagne. 2018. "Oxytocin and brain activity in humans: a systematic review and coordinate-based meta-analysis of functional MRI studies." *Psychoneuroendocrinology* 96:6-24.

Grafen, Alan. 1991. "Modelling in behavioural ecology." *Behavioural ecology: an evolutionary approach* 3:5-31.

Graham, Jesse, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. "Mapping the moral domain." *Journal of personality and social psychology* 101 (2):366.

Grandey, Alicia A, Glenda M Fisk, Anna S Mattila, Karen J Jansen, and Lori A Sideman. 2005. "Is “service with a smile” enough? Authenticity of positive displays during service encounters." *Organizational Behavior and Human Decision Processes* 96 (1):38-55.

Greene, Joshua D, Sylvia A Morelli, Kelly Lowenberg, Leigh E Nystrom, and Jonathan D Cohen. 2008. "Cognitive load selectively interferes with utilitarian moral judgment."

Cognition 107 (3):1144-1154.
Greene, Joshua D, Leigh E Nystrom, Andrew D Engell, John M Darley, and Jonathan D Cohen. 2004. "The neural bases of cognitive conflict and control in moral judgment." *Neuron* 44 (2):389-400.
Greene, Joshua D, R Brian Sommerville, Leigh E Nystrom, John M Darley, and Jonathan D Cohen. 2001. "An fMRI investigation of emotional engagement in moral judgment." *Science* 293 (5537):2105-2108.
Greene, Joshua, and Jonathan Haidt. 2002. "How (and where) does moral judgment work?" *Trends in cognitive sciences* 6 (12):517-523.
Greve, Douglas N, and Bruce Fischl. 2009. "Accurate and robust brain image alignment using boundary-based registration." *Neuroimage* 48 (1):63-72.
Grinevich, Valery, and Inga D Neumann. 2021. "Brain oxytocin: how puzzle stones from animal studies translate into psychiatry." *Molecular psychiatry* 26 (1):265-279.
Groppe, Sarah E, Anna Gossen, Lena Rademacher, Alexa Hahn, Luzie Westphal, Gerhard Gründer, and Katja N Spreckelmeyer. 2013. "Oxytocin influences processing of socially relevant cues in the ventral tegmental area of the human brain." *Biological psychiatry* 74 (3):172-179.
Guastella, Adam J, Philip B Mitchell, and Mark R Dadds. 2008. "Oxytocin increases gaze to the eye region of human faces." *Biological psychiatry* 63 (1):3-5.
Gunnery, Sarah D, and Mollie A Ruben. 2016. "Perceptions of Duchenne and non-Duchenne smiles: A meta-analysis." *Cognition and Emotion* 30 (3):501-515.
Gweon, Hyowon, and Rebecca Saxe. 2013. "Developmental cognitive neuroscience of theory of mind." *Neural circuit development and function in the brain* 3:367-377.
Haidt, Jonathan. 2007. "The new synthesis in moral psychology." *science* 316 (5827):998-1002.
Haidt, Jonathan. 2008. "Morality." *Perspectives on psychological science* 3 (1):65-72.
Haidt, Jonathan, Fredrik Bjorklund, and Scott Murphy. 2000. "Moral dumbfounding: When intuition finds no reason." *Unpublished manuscript, University of Virginia*:191-221.
Haidt, Jonathan, and Craig Joseph. 2004. "Intuitive ethics: How innately prepared intuitions generate culturally variable virtues." *Daedalus* 133 (4):55-66.
Han, Hyemin, Gary H Glover, and Changwoo Jeong. 2014. "Cultural influences on the neural correlate of moral decision making processes." *Behavioural brain research* 259:215-228.
Han, Shihui, and Yina Ma. 2015. "A culture–behavior–brain loop model of human development." *Trends in Cognitive Sciences* 19 (11):666-676.
Harenski, Carla L, and Stephan Hamann. 2006. "Neural correlates of regulating negative emotions related to moral violations." *Neuroimage* 30 (1):313-324.
Harenski, Carla L, Keith A Harenski, Matthew S Shane, and Kent A Kiehl. 2012. "Neural development of mentalizing in moral judgment from adolescence to adulthood." *Developmental cognitive neuroscience* 2 (1):162-173.
Haxby, James V, Elizabeth A Hoffman, and M Ida Gobbini. 2000. "The distributed human neural system for face perception." *Trends in cognitive sciences* 4 (6):223-233.
Hayward, Dana A, Effie J Pereira, A Ross Otto, and Jelena Ristic. 2018. "Smile! Social reward drives attention." *Journal of Experimental Psychology: Human Perception and Performance* 44 (2):206.
Hecht, Erin E, David A Gutman, Todd M Preuss, Mar M Sanchez, Lisa A Parr, and James K Rilling. 2013. "Process versus product in social learning: comparative diffusion tensor imaging of neural systems for action execution–observation matching in macaques, chimpanzees, and humans." *Cerebral Cortex* 23 (5):1014-1024.
Heekeren, Hauke R, Isabell Wartenburger, Helge Schmidt, Hans-Peter Schwintowski, and Arno

Villringer. 2003. "An fMRI study of simple ethical decision-making." *Neuroreport* 14 (9):1215-1219.

Hein, Grit, and Tania Singer. 2010. "Neuroscience meets social psychology: An integrative approach to human empathy and prosocial behavior."

Henrich, Joseph. 2015. *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*: Princeton University Press.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. "In search of homo economicus: behavioral experiments in 15 small-scale societies." *American Economic Review* 91 (2):73-78.

Ho, Mark K, James MacGlashan, Michael L Littman, and Fiery Cushman. 2017. "Social is special: A normative framework for teaching with and learning from evaluative feedback." *Cognition* 167:91-106.

Hofmann, Hans A, Annaliese K Beery, Daniel T Blumstein, Iain D Couzin, Ryan L Earley, Loren D Hayes, Peter L Hurd, Eileen A Lacey, Steven M Phelps, and Nancy G Solomon. 2014. "An evolutionary framework for studying mechanisms of social behavior." *Trends in ecology & evolution* 29 (10):581-589.

Hoppitt, Will, and Kevin N Laland. 2008. "Social processes influencing learning in animals: a review of the evidence." *Advances in the Study of Behavior* 38:105-165.

House, Bailey R, Joan B Silk, Joseph Henrich, H Clark Barrett, Brooke A Scelza, Adam H Boyette, Barry S Hewlett, Richard McElreath, and Stephen Laurence. 2013. "Ontogeny of prosocial behavior across diverse societies." *Proceedings of the National Academy of Sciences* 110 (36):14586-14591.

Hu, Jiehui, Song Qi, Benjamin Becker, Lizhu Luo, Shan Gao, Qiyong Gong, René Hurlemann, and Keith M Kendrick. 2015. "Oxytocin selectively facilitates learning with social feedback and increases activity and functional connectivity in emotional memory and reward processing regions." *Human brain mapping* 36 (6):2132-2146.

Huang, Yi, Keith M Kendrick, and Rongjun Yu. 2014. "Conformity to the opinions of other people lasts for no more than 3 days." *Psychological science* 25 (7):1388-1393.

Hung, Lin W, Sophie Neuner, Jai S Polepalli, Kevin T Beier, Matthew Wright, Jessica J Walsh, Eastman M Lewis, Liqun Luo, Karl Deisseroth, and Gül Dölen. 2017. "Gating of social reward by oxytocin in the ventral tegmental area." *Science* 357 (6358):1406-1411.

Hurlemann, René, Alexandra Patin, Oezguer A Onur, Michael X Cohen, Tobias Baumgartner, Sarah Metzler, Isabel Dziobek, Juergen Gallinat, Michael Wagner, and Wolfgang Maier. 2010. "Oxytocin enhances amygdala-dependent, socially reinforced learning and emotional empathy in humans." *Journal of neuroscience* 30 (14):4999-5007.

Hurley, Carolyn M, Ashley E Anker, Mark G Frank, David Matsumoto, and Hyisung C Hwang. 2014. "Background factors predicting accuracy and improvement in micro expression recognition." *Motivation and emotion* 38 (5):700-714.

Hussey, Elizabeth, and Ashley Safford. 2009. "Perception of facial expression in somatosensory cortex supports simulationist models." *Journal of Neuroscience* 29 (2):301-302.

Hutcherson, Cendri A, and James J Gross. 2011. "The moral emotions: A social–functionalist account of anger, disgust, and contempt." *Journal of personality and social psychology* 100 (4):719.

Hutcherson, Cendri A, Leila Montaser-Kouhsari, James Woodward, and Antonio Rangel. 2015. "Emotional and utilitarian appraisals of moral dilemmas are encoded in separate areas and integrated in ventromedial prefrontal cortex." *Journal of Neuroscience* 35 (36):12593-12605.

Insel, Thomas R, and Lawrence E Shapiro. 1992. "Oxytocin receptor distribution reflects social organization in monogamous and polygamous voles." *Proceedings of the National Academy of Sciences* 89 (13):5981-5985.

Ishii, Keiko, Heejung S Kim, Joni Y Sasaki, Mizuho Shinada, and Ichiro Kusumi. 2014. "Culture modulates sensitivity to the disappearance of facial expressions associated with serotonin transporter polymorphism (5-HTTLPR)." *Culture and Brain* 2 (1):72-88.

Ito, Takayuki, Emi Z Murano, and Hiroaki Gomi. 2004. "Fast force-generation dynamics of human articulatory muscles." *Journal of applied physiology* 96 (6):2318-2324.

Ito, Takehito, Keita Yokokawa, Noriaki Yahata, Ayako Isato, Tetsuya Suhara, and Makiko Yamada. 2017. "Neural basis of negativity bias in the perception of ambiguous facial expression." *Scientific reports* 7 (1):1-9.

Iwase, Masao, Yasuomi Ouchi, Hiroyuki Okada, Chihiro Yokoyama, Shuji Nobezawa, Etsuji Yoshikawa, Hideo Tsukada, Masaki Takeda, Ko Yamashita, and Masatoshi Takeda. 2002. "Neural substrates of human facial expression of pleasant emotion induced by comic films: a PET study." *Neuroimage* 17 (2):758-768.

Izuma, Keise. 2013. "The neural basis of social influence and attitude change." *Current opinion in neurobiology* 23 (3):456-462.

Izuma, Keise, and Ralph Adolphs. 2013. "Social manipulation of preference in the human brain." *Neuron* 78 (3):563-573.

Jaeggi, Adrian V, and Carel P Van Schaik. 2011. "The evolution of food sharing in primates." *Behavioral Ecology and Sociobiology* 65 (11):2125-2140.

Jenkinson, Mark, Peter Bannister, Michael Brady, and Stephen Smith. 2002. "Improved optimization for the robust and accurate linear registration and motion correction of brain images." *Neuroimage* 17 (2):825-841.

Johnson, Addie, and Robert W Proctor. 2004. *Attention: Theory and practice*: Sage.

Johnson, Zachary V, Hasse Walum, Yao Xiao, Paula C Riefkohl, and Larry J Young. 2017. "Oxytocin receptors modulate a social salience neural network in male prairie voles." *Hormones and behavior* 87:16-24.

Johnson, Zachary V, and Larry J Young. 2017. "Oxytocin and vasopressin neural networks: implications for social behavioral diversity and translational neuroscience." *Neuroscience & Biobehavioral Reviews* 76:87-98.

Johnston, Patrick, Angela Mayes, Matthew Hughes, and Andrew W Young. 2013. "Brain networks subserving the evaluation of static and dynamic facial expressions." *Cortex* 49 (9):2462-2472.

Joiner, Jessica, Matthew Piva, Courtney Turrin, and Steve WC Chang. 2017. "Social learning through prediction error in the brain." *npj Science of Learning* 2 (1):8.

Jones, Rebecca M, Leah H Somerville, Jian Li, Erika J Ruberry, Victoria Libby, Gary Glover, Henning U Voss, Douglas J Ballon, and BJ Casey. 2011. "Behavioral and neural properties of social reinforcement learning." *Journal of Neuroscience* 31 (37):13039-13045.

Jurek, Benjamin, and Inga D Neumann. 2018. "The oxytocin receptor: from intracellular signaling to behavior." *Physiological reviews* 98 (3):1805-1908.

Kable, Joseph W. 2011. "The cognitive neuroscience toolkit for the neuroeconomist: A functional overview." *Journal of Neuroscience, Psychology, and Economics* 4 (2):63.

Kahneman, Daniel. 2011. *Thinking, fast and slow*: Macmillan.

Kanat, Manuela, Markus Heinrichs, Irina Mader, Ludger Tebartz Van Elst, and Gregor Domes. 2015. "Oxytocin modulates amygdala reactivity to masked fearful eyes." *Neuropsychopharmacology* 40 (11):2632.

Keane, Webb. 2015. *Ethical life: Its natural and social histories*: Princeton University Press.

Keech, Britney, Simon Crowe, and Darren R Hocking. 2018. "Intranasal oxytocin, social cognition and neurodevelopmental disorders: a meta-analysis." *Psychoneuroendocrinology* 87:9-19.

Keuken, Max C, A Hardie, BT Dorn, S Dev, MP Paulus, KJ Jonas, WPM Van Den Wildenberg, and JA Pineda. 2011. "The role of the left inferior frontal gyrus in social perception: an rTMS study." *Brain research* 1383:196-205.

Keysers, Christian. 2009. "Mirror neurons." *Current Biology* 19 (21):R971-R973.

Kilts, Clinton D, Glenn Egan, Deborah A Gideon, Timothy D Ely, and John M Hoffman. 2003. "Dissociable neural pathways are involved in the recognition of emotion in static and dynamic facial expressions." *Neuroimage* 18 (1):156-168.

Kim, Elizabeth B, Chuansheng Chen, Judith G Smetana, and Ellen Greenberger. 2016. "Does children's moral compass waver under social pressure? Using the conformity paradigm to test preschoolers' moral and social-conventional judgments." *Journal of experimental child psychology* 150:241-251.

Kim, Hackjin, Shinsuke Shimojo, and John P O'doherty. 2010. "Overlapping responses for the expectation of juice and money rewards in human ventromedial prefrontal cortex." *Cerebral cortex* 21 (4):769-776.

Kim, Heejung S, David K Sherman, Taraneh Mojaverian, Joni Y Sasaki, Jinyoung Park, Eunkook M Suh, and Shelley E Taylor. 2011. "Gene–culture interaction: Oxytocin receptor polymorphism (OXTR) and emotion regulation." *Social Psychological and Personality Science* 2 (6):665-672.

Kim, Heejung S, David K Sherman, Joni Y Sasaki, Jun Xu, Thai Q Chu, Chorong Ryu, Eunkook M Suh, Kelsey Graham, and Shelley E Taylor. 2010. "Culture, distress, and oxytocin receptor polymorphism (OXTR) interact to influence emotional support seeking." *Proceedings of the National Academy of Sciences* 107 (36):15717-15721.

Kim-Cohen, Julia, Avshalom Caspi, Alan Taylor, Brenda Williams, Rhiannon Newcombe, Ian W Craig, and Terrie E Moffitt. 2006. "MAOA, maltreatment, and gene–environment interaction predicting children's mental health: new evidence and a meta-analysis." *Molecular psychiatry* 11 (10):903-913.

King, Jim. 2016. "'It's time, put on the smile, it's time!': The emotional labour of second language teaching within a Japanese university." In *New directions in language learning psychology*, 97-112. Springer.

King, Lanikea B, Hasse Walum, Kiyoshi Inoue, Nicholas W Eyrich, and Larry J Young. 2016. "Variation in the oxytocin receptor gene predicts brain region–specific expression and social attachment." *Biological psychiatry* 80 (2):160-169.

Kitayama, Shinobu, Anthony King, Ming Hsu, Israel Liberzon, and Carolyn Yoon. 2016. "Dopamine-system genes and cultural acquisition: the norm sensitivity hypothesis." *Current opinion in psychology* 8:167-174.

Kitayama, Shinobu, Anthony King, Carolyn Yoon, Steve Tompson, Sarah Huff, and Israel Liberzon. 2014. "The dopamine D4 receptor gene (DRD4) moderates cultural difference in independent versus interdependent social orientation." *Psychological science* 25 (6):1169-1177.

Klucharev, Vasily, Kaisa Hytönen, Mark Rijpkema, Ale Smidts, and Guillén Fernández. 2009. "Reinforcement learning signal predicts social conformity." *Neuron* 61 (1):140-151.

Knudsen, Eric I. 2004. "Sensitive periods in the development of the brain and behavior." *Journal of cognitive neuroscience* 16 (8):1412-1425.

Kobayashi, Chiyoko, Gary H Glover, and Elise Temple. 2006. "Cultural and linguistic influence on neural bases of 'Theory of Mind': an fMRI study with Japanese bilinguals." *Brain and language* 98 (2):210-220.

Koenigs, Michael, Liane Young, Ralph Adolphs, Daniel Tranel, Fiery Cushman, Marc Hauser, and Antonio Damasio. 2007. "Damage to the prefrontal cortex increases utilitarian moral judgements." *Nature* 446 (7138):908-911.

Kohlberg, Lawrence, and Richard Kramer. 1969. "Continuities and discontinuities in childhood

and adult moral development." *Human development* 12 (2):93-120.
Konner, Melvin. 2010. *The evolution of childhood: Relationships, emotion, mind*: Harvard University Press.
Korb, Sebastian, Jennifer Malsert, Lane Strathearn, Patrik Vuilleumier, and Paula Niedenthal. 2016. "Sniff and mimic—intranasal oxytocin increases facial mimicry in a sample of men." *Hormones and Behavior* 84:64-74.
Korb, Sebastian, Stéphane With, Paula Niedenthal, Susanne Kaiser, and Didier Grandjean. 2014. "The perception and mimicry of facial movements predict judgments of smile authenticity." *PLoS One* 9 (6):e99194.
Kosfeld, Michael, Markus Heinrichs, Paul J Zak, Urs Fischbacher, and Ernst Fehr. 2005. "Oxytocin increases trust in humans." *Nature* 435 (7042):673.
Kou, Juan, Yingying Zhang, Feng Zhou, Cornelia Sindermann, Christian Montag, Benjamin Becker, and Keith M Kendrick. 2020. "A randomized trial shows dose-frequency and genotype may determine the therapeutic efficacy of intranasal oxytocin." *Psychological Medicine*:1-10.
Krall, Sarah Constance, Claudia Rottschy, Eileen Oberwelland, Danilo Bzdok, Peter T Fox, Simon B Eickhoff, Gereon R Fink, and Kerstin Konrad. 2015. "The role of the right temporoparietal junction in attention and social interaction as revealed by ALE meta-analysis." *Brain Structure and Function* 220 (2):587-604.
Kundu, Payel, and Denise Dellarosa Cummins. 2013. "Morality and conformity: The Asch paradigm applied to moral decisions." *Social Influence* 8 (4):268-279.
Laland, Kevin N. 2004. "Social learning strategies." *Animal Learning & Behavior* 32 (1):4-14.
Lambek, Michael. 2010. *Ordinary ethics: Anthropology, language, and action*: Fordham Univ Press.
Laursen, Helle Ruff, Hartwig Roman Siebner, Tina Haren, Kristoffer Madsen, Rikke Grønlund, Oliver Hulme, and Susanne Henningsson. 2014. "Variation in the oxytocin receptor gene is associated with behavioral and neural correlates of empathic accuracy." *Frontiers in behavioral neuroscience* 8:423.
Leary, Mark R. 1983. "A brief version of the Fear of Negative Evaluation Scale." *Personality and social psychology bulletin* 9 (3):371-375.
Lebreton, Maël, Soledad Jorge, Vincent Michel, Bertrand Thirion, and Mathias Pessiglione. 2009. "An automatic valuation system in the human brain: evidence from functional neuroimaging." *Neuron* 64 (3):431-439.
LeClair, Jessica, Skirmantas Janusonis, and Heejung S Kim. 2014. "Gene–culture interactions: a multi-gene approach." *Culture and Brain* 2 (2):122-140.
LeClair, Jessica, Joni Y Sasaki, Keiko Ishii, Mizuho Shinada, and Heejung S Kim. 2016. "Gene–culture interaction: influence of culture and oxytocin receptor gene (OXTR) polymorphism on loneliness." *Culture and Brain* 4 (1):21-37.
Lee, Daeyeol. 2008. "Game theory and neural basis of social decision making." *Nature neuroscience* 11 (4):404-409.
Lee, Daeyeol, Hyojung Seo, and Min Whan Jung. 2012. "Neural basis of reinforcement learning and decision making." *Annual review of neuroscience* 35:287.
Lennox, Richard D, and Raymond N Wolfe. 1984. "Revision of the self-monitoring scale."
Leppanen, Jenni, Kah Wee Ng, Kate Tchanturia, and Janet Treasure. 2017. "Meta-analysis of the effects of intranasal oxytocin on interpretation and expression of emotions." *Neuroscience & Biobehavioral Reviews* 78:125-144.
Leppänen, Jukka M, and Jari K Hietanen. 2004. "Positive facial expressions are recognized faster than negative facial expressions, but why?" *Psychological research* 69 (1):22-29.

Levorsen, Marie, Ayahito Ito, Shinsuke Suzuki, and Keise Izuma. 2021. "Testing the reinforcement learning hypothesis of social conformity." *Human Brain Mapping* 42 (5):1328-1342.

Levy, Dino J, and Paul W Glimcher. 2012. "The root of all value: a neural common currency for choice." *Current opinion in neurobiology* 22 (6):1027-1038.

Lew-Levy, Sheina, Rachel Reckin, Noa Lavi, Jurgi Cristóbal-Azkarate, and Kate Ellis-Davies. 2017. "How do hunter-gatherer children learn subsistence skills?" *Human Nature* 28 (4):367-394.

Li, Jingguang, Yajun Zhao, Rena Li, Lucas S Broster, Chenglin Zhou, and Suyong Yang. 2015. "Association of oxytocin receptor gene (OXTR) rs53576 polymorphism with sociality: a meta-analysis." *PLoS One* 10 (6):e0131820.

Li, Yun, Wenjuan Li, Tingting Zhang, Junjun Zhang, Zhenlan Jin, and Ling Li. 2021. "Probing the role of the right inferior frontal gyrus during Pain-Related empathy processing: Evidence from fMRI and TMS." *Human brain mapping* 42 (5):1518-1531.

Likowski, Katja U, Andreas Mühlberger, Antje Gerdes, Matthias J Wieser, Paul Pauli, and Peter Weyers. 2012. "Facial mimicry and the mirror neuron system: simultaneous acquisition of facial electromyography and functional magnetic resonance imaging." *Frontiers in human neuroscience* 6:214.

Lin, Alice, Ralph Adolphs, and Antonio Rangel. 2011. "Social and monetary reward learning engage overlapping neural substrates." *Social cognitive and affective neuroscience* 7 (3):274-281.

Lin, Alice, Ralph Adolphs, and Antonio Rangel. 2012. "Social and monetary reward learning engage overlapping neural substrates." *Social cognitive and affective neuroscience* 7 (3):274-281.

Liu, Xun, Jacqueline Hairston, Madeleine Schrier, and Jin Fan. 2011. "Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies." *Neuroscience & Biobehavioral Reviews* 35 (5):1219-1236.

Lombardo, Michael V, Bhismadev Chakrabarti, Edward T Bullmore, Simon Baron-Cohen, and MRC AIMS Consortium. 2011. "Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism." *Neuroimage* 56 (3):1832-1838.

LoParo, Devon, and ID Waldman. 2015. "The oxytocin receptor gene (OXTR) is associated with autism spectrum disorder: a meta-analysis." *Molecular psychiatry* 20 (5):640-646.

Loth, Eva, Jean-Baptiste Poline, Benjamin Thyreau, Tianye Jia, Chenyang Tao, Anbarasu Lourdusamy, David Stacey, Anna Cattrell, Sylvane Desrivières, and Barbara Ruggeri. 2014. "Oxytocin receptor genotype modulates ventral striatal activity to social cues and response to stressful life events." *Biological psychiatry* 76 (5):367-376.

Lucht, Michael J, Sven Barnow, Christine Sonnenfeld, Ines Ulrich, Hans Joergen Grabe, Winnie Schroeder, Henry Völzke, Harald J Freyberger, Ulrich John, and Falko H Herrmann. 2013. "Associations between the oxytocin receptor gene (OXTR) and “mind-reading” in humans—an exploratory study." *Nordic Journal of Psychiatry* 67 (1):15-21.

Luhrmann, Tanya. 2011. "Toward an anthropological theory of mind." *Suomen Antropologi: Journal of the Finnish Anthropological Society* 36 (4):5-69.

Luo, Lizhu, Benjamin Becker, Yayuan Geng, Zhiying Zhao, Shan Gao, Weihua Zhao, Shuxia Yao, Xiaoxiao Zheng, Xiaole Ma, and Zhao Gao. 2017. "Sex-dependent neural effect of oxytocin during subliminal processing of negative emotion faces." *Neuroimage* 162:127-137.

Luo, Siyang, Bingfeng Li, Yina Ma, Wenxia Zhang, Yi Rao, and Shihui Han. 2015. "Oxytocin receptor gene and racial ingroup bias in empathy-related brain activity." *NeuroImage* 110:22-31.

Luo, Siyang, Yina Ma, Yi Liu, Bingfeng Li, Chenbo Wang, Zhenhao Shi, Xiaoyang Li, Wenxia Zhang, Yi Rao, and Shihui Han. 2015. "Interaction between oxytocin receptor polymorphism and interdependent culture values on human empathy." *Social cognitive and affective neuroscience* 10 (9):1273-1281.

Ma, Yina, Simone Shamay-Tsoory, Shihui Han, and Caroline F Zink. 2016. "Oxytocin and social adaptation: insights from neuroimaging studies of healthy and clinical populations." *Trends in cognitive sciences* 20 (2):133-145.

MacDonald, Kai, and Tina Marie MacDonald. 2010. "The peptide that binds: a systematic review of oxytocin and its prosocial effects in humans." *Harvard review of psychiatry* 18 (1):1-21.

Mahajan, Neha, Margaret A Martinez, Natashya L Gutierrez, Gil Diesendruck, Mahzarin R Banaji, and Laurie R Santos. 2011. "The evolution of intergroup bias: perceptions and attitudes in rhesus macaques." *Journal of personality and social psychology* 100 (3):387.

Mahmood, Saba. 2011. "Politics of piety." In *Politics of Piety*. Princeton University Press.

Marlin, Bianca J, Mariela Mitre, James A D'amour, Moses V Chao, and Robert C Froemke. 2015. "Oxytocin enables maternal behaviour by balancing cortical inhibition." *Nature* 520 (7548):499.

Marlowe, Frank W, J Colette Berbesque, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Jean Ensminger, Michael Gurven, Edwins Gwako, and Joseph Henrich. 2008. "More 'altruistic'punishment in larger societies." *Proceedings of the Royal Society B: Biological Sciences* 275 (1634):587-592.

Marsh, Abigail A, H Yu Henry, Daniel S Pine, Elena K Gorodetsky, David Goldman, and RJR Blair. 2012. "The influence of oxytocin administration on responses to infant faces and potential moderation by OXTR genotype." *Psychopharmacology* 224 (4):469-476.

Marsh, Abigail A, Henry H Yu, Daniel S Pine, Elena K Gorodetsky, David Goldman, and RJR Blair. 2012. "The influence of oxytocin administration on responses to infant faces and potential moderation by OXTR genotype." *Psychopharmacology* 224 (4):469-476.

Martin, Jared, Magdalena Rychlowska, Adrienne Wood, and Paula Niedenthal. 2017. "Smiles as multipurpose social signals." *Trends in cognitive sciences* 21 (11):864-877.

Marusak, Hilary A, Daniella J Furman, Nisha Kuruvadi, David W Shattuck, Shantanu H Joshi, Anand A Joshi, Amit Etkin, and Moriah E Thomason. 2015. "Amygdala responses to salient social cues vary with oxytocin receptor genotype in youth." *Neuropsychologia* 79:1-9.

Mathersul, Danielle, Skye McDonald, and Jacqueline A Rushby. 2013. "Understanding advanced theory of mind and empathy in high-functioning adults with autism spectrum disorder." *Journal of clinical and experimental neuropsychology* 35 (6):655-668.

Matsumoto, David, Seung Hee Yoo, and Johnny Fontaine. 2008. "Mapping expressive differences around the world: The relationship between emotional display rules and individualism versus collectivism." *Journal of cross-cultural psychology* 39 (1):55-74.

McAuliffe, Katherine, Peter R Blake, Nikolaus Steinbeis, and Felix Warneken. 2017. "The developmental foundations of human fairness." *Nature Human Behaviour* 1 (2):0042.

McCullough, Michael E, Patricia Smith Churchland, and Armando J Mendez. 2013. "Problems with measuring peripheral oxytocin: can the data on oxytocin and human behavior be trusted?" *Neuroscience & Biobehavioral Reviews* 37 (8):1485-1492.

McElreath, Richard, Robert Boyd, and PeterJ Richerson. 2003. "Shared norms and the

evolution of ethnic markers." *Current anthropology* 44 (1):122-130.
McGettigan, Carolyn, Eamonn Walsh, R Jessop, ZK Agnew, DA Sauter, JE Warren, and SK Scott. 2015. "Individual differences in laughter perception reveal roles for mentalizing and sensorimotor systems in the evaluation of emotional authenticity." *Cerebral cortex* 25 (1):246-257.
McLellan, TL, JC Wilcke, Lucy Johnston, Richard Watts, and LK Miles. 2012. "Sensitivity to posed and genuine displays of happiness and sadness: A fMRI study." *Neuroscience letters* 531 (2):149-154.
McLellan, Tracey, Lucy Johnston, John Dalrymple-Alford, and Richard Porter. 2010. "Sensitivity to genuine versus posed emotion specified in facial displays." *Cognition and Emotion* 24 (8):1277-1292.
Mega, Laura F, Gerd Gigerenzer, and Kirsten G Volz. 2015. "Do intuitive and deliberate judgments rely on two distinct neural systems? A case study in face processing." *Frontiers in Human Neuroscience* 9:456.
Meissner, Christian A, and John C Brigham. 2001. "Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review." *Psychology, Public Policy, and Law* 7 (1):3.
Melchers, Martin, Christian Montag, Sebastian Markett, and Martin Reuter. 2013. "Relationship between oxytocin receptor genotype and recognition of facial emotion." *Behavioral neuroscience* 127 (5):780.
Menon, Vinod, and Lucina Q Uddin. 2010. "Saliency, switching, attention and control: a network model of insula function." *Brain structure and function* 214 (5):655-667.
Mesoudi, Alex, Lei Chang, Sasha RX Dall, and Alex Thornton. 2016. "The evolution of individual and cultural variation in social learning." *Trends in ecology & evolution* 31 (3):215-225.
Michalska, Kalina J, Jean Decety, Chunyu Liu, Qi Chen, Meghan Elizabeth Martz, Suma Jacob, Alison Hipwell, Steve S Lee, Andrea Chronis-Tuscano, and Irwin D Waldman. 2014. "Genetic imaging of the association of oxytocin receptor gene (OXTR) polymorphisms with positive maternal parenting." *Frontiers in behavioral neuroscience* 8:21.
Mikhail, John. 2007. "Universal moral grammar: Theory, evidence and the future." *Trends in cognitive sciences* 11 (4):143-152.
Milgram, Stanley. 1974. "The dilemma of obedience." *The Phi Delta Kappan* 55 (9):603-606.
Miller, Arlene Michaels, Edward Wang, Laura A Szalacha, and Olga Sorokin. 2009. "Longitudinal changes in acculturation for immigrant women from the former Soviet Union." *Journal of cross-cultural psychology* 40 (3):400-415.
Mitchell, Rachel LC, and Louise H Phillips. 2015. "The overlapping relationship between emotion perception and theory of mind." *Neuropsychologia* 70:1-10.
Molenberghs, Pascal, Ross Cunnington, and Jason B Mattingley. 2009. "Is the mirror neuron system involved in imitation? A short review and meta-analysis." *Neuroscience & biobehavioral reviews* 33 (7):975-980.
Moll, Jorge, Ricardo de Oliveira-Souza, Ivanei E Bramati, and Jordan Grafman. 2002. "Functional networks in emotional moral and nonmoral social judgments." *Neuroimage* 16 (3):696-703.
Moll, Jorge, Ricardo de Oliveira-Souza, and Paul J Eslinger. 2003. "Morals and the human brain: a working model." *Neuroreport* 14 (3):299-305.
Moll, Jorge, Paul J Eslinger, and Ricardo de Oliveira-Souza. 2001. "Frontopolar and anterior temporal cortex activation in a moral judgment task: preliminary functional MRI results in normal subjects." *Arquivos de neuro-psiquiatria* 59 (3B):657-664.
Moll, Jorge, Roland Zahn, Ricardo de Oliveira-Souza, Frank Krueger, and Jordan Grafman. 2005. "The neural basis of human moral cognition." *Nature Reviews Neuroscience* 6

(10):799.
Mollick, Jessica A, and Hedy Kober. 2020. "Computational models of drug use and addiction: A review." *Journal of abnormal psychology* 129 (6):544.
Moran, Joseph M, Liane L Young, Rebecca Saxe, Su Mei Lee, Daniel O'Young, Penelope L Mavros, and John D Gabrieli. 2011. "Impaired theory of mind for moral judgment in high-functioning autism." *Proceedings of the National Academy of Sciences* 108 (7):2688-2692.
Mu, Yan, Chunyan Guo, and Shihui Han. 2016. "Oxytocin enhances inter-brain synchrony during social coordination in male adults." *Social cognitive and affective neuroscience* 11 (12):1882-1893.
Müller-Bardorff, Miriam, Maximilian Bruchmann, Martin Mothes-Lasch, Pienie Zwitserlood, Insa Schlossmacher, David Hofmann, Wolfgang Miltner, and Thomas Straube. 2018. "Early brain responses to affective faces: a simultaneous EEG-fMRI study." *NeuroImage* 178:660-667.
Mustoe, Aaryn, Jack H Taylor, and Jeffrey A French. 2018. "Oxytocin structure and function in new world monkeys: From pharmacology to behavior." *Integrative zoology*.
Narumoto, Jin, Tomohisa Okada, Norihiro Sadato, Kenji Fukui, and Yoshiharu Yonekura. 2001. "Attention to emotion modulates fMRI activity in human right superior temporal sulcus." *Cognitive Brain Research* 12 (2):225-231.
Neumann, Roland, and Fritz Strack. 2000. "" Mood contagion": the automatic transfer of mood between persons." *Journal of personality and social psychology* 79 (2):211.
Nohlen, Hannah U, Frenk van Harreveld, and William A Cunningham. 2019. "Social evaluations under conflict: negative judgments of conflicting information are easier than positive judgments." *Social cognitive and affective neuroscience* 14 (7):709-718.
Northoff, Georg, Alexander Heinzel, Moritz De Greck, Felix Bermpohl, Henrik Dobrowolny, and Jaak Panksepp. 2006. "Self-referential processing in our brain—a meta-analysis of imaging studies on the self." *Neuroimage* 31 (1):440-457.
O'doherty, John, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J Dolan. 2004. "Dissociable roles of ventral and dorsal striatum in instrumental conditioning." *science* 304 (5669):452-454.
O'Doherty, John P, Peter Dayan, Karl Friston, Hugo Critchley, and Raymond J Dolan. 2003. "Temporal difference models and reward-related learning in the human brain." *Neuron* 38 (2):329-337.
O'Doherty, John P. 2014. "The problem with value." *Neuroscience & Biobehavioral Reviews* 43:259-268.
O'Doherty, John, Joel Winston, Hugo Critchley, David Perrett, D Michael Burt, and Raymond J Dolan. 2003. "Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness." *Neuropsychologia* 41 (2):147-155.
Olderbak, Sally, Oliver Wilhelm, Andrea Hildebrandt, and Jordi Quoidbach. 2019. "Sex differences in facial emotion perception ability across the lifespan." *Cognition and Emotion* 33 (3):579-588.
Ottman, Ruth. 1996. "Gene–environment interaction: definitions and study design." *Preventive medicine* 25 (6):764-770.
Pan, Wei-Xing, Robert Schmidt, Jeffery R Wickens, and Brian I Hyland. 2005. "Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network." *Journal of Neuroscience* 25 (26):6235-6242.
Paracampo, Riccardo, Emmanuele Tidoni, Sara Borgomaneri, Giuseppe Di Pellegrino, and Alessio Avenanti. 2017. "Sensorimotor network crucial for inferring amusement from smiles." *Cerebral cortex* 27 (11):5116-5129.
Pavarini, Gabriela, Rui Sun, Marwa Mahmoud, Ian Cross, Simone Schnall, Agneta Fischer,

Julia Deakin, Hisham Ziauddeen, Aleksandr Kogan, and Laura Vuillier. 2019. "The role of oxytocin in the facial mimicry of affiliative vs. non-affiliative emotions." *Psychoneuroendocrinology* 109:104377.

Petrovic, Predrag, Raffael Kalisch, Tania Singer, and Raymond J Dolan. 2008. "Oxytocin attenuates affective evaluations of conditioned faces and amygdala activity." *Journal of Neuroscience* 28 (26):6607-6615.

Pfundmair, Michaela, Wiebke Erk, and Annika Reinelt. 2017. "“Lie to me”—Oxytocin impairs lie detection between sexes." *Psychoneuroendocrinology* 84:135-138.

Piaget, Jean. 1932. "The moral development of the child." *Kegan Paul, London* 418.

Pineda, Jaime A. 2008. "Sensorimotor cortex as a critical component of an'extended'mirror neuron system: Does it solve the development, correspondence, and control problems in mirroring?" *Behavioral and Brain Functions* 4 (1):1-16.

Pitcher, David, Lúcia Garrido, Vincent Walsh, and Bradley C Duchaine. 2008. "Transcranial magnetic stimulation disrupts the perception and embodiment of facial expressions." *Journal of Neuroscience* 28 (36):8929-8933.

Piva, Matthew, and Steve WC Chang. 2018. "An integrated framework for the role of oxytocin in multistage social decision-making." *American Journal of Primatology* 80 (10):e22735.

Poldrack, Russell A. 2007. "Region of interest analysis for fMRI." *Social cognitive and affective neuroscience* 2 (1):67-70.

Porter, Stephen, Leanne Ten Brinke, and Brendan Wallace. 2012. "Secrets and lies: Involuntary leakage in deceptive facial expressions as a function of emotional intensity." *Journal of Nonverbal Behavior* 36 (1):23-37.

Powell, Nicole D, Erica K Sloan, Michael T Bailey, Jesusa MG Arevalo, Gregory E Miller, Edith Chen, Michael S Kobor, Brenda F Reader, John F Sheridan, and Steven W Cole. 2013. "Social stress up-regulates inflammatory gene expression in the leukocyte transcriptome via β-adrenergic induction of myelopoiesis." *Proceedings of the National Academy of Sciences* 110 (41):16574-16579.

Prenger, Margaret, and Penny A MacDonald. 2018. "Problems with facial mimicry might contribute to emotion recognition impairment in Parkinson’s disease." *Parkinson’s Disease* 2018.

Preston, Stephanie D, and Frans de Waal. 2002a. "The communication of emotions and the possibility of empathy in animals."

Preston, Stephanie D, and Frans BM De Waal. 2002b. "Empathy: Its ultimate and proximate bases." *Behavioral and brain sciences* 25 (1):1-20.

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly. 2000. "Inference of population structure using multilocus genotype data." *Genetics* 155 (2):945-959.

Prochnow, D, H Kossack, S Brunheim, K Müller, H-J Wittsack, H-J Markowitsch, and RJ Seitz. 2013. "Processing of subliminal facial expressions of emotion: a behavioral and fMRI study." *Social neuroscience* 8 (5):448-461.

Puglia, Meghan H, Jessica J Connelly, and James P Morris. 2018. "Epigenetic regulation of the oxytocin receptor is associated with neural response during selective social attention." *Translational psychiatry* 8 (1):1-10.

Quintana, Daniel S, and Adam J Guastella. 2020. "An allostatic theory of oxytocin." *Trends in Cognitive Sciences* 24 (7):515-528.

Quintana, Daniel S, Alexander Lischke, Sally Grace, Dirk Scheele, Yina Ma, and Benjamin Becker. 2021. "Advances in the field of intranasal oxytocin research: lessons learned and future directions for clinical research." *Molecular Psychiatry* 26 (1):80-91.

Quintana, Daniel S, Jaroslav Rokicki, Dennis van der Meer, Dag Alnæs, Tobias Kaufmann, Aldo Córdova-Palomera, Ingrid Dieset, Ole A Andreassen, and Lars T Westlye. 2019.

"Oxytocin pathway gene networks in the human brain." *Nature communications* 10.

Raghanti, Mary Ann, Melissa K Edler, Alexa R Stephenson, Emily L Munger, Bob Jacobs, Patrick R Hof, Chet C Sherwood, Ralph L Holloway, and C Owen Lovejoy. 2018. "A neurochemical hypothesis for the origin of hominids." *Proceedings of the National Academy of Sciences* 115 (6):E1108-E1116.

Rand, David G, Alexander Peysakhovich, Gordon T Kraft-Todd, George E Newman, Owen Wurzbacher, Martin A Nowak, and Joshua D Greene. 2014. "Social heuristics shape intuitive cooperation." *Nature communications* 5 (1):1-12.

Rangel, Antonio, Colin Camerer, and P Read Montague. 2008. "A framework for studying the neurobiology of value-based decision making." *Nature reviews neuroscience* 9 (7):545.

Read, Kenneth E. 1955. "Morality and the concept of the person among the Gahuku-Gama." *Oceania* 25 (4):233-282.

Reader, Simon M, and Kevin N Laland. 2002. "Social intelligence, innovation, and enhanced brain size in primates." *Proceedings of the National Academy of Sciences* 99 (7):4436-4441.

Reuter, Martin, Christian Montag, Steffen Altmann, Fabian Bendlow, Christian Elger, Peter Kirsch, Albert Becker, Susanne Schoch-McGovern, Matthias Simon, and Bernd Weber. 2017. "Functional characterization of an oxytocin receptor gene variant (rs2268498) previously associated with social cognition by expression analysis in vitro and in human brain biopsy." *Social neuroscience* 12 (5):604-611.

Riedl, Katrin, Keith Jensen, Josep Call, and Michael Tomasello. 2012. "No third-party punishment in chimpanzees." *Proceedings of the national academy of sciences* 109 (37):14824-14829.

Rilling, James K, Xiangchuan Chen, Xu Chen, and Ebrahim Haroon. 2018. "Intranasal oxytocin modulates neural functional connectivity during human social interaction." *American journal of primatology*:e22740.

Rilling, James K, Ashley C DeMarco, Patrick D Hackett, Xu Chen, Pritam Gautam, Sabrina Stair, Ebrahim Haroon, Richmond Thompson, Beate Ditzen, and Rajan Patel. 2014. "Sex differences in the neural and behavioral response to intranasal oxytocin and vasopressin during human social interaction." *Psychoneuroendocrinology* 39:237-248.

Rilling, James K, Ashley C DeMarco, Patrick D Hackett, Richmond Thompson, Beate Ditzen, Rajan Patel, and Giuseppe Pagnoni. 2012. "Effects of intranasal oxytocin and vasopressin on cooperative behavior and associated brain activity in men." *Psychoneuroendocrinology* 37 (4):447-461.

Rilling, James K, David A Gutman, Thorsten R Zeh, Giuseppe Pagnoni, Gregory S Berns, and Clinton D Kilts. 2002. "A neural basis for social cooperation." *Neuron* 35 (2):395-405.

Rilling, James K, and Jennifer S Mascaro. 2017. "The neurobiology of fatherhood." *Current opinion in psychology* 15:26-32.

Rilling, James K, Alan G Sanfey, Jessica A Aronson, Leigh E Nystrom, and Jonathan D Cohen. 2004. "The neural correlates of theory of mind within interpersonal interactions." *Neuroimage* 22 (4):1694-1703.

Rittschof, Clare C, and Gene E Robinson. 2014. "Genomics: moving behavioural ecology beyond the phenotypic gambit." *Animal Behaviour* 92:263-270.

Rizzolatti, Giacomo, and Laila Craighero. 2004. "The mirror-neuron system." *Annu. Rev. Neurosci.* 27:169-192.

Robinson, Gene E, Russell D Fernald, and David F Clayton. 2008. "Genes and social behavior." *science* 322 (5903):896-900.

Rodrigues, Sarina M, Laura R Saslow, Natalia Garcia, Oliver P John, and Dacher Keltner. 2009.

"Oxytocin receptor genetic variation relates to empathy and stress reactivity in humans." *Proceedings of the National Academy of Sciences* 106 (50):21437-21441.

Rogers, Forrest D, Sara M Freeman, Marina Anderson, Michelle C Palumbo, and Karen L Bales. 2021. "Compositional variation in early-life parenting structures alters oxytocin and vasopressin 1a receptor development in prairie voles (Microtus ochrogaster)." *Journal of neuroendocrinology* 33 (8):e13001.

Rokicki, Jaroslav, Tobias Kaufmann, Ann-Marie G De Lange, Dennis van der Meer, Shahram Bahrami, Alina M Sartorius, Unn K Haukvik, Nils Eiel Steen, Emanuel Schwarz, and Dan J Stein. 2022. "Oxytocin receptor expression patterns in the human brain across development." *Neuropsychopharmacology*:1-11.

Rudolf von Rohr, Claudia, Judith M Burkart, and Carel P Van Schaik. 2011. "Evolutionary precursors of social norms in chimpanzees: a new approach." *Biology & Philosophy* 26 (1):1-30.

Rychlowska, Magdalena, Elena Cañadas, Adrienne Wood, Eva G Krumhuber, Agneta Fischer, and Paula M Niedenthal. 2014. "Blocking mimicry makes true and false smiles look the same." *PLoS One* 9 (3):e90876.

Rymarczyk, Krystyna, Łukasz Żurawski, Kamila Jankowiak-Siuda, and Iwona Szatkowska. 2019. "Empathy in facial mimicry of fear and disgust: simultaneous EMG-fMRI recordings during observation of static and dynamic facial expressions." *Frontiers in psychology* 10:701.

Sanfey, Alan G, James K Rilling, Jessica A Aronson, Leigh E Nystrom, and Jonathan D Cohen. 2003. "The neural basis of economic decision-making in the ultimatum game." *Science* 300 (5626):1755-1758.

Sasaki, Joni Y. 2013. "Promise and challenges surrounding culture–gene coevolution and gene–culture interactions." *Psychological inquiry* 24 (1):64-70.

Sasaki, Joni Y, Jessica LeClair, Alexandria West, and Heejung S Kim. 2016. "The gene-culture interaction framework and implications for health." *The oxford handbook of cultural neuroscience*:279.

Sato, Wataru, Takanori Kochiyama, Shota Uono, and Sakiko Yoshikawa. 2008. "Time course of superior temporal sulcus activity in response to eye gaze: a combined fMRI and MEG study." *Social cognitive and affective neuroscience* 3 (3):224-232.

Saxe, Rebecca. 2006. "Uniquely human social cognition." *Current opinion in neurobiology* 16 (2):235-239.

Saxe, Rebecca, and Lindsey J Powell. 2006. "It's the thought that counts: specific brain regions for one component of theory of mind." *Psychological science* 17 (8):692-699.

Schaafsma, Sara M, Donald W Pfaff, Robert P Spunt, and Ralph Adolphs. 2015. "Deconstructing and reconstructing theory of mind." *Trends in cognitive sciences* 19 (2):65-72.

Scheele, Dirk, Keith M Kendrick, Christoph Khouri, Elisa Kretzer, Thomas E Schläpfer, Birgit Stoffel-Wagner, Onur Güntürkün, Wolfgang Maier, and René Hurlemann. 2014. "An oxytocin-induced facilitation of neural and emotional responses to social touch correlates inversely with autism traits." *Neuropsychopharmacology* 39 (9):2078.

Scheele, Dirk, Andrea Wille, Keith M Kendrick, Birgit Stoffel-Wagner, Benjamin Becker, Onur Güntürkün, Wolfgang Maier, and René Hurlemann. 2013. "Oxytocin enhances brain reward system responses in men viewing the face of their female partner." *Proceedings of the National Academy of Sciences* 110 (50):20308-20313.

Schnuerch, Robert, Martin Schnuerch, and Henning Gibbons. 2015. "Assessing and correcting for regression toward the mean in deviance-induced social conformity." *Frontiers in Psychology* 6:669.

Schultz, Wolfram. 2000. "Multiple reward signals in the brain." *Nature reviews neuroscience*

1 (3):199.
Schultz, Wolfram. 2016. "Dopamine reward prediction-error signalling: a two-component response." *Nature reviews neuroscience* 17 (3):183-195.
Schultz, Wolfram, Paul Apicella, and Tomas Ljungberg. 1993. "Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task." *Journal of neuroscience* 13 (3):900-913.
Schultz, Wolfram, Peter Dayan, and P Read Montague. 1997. "A neural substrate of prediction and reward." *Science* 275 (5306):1593-1599.
Schurz, Matthias, Joaquim Radua, Markus Aichhorn, Fabio Richlan, and Josef Perner. 2014. "Fractionating theory of mind: a meta-analysis of functional brain imaging studies." *Neuroscience & Biobehavioral Reviews* 42:9-34.
Seligman, Rebecca, and Ryan A Brown. 2009. "Theory and method at the intersection of anthropology and cultural neuroscience." *Social cognitive and affective neuroscience* 5 (2-3):130-137.
Seo, Hyojung, Xinying Cai, Christopher H Donahue, and Daeyeol Lee. 2014. "Neural correlates of strategic reasoning during competitive games." *Science* 346 (6207):340-343.
Shahrestani, Sara, Andrew H Kemp, and Adam J Guastella. 2013. "The impact of a single administration of intranasal oxytocin on the recognition of basic emotions in humans: a meta-analysis." *Neuropsychopharmacology* 38 (10):1929-1936.
Shamay-Tsoory, Simone G. 2011. "The neural bases for empathy." *The Neuroscientist* 17 (1):18-24.
Shamay-Tsoory, Simone G, and Ahmad Abu-Akel. 2016. "The social salience hypothesis of oxytocin." *Biological psychiatry* 79 (3):194-202.
Shamay-Tsoory, Simone G, Meytal Fischer, Jonathan Dvash, Hagai Harari, Nufar Perach-Bloom, and Yechiel Levkovitz. 2009. "Intranasal administration of oxytocin increases envy and schadenfreude (gloating)." *Biological psychiatry* 66 (9):864-870.
Shamay-Tsoory, Simone G, Nira Saporta, Inbar Z Marton-Alper, and Hila Z Gvirts. 2019. "Herding brains: a core neural mechanism for social alignment." *Trends in cognitive sciences* 23 (3):174-186.
Shen, Xun-bing, Qi Wu, and Xiao-lan Fu. 2012. "Effects of the duration of expressions on the recognition of microexpressions." *Journal of Zhejiang University Science B* 13 (3):221-230.
Shen, Xunbing, Wenfeng Chen, Guoying Zhao, and Ping Hu. 2019. "Recognizing Microexpression: An Interdisciplinary Perspective." *Frontiers in Psychology* 10:1318.
Shenhav, Amitai, and Joshua D Greene. 2010. "Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude." *Neuron* 67 (4):667-677.
Shenhav, Amitai, and Joshua D Greene. 2014. "Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex." *Journal of Neuroscience* 34 (13):4741-4749.
Sheskin, Mark, and Laurie Santos. 2012. "The evolution of morality: Which aspects of human moral concerns are shared with nonhuman primates." *The Oxford handbook of comparative evolutionary psychology* 13:434-49.
Shinotoh, Hitoshi. 2020. Oxytocin and mimicry enhance brain activity during social cognition in frontotemporal dementia. AAN Enterprises.
Shore, Danielle M, and Erin A Heerey. 2011. "The value of genuine and polite smiles." *Emotion* 11 (1):169.
Shweder, R, N Much, Manamohan Mahapatra, and Lawrence Park. 1997. "Divinity and the "big three" explanations of suffering." *Morality and health* 119:119-169.
Silani, Giorgia, Geoffrey Bird, Rachel Brindley, Tania Singer, Chris Frith, and Uta Frith. 2008.

"Levels of emotional awareness and autism: an fMRI study." *Social neuroscience* 3 (2):97-112.

Silk, Joan B, and Bailey R House. 2011. "Evolutionary foundations of human prosocial sentiments." *Proceedings of the National Academy of Sciences* 108 (Supplement 2):10910-10917.

Simon, Gregory M. 2009. "The soul freed of cares? Islamic prayer, subjectivity, and the contradictions of moral selfhood in Minangkabau, Indonesia." *American Ethnologist* 36 (2):258-275.

Singer, Tania, Ben Seymour, John O'doherty, Holger Kaube, Raymond J Dolan, and Chris D Frith. 2004. "Empathy for pain involves the affective but not sensory components of pain." *Science* 303 (5661):1157-1162.

Skuse, David H, Adriana Lori, Joseph F Cubells, Irene Lee, Karen N Conneely, Kaija Puura, Terho Lehtimäki, Elisabeth B Binder, and Larry J Young. 2014. "Common polymorphism in the oxytocin receptor gene (OXTR) is associated with human social recognition skills." *Proceedings of the National Academy of Sciences* 111 (5):1987-1992.

Sliwinska, Magdalena W, and David Pitcher. 2018. "TMS demonstrates that both right and left superior temporal sulci are important for facial expression recognition." *NeuroImage* 183:394-400.

Smith, Karen E, Eric C Porges, Greg J Norman, Jessica J Connelly, and Jean Decety. 2014. "Oxytocin receptor gene variation predicts empathic concern and autonomic arousal while perceiving harm to others." *Social neuroscience* 9 (1):1-9.

Staes, Nicky, Jeroen MG Stevens, Philippe Helsen, Mia Hillyer, Marisa Korody, and Marcel Eens. 2014. "Oxytocin and vasopressin receptor gene variation as a proximate base for inter-and intraspecific behavioral differences in bonobos and chimpanzees." *PLoS One* 9 (11):e113364.

Stallen, Mirre, Carsten KW De Dreu, Shaul Shalvi, Ale Smidts, and Alan G Sanfey. 2012. "The herding hormone: oxytocin stimulates in-group conformity." *Psychological science* 23 (11):1288-1292.

Stallen, Mirre, and Alan G Sanfey. 2015. "The neuroscience of social conformity: Implications for fundamental and applied research." *Frontiers in neuroscience* 9:337.

Stallen, Mirre, Ale Smidts, and Alan Sanfey. 2013. "Peer influence: neural mechanisms underlying in-group conformity." *Frontiers in human neuroscience* 7:50.

Stanković, Miloš, Jelena Bašić, Vuk Milošević, and Milkica Nešić. 2019. "Oxytocin receptor (OXTR) gene polymorphisms and recognition memory for emotional and neutral faces: a pilot study." *Learning and Motivation* 67:101577.

Stolier, Ryan M, and Jonathan B Freeman. 2017. "A neural mechanism of social categorization." *Journal of Neuroscience* 37 (23):5711-5721.

Stone, Arthur A, Christine A Bachrach, Jared B Jobe, Howard S Kurtzman, and Virginia S Cain. 1999. *The science of self-report: Implications for research and practice*: Psychology Press.

Strathearn, Lane, Jian Li, Peter Fonagy, and P Read Montague. 2008. "What's in a smile? Maternal brain responses to infant facial cues." *Pediatrics* 122 (1):40-51.

Striepens, Nadine, Keith M Kendrick, Vanessa Hanking, Rainer Landgraf, Ullrich Wüllner, Wolfgang Maier, and René Hurlemann. 2013. "Elevated cerebrospinal fluid and blood concentrations of oxytocin following its intranasal administration in humans." *Scientific reports* 3:3440.

Strobel, Alexander, Jan Zimmermann, Anja Schmitz, Martin Reuter, Stefanie Lis, Sabine Windmann, and Peter Kirsch. 2011. "Beyond revenge: neural and genetic bases of altruistic punishment." *Neuroimage* 54 (1):671-680.

Sutton, Richard S, and Andrew G Barto. 1998. *Introduction to reinforcement learning*. Vol. 135: MIT press Cambridge.

Taylor, Shelley E, Baldwin M Way, William T Welch, Clayton J Hilmert, Barbara J Lehman, and Naomi I Eisenberger. 2006. "Early family environment, current adversity, the serotonin transporter promoter polymorphism, and depressive symptomatology." *Biological psychiatry* 60 (7):671-676.

Ten Brinke, Leanne, Stephen Porter, Natasha Korva, Katherine Fowler, Scott O Lilienfeld, and Christopher J Patrick. 2017. "An examination of the communication styles associated with psychopathy and their influence on observer impressions." *Journal of Nonverbal Behavior* 41 (3):269-287.

Tollenaar, Marieke S, Michaela Chatzimanoli, Nic JA van der Wee, and Peter Putman. 2013. "Enhanced orienting of attention in response to emotional gaze cues after oxytocin administration in healthy young men." *Psychoneuroendocrinology* 38 (9):1797-1802.

Tost, Heike, Bhaskar Kolachana, Shabnam Hakimi, Herve Lemaitre, Beth A Verchinski, Venkata S Mattay, Daniel R Weinberger, and Andreas Meyer–Lindenberg. 2010. "A common allele in the oxytocin receptor gene (OXTR) impacts prosocial temperament and human hypothalamic-limbic structure and function." *Proceedings of the National Academy of Sciences* 107 (31):13936-13941.

Tung, Jenny, Luis B Barreiro, Zachary P Johnson, Kasper D Hansen, Vasiliki Michopoulos, Donna Toufexis, Katelyn Michelini, Mark E Wilson, and Yoav Gilad. 2012. "Social environment is associated with gene regulatory variation in the rhesus macaque immune system." *Proceedings of the National Academy of Sciences* 109 (17):6490-6495.

Twomey, Steve. 2010. "Phineas Gage: Neuroscience's most famous patient." *Smithsonian*.

Underwood, Bill, and Bert Moore. 1982. "Perspective-taking and altruism." *Psychological bulletin* 91 (1):143.

Vaidyanathan, Radhika, and Elizabeth AD Hammock. 2017. "Oxytocin receptor dynamics in the brain across development and species." *Developmental neurobiology* 77 (2):143-157.

Valentin, Vivian V, and John P O'Doherty. 2009. "Overlapping prediction errors in dorsal striatum during instrumental learning with juice and money reward in the human brain." *Journal of neurophysiology* 102 (6):3384-3391.

Valstad, Mathias, Gail A Alvares, Maiken Egknud, Anna Maria Matziorinis, Ole A Andreassen, Lars T Westlye, and Daniel S Quintana. 2017. "The correlation between central and peripheral oxytocin concentrations: a systematic review and meta-analysis." *Neuroscience & Biobehavioral Reviews* 78:117-124.

Van Rijn, Sophie, André Aleman, Eric Van Diessen, Celine Berckmoes, Guy Vingerhoets, and René S Kahn. 2005. "What is said or how it is said makes a difference: role of the right fronto-parietal operculum in emotional prosody as revealed by repetitive TMS." *European Journal of Neuroscience* 21 (11):3195-3200.

van Schaik, Carel, Judith M Burkart, Adrian V Jaeggi, and Claudia Rudolf von Rohr. 2014. "Morality as a biological adaptation–an evolutionary model based on the lifestyle of human foragers." In *Empirically informed ethics: Morality between facts and norms*, 65-84. Springer.

van Schaik, Carel P, and Judith M Burkart. 2011. "Social learning and evolution: the cultural intelligence hypothesis." *Philosophical Transactions of the Royal Society B: Biological Sciences* 366 (1567):1008-1016.

Völlm, Birgit A, Alexander NW Taylor, Paul Richardson, Rhiannon Corcoran, John Stirling, Shane McKie, John FW Deakin, and Rebecca Elliott. 2006. "Neuronal correlates of theory of mind and empathy: a functional magnetic resonance imaging study in a nonverbal task." *Neuroimage* 29 (1):90-98.

Von Rohr, Claudia Rudolf, Sonja E Koski, Judith M Burkart, Clare Caws, Orlaith N Fraser, Angela Ziltener, and Carel P Van Schaik. 2012. "Impartial third-party interventions in captive chimpanzees: a reflection of community concern." *PloS one* 7 (3):e32494.

Waller, Rebecca, Nadia S Corral-Frías, Bianca Vannucci, Ryan Bogdan, Annchen R Knodt, Ahmad R Hariri, and Luke W Hyde. 2016. "An oxytocin receptor polymorphism predicts amygdala reactivity and antisocial behavior in men." *Social cognitive and affective neuroscience* 11 (8):1218-1226.

Walum, Hasse, Irwin D Waldman, and Larry J Young. 2016. "Statistical and methodological considerations for the interpretation of intranasal oxytocin studies." *Biological psychiatry* 79 (3):251-257.

Walum, Hasse, and Larry J Young. 2018. "The neural mechanisms and circuitry of the pair bond." *Nature Reviews Neuroscience* 19 (11):643-654.

Wang, Danyang, Xinyuan Yan, Ming Li, and Yina Ma. 2017. "Neural substrates underlying the effects of oxytocin: a quantitative meta-analysis of pharmaco-imaging studies." *Social Cognitive and Affective Neuroscience* 12 (10):1565-1573.

Watanabe, Takamitsu, Takeshi Otowa, Osamu Abe, Hitoshi Kuwabara, Yuta Aoki, Tatsunobu Natsubori, Hidemasa Takao, Chihiro Kakiuchi, Kenji Kondo, and Masashi Ikeda. 2017. "Oxytocin receptor gene variations predict neural and behavioral response to oxytocin in autism." *Social cognitive and affective neuroscience* 12 (3):496-506.

Whalen, Paul J, Jerome Kagan, Robert G Cook, F Caroline Davis, Hackjin Kim, Sara Polis, Donald G McLaren, Leah H Somerville, Ashly A McLean, and Jeffrey S Maxwell. 2004. "Human amygdala responsivity to masked fearful eye whites." *Science* 306 (5704):2061-2061.

Wilson, Edward O. 1975. "Some central problems of sociobiology." *Social Science Information* 14 (6):5-18.

Wittfoth, Matthias, Dina M Schardt, Manfred Fahle, and Manfred Herrmann. 2009. "How the brain resolves high conflict situations: double conflict involvement of dorsolateral prefrontal cortex." *Neuroimage* 44 (3):1201-1209.

Wojciszke, Bogdan, Roza Bazinska, and Marcin Jaworski. 1998. "On the dominance of moral categories in impression formation." *Personality and Social Psychology Bulletin* 24 (12):1251-1263.

Wood, Adrienne, Magdalena Rychlowska, Sebastian Korb, and Paula Niedenthal. 2016. "Fashioning the face: sensorimotor simulation contributes to facial expression recognition." *Trends in cognitive sciences* 20 (3):227-240.

Wu, Haiyan, Yi Luo, and Chunliang Feng. 2016. "Neural signatures of social conformity: A coordinate-based activation likelihood estimation meta-analysis of functional brain imaging studies." *Neuroscience & Biobehavioral Reviews* 71:101-111.

Wu, Nan, Zhi Li, and Yanjie Su. 2012. "The association between oxytocin receptor gene polymorphism (OXTR) and trait empathy." *Journal of affective disorders* 138 (3):468-472.

Xu, Lei, Benjamin Becker, and Keith M Kendrick. 2019. "Oxytocin facilitates social learning by promoting conformity to trusted individuals." *Frontiers in Neuroscience*:56.

Yan, Wen-Jing, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. 2013. "How fast are the leaked facial expressions: The duration of micro-expressions." *Journal of Nonverbal Behavior* 37 (4):217-230.

Yoder, Keith J, Eric C Porges, and Jean Decety. 2015. "Amygdala subnuclei connectivity in response to violence reveals unique influences of individual differences in psychopathic traits in a nonforensic sample." *Human brain mapping* 36 (4):1417-1428.

Yoder, KJ, C Harenski, KA Kiehl, and J Decety. 2015. "Neural networks underlying implicit and explicit moral evaluations in psychopathy." *Translational psychiatry* 5 (8):e625-

e625.
Young, Liane, Antoine Bechara, Daniel Tranel, Hanna Damasio, Marc Hauser, and Antonio Damasio. 2010. "Damage to ventromedial prefrontal cortex impairs judgment of harmful intent." *Neuron* 65 (6):845-851.
Young, Liane, Joan Albert Camprodon, Marc Hauser, Alvaro Pascual-Leone, and Rebecca Saxe. 2010. "Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments." *Proceedings of the National Academy of Sciences* 107 (15):6753-6758.
Young, Liane, Fiery Cushman, Marc Hauser, and Rebecca Saxe. 2007. "The neural basis of the interaction between theory of mind and moral judgment." *Proceedings of the National Academy of Sciences* 104 (20):8235-8240.
Young, Liane, David Dodell-Feder, and Rebecca Saxe. 2010. "What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind." *Neuropsychologia* 48 (9):2658-2664.
Young, Liane, and James Dungan. 2012. "Where in the brain is morality? Everywhere and maybe nowhere." *Social neuroscience* 7 (1):1-10.
Young, Liane, and Rebecca Saxe. 2009. "An FMRI investigation of spontaneous mental state inference for moral judgment." *Journal of cognitive neuroscience* 21 (7):1396-1405.
Yu, Rongjun, and Li Chen. 2015. "The need to control for regression to the mean in social psychology studies." *Frontiers in Psychology* 5:1574.
Zak, Paul J. 2013. *The moral molecule: how trust works*: Penguin.
Zak, Paul J, Angela A Stanton, and Sheila Ahmadi. 2007. "Oxytocin increases generosity in humans." *PloS one* 2 (11):e1128.
Zaki, Jamil, Jessica Schirmer, and Jason P Mitchell. 2011. "Social influence modulates the neural computation of value." *Psychological science* 22 (7):894-900.
Zehetleitner, Michael, and Hermann J Mueller. 2010. "Salience from the decision perspective: You know where it is before you know it is there." *Journal of Vision* 10 (14):35-35.
Zhang, Hejing, Jörg Gross, Carsten De Dreu, and Yina Ma. 2019. "Oxytocin promotes coordinated out-group attack during intergroup conflict in humans." *elife* 8:e40698.
Zhang, Ming, Qiufang Fu, Yu-Hsin Chen, and Xiaolan Fu. 2014. "Emotional context influences micro-expression recognition." *PloS one* 9 (4):e95018.
Zhang, Ming, Qiufang Fu, Yu-Hsin Chen, and Xiaolan Fu. 2018. "Emotional context modulates micro-expression processing as reflected in event-related potentials." *PsyCh journal* 7 (1):13-24.
Zhang, Ming, Ke Zhao, Fangbing Qu, Kaiyun Li, and Xiaolan Fu. 2020. "Brain Activation in Contrasts of Microexpression Following Emotional Contexts." *Frontiers in neuroscience*:329.
Zheng, Jing-Jing, Shu-Jing Li, Xiao-Di Zhang, Wan-Ying Miao, Dinghong Zhang, Haishan Yao, and Xiang Yu. 2014. "Oxytocin mediates early experience–dependent cross-modal plasticity in the sensory cortices." *Nature neuroscience* 17 (3):391-399.
Zhu, Lusha, Kyle E Mathewson, and Ming Hsu. 2012. "Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning." *Proceedings of the National Academy of Sciences* 109 (5):1419-1424.
Zimbardo, Philip G, Craig Haney, W Curtis Banks, and David Jaffe. 1971. *The Stanford prison experiment*: Zimbardo, Incorporated.
Zinchenko, Oksana, and Marie Arsalidou. 2018. "Brain responses to social norms: Meta-analyses of f MRI studies." *Human brain mapping* 39 (2):955-970.