2021

Learning About Others Dynamically Changes Behavior And The Brain

Ariana Familiar
University of Pennsylvania

Learning About Others Dynamically Changes Behavior And The Brain

Abstract

Humans are social beings. The ability to interact socially requires associating perceptual and social information about other people. While prior work has elucidated the cognitive and neural basis of general social knowledge, less is known about how person-specific information is learned and remembered. The goal of this dissertation was to explore how learning associations between visual and abstract information could influence conceptual representations of specific individuals. Across three studies people learned social and reward values associated with different faces. Chapter 2 examined how the learned values influenced explicit judgments by measuring behavioral face similarity spaces before and after learning. While pre-learning spaces were structured by the visual similarity of the faces, social values selectively determined the post-learning spatial organization, and generalized to expectations of behavior in a future social context. Chapter 3 investigated the neural correlates of the face-value associations. Using functional magnetic resonance imaging (fMRI), brain activity patterns were measured while participants viewed faces and performed a task unrelated to the values, once before and once after learning. A region in the left anterior temporal lobe (ATL) had activity patterns that were biased by the social values after learning, such that faces of more similar social values evoked more similar activity patterns, and the magnitude of these learning-induced changes was directly related to an individual's learning performance as a function of value type. Additionally, activity pattern similarity in the left inferior parietal lobe (IPL) tracked the spatial organization of individual behavioral similarity spaces after learning. Chapter 4 assessed whether there were perceptual consequences of such behavioral and neural modulations and whether effects were domain-general. A categorical perception paradigm was used to test whether learned values implicitly influenced face discrimination. Preliminary evidence indicated that both social and reward values affected discrimination performance for face and flower stimuli, however the effect of social values did not persist over a long-term delay and was susceptible to task order effects. Together, this work indicates that learned associations between visual and social attributes of other people can warp behavioral and neural representations, and such changes have downstream consequences on face perception and social preferences.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Psychology

First Advisor

Sharon L. Thompson-Schill

Keywords

conceptual knowledge, face perception, multi-voxel pattern analysis, personality traits, representational similarity analysis, social perception

Subject Categories

Cognitive Psychology | Neuroscience and Neurobiology

LEARNING ABOUT OTHERS DYNAMICALLY CHANGES BEHAVIOR AND THE BRAIN

Ariana Familiar

A DISSERTATION

in

Psychology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation	Graduate Group Chairperson
________________________	________________________
Sharon Thompson-Schill, Ph.D.	Russell Epstein, Ph.D.
Christopher H. Browne Distinguished Professor of Psychology	Professor of Psychology

Dissertation Committee

Nicole Rust, Ph.D., Associate Professor of Psychology

Russell Epstein, Ph.D., Professor of Psychology

ACKNOWLEDGMENT

If it were not for the family, friends, co-graduate students, mentors, and animals I have in my life, this work would not have been possible.

First and foremost, I thank my parents, Bob and Mandy Familiar. You have supported and guided me unconditionally.

I express my sincere gratitude for my advisor, Sharon Thompson-Schill. Sharon's tutelage has not only guided the scientific research presented in this document but has shaped my professional and personal self. Many thanks to the members of the Thompson-Schill lab, past and present, who provided me with impactful direction, feedback, and support in my scientific endeavors and otherwise, and particular thanks to Alice Xie, Heath Matheson, and Siqi Lin. Won Mok Shim was my first academic mentor who played an important role in my development before graduate school. I am very thankful for the Penn Psychology faculty, especially Nicole Rust, Russell Epstein, David Brainard, and Sara Jaffee, for their invaluable conversations and guidance throughout my graduate career. I also appreciate the Penn Psychology graduate students I got to know. There were many conversations shared, oftentimes over food, from which I learned and laughed a lot. I am especially grateful for Michael Barnett and Benjamin Chin.

I thank Bobby Familiar, who I'm lucky to call my brother, as well as my friends, short and long, near and far.

Two of the most important beings to me are my partner, Anderson Lebbad, and my dog, Ori. Anderson, you have been at my side through all the ups and downs, thank you for being my support system both emotionally and at the crag. Ori, you are the light of my life.

Last, this work would not have been possible without support from the Social and Behavioral Science Initiative (SBSI) of the MindCORE center at Penn, the National Sciences Foundation (NSF) IGERT Traineeship in Complex Scene Perception, and the National Institutes of Health (NIH).

ABSTRACT

LEARNING ABOUT OTHERS DYNAMICALLY CHANGES BEHAVIOR AND THE BRAIN

Ariana Familiar

Sharon Thompson-Schill

Humans are social beings. The ability to interact socially requires associating perceptual and social information about other people. While prior work has elucidated the cognitive and neural basis of general social knowledge, less is known about how person-specific information is learned and remembered. The goal of this dissertation was to explore how learning associations between visual and abstract information could influence conceptual representations of specific individuals. Across three studies people learned social and reward values associated with different faces. Chapter 2 examined how the learned values influenced explicit judgments by measuring behavioral face similarity spaces before and after learning. While pre-learning spaces were structured by the visual similarity of the faces, social values selectively determined the post-learning spatial organization, and generalized to expectations of behavior in a future social context. Chapter 3 investigated the neural correlates of the face-value associations. Using functional magnetic resonance imaging (fMRI), brain activity patterns were measured while participants viewed faces and performed a task unrelated to the values, once before and once after learning. A region in the left anterior temporal lobe (ATL) had activity patterns that were biased by the social values after learning, such that faces of more similar social values evoked more similar activity patterns, and the magnitude of these learning-induced changes was directly related to an individual's learning performance as a function of value type. Additionally, activity pattern similarity in the left inferior parietal lobe (IPL) tracked the spatial organization of individual behavioral similarity spaces after learning. Chapter 4 assessed whether there were perceptual consequences of such behavioral and neural modulations and whether effects were domain-general. A categorical perception paradigm was used to test whether learned values implicitly influenced face discrimination. Preliminary evidence indicated that both social and reward values affected discrimination performance for face and flower stimuli, however the effect of social values did not persist over a long-term delay and was susceptible to task order effects. Together, this work indicates that learned associations between visual and social attributes of other people can warp behavioral and neural representations, and such changes have downstream consequences on face perception and social preferences.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. GENERAL INTRODUCTION

To interact with one's environment and other agents, people must flexibly incorporate knowledge gleaned from past experience with the information available in their surroundings. Conceptual knowledge scaffolds this ability, and is structured by concrete information derived from sensory and motor sources (e.g., knowledge of objects, places, people, animals and their properties) as well as abstract information that cannot be directly tied to a physical state or referent (e.g., justice, love, consciousness, selfishness). It is perhaps unsurprising that recent advances in research on conceptual knowledge in the mind and brain have mainly focused on concrete concepts, as their properties are based in sensorial experience and in turn are quantifiable and comparable by their physical attributes. Abstract concepts are no less relevant to human cognition than concrete ones, however they can be harder to pin down and their format remains heavily debated (Borghi et al., 2017; Kiefer & Pulvermüller, 2012).

One avenue for studying the cognitive and neural basis of abstract concepts is to examine the ways in which they can be tied to perceivable information. Some would even go as far as claiming that all concepts are fundamentally grounded in and depend on perceptual and motor systems (Matheson & Barsalou, 2018; but see Mahon, 2015). On the flip side of the same coin, perceptual representations are often imbued with conceptual associations acquired through experience, and elucidating interactions between perceptually-derived and abstract information is crucial for understanding the organization of conceptual knowledge and how it may change over time. Focusing on conceptual structures that involve visual attributes is a particularly intriguing way to probe such associations, as we typically must parse and make meaning of the rich and varied visual input relayed from the retina. Moreover, a large body of literature has established the cognitive and neural organization of hierarchically constructed visual representations (Hochstein & Ahissar, 2002; Cavanagh, 2011), providing the groundwork for examining how such visual information can be integrated with other information types.

The overarching motivation of this dissertation is to explore how learning associations between visual and abstract information can influence representations initially organized by perceptual information. To examine this, I focus on person-specific knowledge. Social behavior necessitates the inference of abstract social properties (e.g., personality traits, social categories, mental states of others) and the association of these attributes with perceptual information about others (e.g., face identity, voice), which people are typically able to do flexibly and with ease. That said, it remains to be established how such distinct pieces of information about other individuals become linked.

This is not to ignore the past several decades of research on social cognition, social perception, and social neuroscience. Indeed, significant advancements have been made in the study of social behaviors, such as inferring the mental states of others, the influence of social stereotypes on behavior, social impressions of others' faces, as well as the relevant brain systems and downstream consequences at a societal-level (Adolphs, 2006; Frith & Frith, 2012; Frith, 2007; Rule et al., 2013). Work in the realm of social perception has examined the properties and categories by which judgments of social attributes, such as personality traits and group membership, are made based on visual information. Research in this area has found remarkable consistency in observers' evaluations of faces along social trait dimensions (e.g., judging how kind a face appears), even when presented with an image of a face for only a fraction of a second (Todorov et al., 2015). Perceptual features based on facial morphology have been found to predict social impressions of faces, and interestingly the underlying dimensions that explain a large amount of variance in these judgments (i.e., dominance and trustworthiness; Oosterhof & Todorov, 2008) are comparable to those that structure more abstract, higher-level social judgments such as group categories and associated stereotypes (i.e., competence and warmth; Fiske et al., 2007). Recent studies have built on these findings by showing how social perception can be dependent on the stimulus, context, and perceiver (Todorov & Porter, 2014; Sutherland et al., 2017; Mileva et al., 2019) and depend on the general conceptual beliefs of the observer (Stolier et al., 2018a,b; Stolier et al., 2020).

The vast majority of these studies, however, have focused on categorical judgments of unfamiliar others. Arguably more important in daily life and more important for understanding person knowledge is our memory for specific individuals with whom we are familiar with compared to those who are strangers. There is some evidence that social perception differs when we have additional knowledge about others. For example, one recent study found that trait judgments of famous faces were less influenced by image-based variation than unfamiliar faces (Mileva et al., 2019). At the same time, studies on memory for familiar others have been primarily restricted to episodic memory (Maddock et al., 2001), mental imagery (Ishai et al., 2002; Thornton & Mitchell, 2017), or more fundamental aspects of face perception (Burton et al., 2011; Visconti di Oleggio Castello et al., 2017; Elfgren et al., 2006). Neuropsychological studies on patients with progressive brain damage have implicated a dissociation of person-specific and more general semantic knowledge (Kay & Hanley, 1999; Thompson et al., 2004; Predovan et al., 2014), and some neuroimaging studies have studied the neural correlates of person-specific knowledge but have focused on associations between faces and biographical information such as name, job, or age (Tsukiura et al., 2010; Wang et al., 2017).

1.1 Current approach

It remains to be established how abstract social properties are associated with perceptual information about specific individuals. Across three chapters, I investigate how learned associations between social traits and face identities can modulate the structure of cognitive and neural representations of faces and related influences on behavior.

Cognitive representations of information can be defined as multi-dimensional spaces (Shepard, 1980). Distinct points in this space correspond to unique items, and the distance between items is determined by the similarity of their attributes. Representational spaces can be constructed based on explicit judgements of similarity made by subjects. Similarity judgments have been used in many areas of psychology and can be informative in describing how people perceive and assess items and make decisions (Goldstone & Son, 2012). Similarity spaces can also be defined with neural or model-derived responses to items, allowing for comparisons of information structures between seemingly disparate measurement techniques. The validity and relevance of similarity in defining models of cognitive processes has been debated (Murphy & Medin, 1985; Goldstone, 1994a; Goldstone & Son, 2012), however, research on the mental and neural representations of conceptual knowledge have benefited from comparing behavioral, neural, and model-based similarity spaces and their underlying dimensions (Kriegeskorte et al., 2008, Kriegeskorte & Kievit, 2013). Here, I intend to use similarity measures as a tool to probe people's mental structures of information.

In Chapter 1, I use similarity judgments among a set of faces to test whether and how learned social traits (social values) and task rewards (reward values) can change an observer's mental 'face space' representation and the spatial relationships between faces. Furthermore, I examine how changes in behavioral similarity are related to learning, and how they generalize to participants' expectations in a future social context. This study establishes how learning visual-abstract associations can bias representations of face identities and how these biases may differ across perceivers.

In Chapter 2, I study the relationship between behavioral and neural face spaces and how they are influenced by learning face-value associations. Using functional magnetic resonance imaging (fMRI) and multi-voxel pattern analysis (MVPA), I examine whether learning associations between faces, social, and reward information influences the similarity of responses to different faces in the visual system, and test recent proposals that social information biases responses to faces at early (Olson et al., 2013) and late (Freeman & Ambady, 2011; Stolier & Freeman, 2017) stages of the ventral visual processing pathway. This elucidates whether learned associations can influence conceptual structures of visual information in the brain.

In Chapter 3, implicit behavioral changes beyond the value learning task are assessed. Specifically, I examine whether the re-organization of face spaces due to learned associations has an impact on perceptual discrimination. I utilize a paradigm from the categorical perception literature, which has established that learned categorical groupings of stimuli influences discrimination performance as a function of group membership, suggestive of changes in the underlying representational space for those items. This study investigates whether representational changes due to value learning influence behavior in a value-irrelevant perceptual discrimination task. Moreover, the domain-generality of these effects are examined by testing an inanimate object in addition to the face stimuli.

The present work builds on existing literature in cognition and perception by focusing on how learning about other people can dynamically change underlying information structures. Each chapter assesses modulations at both the group-level, as well as at the individual-level, to better understand the heterogeneity in how people acquire and utilize different sources of information. Additionally, in each chapter I use a paradigm in which people interactively learn about others, as opposed to merely providing them explicit labels, to implement a learning context in which participants must infer social properties about other people over multiple events. In sum, this research contributes to our understanding of the behavioral and neural underpinnings of visual-abstract associations in person knowledge.

II. LEARNING SOCIAL VALUES MODULATES FACE SIMILARITY AND SOCIAL EXPECTATIONS

1. Introduction

Humans possess the remarkable ability to learn about hundreds of other people and objects. One important piece of information about encountered items is their value, which often becomes associated through positive and negative experiences. Values can be defined as concepts or beliefs that guide human perception and behavior (Allport, 1961; Schwartz, 1992). In this way, values are motivational constructs that can lead one to obtain desirable goals and behavioral outcomes. Remembering and comparing values of alternate choices is crucial to making decisions. While values of items can be monetary, they are often more abstract in nature. Arguably one of the most biologically relevant values are social values that are associated with other people, such as personality traits that inform social behavior. For instance, a high social value would correspond to individuals that have positive prosocial traits (e.g., generous), and low social value to those with negative prosocial traits (e.g., selfish). Extracting and encoding social values based on interactions with others is essential for social behavior across different environmental situations, as they can be used to assess and predict the behavior of others (Frith & Frith, 2012; Rilling & Sanfey, 2011; Ruff & Fehr, 2014).

Importantly, in order to utilize social values associated with other people, one must link learned values with additional knowledge about a person. This can consist of perceptual information, such as their visual appearance and the sound of their voice, as well as semantic information like names and personal characteristics. Learning, remembering, and updating information about others is essential for understanding past behavior and using it to generate expectations of future actions. Notably, this involves interactions between the cognitive processes that allow us to recognize someone, and those that support memory of the person and our past interactions with them. Here, we focus on associations between visual face identity, as it is a readily available perceptual cue to recognize an individual, and social traits, as they can inform realistic evaluations of other people and guide interactions with them (Lee, 2008).

In the realm of face processing, there has been ample research on social impressions of unfamiliar faces. It is well-established that social trait categories can be perceived rapidly during face perception (Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015; Todorov, 2017). For example, judgments of race, gender, emotional expressions, and personality traits can occur at a fraction of a second (Adolphs, 2006; Willis & Todorov, 2006; Kubota & Ito, 2016). Moreover, people infer personality traits of unfamiliar faces, such as competence, intelligence, aggressiveness, and trustworthiness, with significant consensus, even when presented with a face for only 100 ms (Willis & Todorov, 2006; Todorov, 2008; Bar, Neta, & Linz, 2006; Hassin & Trope, 2000). It has been found that such trait impressions are linked to physical facial features, such as the shape of a face's mouth or brow (Oosterhof & Todorov, 2008; Todorov & Oosterhof, 2011; Berry & McArthur, 1986; Montepare & Zebrowitz, 1998; Todorov et al., 2015; Zebrowitz, 2017). Importantly, these impressions have been shown to be related to social outcomes, such as how we evaluate and treat others, and how we expect them to act (Hassin & Trope, 2000; Hehman, Stolier, Freeman, Flake, & Xie, 2019; Todorov et al., 2015). For example, perceived competence and dominance have been shown to predict behavior in political elections and leadership selection (Todorov, Mandisodza, Goren, & Hall, 2005; Ballew & Todorov, 2007; Lenz & Lawson, 2011; Antonakis & Eubanks, 2017), criminal sentencing (Zebrowitz & McDonald, 1991), as well as success in one's career (Graham, Harvey, & Puri, 2017; Rule & Ambady, 2011; but see, Stoker, Garretsen, & Spreeuwers, 2016). Taken together, social information can be gleaned rapidly during perception based on facial appearance, and resulting social trait impressions have significant behavioral consequences.

One method of operationalizing the behaviorally relevant information associated with visual stimuli such as faces is by constructing spaces that represent relationships between them (Goldstone, 1994b). Distances in representational spaces reflect the similarity between items, such that more similar stimuli along a given dimension are closer together. The dimensions of these spaces can be defined by properties relevant to distinguishing between items and to quantifying meaningful groupings (categories). Advantages of using such an approach include the ability to define the relationships of a set of stimuli to one another using multiple dimensions, as well as to compare representational spaces across measurement modalities (e.g., model-defined and behaviorally-defined spaces; Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013; Groen et al., 2018). In the case of social impressions, spatial dimensions have typically been defined by facial features that vary continuously and correlate with perceived social traits (Oosterhof & Todorov, 2008; Todorov & Oosterhof, 2011). In this way, face spaces defined by former studies have been structured by perceptual information, as the variations between faces are the visual properties related to social judgments.

Previous work in social perception has linked perceptual face spaces to first impressions of unfamiliar faces and has found that resulting trait judgments are directly predicted by the facial features of a face. At the same time, social information associated with face identities is influenced by experience. It is certainly not the case that a co-worker is forever perceived as dominant and aggressive merely due to their prominent brow; instead, trait judgements are adaptively updated through learning contexts such as firsthand interactions or knowledge of other social categories to which they belong (e.g., their profession or nationality; Jenkins et al., 2018; Rosenberg et al., 1968). This begs the question of whether perceptual face spaces can be modulated by learning such traits. It remains to be established how such spaces may become re-organized when associated social traits are learned and how they are related to behavior beyond initial impressions of unfamiliar faces.

In the present study, we examine whether and how learned values can influence face similarity spaces. Participants performed a value learning task over the course of four days in which they learned social (generosity) and reward (point) values associated with different faces. Additionally, they made similarity judgments once before and once after learning. Although explicitly told to judge similarity, participants were free to use whatever information they believed relevant for making their judgments. After value learning, we also measured participants' preferences for being paired with each person (face) on a separate cooperative task if invited back for another study, and examined whether any face space changes due to the learned values generalized to such future expectations of social behavior. Specifically, we examined the types of information participants used to make similarity comparisons (perceptual, social value, and/or reward value), how they generalized to social preferences, and whether individual differences in learning for each value type was related to behavioral similarity and preference judgments.

We focus on the social trait of generosity, as it can communicate trustworthiness (Klapwijk & Van Lange, 2009), and trustworthiness has been shown to be a particularly relevant cognitive dimension underlying trait inferences during face perception (Oosterhof & Todorov, 2008; Willis & Todorov, 2006; Todorov, Pakrashi, & Oosterhof, 2009). Instead of having participants judge traits based on facial features, they learned about traits associated with stimuli through a cover task. They played a game in which they learned about point allocations made by other players. The proportion of point pools donated on average, as well as the magnitude of points donated on average, were varied across players (modified from Hackel, Doll, & Amodio, 2015). In other words, participants learned that different players were more or less generous (social values) and had more or less points to give (reward values). Participants were instructed to maximize the points the received from the players by learning these values. Importantly, the social and reward values were orthogonally assigned to the stimuli, so that social values were not confounded with behavioral task salience and the effects of each value type could be examined separately.

The goal of this study is to establish whether behavioral face spaces are modulated by learning associations between faces and values. If the organization of face space is influenced by learned values, it would show that this information can influence face judgments when values are learned through experience and are not based solely on initial impressions. This would support the notion that representational spaces can be dynamically modulated by experience and would underscore the malleability of face-trait representations.

2. Experiment 1

2.1 Materials and Methods

2.1.1 Participants

Thirty-four subjects were recruited from the University of Pennsylvania, four were excluded due to low accuracy on the learning task, resulting in thirty subjects in the dataset (6 male; ages 18-30 years). All subjects gave written informed consent within a protocol approved by the IRB at the University of Pennsylvania and were financially compensated for their time.

2.1.2 Stimuli

Nine monochrome face images were chosen from the Psychological Image Collection at Stirling database (pics.stir.ac.uk), all consisting of a single front-facing face with a neutral expression. Images were cropped with an oval mask to only include the center of each face (minimizing details such as hair, ears, etc.), and matched in mean luminance using the SHINE Toolbox (Willenbockel et al., 2010). We chose faces of the same race and gender (Caucasian male) and of similar ages in order to minimize any effects of these socially relevant dimensions. That said, it is still possible for the faces to vary by these and other perceived characteristic traits, such as friendliness, attractiveness, etc. As perceptual aspects of facial appearance have been shown to predict initial trait judgments across individuals (see Introduction), we aimed to control for these factors by collecting perceptual similarity judgments from a separate group of participants (see description of procedure below). We used these judgments to orthogonalize the measured perceptual similarity with the values assigned to the faces, and to control for perceptual similarity in the data analysis.

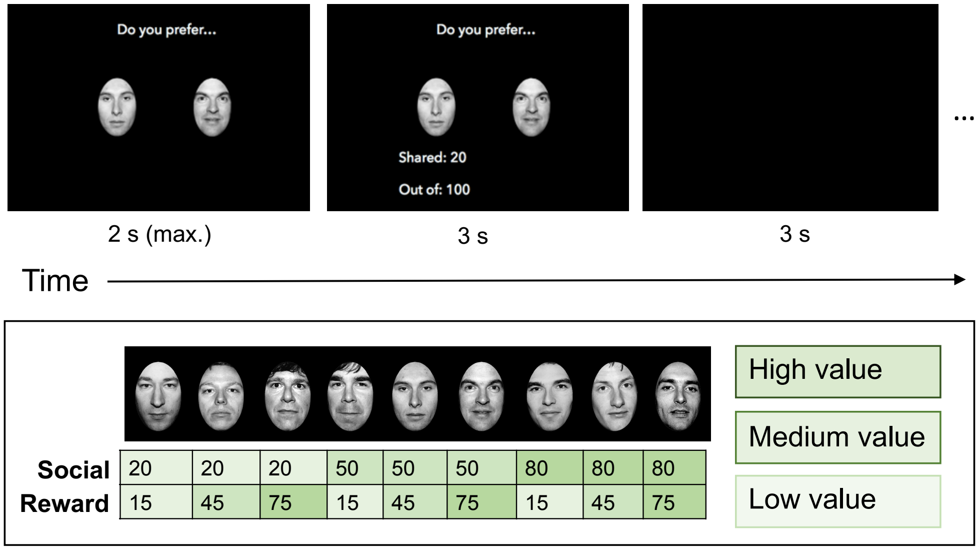
2.1.3 Value Learning Procedure

A modified version of the task used in Hackel, Doll, & Amodio (2015) was conducted online using PsyToolkit (Stoet, 2017) over four consecutive days. Participants were recruited under the impression that they would be randomly assigned to either a social choice or a social learning role, however all were assigned to a social learning role. After this assignment, participants were told that during training, they would learn about the actions of other ‘players’ assigned to the social choice role, who allocated a pool of points between themselves and a future player (the participant). On a given trial, participants chose one of two players, presented side by side on a computer screen. To make a choice, participants pressed one of two keys to indicate the player on the left (F-key) or the player on the right (J-key); they had 2 seconds to respond before the next trial commenced (inter-trial-interval of 3 seconds). Upon choosing, they were presented with feedback for 3 seconds about how many points that player gave them and the point pool the player was allocated on that trial (Figure 2.1). If no choice was made, no feedback was presented.

Participants were instructed to maximize their accrued points, as the total number of points they earn amounted to a cash bonus. Moreover, they were told that some players were given more points on average to allocate (reward value), and some gave a higher proportion of their point pool on average (social value), so they would have to learn about both sources of information over the course of training in order to maximize their total points.

On average, players shared 20, 50, or 80% of their point pool, which was either 15, 45, or 75 points. On a given trial, noise was added to these values by randomly selecting a value from a normal distribution centered on zero (standard deviation of 5 for point values, 5% for generosity values) and adding it to the average value for that face. Point pools were calculated by dividing the point value by the generosity value for that trial. Participants completed about an hour of training each day (288 trials per session), for four days, and were shown their accrued number of points at the end of each session.

Social values (average proportion of point pool shared) and reward (average magnitude of points shared) values were orthogonally assigned to the faces as follows. Pairwise differences in perceptual similarity were calculated (described below), and combinations of social and reward values were assigned to the faces such that they were orthogonal to perceptual similarity, and orthogonal to one another (Fig. 2.S1-S2). This allowed us to examine the effect of each value type separately, while controlling for perceptual similarity.



Social	20	20	20	50	50	50	80	80	80
Reward	15	45	75	15	45	75	15	45	75

Figure 2.1

Value learning paradigm. (Top) Example of one trial in the learning task, participants saw two faces simultaneously and chose to play with one of the two players. If they made a choice within the allotted time (max. 2 seconds), they received feedback about how many points they received on that trial from the chosen player (labeled “Shared”) and how many points in that players point pool for that trial (labeled “Out of”). (Bottom) value assignments for the nine faces. Each player had one social value (average percent of point pool shared) and one reward value (average point magnitude shared), and values were orthogonalized across the faces.

2.1.4 Free Sorting Similarity Tasks

2.1.4.1 Behavioral Similarity. A free sorting task was used to quantify conceptual similarity spaces before and after learning (Goldstone, 1994b). Participants were shown the nine face images on a white background and were instructed to organize the images in a spatial manner that reflected their similarity. The closer together the images were in space, the more similar the people depicted were, and the farther apart, the more dissimilar. There were no time constraints on the completion of the task, and people were instructed to use whatever information they thought was relevant to make their judgements. Participants performed the task once before the learning task on the first day of learning, and once after completing the value learning task on the last day of learning.

2.1.4.2 Perceptual Similarity. A separate group of 20 participants performed the same free sorting task for course credit, however they were explicitly instructed to organize the images in a manner that reflected the perceptual similarity of the faces.

2.1.5 Post-Learning Ratings

At the end of the last learning session, participants completed the following ratings, conducted using Qualtrics.

2.1.5.1 Social Value Ranking. Participants were instructed to rank the players in order of their overall generosity in the social choice role (i.e., the average proportion of their point pool shared). They were presented with the nine faces in a random order, and then clicked and dragged the images to reposition them until they were satisfied with the ranking.

2.1.5.2 Reward Value Ranking. This task was the same as the social value ranking except the participants were instructed to rank the players in order of their overall points donated in the social choice role (i.e., how many points the players gave to the participant on average). Data from two participants did not save properly and so all reported analyses involving this measure are based on the remaining twenty-eight subjects.

2.1.5.3 Social Preferences. Participants were told they may be invited back for an additional study involving a cooperative non-monetary task and we would accommodate preferences for people to be paired together. They were instructed to indicate their preference to be paired with the other players on that task (based on Hackel, Doll, & Amodio, 2015). For each face image, participants rated on a scale of 1-7 how much they preferred to be paired with that player (1 = not at all; 7 = definitely yes).

2.2 Results

2.2.1 Value learning performance

In order to have participants learn associations between values and faces, they completed four days of a value learning task for about one hour per day (modified from Hackel, Doll, & Amodio, 2015). Over the four days, participants learned social values (average percentage of points donated) and reward values (average magnitude of points donated) associated with nine faces (Fig. 2.1). For each day, accuracy for a value type was calculated based on whether participants chose higher value faces on trials in which the two faces differed on that value type but not on the other type (chance = 50%). For example, social accuracy for a trial was determined by whether the higher social value face was chosen, for two faces that had different social values (e.g., 20% and 80%) but equal reward values (e.g., 15 points). A repeated measures ANOVA of accuracy (value type x day) showed a significant effect of day ($F(3, 87) = 4.17$, $p = .008$) and interaction between day and value type ($F(3, 87) = 6.93$, $p < .001$), with no effect of value type ($F(3, 87) = 1.17$, $p = .289$; Fig. 2.2, left). Paired t-tests indicated a significant increase in reward accuracy between Day 1 ($M = .59$, $SEM = .02$) and Day 4 ($M = .72$, $SEM = .04$; $t(29) = 4.2$, $p < .001$), but no change in social accuracy between these days (Day 1: $M = .76$, $SEM = .03$; Day 4: $M = .7$, $SEM = .03$; $t(29) = -1.54$, $p = .134$). Additionally, social accuracy was significantly higher than reward accuracy on Day 1 ($t(29) = 4.36$, $p < .001$), but there was no difference between value types on Day 4 ($t(29) = 0.42$, $p = .679$).

These results indicate that accuracy for choosing faces based on their associated reward values increased over the four days of learning, and across participants accuracy was equivalent for the two value types at the end of learning. Interestingly, accuracy for social values did not change across the four days at a group-level. In fact, choosing faces based on their social values can be counter-productive to the goal of maximizing points for the cash bonus if primarily prioritizing social over reward information (e.g., choosing a high-social/low-reward over a low-social/high-reward face). Furthermore, it might be expected that social values would be harder to learn than reward values, as reward values are explicitly presented to participants in the trial feedback while social values have to be calculated based on the reward value and point pool information. Nonetheless, people learned the social values and used them to guide their choices on the learning task.

Additionally, there were individual differences in accuracy for both value types on Day 4, with some people having higher social than reward accuracy and others having higher reward than social accuracy (Fig. 2.2, right). The group became more polarized in these tendencies over the course of learning, with 24 of the 30 participants having higher reward accuracy than social accuracy on Day 1, but only 15 on Day 4 (Fig. 2.S3). Participants were also generally consistent in which value type they had higher accuracy for on Days 2 – 4 (Fig. 2.S3).

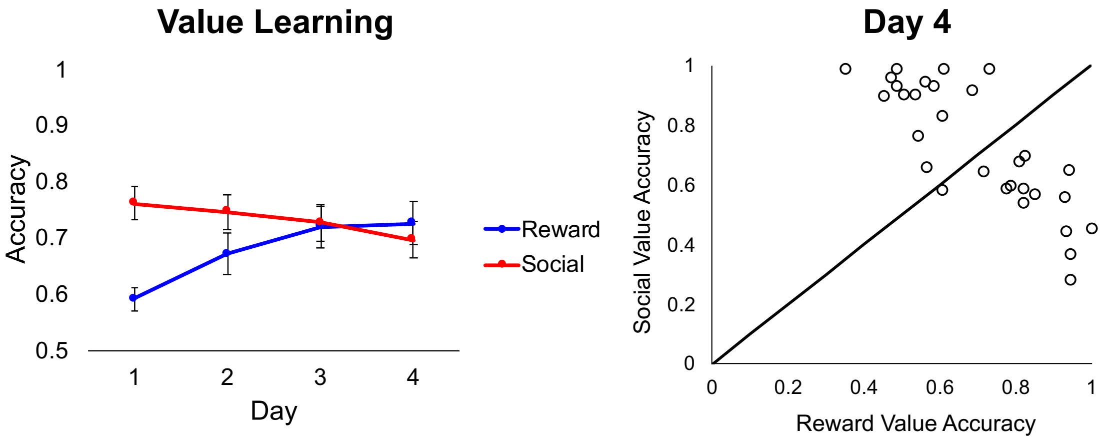


Figure 2.2

Value learning performance ($N = 30$). (Left) accuracy across participants for social (red) and reward (blue) values separately, on each day of learning; error bars indicate +/- *SEM*. (Right) accuracy for each value type on the last day of learning (Day 4), each point represents one participant.

2.2.2 Behavioral judgments of face and value similarity

Having established that participants learned social and reward values associated with the faces, we next examined whether and how value learning influenced the type of information used to judge similarity of the faces. To measure this, we had participants complete a free sorting task in which they generated similarity spaces with the face images. Participants simultaneously viewed the nine faces on a computer and arranged the faces such that the spatial organization (distance) of the images reflected their similarity (similar faces would be placed closer together, dissimilar faces would be placed farther apart). There was no time constraint on this task, and participants completed it once before the value learning task on the first day of learning (pre-learning), and once after the learning task on the last day of learning (post-learning). We did not instruct participants to use any specific types of information to make their similarity judgments (e.g., perceptual or value-based), thus they were free to choose the property(-ies) by which to make their responses.

To calculate perceived similarity of the faces, we computed the distance between each pair of faces as the pixel-wise Euclidean distance between their centers. A higher value indicates that the faces were perceived as more dissimilar. We then averaged the distances for each pair of faces across participants, to construct a behavioral dissimilarity matrix (DM) for pre- and post-learning spaces separately (Fig. 2.3). In order to examine what information participants were using to structure their face spaces, we correlated the pre- and post-learning behavioral DMs with DMs derived from other measures[1]. If the DMs are positively correlated, then they are similarly structured. In other words, we can conclude that the given type of information is related to how participants judged similarity.

We first tested whether perceptual information, or physical face appearance, was related to the behavioral similarity judgments. A separate group of participants (*N* = 20) were explicitly told to complete the free sorting task using perceptual information to judge the similarity of the faces. A perceptual DM was constructed using the pixel-wise distance between faces, and averaging the pairwise distances across participants (note that these results were used to orthogonalize perceptual similarity with the objective social and reward similarity in the experimental design; Fig. 2.S1-S2). The pre-learning behavioral DM was correlated with the perceptual DM ($r(34) = .73$, $p < .001$), but the

[1] Spearman correlations were used when similarity values for at least one measure were not normally distributed (determined by Shapiro-Wilk normality tests), otherwise Pearson correlations were utilized. Partial correlations were performed on rank-transformed data due to non-normality.

post-learning DM was not (although marginally related; $r(34) = .31$, $p = .065$; Fig. 2.S4), and the difference between the correlations was significant ($z = 2.76$; Steiger, 1980). This indicates that before value learning, participants used the visual appearance of the faces to make their similarity judgments, but after learning perceptual similarity was not determinate of the behavioral similarity space organization.

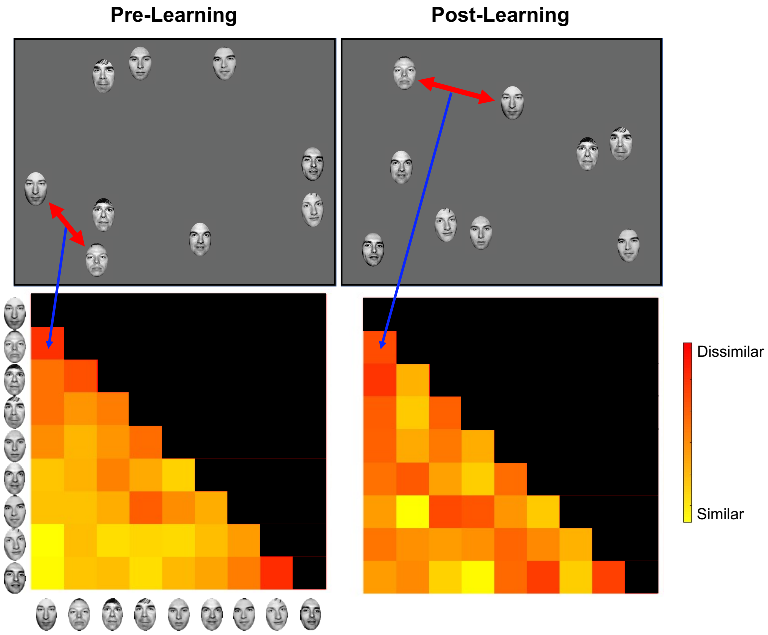


Figure 2.3

Behavioral similarity spaces. Illustration of dissimilarity matrix construction based on distance between face pairs in similarity space, separately for pre-value learning (left) and post-learning (right). Dissimilarity matrices (bottom) show the group-level results (average across subject-level DMs).

Next, we compared the behavioral DM with DMs based explicit value rankings that participants performed on the last day of value learning. Participants ranked the faces from lowest to highest value, once for social values and once for reward values. DMs were constructed based on the average difference in rank for face pairs across participants. Subjective value DMs were correlated with objective (experimentally-defined) value DMs (social: $r_s(34) = .9$, $p < .001$; reward: $r_s(34) = .68$, $p < .001$; Fig. 2.S5) as well as one another ($r_s(34) = .47$, $p = .004$). The subjective social DM was not correlated with the objective reward DM (subj-social/obj-reward: $r_s(34) = -.12, p = .475$); however, the subjective reward DM was correlated with both objective value DMs (subj-reward/obj-social: $r_s(34) = .42, p = .01$). An additional partial correlation analysis showed the subjective and objective social value DMs were related even when controlling for subjective reward similarity ($r(33) = .88, p < .001$). Subjective value DMs were not related to perceptual similarity (social: $r_s(34) = -.16, p = .362$; reward: $r_s(34) = -.22, p < .192$). Overall, this indicates that subjective value DMs were related to the objective values associated with the faces.

The post-learning behavioral DM was positively correlated with the subjective social value DM (partial $r(33) = .47, p = .004$, controlling for reward) but not reward value DM (partial $r(33) = .19, p = .265$, controlling for social; Fig. 2.4). The pre-learning behavioral DM was not correlated with either value DMs (social: $r(33) = -.02, p = .929$; reward: $r(33) = -.13, p = .457$), and a test of the difference between dependent correlations confirmed the correlations between social and post-learning DMs was greater than the correlation between social and pre-learning DMs ($z = 3.44$). The same pattern of results was found for the objective value DMs (Fig. 2.S6), with post-learning behavioral similarity spaces relating to social ($r_s(34) = .53, p = .001$) but not reward ($r_s(34) = .08, p = .645$) values, and neither value relating to pre-learning (social: $r_s(34) = -.19, p = .26$; reward: $r_s(34) = .1, p = .568$). This shows that participants' behavioral face spaces were modulated by the learned social values, and not the reward values, such that faces of similar social values became grouped together.

Last, we examined individual differences in behavioral space modulations and value learning performance. Specifically, we tested whether an individual's accuracy for each value type on the last day of value learning was related to the influence of each value on their behavioral similarity space judgments. We defined value learning sensitivity as a participant's social value accuracy minus their reward value accuracy. In this way, a higher learning sensitivity corresponds to being more accurate at choosing faces based on their social values compared to reward values (and vice versa). Behavioral sensitivity was calculated as the difference in correlations between a participant's value rankings and the change in their behavioral similarity scores (for each face pair, the post-learning distance minus pre-learning distance). A higher behavioral sensitivity value corresponds to a greater relationship between the change in their similarity ratings and their social value rankings, compared to their reward value rankings. The correlation between learning sensitivity and behavioral sensitivity was marginally significant (*r*(28) = .32, *p* = .1; Fig. 2.S7).

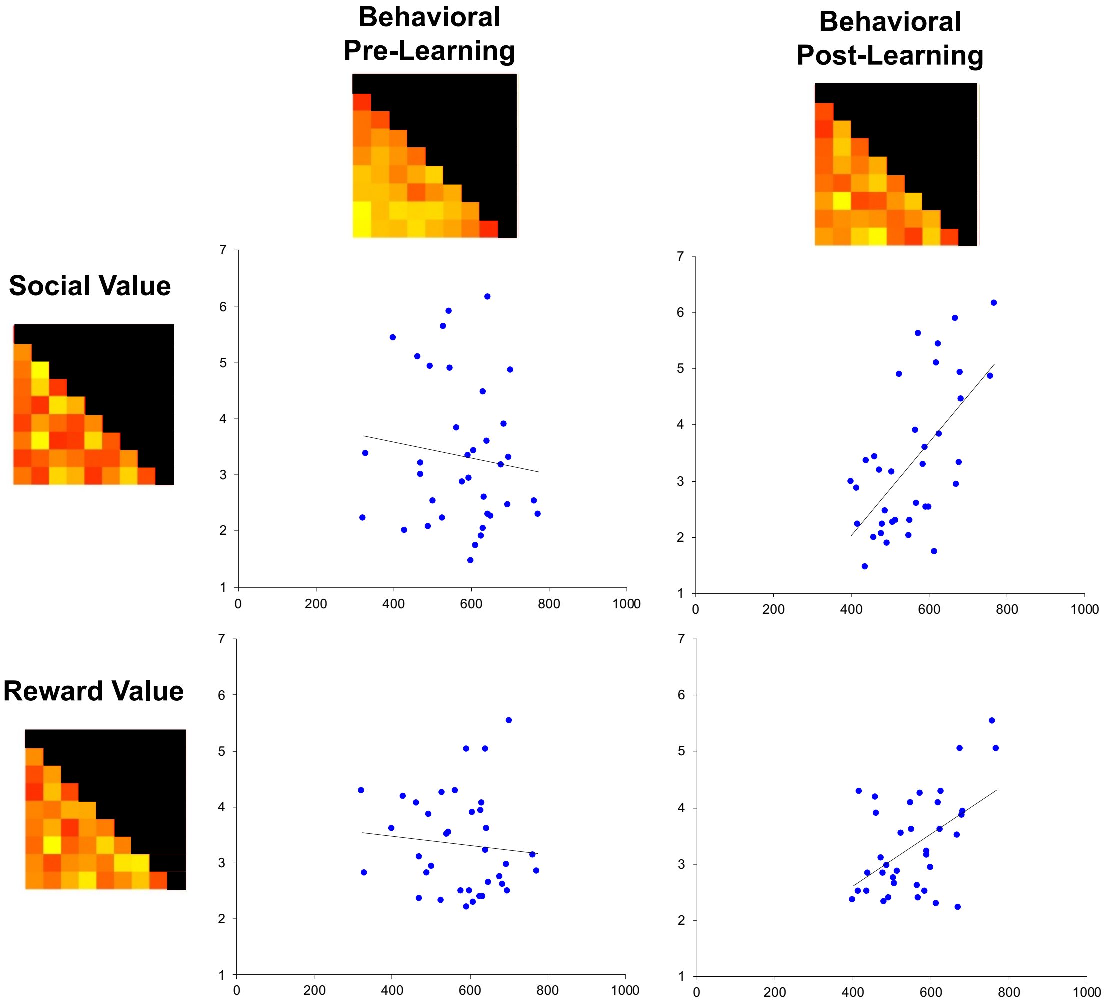


Figure 2.4

Similarity comparisons. Group-level dissimilarity matrices based on behavioral free sorting (upper) and value ranking (left) tasks (each cell corresponds to one face pair). Scatter plots show relationship between DMs (each point represents one face pair, linear trend line shown in black for visualization). The post-learning (right), but not pre-learning (left), DM was related to the subjective social value DM (upper) but not reward value (lower) DM.

2.2.3 Generalization of learned value information to social preferences

To examine whether the learned value information generalized to a separate context, we measured participants' expectations of social behavior. At the end of learning (after the value learning task on day 4), participants completed an additional task in which they rated each face based on how much they would like to be paired with that person if they were invited back for another study on cooperative problem-solving (based on Hackel, Doll, & Amodio, 2015). This allowed us to study what types of information participants used in their preferences in a future social context.

Value	**Level**	***M* (*SEM*)**
Social	*High*	6.14 (0.16)
	Medium	3.52 (0.18)
	Low	2.16 (0.17)
Reward	*High*	4.73 (0.16)
	Medium	3.76 (0.15)
	Low	3.33 (0.16)

Table 2.1

Preference ratings for each value type and value level (averaged over the other value type; $N = 30$).

A repeated-measures ANOVA on preference ratings for each face showed significant effects of the objective social ($F(2, 58) = 112.6$, $p < .001$) and reward values ($F(1.6, 46.51) = 19.32$, $p < .001$, Greenhouse-Giesser corrected) on ratings, with no interaction between the value types ($F(2.29, 66.29) = 0.75$, $p = .493$, Greenhouse-Giesser corrected; Fig. 2.5, left)[2]. Averaging over the different reward values, high social value faces were the most preferred, and low social values the least (Table 2.1).

A preference similarity measure was calculated as the difference in preference ratings for each pair of faces. The corresponding preference rating DM was strongly correlated with the DM based on the social value rankings (partial $r(33) = .93$, $p < .001$

[2] Each subjects' ratings were normalized to the range of 1 – 7 prior to all analyses.

controlling for reward; Fig. 2.S8, upper), and marginally correlated with the reward value DM (partial $r(33) = .28$, $p = .1$ controlling for social). The same trend was found when comparing preference and objective value DMs (social: $r_s(34) = .86$, $p < .001$; reward: $r_s(34) = -.02$, $p = .896$; Fig. 2.S8, lower). There was also a positive relationship between preference and post-learning DMs ($r(34) = .61$, $p < .001$), but not pre-learning distances ($r(34) = -.05$, $p = .754$), and there was a significance difference between these correlations ($z = 3.6$; Fig. 2.S8, middle). Preference similarity was not related to perceptual similarity ($r(34) = -.04$, $p = .838$). Together, this indicates that at a group-level faces of more similar social values shared similar preference scores.

Next, we sought to compare individual differences in sensitivity to social and reward information in the preference ratings to individual biases for these values during the learning task. To test whether participants made their preference ratings based on social and/or reward information, we derived a preference sensitivity measure as follows. For each participant and value type, the average rating for low value faces was subtracted from the average rating for high value faces (social/reward sensitivity). Then the reward sensitivity value was subtracted from the social sensitivity value.

$$\begin{aligned}\textit{Preference sensitivity} = {} & \textit{(mean(high-social) – mean(low-social))} – \\ & \textit{(mean(high-reward) – mean(low-reward))}\end{aligned}$$

This resulted in a preference sensitivity value corresponding to how much a participant differentiated between faces based on each value type in their ratings. A positive value corresponds to greater sensitive to social values than reward values. Across participants, preference sensitivity was positive on average ($M = 2.58$, $SEM = 0.5$). Comparing preference sensitivity with Day 4 learning sensitivity, there was a marginally significant positive correlation ($r_s(28) = .35$, $p = .055$; Fig. 2.5, right).

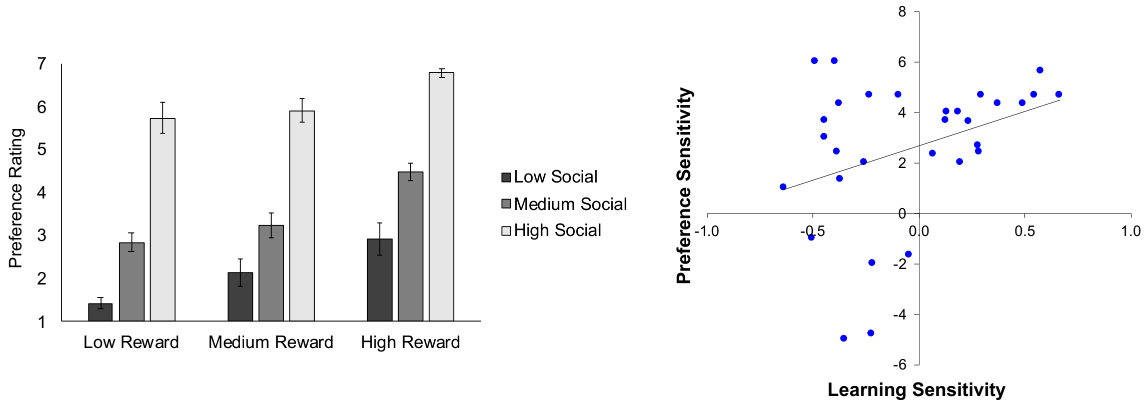


Figure 2.5

Social preference rating results. Across participants, preference ratings were highest for high value faces and lowest for low value faces (left; error bars indicate +/- SEM), there was a significant effect of each value type on ratings. Individual sensitivity to social and reward values in the preference ratings was positively related to accuracy for each value type on day 4 of learning (right; each point represents one subject, black line indicates linear trend).

3. Experiment 2

In order to validate the results from Experiment 1, a separate group of participants completed the same study in order to test reproducibility. All aspects of this experiment were the same with the following exceptions: (1) prior to completing any of the behavioral tasks Experiment 1 subjects also participated in a neuroimaging (fMRI) portion of the study in which they viewed the face images and performed a repeat detection task; and (2) Experiment 1 subjects were paid for their participation and were given a cash bonus based on their performance during value learning, whereas Experiment 2 subjects received no monetary compensation for their participation or task performance.

3.1 Methods

3.1.1 Participants

Twenty participants (three males, ages 18-22) were recruited from the University of Pennsylvania. All subjects gave written informed consent within a protocol approved by the IRB at the University of Pennsylvania and received course credit for their time.

3.1.2 Stimuli & Procedure

All stimuli and procedures were the same as in Experiment 1.

3.2 Results

3.2.1 Value learning performance

Just as in Experiment 1, a correct response on a given trial was choosing the player with the higher assigned value, when the other value types were equal. Accuracy (percent correct) was computed across trials within a session. If no response was made during a trial, that trial was not included in the analysis (2% excluded trials, across participants and across sessions).

A repeated measures ANOVA of accuracy (value type x day) showed a significant effect of day (*F*(3, 57) = 10.45, *p* < .001) and interaction between day and value type (*F*(3, 57) = 7.75, *p* < .001), with no effect of value type (*F*(1, 19) = 0.15, *p* = .706). Paired t-tests indicated a significant increase in reward accuracy between Day 1 (*M* = .59, *SEM* = .03) and Day 4 (*M* = .77, *SEM* = .05; *t*(19) = 4.76, *p* < .001; Fig. 2.6, left), but no change in social accuracy between these days (Day 1: *M* = .74, *SEM* = .04; Day 4: *M* = .68, *SEM* = .04; *t*(19) = -1.49, *p* = .152). Additionally, social accuracy was significantly higher than reward accuracy on Day 1 (*t*(19) = 3.4, *p* = .0013, but there was no difference between value types on Day 4 (*t*(19) = -1.24, *p* = .229). Together, these results replicate Experiment 1.

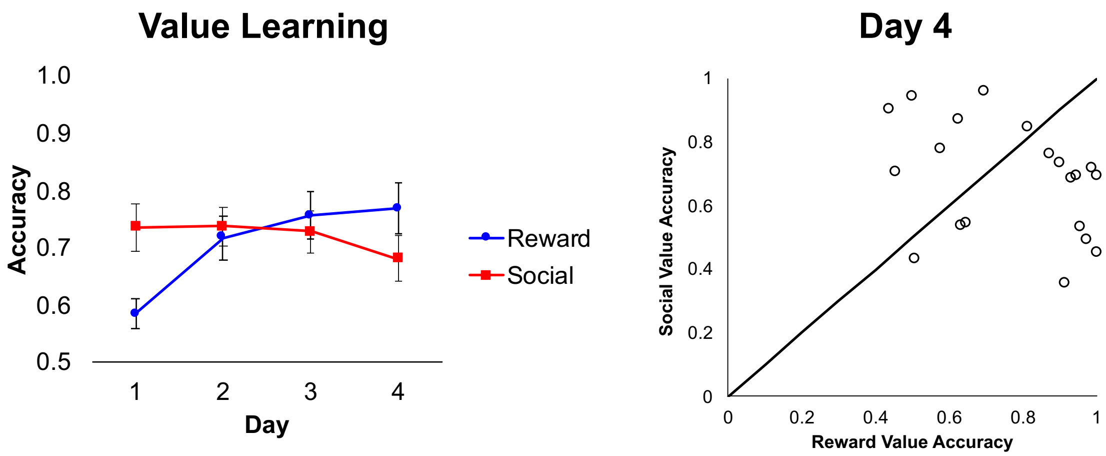


Figure 2.6

Value learning performance. Average accuracy for social (red) and reward (blue) value on each day of learning (left; error bars indicate +/- SEM). Individual learning accuracy on day 4 for each value type (right; each point corresponds to one subject).

3.2.2 Behavioral similarity

Subjective and objective social DMs were positively correlated ($r_s(34) = .43$, $p = .009$). Unlike in Experiment 1, the subjective reward DM did not correlate with objective reward ($r_s(34) = -.15$, $p = .39$), but like Experiment 1 it correlated with the subjective ($r(34) = .35$, $p = .008$) and objective ($r_s(34) = .45$, $p = .006$) social DMs.

The post-learning behavioral DM was positively correlated with the social value DMs (subjective: partial $r(34) = .65$, $p < .001$, controlling for reward; objective: $r_s(34) = .45$, $p = .006$; Fig. 2.7, right) but not the reward value DMs (subjective: partial $r(34) = -.17$, $p = .341$; controlling for social; objective: $r_s(34) = .15$, $p = .391$), while the pre-learning DM did not correlate with DMs for either value type (subjective social: $r(34) = .19$, $p = .281$; objective social: $r_s(34) = .026$, $p = .878$; subjective reward: $r(34) = .02$, $p = .887$; objective reward: $r_s(34) = -.16$, $p = .355$), and the difference in correlations between post- and pre-learning with subjective social similarity was significant ($z = 2.62$).

The perceptual DM was only correlated with the pre-learning behavioral DM ($r(34) = .59$, $p < .001$; Fig. 2.7, left) and no other DM (Table 2.S1). The relationship between individual behavioral sensitivity and day 4 learning sensitivity was not significant (*r*(18) = -.25, *p* = .285).

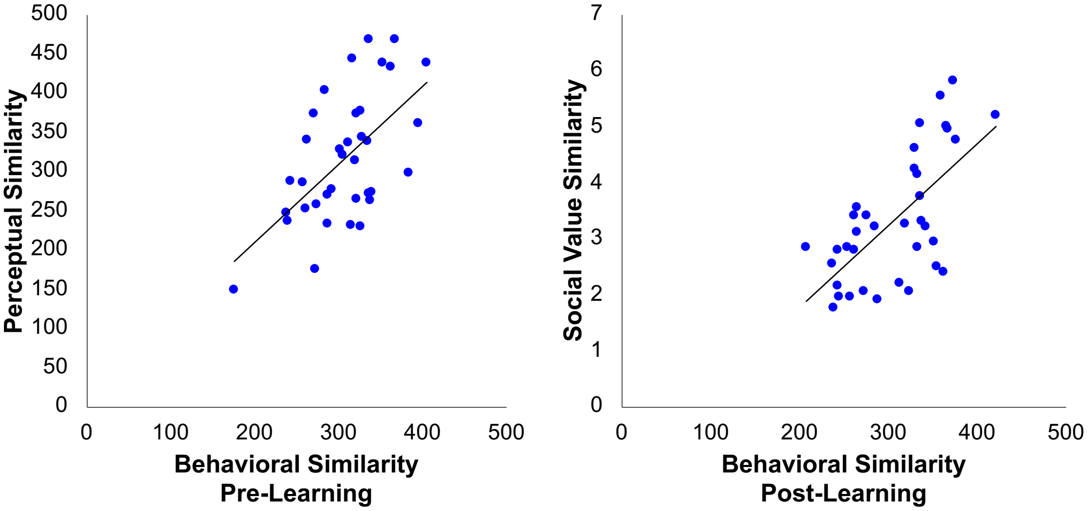


Figure 2.7

Modulation of behavioral face similarity (free sorting) by learned social values. Before value learning, behavioral similarity was related to the perceptual similarity of the faces (left), while after learning similarity was related to social (right), but not reward (not shown), similarity. Each point represents one face pair, black line indicates linear trend.

3.2.3 Social preferences

Social preference ratings also followed a similar trend as in Experiment 1, with high social value faces being the most preferred and low social value faces the least (Fig. 2.8, left; Table 2.S2). A repeated-measures ANOVA of preference ratings (value type x value level) showed significant effects of social (*F*(1.39, 26.37) = 25.53, *p* < .001, Greenhouse-Giesser corrected) and reward (*F*(1.36, 25.77) = 13.28, *p* < .001, Greenhouse-Giesser corrected) values, but no interaction (*F*(4, 76) = .51, *p* = .729), replicating the results found in Experiment 1.

The social preference DM was correlated with the social (subjective: partial *r*(33) = .71, *p* < .001; objective: $r_s(34) = .54$, *p* = .001), but not reward (subjective: partial *r*(33) = -.22, *p* = .204; objective $r_s(34) = -.22$, *p* = .203), DMs.

Like in Experiment 1, the individual differences analysis comparing learning sensitivity on day 4 with preference sensitivity showed a positive correlation that was marginally significant (*r*(18) = .37, *p* = .106; Fig. 2.8, right).

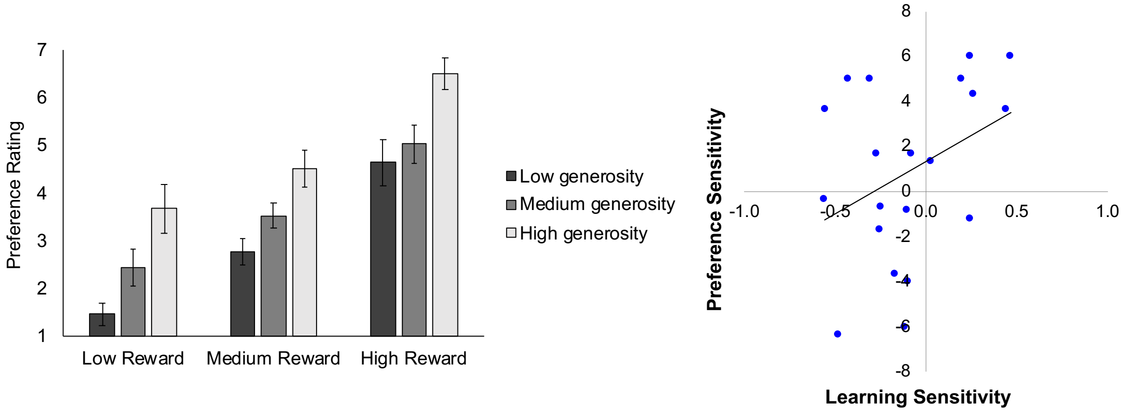


Figure 2.8

Social preference results. Group-level results as a function of value type (left; error bars represent +/- SEM), and individual differences in sensitivity to each value type in the preference ratings and learning accuracy on day 4 (each point represents one subject; black line indicates linear trend).

4. General Discussion

In this study, we examined how learned values (social traits and reward points) influence similarity space representations of facial identities. We were particularly interested in testing whether perceptually organized spaces could be re-organized by learning associations between visual and non-visual information. While previous studies had established that perceptually defined face spaces are related to initial trait impressions of unfamiliar faces, it remained to be examined whether learning trait information associated with faces could re-organize perceptual spaces.

We found that before learning, the organization of similarity spaces was determined by the perceptual similarity of the faces, such that faces that were more perceptually similar were closer together. After learning, perceptual similarity no longer determined the distances between faces, and instead social values, but not reward values, influenced the structure of face space. Notably the task instructions were to arrange the faces such that their spatial organization reflected their similarity, and subjects were not explicitly told to use any specific piece of information to make their judgments. These results replicated across two separate groups of subjects, and show how learned social information about a person can modulate representations of their facial identity.

That social, and not reward values, influenced behavioral similarity result was surprising given the design of the value learning task. In order to maximize accrued points over the course of learning, which in Experiment 1 amounted to a cash bonus at the end of the experiment, it was in the participants' interest to focus on the reward values. In other words, by choosing faces based on the average points they were donating to the participant, they could choose the faces that would give them more points on average. Instead, by the end of learning about half of the participants were better at choosing faces based on their social values compared to reward values. This was not likely due to an inability to learn the reward values, as group-level reward value accuracy increased over the four days, was significantly above chance, and participants were able to correctly rank the faces based on their reward values on the last day of learning in Experiment 1 (although this was not found in Experiment 2). Additionally, participants could have used either or both value types to judge similarity, but instead they used social value alone. That these learned social traits were selectively integrated into face spaces suggests that the social value information was more behaviorally relevant to the participants' judgements of similarity. Put alternatively, when determining what properties to use to discern between the faces, participants prioritized social over reward information. It mattered less to them that a person gave them more points than another, but more so whether they were more or less generous with their point pools, emphasizing the salience of prosocial behavior even when there is monetary incentive to ignore it.

Participant's perceived social similarity was also related to prospective social behavior. Specifically, learned social values were related to preferences of interacting with each person in the future. Faces with higher social values (more generous people) were more preferred than those with lower values. This suggests that the traits that subjects learned through the value learning task generalized to a separate context, namely their expectations of future social behavior based on this acquired knowledge.

The analysis of individual differences indicated one trend that replicated across the two experiments, but only reached marginal significance in each. Namely, there was a correlation between sensitivity to value type during learning and in the social preference ratings. Subjects that were more accurate at choosing faces based on their social values (when reward values were equal) on the last day of learning compared to choosing faces based on their reward values (when social values were equal) tended to be more sensitive to social values than rewards values in their future preference ratings, such that there was a larger differential between ratings for high and low social value faces compared to that between high and low reward value faces. Additionally, subjects that were more accurate at reward value choices were more sensitive to reward value information in the preference ratings, and subjects that were about equally accurate were equally sensitive to each value type.

The main implication of these findings is that face spaces initially organized by perceptual features can be warped by learned abstract information. For familiar others, information beyond that which is readily available in the sensory input, namely that acquired through former experience, can influence the representational structure of person-specific knowledge. Moreover, people use the learned social traits to inform their expectations of an individual's social behavior. Our study cannot determine whether the visual face and abstract social information becomes integrated into a single multi-dimensional space, as perceptual similarity was orthogonal to value similarity by design. That said, the integration of social value and facial identity information would allow for efficient recognition of people and associated traits, which could be used to guide further social interactions and decisions.

5. Supplemental Material

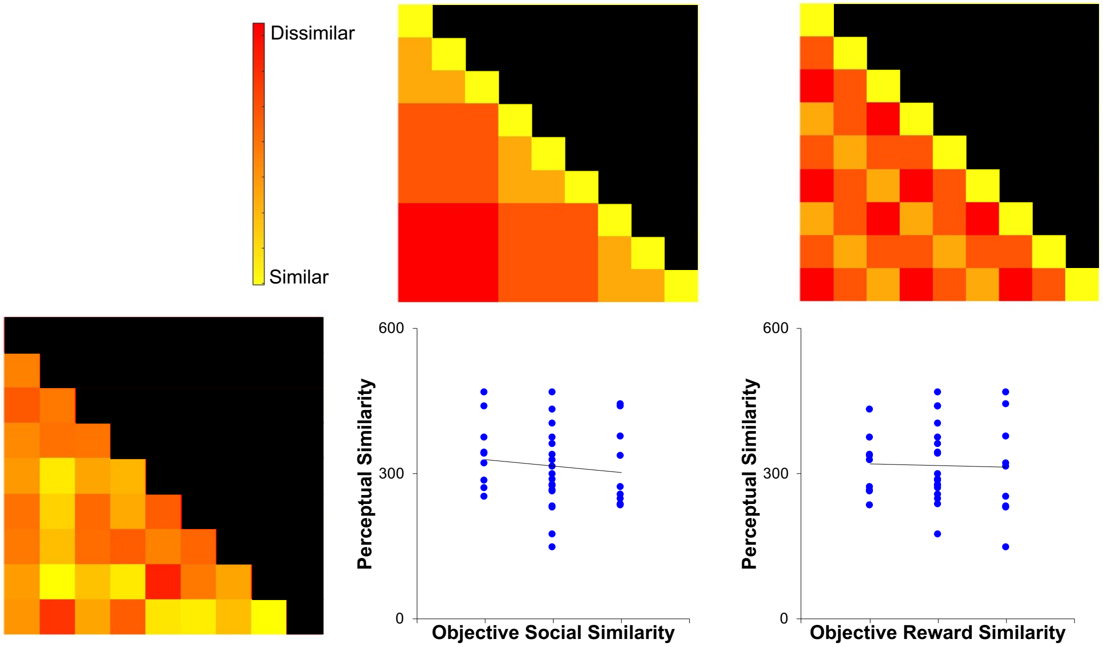


Figure 2.S1

Correlation between perceptual and objective value DMs. By design, perceptual similarity was orthogonal to the experimentally-defined (objective) values associated with each face. Each point represents one face pair, black line indicates linear trend.

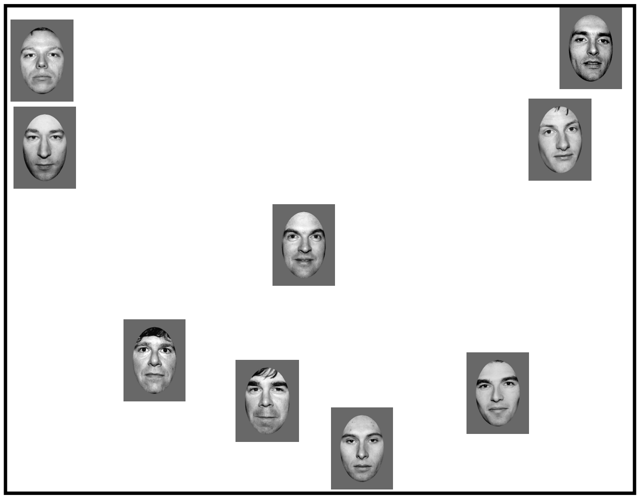


Figure 2.S2

MDS solution of pairwise similarity based on perceptual similarity ratings (*N* = 20).

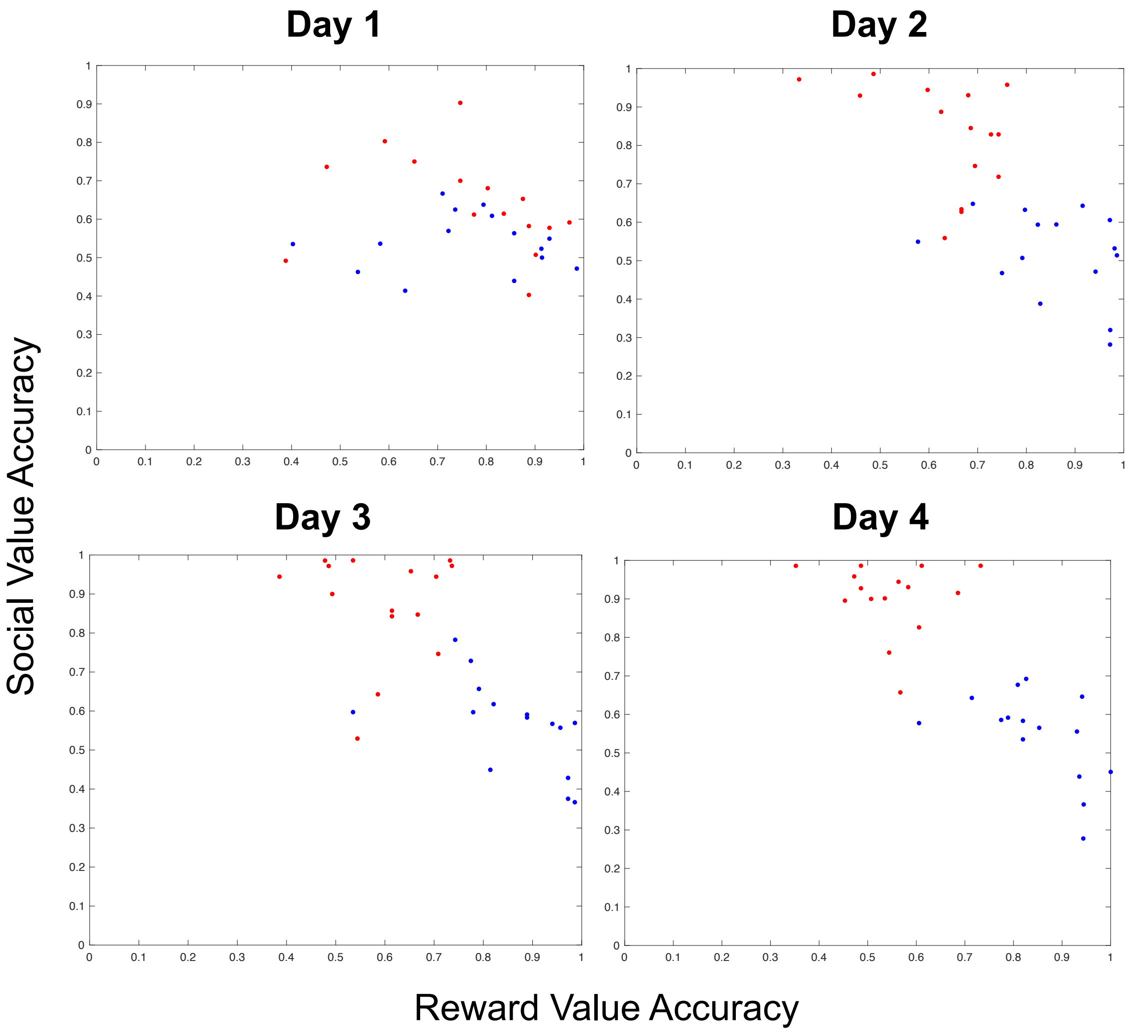


Figure 2.S3

Individual learning accuracy for reward (x-axis) and social (y-axis) value learning on each day of learning (each point represents one subject). Colors in all plots indicate k-means cluster grouping on Day 4, and is intended to visually depict the general consistency across days 2 – 4 for subjects to have higher accuracy for a given value type compared to the other.

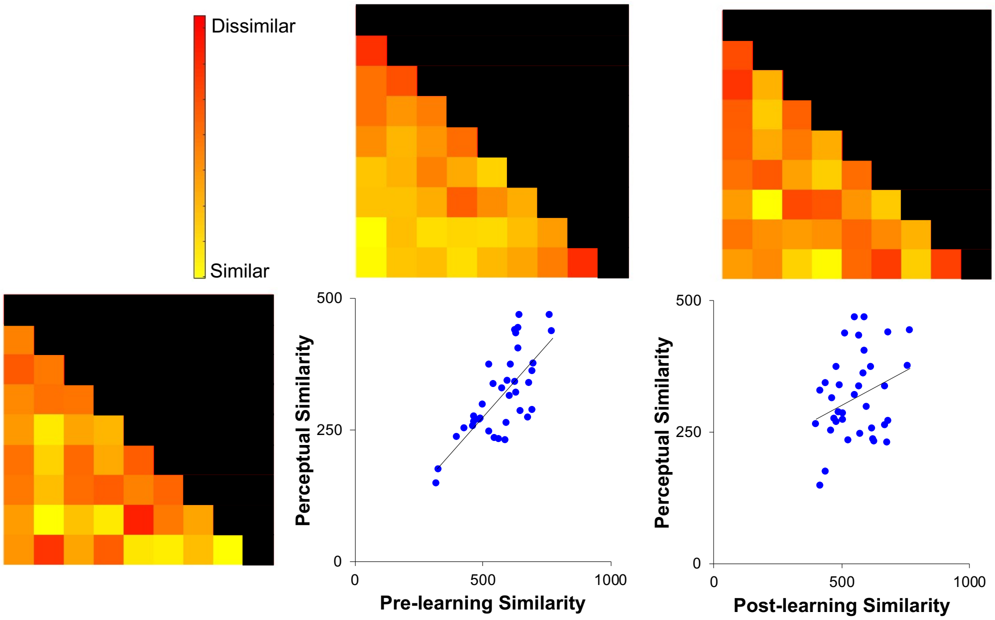


Figure 2.S4

Relationship between perceptual, pre-learning behavioral, and post-learning behavioral similarity judgments across participants. Pre-learning similarity (center) was related to perceptual similarity, but post-learning (right) was not. Each point represents one face pair.

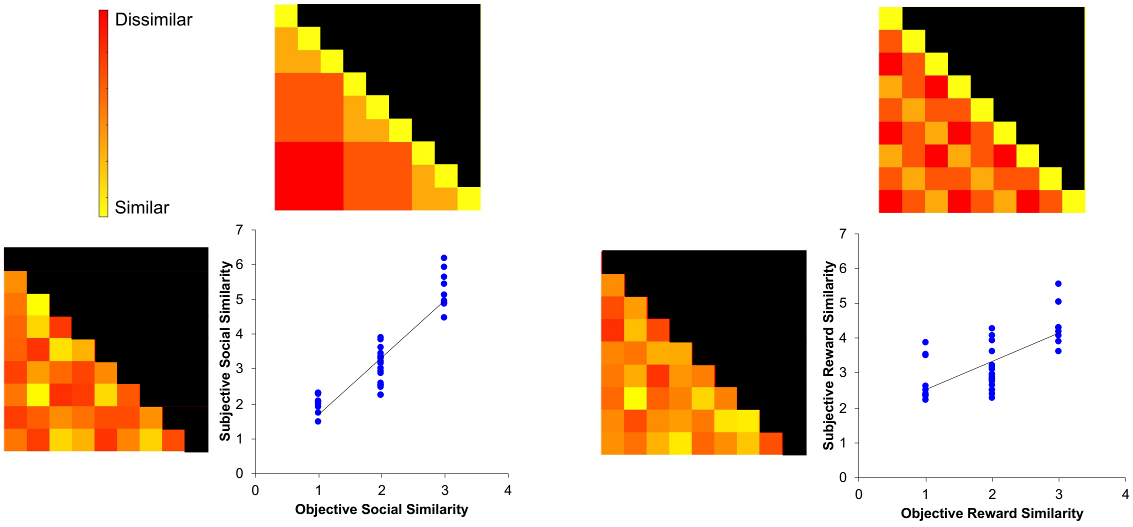


Figure 2.S5

Relationship between subjective value rankings and objective values across participants (each point represents one face pair).

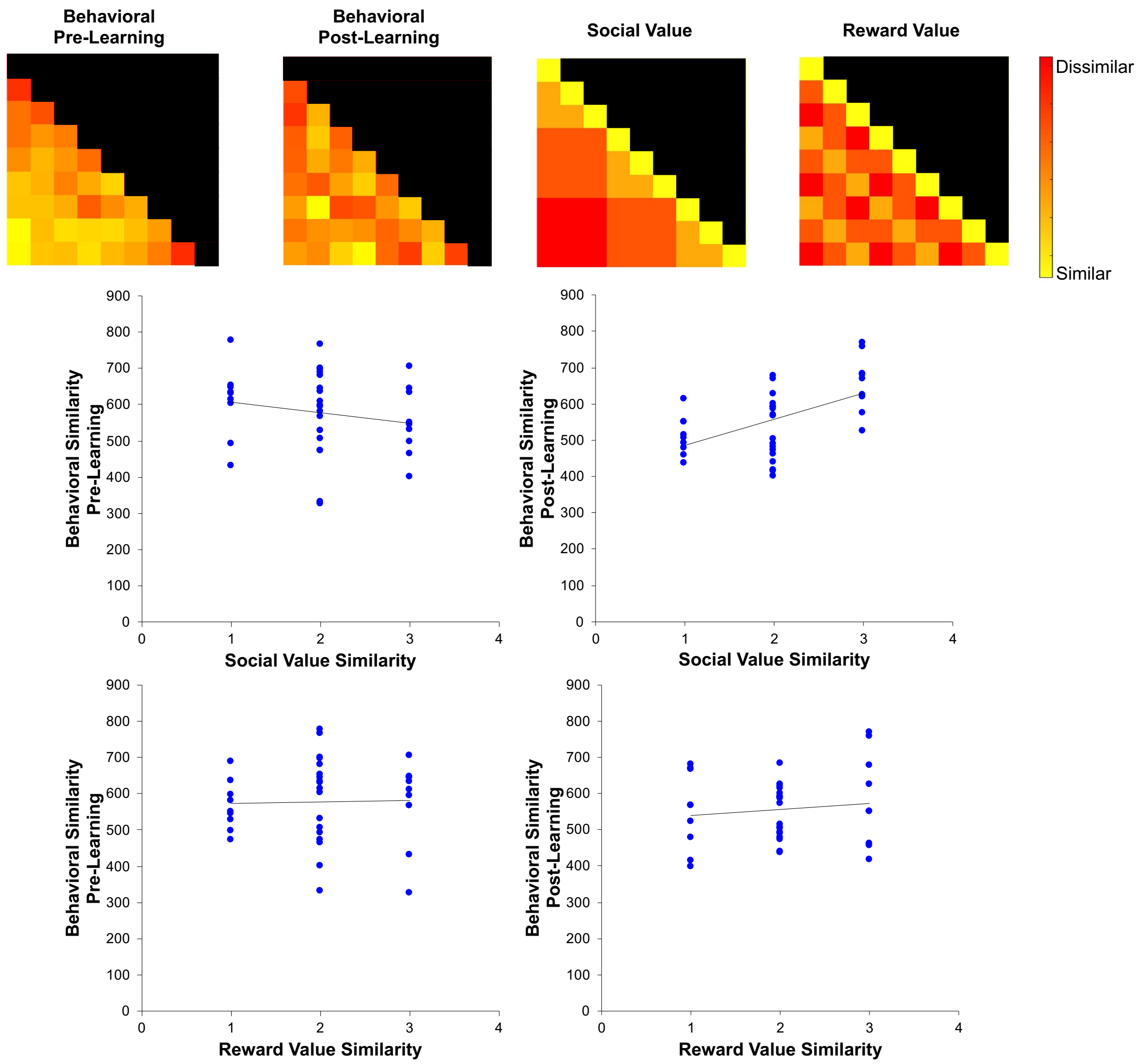


Figure 2.S6

Relationship between objective value DMs and behavioral similarity DMs. Post-learning behavioral similarity was related to object social values, but not reward values. Pre-learning behavioral similarity was not related to either value type. Each point represents one face pair, linear trend (black line) shown for visualization purposes.

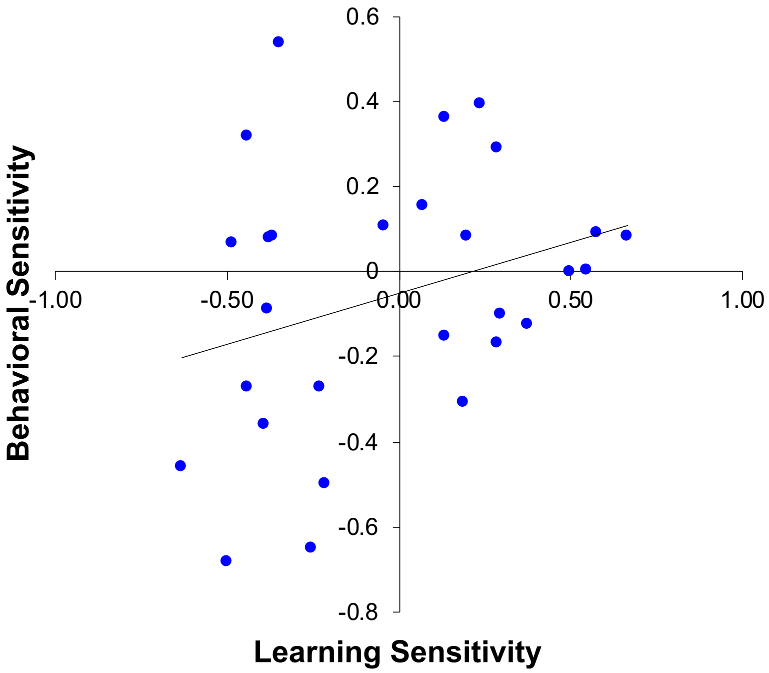


Figure 2.S7

Individual differences in value learning (on Day 4) and changes in behavioral similarity judgements (free sorting task) as a function of value type in Experiment 1. Positive values correspond to a tendency towards social value information compared to reward value information. The positive correlation was marginally significant, but this trend did not replicate in Experiment 2. Each point represents one subject, black line corresponds to linear trend.

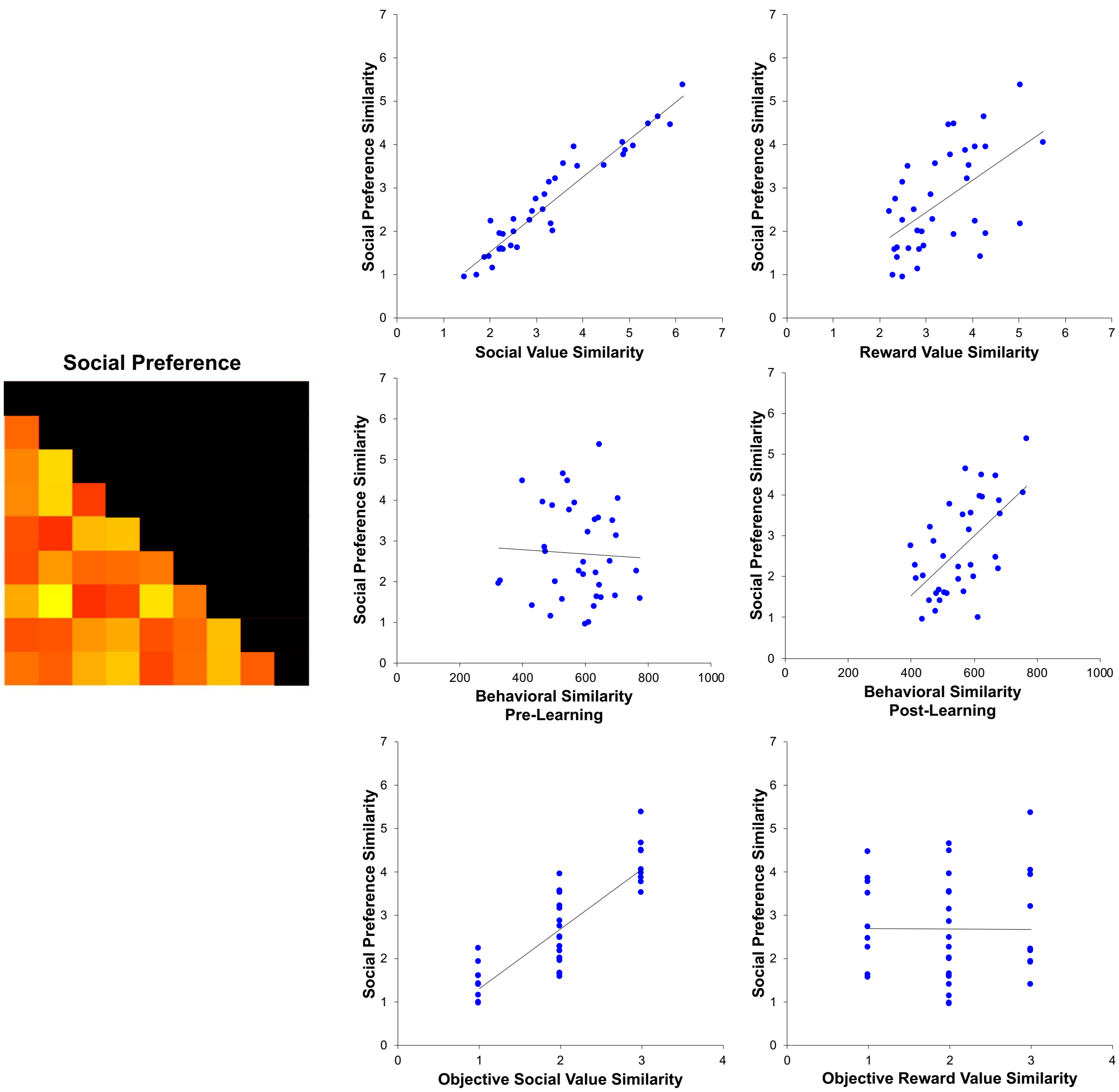


Figure 2.S8

Social preference similarity results (*N* = 30). Group-level similarity based on social preference ratings was related to subjective social value (upper left plot), objective social values (lower left), and post-learning behavioral similarity (middle right). There was no relationship between social preference similarity and subjective reward value (upper right), objective reward value (lower right), or pre-learning behavioral similarity (middle left). Each point represents one face pair, black line indicates linear trend.

Similarity Measure	***r*(34)**	***p***
Post-learning behavioral	.03	.853
Subjective social	-.03	.85
Subjective reward	-.03	.887
Social preferences	.03	.865

Table 2.S1

Correlation between perceptual similarity and other measures.

Value		***M* (*SEM*)**
Social	*High*	5.39 (0.45)
	Medium	3.6 (0.39)
	Low	2.53 (0.4)
Reward	*High*	4.89 (0.48)
	Medium	3.67 (0.39)
	Low	2.96 (0.49)

Table 2.S2

Social preference group-level results. Average ratings for each value type and level shown, note that results for one value type are averaged over the other (e.g., each social value is averaged over low, medium, and high reward values).

III. NEURAL REPRESENTATIONS OF FACE-VALUE ASSOCIATIONS

1. Introduction

In Chapter 2 it was found that value learning modulated behavioral similarity face spaces, and social values in particular determined how spaces were re-organized after learning. Although central to social behaviors, how such abstract social information is associated with individual visual face identities in the brain is not well-understood. What is the neural organization of abstract and perceptual person-specific knowledge?

Previous neuroimaging research on social cognition has focused on general aspects of person information, such as localizing regions involved in processing social conceptual knowledge (Zahn et al., 2007, 2009, 2017) or those involved when thinking about others compared to thinking about the self (Amodio & Frith, 2006; Bzdok et al., 2012; Schurz et al., 2014, 2021). For example, studies using functional magnetic resonance imaging (fMRI) have examined brain activity during judgments of person-related properties (e.g. “brave”, “assertive”, “tactless”), and compared resulting regions with those evoked during object- or animal-related property judgments (Mitchell, Heatherton, & Macrae, 2002; Zahn et al. 2007). Together, this work has pointed to a network of brain regions that underlie learning and memory of such social information, including the anterior temporal lobe (ATL), prefrontal cortex, orbitofrontal cortex, and temporoparietal junction (Saxe, 2006; Frith, 2007). Converging evidence from studies on patients with frontotemporal dementia (FTD) has indicated a dissociation between general knowledge of social concepts and societal norms in the lateral, superior ATL and motivation to act in accordance with those norms and understanding the consequences of inappropriate social conduct in the orbitofrontal and prefrontal cortices (Zahn et al., 2009, 2017).

On the other hand, research on visual face processing has established a network of regions in the occipital, temporal, and frontal cortices that contain neuronal populations selective for face information (Grill-Spector, Weiner, Kay, & Gomez, 2017; Rapcsak, 2019). Neuroimaging studies have provided evidence that when viewing faces compared to other visual objects, including face-like non-face stimuli, there are higher neural responses in these regions (Kanwisher et al., 1997; Haxby et al. 1999, Gauthier et al. 2000; Hoffman & Haxby 2000; Weiner & Grill-Spector, 2010). Moreover, damage to these regions can impair face detection and recognition, such as prosopagnosia (Barton, Press, Keenan, & O'Connor, 2002). While ventral areas such as the occipital face area (OFA) and fusiform face area (FFA) have been implicated in the detection and recognition of faces, dorsal regions such as the posterior superior temporal sulcus (pSTS) are thought to process more dynamic face information, such as emotional expressions (Fox et al. 2009; Duchaine & Yovel, 2015). Face identity in particular has been linked to ventral face-selective regions in fMRI studies with adaptation paradigms (Grill-Spector, Weiner, Kay, & Gomez, 2017).

Recently it has been proposed that the association of face identity with other types of person knowledge in memory occurs at late stages of visual processing, particularly in face-selective parts of the medial and ventral ATL (Olson, McCoy, Klobusicky, & Ross, 2013; Collins & Olson, 2014). The medial and ventral ATL is anatomically well situated to link information between earlier face-processing regions and memory for abstract social knowledge, as it has reciprocal connections with more posterior parts of the fusiform gyrus involved in face processing, as well as memory-related regions such as the superior, lateral ATL, temporal pole, orbitofrontal cortex, and entorhinal and perirhinal cortices connected to the hippocampus. Although the human homologue of a ventro-medial ATL face-selective patch found in macaque monkeys has been established (Tsao et al., 2008; Rajimehr et al., 2009; Jonas et al., 2016), it can be difficult to identify using functional neuroimaging due to signal susceptibility artifacts and drop-out due to proximity to the sinuses (Devlin et al., 2000; Jezzard & Clare, 1999).

At the same time, neurophysiological and neuroimaging studies have provided evidence that responses in this area to faces are not solely driven by visual input and are sensitive to familiarity manipulations. Single-unit neuronal recordings in the ventral anterior inferior temporal cortex of macaques have shown view-point invariant representations of face identities (Eifuku et al., 2011) and selectivity for learned associations between faces and abstract patterns (Eifuku et al., 2010). fMRI studies using both face images and proper name presentation have found greater activation for famous and non-famous familiar individuals compared to unfamiliar people in the ATL and medial temporal lobe regardless of the stimulus type (Ross & Olson, 2012; Elfgren et al., 2006). Moreover, multi-voxel pattern methods have revealed successful classification of face identity (Kriegeskorte, Formisano, Sorger, & Goebel, 2007; Nestor, Plaut, & Behrmann, 2011) and of learned associations between biographical information and faces in the ATL (Tsukiura et al., 2010; Wang et al., 2017). Neuropsychological studies on patients with associative prosopagnosia have shown that ATL damage can result in deficits in remembering information about friends and celebrities even though face perception is intact (Rice et al., 2018), and right ATL lesions in particular have been associated with impaired knowledge of biographical information as well as feelings of familiarity when re-meeting others (Gainotti, 2007a,b; Borghesani et al., 2019). Notably, ATL damage impairs face memory and identification, not face perception, as well as the ability to form new word-picture associations (Sharon et al., 2011; Gainotti & Marra, 2011; Olson et al., 2013). Taken together, face-selective parts of the ATL may be responsible for integrating person-specific visual and social information between perception and memory.

That said, fMRI studies examining social category perception of faces have found some evidence that social impressions linked to perceptual features can bias activity in earlier visual regions than the ATL. Specifically, social categories cued by visual face features (e.g. age, race, gender) have been related to neural responses in face-selective regions along the fusiform gyrus (Hughes et al., 2019, Freeman & Johnson, 2016; Reggev et al., 2020). For example, one study found that faces of stereotypically overlapping social categories (e.g., happy/female, angry/male) evoked more similar multi-voxel activity patterns in a region of the right fusiform gyrus (Stolier & Freeman, 2016), suggesting that face representations in this area were influenced by stereotype associations. At a behavioral-level, these stereotypes were related to biases in mouse-trajectories during active social categorization of these faces. Thus, associations of social and visual face information are related to responses in early to mid-stages of the face processing system, at least when the social categories are cued by visual facial features.

In the present study, we examine which neural systems are involved in processing social information tied to specific individuals. Specifically, we had people learn personality traits (generosity) associated with different faces through an interactive task and tested whether there were brain regions whose activity patterns were modulated by learning. As opposed to former studies that have primarily looked at pre-existing person knowledge (e.g., famous faces) and social information derived from visual features (e.g., ingroup/outgroup stereotypes), we examine how learning associations between abstract social values and face identities can dynamically influence neural responses.

To examine changes in the representational structure of different brain regions we utilize multi-voxel activity pattern similarity analyses in conjunction with a whole-brain roaming searchlight method (Kriegeskorte, Goebel, & Bandettini, 2006; Kriegeskorte, Mur, & Bandettini, 2008) as well as independently localized face-selective regions-of-interest (ROIs). We had people learn associations between faces, social values (generosity), and reward values (points) via a multi-day learning task. Before and after learning, fMRI data was collected as participants viewed the faces and performed a repetition detection task that was independent of the learned value information. The similarity of neural activity patterns for each pair of faces was calculated and compared to social, reward, perceptual, and behavioral similarity measures. This allowed us to locate regions whose activity patterns tracked the learned values after learning, even during a value-unrelated perceptual task, as well as those that were related to the behavioral and perceptual similarity judgments. Additionally, we examined individual differences in learning behavior, social preference ratings, and neural changes in resulting regions.

The results of this study elucidate the neural basis of person-specific social traits when face-value associations are learned. Crucially, the experimental design involved characterizing face spaces before and after learning, which allowed us to measure the properties by which they were warped and the neural structures whose population responses were related to this re-organization.

2. Methods

All subjects, stimuli, and behavioral procedures were those described in Chapter 2, Section 2 (Experiment 1).

2.1 Scanning Procedures

Participants completed one neuroimaging (fMRI) scan session the day before the first day of learning (pre-learning) and one scan session the day after the last day of learning (post-learning). Each scan session consisted of five main functional runs and were equivalent with the exception of three additional localizer runs collected in the pre-learning session (except for two subjects who completed them in the post-learning session). The session including the functional localizers was about 1.5 hours long, while the other session lasted about 1 hour. During scanning, stimuli were presented using a Dell Latitude laptop and custom Matlab scripts with Psychtoolbox (Brainard, 1997; Pelli, 1997).

2.1.1 Main experiment. Across the five main runs subjects performed 1,000 trials of a one-back repeat detection task (200 trials per run). On each trial, participants viewed one face at the center of the screen with a gray background and pressed two buttons (one with each thumb) simultaneously if the face was identical to the one that immediately preceded it (10% of trials). The face image was presented for 1.5 seconds, followed by a 0.5 sec blank screen. On 90 trials, a blank gray screen was shown instead of a face; baseline activity was modeled using these trials. The onset of each trial was time-locked with each TR (2 sec). The order of stimuli was determined by a de Bruijn sequence in which each image follows every image at the same frequency, and each subject viewed the same pseudo-randomized order before and after learning. 90 trials were used to model activity patterns for each face within a scan session.

Crucially, this task was unrelated to the learned values. In order to perform the task, people simply needed to recognize each face and compare it to the one on the former trial. Thus, any neural signature related to the learned values would be activated when viewing and comparing the faces, due to the associations formed during the learning task.

2.1.2 Independent face localizer. To functionally localize face-selective regions in the visual system we utilized the fLoc functional localizer package (Stigliani, Weiner, & Grill-Spector, 2015). This included three runs in which participants viewed blocks of 8 gray scale images within a given category (faces, places, objects, or scrambled), and performed a one-back repeat detection task (10 second blocks, 0.75 sec image presentation & 0.25 sec ISI; 146 TRs per run with 2 sec TRs). Face images included both adult and child portraits, in various head viewpoints and screen locations within a bounded square in the center of the screen with a phase-scrambled background; the remaining screen area was a uniform gray background.

2.2 MRI Scan Acquisition

Subjects were scanned on a 3T Siemens Prisma MRI scanner with a 32-channel head coil. Anatomical scans were obtained using a high-resolution T1-weighted MPRAGE sequence (voxel size = 1.0 x 1.0 x 1.0 mm). Functional runs were obtained with a T2-weighted gradient-echo echo-planar imaging (EPI) multi-band sequence (TR = 2 s, TE = 30 ms, FOV = 19.8 x 19.8 cm, voxel size = 2.0 x 2.0 x 2.0 mm, flip angle = 75°). B0 fieldmap images were constructed using magnitude (TR = 0.6 s, TE = 4.1 ms, FOV = 2.4 x 2.4 cm, voxel size = 3.0 x 3.0 x 3.0 mm, flip angle = 45°) and phase (TR = 0.6 s, TE = 6.6 ms, FOV = 2.4 x 2.4 cm, voxel size = 3.0 x 3.0 x 3.0 mm, flip angle = 45°) images collected at the end of the session.

2.3 Data Analysis

2.3.1 Pre-processing. Functional scan data was analyzed using FSL (Smith et al., 2004) and in-house MATLAB scripts. Data were preprocessed to correct for slice-timing and head motion (motion-censored TRs across subjects: *M* = 3.7%, *SEM* = 0.35%), high-pass filtered to reduce low frequency noise, distortion corrected with the B0 fieldmap image (except for one subject who did not have B0 scans), and co-registered to an anatomical scan collected at the beginning of the scan session. To normalize the data, each volume was scaled by a constant factor such that the mean of each voxel's time series was equal.

2.3.2 First-level analysis. After pre-processing, data from the five main runs were each analyzed with a general linear model (GLM) that included one regressor for each face condition, one for repeat trials (trials in which the face matched the one that preceded it), and motion parameters as covariates of no-interest, in order to estimate the activity patterns (beta weights) for each face. Each participant's data was modeled individually. Pattern similarity analyses were conducted on unsmoothed data to retain the highest level of spatial resolution available.

2.3.3 Localizer runs. For the localizer data, the same pre-processing steps as the main experimental runs were used, with the addition of spatial smoothing (5mm Gaussian kernel). For each run, a GLM was used to model responses to each stimulus category (adult faces, child faces, car objects, instrument objects, corridor places, house places, scrambled) and blank trials separately (with additional motion parameters), and a contrast was used to find voxels whose responses were significantly higher for faces compared to objects, places, and scrambled images. Results from the three runs were combined within-subject using a fixed-effects model.

2.3.4 Cortical surface reconstruction. Within scan sessions, each subject's cortical surface was reconstructed based on their anatomical scan using Freesurfer (Fischl, 2012). For each subject this yielded an outer surface (pial surface – grey matter boundary) and an inner surface (grey matter – white matter boundary), consisting of vertices and edges. These surfaces were aligned and resampled to a standard topology using AFNI's MapIcosahedron (implemented with PyMVPA; Hanke et al., 2009). This resulted in surfaces for each subject and session with vertices (nodes) that correspond to the same surface locations across sessions and people. Pattern similarity analyses could then be performed on a node-by-node basis. To present group results, intermediate and inflated surfaces were averaged across participants and visualized with AFNI and SUMA (Cox, 1996; Saad et al., 2004).

2.4 Whole-brain searchlight analyses

2.4.1 Voxel selection. All searchlight analyses were conducted using a surface-based approach with the CoSMO-MVPA toolbox (Oosterhof et al. 2016) in MATLAB. Surfaces with 64 linear divisions (40,962 nodes per hemisphere) were used. For each surface node, grey matter voxels surrounding the node, constrained by the outer and inner surfaces, were selected using a geodesic distance metric. This resulted in one searchlight per node of approximately 100 neighboring voxels that followed the cortical anatomy of the given participant. Unlike volume-based searchlights, surface-based definitions have higher spatial sensitivity due to searchlights being constrained to grey matter voxels that are anatomically adjacent, and as a result have been shown to be more sensitive to measuring information in multi-voxel activity patterns (Oosterhof, Wiestler, Downing, & Diedrichsen, 2011).

2.4.2 Pattern similarity searchlight. We first searched for brain regions that contained information about the social or reward values after, and not before, learning. In order to locate areas whose activity patterns tracked value similarity, a pattern similarity analysis with a whole-brain roaming searchlight (Kriegeskorte et al., 2006, 2008) was performed for each subject, scan session, and value type separately. Activity patterns were averaged across the five runs for each face separately, and mean-centered by subtracting the average activity pattern across face conditions from each face's activity pattern in order to remove a main effect common to all conditions (Diedrichsen & Kriegeskorte, 2017). For each node/searchlight, neural dissimilarity matrices (DMs) were constructed with a neural dissimilarity value for each face pair, computed as the pairwise Euclidean distances of their corresponding activity patterns. Resulting neural DMs were correlated with behavioral DMs derived from the subject's post-learning value rankings (pairwise differences between rank positions) using partial correlations to regress out the other value type. The resulting correlation value was assigned to the given node, and this was procedure performed for every node in the brain. The searchlight results were Fisher normalized and then submitted to a group-level analysis to find clusters of nodes whose positive correlation values post-learning were significantly higher than pre-learning, corrected for multiple comparisons with a permutation approach (see below).

Additional searchlight analyses were run for each value type in which behavioral DMs based on the perceptual similarity judgments (performed by a separate group of participants) were regressed out using a partial correlation method. This allowed us to compare the original searchlight maps with those that controlled for any effect of the perceptual similarity of the faces. In other words, this tested for the unique influence of value information on changes in neural patterns apart from the faces' visual similarity.

The same approach was used to locate regions whose activity pattern similarity was related to the free sorting similarity results. For each subject, their behavioral DMs were compared with their neural similarity within each session separately as well as between pre- and post-learning sessions.

2.4.3 Permutation procedure to assess statistical significance. To locate clusters of significant nodes across subjects and correct for multiple comparisons, we performed a Monte Carlo permutation procedure with Threshold-Free Cluster Enhancement (TFCE; Nichols & Smith, 2009) for each test separately (using CoSMo-MVPA). First, null datasets were generated by randomly shuffling the cells of the given target behavioral DM and then running the same searchlight analyses described above (repeated 100 times for each subject separately). Then, the original statistics from a paired t-test (between sessions) or a one-sample t-test against 0 (within session) were weighted with TFCE (based on cluster extent and effect size) and compared with a null distribution of TFCE values based on 10,000 iterations of randomly selecting from the null datasets and performing the same test, producing a corrected map. All reported statistics for a given analysis reflect this value.

2.5 Region-of-interest analyses

In order to examine neural activity patterns within face-selective regions of the visual system, the independent localizer data was used to define three regions in each hemisphere for each subject separately: the occipital face area (OFA), posterior fusiform face area (FFA), and mid-fusiform face area (FFA-2; Fig. 3.S1). These regions comprise the ventral ‘core’ face network (Grill-Spector, Weiner, Kay, & Gomez 2017). For each subject, the resulting t-statistic map from the contrast of faces to other visual conditions (objects, places, and scrambled images) were projected onto their surface models using AFNI and SUMA, and each region-of-interest (ROI) was drawn by selecting corresponding nodes on the surface using SUMA. The number of nodes in each ROI were approximately equal across subjects (Table 3.S1).

Pattern similarity analyses were run for each subject and ROI separately. Surface-based searchlights were defined for each node within a given ROI as a set of neighboring gray matter voxels surrounding the node, and the same similarity analyses described above were performed. Five searchlight sizes were used (100, 80, 60, 40, or 20 voxels), to address the possibility that any failure to find an effect of value was not due to the spatial resolution of the searchlights. Similarity results were averaged across searchlights within an ROI, and a paired t-test was used to assess differences between pre- and post-learning across subjects.

2.6 Partial correlation

For several searchlight analyses, additional partial correlations were used to control for the effects of co-varying measures (also using CoSMo-MVPA). Each of these utilized the same approaches described above (i.e., Euclidean distance metric to compute the neural DM, and Spearman correlations to assess the relationship between neural and behavioral DMs).

3. Results

3.1 Whole-brain pattern similarity analyses to locate regions with neural activity patterns modulated by learned values

To address the question of whether learned associations between visual face and abstract social information were reflected in modulations of neural responses in the visual system, we used multi-voxel pattern similarity analyses combined with whole-brain, surface-based searchlight methods to analyze fMRI data collected while participants viewed face images and performed a one-back repeat detection task of face identity. fMRI data was collected once before and once after participants completed the value learning task, and the pre- and post-learning scan sessions were identical with the exception of additional localizer scans in one of the two sessions (pre-learning session for all but two subjects). Notably, the main scan task did not require any learned value information to perform, thus any presence of value information in the evoked activity would be independent of an explicit value-based task.

Subjective value similarity, based on post-learning behavioral value rankings, was used to locate regions whose activity patterns were related to social or reward values after, and not before, learning. For each participant, a behavioral dissimilarity matrix (DM) was constructed from their rankings, such that each cell contained a dissimilarity value for a pair of faces (the difference between their rank; Chapter 2). Resulting behavioral DMs were used to predict neural similarity for pairs of faces within pre- and post-learning scan sessions (i.e., the neural DM for a given searchlight was correlated with the participant's behavioral DM), for social and reward values separately. A paired t-test between resulting pre- and post-learning pattern similarity maps (Fisher-transformed, Spearman correlation values for each surface node) with permutation-based multiple comparisons correction (threshold-free cluster enhancement, TFCE) was used to assess which regions at a group-level were modulated by the learned value information.

Additional searchlight similarity analyses were performed to compare neural activity patterns with behavioral similarity (free sorting task) and perceptual similarity, within and between scan sessions. Within session comparisons involved a t-test against 0 with TFCE.

3.1.1 Social Value

We first examined regions in which the similarity of activity patterns for pairs of faces was influenced by their social value similarity. Across subjects, clusters within the left ventro-medial ATL contained multi-voxel patterns that had a significant, positive increase in their correlation with subjective social values as indicated by a paired t-test (Peak node: post > pre-learning *M* = .12, *SEM* = .05; Table 3.1; Fig. 3.1, left; Fig. 3.S2).

In this area, face pairs with more similar social values evoked more similar activity patterns after learning. This was the only region in the visual system that incorporated the learned social information, and extended between the anterior fusiform gyrus and the anterior collateral sulcus, consistent with the location of the anterior temporal face patch as found by previous functional neuroimaging studies (Weiner et al., 2010; Tsao, Moeller, & Freiwald, 2008; Goesaert & Op de Beeck, 2013; Nasr & Tootell, 2012; Pinsk et al., 2009; Von Der Heide, Skipper, & Olson 2013; for review, see Collins & Olson, 2014), as well as the medial ATL. There was no relationship between reward value similarity and neural activity patterns at the peak node in this region after learning (post > pre-learning *M* = .01, *SEM* = .05). Additionally, no significant regions were found in a whole-brain searchlight using the objective social value DM.

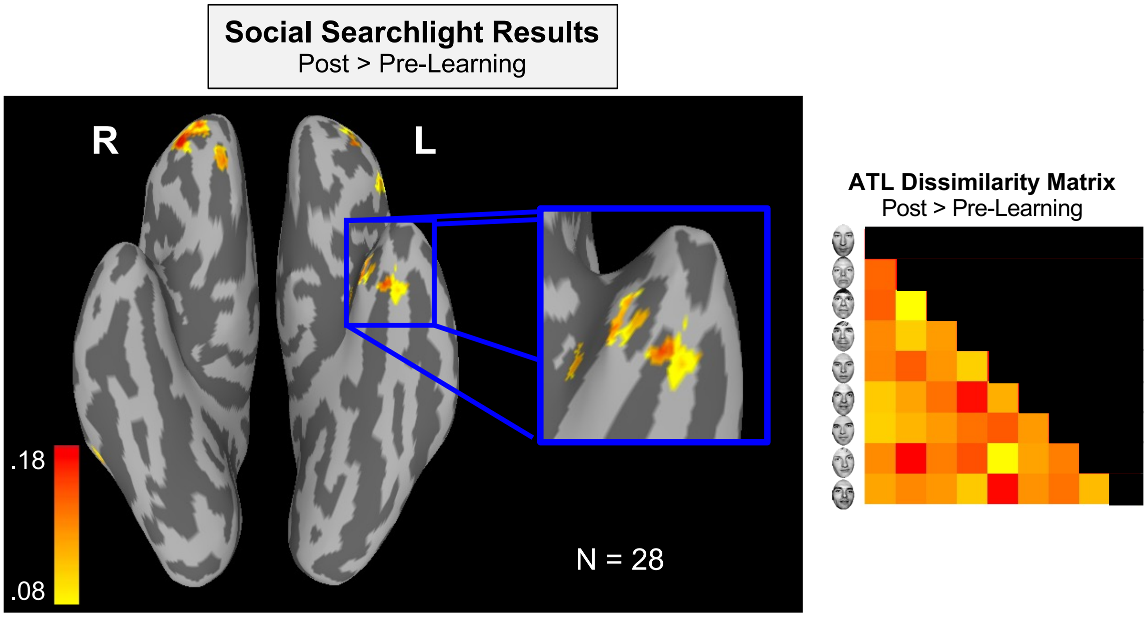


Figure 3.2

Social similarity searchlight results. Left: A whole-brain searchlight analysis found that a region in the left anterior temporal lobe contained activity patterns whose similarity after learning was related to social value similarity, compared to before learning. Right: the group-level dissimilarity matrix is based on activity patterns across nodes in the region.

Location	*Surface area (mm)*	*Post-Pre M (SEM)*	*Post-Pre Max*	*p-value M*
Anterior fusiform gyrus + collateral sulcus	97	.12 (.05)	.15	.023
Medial temporal pole	59	.12 (.05)	.15	.036

Table 3.2

Social similarity cluster results in the left ATL (*N* = 28). Aggregate results calculated across searchlights.

To test for perceptual face information in these regions, a whole-brain searchlight analysis with the perceptual DM was performed (see Section 3.1.4), and results for the peak searchlight node in the left ATL region was examined. Activity pattern similarity was predicted by perceptual similarity in the post-learning (*M* = .05, *SEM* = .04), but not pre-learning (*M* = -.01, *SEM* = .03) session; however the difference between sessions was not significant (*t*(29) = 1.24, *p* = .226).

3.1.2 Reward Value

Whole-brain searchlight similarity analyses based on the post-learning reward value rankings showed one significant region that tracked reward value similarity after versus before learning, located in the left posterior, inferior parietal cortex (Table 3.2; Fig. 3.2). This suggests that the learned social values selectively modulated the activity patterns in the left ATL. An additional whole-brain analysis using the objective value DM showed only one region in the parietal cortex (Table 3.S2; Figure 3.S3).

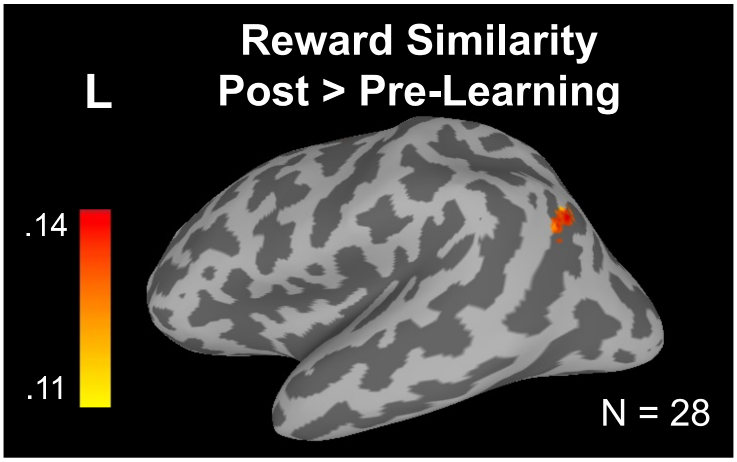


Figure 3.3

Reward similarity searchlight results.

Peak location	*Surface area (mm)*	*Post-Pre M (SEM)*	*Post-Pre Max*	*p-value M*
L. Inferior Parietal Cortex (posterior)	127	.12 (.001)	.14	.029

Table 3.3

Reward similarity cluster results (*N* = 28). Aggregate results calculated across searchlights in the left IPC region.

3.1.3 Behavioral similarity

In order to relate each individual's behavioral similarity judgments with neural activity pattern similarity, we ran additional whole-brain searchlight analyses using the behavioral DMs based on the free sorting task. Within pre- and post-learning sessions, the behavioral DM based on the pixel-wise distance between pairs of faces was correlated with the pairwise neural activity pattern DM. Resulting similarity maps were compared between sessions.

A large cluster in the left inferior parietal cortex was found to have a significantly greater relationship between behavioral and neural similarity in the post-learning, compared to pre-learning session (Fig. 3.3). This extended between the angular gyrus (AG) and the supramarginal gyrus (SMG). Additional clusters were found in the right, posterior middle temporal gyrus, right insular cortex, and right, inferior precentral gyrus.

No significant results were found comparing pre > post-learning (Fig. 3.S4), or within pre- and post-learning sessions.

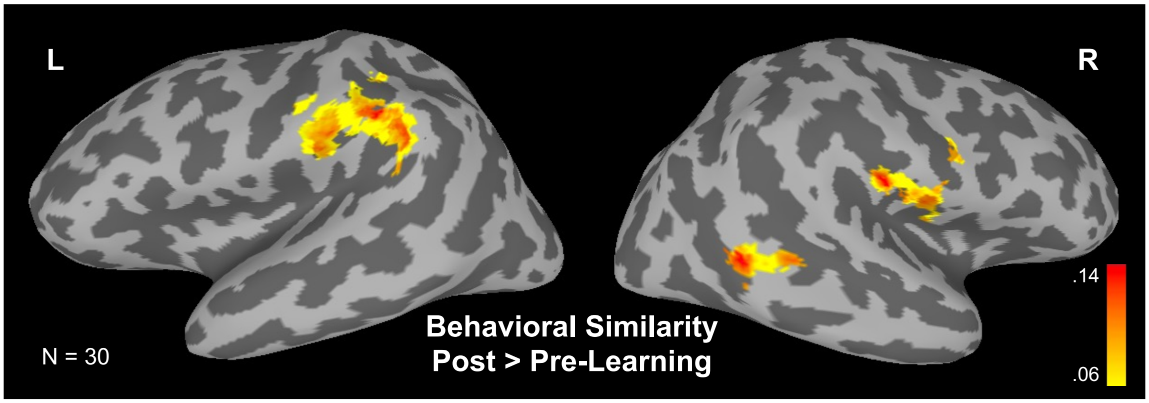


Figure 3.4

Behavioral similarity searchlight results.

Peak location	*Surface area (mm)*	*Post-Pre M (SEM)*	*Post-Pre Max*	*p-value M*
L. Inferior Parietal Cortex	851	.09 (.001)	.14	.022
R. Middle Temporal Gyrus (posterior)	247	.1 (.002)	.14	.02
R. Insula	396	.1 (.001)	.14	.022
R. Precentral Gyrus	54	.1 (.001)	.11	.038

Table 3.4

Behavioral similarity cluster results ($N = 30$). Aggregate results based on all searchlight nodes in the given cluster.

3.1.4 Perceptual similarity

The final whole-brain searchlight similarity analysis examined regions whose activity patterns were related to the perceptual DM (constructed using results from a separate group of subjects who performed the free sorting task explicitly using the physical similarity of the faces). Within pre- and post-learning scan sessions there were several clusters along the ventral visual pathway with positive correlations; however, no region reached significance (Fig. 3.S5).

Comparing post- and pre-learning sessions, there were three regions in the right hemisphere whose activity patterns were positively related to perceptual similarity after learning compared to before (Table 3.4; Fig. 3.4). No region had significantly higher correlations in the pre- > post-learning comparison.

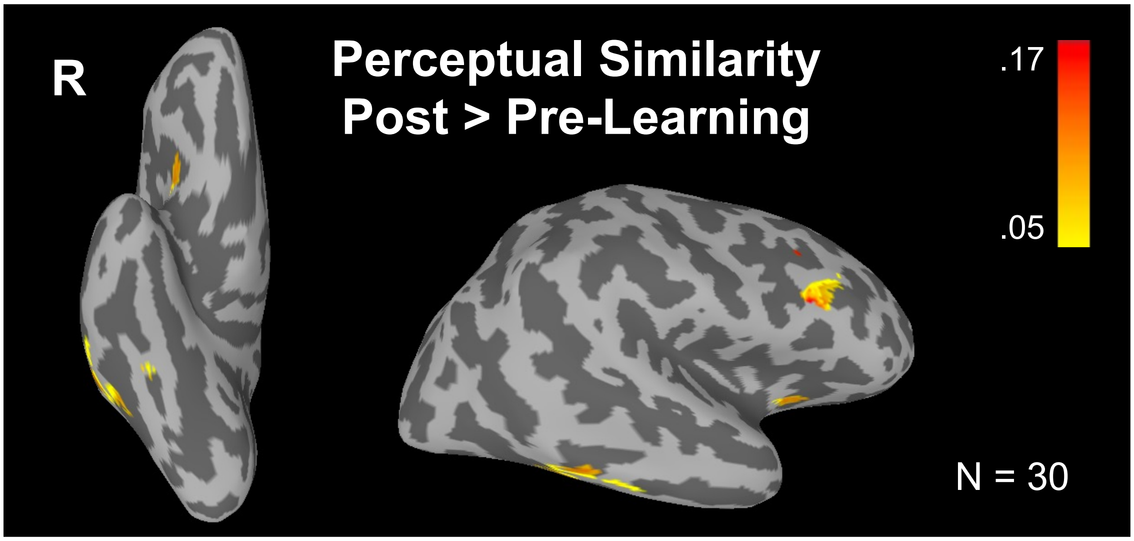


Figure 3.5

Perceptual similarity searchlight results (*N* = 30).

Peak location	*Surface area (mm)*	*Post-Pre M (SEM)*	*Post-Pre Max*	*p-value M*
R. Inferior Temporal Gyrus (anterior)	259	.1 (.002)	.13	.036
R. Inferior Frontal Gyrus	206	.12 (.002)	.17	.014
R. Orbitofrontal Cortex	95	.11 (.001)	.12	.035

Table 3.5

Perceptual similarity cluster results (*N* = 30). Aggregate results based on all searchlight nodes in the given cluster.

3.2 Relationship between neural changes in the ATL and individual differences in learning behavior and preference ratings

We next sought to compare individual differences in sensitivity to social information on the learning task with neural changes in the ATL regions due to learned social values. For each participant, we calculated the change in social value similarity in the left ATL region by subtracting the average correlation value across corresponding nodes in the pre-learning session from that in the post-learning session (Δ similarity).

To quantify behavioral biases, we calculated a sensitivity measure based on performance during the last day of learning (learning sensitivity), as well as in the preference ratings (preference sensitivity). Each participant's reward accuracy was subtracted from their social accuracy, in order to measure an individual's tendency to be more accurate at choosing faces based on social information in the learning task. Consequently, positive sensitivity values correspond to better accuracy at choosing faces based on their associated social values, compared to their reward values. We used this measure, as opposed to using social accuracy alone, in order to differentiate participants who had equivalent performance for each value type (e.g., 100% social, 100% reward), from those who were selectively better for social values (e.g., 100% social, 60% reward).

Preference sensitivity was calculated by first calculating the average ratings for high social value faces subtracted by the average ratings for low value faces (social value sensitivity), as well as the average ratings for high reward faces subtracted by those for the average low reward faces (reward value sensitivity). The reward value sensitivity was then subtracted from the social value sensitivity (preference sensitivity).

To compare behavioral sensitivity measures with neural changes, we correlated participants' sensitivity with the change in neural-social similarity in the left ATL. Spearman correlations were used because the behavioral results were not normally distributed (see Chapter 2), and rank-ordering should be used to compare brain and behavioral measures without assuming a linear relationship (Kriegeskorte et al., 2008). The relationship between preference sensitivity and ATL similarity was significant (r_s = .43, *p* = .018; Fig. 3.5, right), such that the more sensitive a subject was to social values in their preference ratings the greater the ATL was related to their social value DM after compared to before learning. A positive correlation was found between learning sensitivity and ATL similarity; however, this relationship was marginally significant (r_s = .33, *p* = .078; Fig. 3.5, left). Partial correlation analyses controlling for the opposite behavioral sensitivity measure confirmed these results (rank-transformed; preference/ATL: partial *r* = .39, *p* = .039; learning/ATL: partial *r* = .16, *p* = .408).

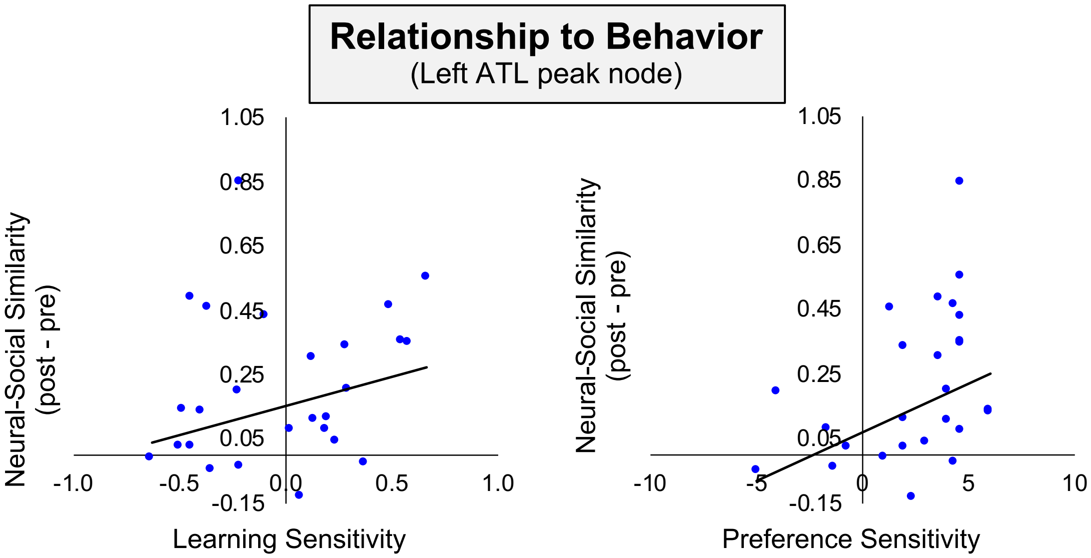


Figure 3.6

Relationship between changes in social similarity in the left ATL (brainand behavior as a function of value type (*N* = 30). There was a positive correlation between the change in neural-social similarity compared to learning sensitivity (left) and preference sensitivity (right); however, only the preference sensitivity comparison reached significance. Each point represents one subject, black line indicates linear trend.

3.3 Testing for learned value information in ventral face-selective regions

Our whole-brain searchlight analyses showed that learned social information influenced activity patterns in the left ventro-medial ATL. While these results support recent proposals of the role of the ATL face patch in memory and perception of other people (see Introduction), they are at odds with findings of social category information in activity patterns of earlier face-selective regions of the ventral visual processing hierarchy evoked during face viewing, particularly in the right fusiform gyrus (see Introduction). One cause of this difference may be due to the spatial resolution of the searchlights employed here (although Stolier and Freeman (2016) used a comparable searchlight size of 123 voxels). If a searchlight is too small or too large, it may lack the ability to detect relevant activity patterns in a region. Relatedly, one previous meta-analysis showed that decoding classification success using multi-voxel patterns depends on the number of voxels included, with earlier visual regions benefitting from a higher voxel count than late visual regions (Coutanche et al., 2016). This suggests that the distinguishability of activity patterns for regions of the visual system depends on the spatial resolution of the searchlight. Additionally, while there is significant overlap in the anatomical locations of ventral, visual face-selective regions across people (Grill-Spector et al., 2017), there are nonetheless idiosyncratic differences in the precise location and extent of these regions within individuals. Thus, the whole-brain analysis may have failed to detect these regions due to the variance in spatial specificity across subjects.

To address these potential issues of spatial resolution (number of voxels) and specificity (anatomical location), we performed ROI-based pattern similarity analyses using individual ROIs derived from independent localizer data and tested multiple searchlight sizes. For each participant, we localized three face-selective regions in each hemisphere, which were manually drawn such that they were approximately the same size across participants (see Methods). These were the occipital face area (OFA), posterior fusiform face area (posFFA), and anterior fusiform face area (antFFA). For each node in an ROI, activity pattern similarity for the set of voxels in the given searchlight were compared to the participant's subjective social or reward value DMs. Resulting correlation values were averaged across searchlights in an ROI, and a paired t-test was used to test for changes in activity patterns related to learned values between pre- and post-learning scan sessions. Searchlight sizes of 20, 40, 60, 80, and 100 voxels were used. The results of this approach confirmed that there was no relationship of social or reward similarity and activity pattern similarity between the pre- and post-learning sessions for any of the ROIs tested (Fig. 3.6).

Next, we tested for perceptual and behavioral similarity with the activity patterns in each ROI. Across voxel sizes for the posterior FFA ROI, there was a positive trend for the correlation of perceptual and neural DMs to be positive and greater after compared to before learning; however, this trend only reached significance in the 20-voxel ROI (Table 3.S3). In this region, faces that were more perceptually similar evoked more similar activity patterns in the post-learning session. No other ROI had activity pattern similarity related to the perceptual DM, and no ROI had a significant relationship between neural and behavioral similarity (Fig. 3.S6).

Together, these results support the notion that the lack of relationship between neural and social similarity in the core face network, namely regions along the fusiform gyrus, found in this study is not due to spatial constraints based on how searchlights were defined. This provides further evidence that associations of abstract social information with faces are encoded in neural responses at late stages of face processing (ATL face patch) and suggests that prior evidence for social information in the posterior fusiform gyrus was primarily driven by perceptual face features related to the social categories studied.

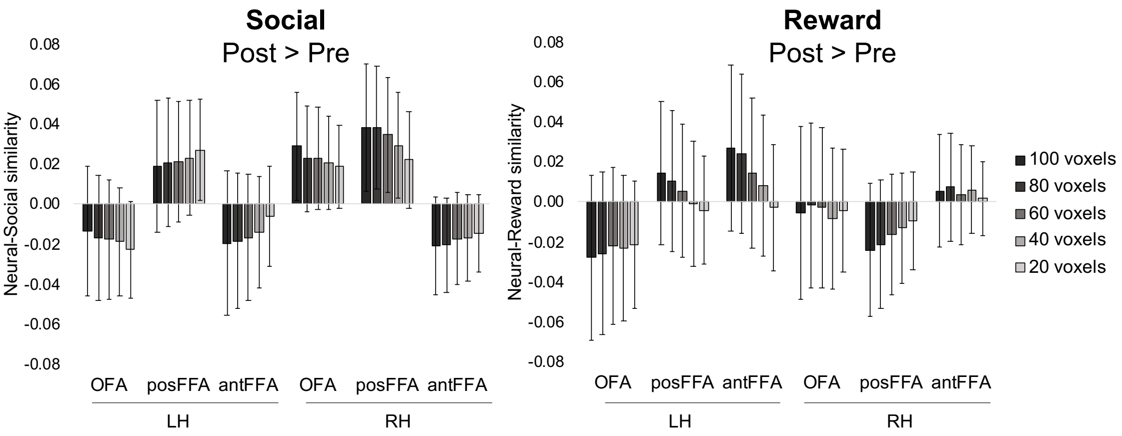


Figure 3.6

Results of ROI similarity analysis. Difference between post- and pre-learning similarity values (averaged across searchlights, and across participants), error bars indicate +/- *SEM*. There was no modulation of learned social values (left) or reward values (right) on activity patterns in any ROI tested (*N* = 28).

4. Discussion

In this study we examined the influence of learned face-value associations on neural activity. Former work had found evidence that learned associations between face identities and person-specific information (e.g., biographical information) recruited face-selective parts of the ventro-medial anterior temporal lobe (vmATL). However, recent functional neuroimaging (fMRI) studies have also found that social category information (e.g., stereotype biases) are related to the similarity of activity patterns in earlier face processing regions along the fusiform gyrus (see Introduction). We tested whether learning associations between faces and abstract social information influenced activity pattern similarity in the visual system. Visual-social associations for individual faces were learned during the experiment and fMRI data was collected before and after learning. This allowed us to examine changes in neural activity evoked by learning unlike studies probing pre-existing knowledge. Moreover, the social values utilized (personality traits) were not defined by perceptual features, and by design perceptual similarity was orthogonal to the similarity of the learned values. Thus, any influence of learned values on neural activity patterns would be separable from visual face information.

We found that activity patterns in the left vmATL when viewing faces were influenced by learned social values, such that after learning faces of more similar social values evoked more similar activity patterns in this region. This was true even though during fMRI scanning participants were performing a task unrelated to the learned values. Furthermore, the magnitude of this change was related to an individual's sensitivity to the social, compared to reward, values in assessing their preference to interact in a future social context. In other words, expectations of social behavior in future interactions were associated with the encoding of social information in the left vmATL. There was a trending relationship between accuracy as a function of value type on the last day of learning and the magnitude of change in neural similarity, but this did not reach statistical significance. These findings support the notion that the left vmATL encodes person-specific information learned through interactive experience.

Using both whole-brain roaming searchlight and localized region-of-interest (ROI) approaches, it was found that social value information was not related to activity in the occipital face area (OFA), posterior fusiform face area (posFFA) or anterior fusiform face area (antFFA), across multiple ROI sizes. This suggests that abstract social information is not incorporated in responses in these 'core' face processing regions, which is contrary to models that posit high-level social cognitive processes (namely in the ATL and orbitofrontal cortex/OFC) bias perceptual representations in these regions (Freeman & Johnson, 2016; Brooks & Freeman, 2019; Freeman, Stolier, & Brooks, 2020). Evidence for these claims come from studies showing differences in fusiform gyrus activity for in-group compared to out-group members based on race (Van Bavel, Packer, & Cunningham, 2008), as well as greater activity pattern similarity for more stereotypically 'overlapping' social categories during both passive viewing and active social categorization (Stolier & Freeman, 2016, 2017). Notably, the social categories in these experiments are depicted by facial features (e.g., race, emotional state, gender), and in the latter cases the implicated fusiform activity is contiguous with a large cluster centered on the early visual cortex (Stolier & Freeman, 2016, 2017). Thus, it could be that biases in face-selective fusiform regions are driven by perceptual features that cue the social categories (Mason, Cloutier, & Macrae, 2006), as opposed to top-down feedback from higher-level stages. In Stolier and Freeman (2016), a similarity analysis was performed while controlling for the visual similarity of the faces based on three models (HMAX layer C2, image silhouette, and pixel-intensity maps), and significant clusters were found in right medial temporal lobe[3] and left OFC, but not in the fusiform gyrus. Another important difference between these studies and the present study is that we examine activity evoked by specific faces, but the former averages activity across unique exemplars for the social categories.

Behavioral similarity judgments of the faces after learning were related to neural similarity in a large region of the left inferior parietal lobe (IPL), inclusive of the angular gyrus (AG) and supramarginal gyrus (SMG). This area has been implicated in encoding psychological distance across domains such as social, spatial, and temporal distance (Parkinson, Liu, & Wheatley, 2014). Interestingly, activity pattern similarity in the IPL has been linked to social distance metrics derived from social networks, such that pattern similarity during tasks such as viewing audiovisual movies and explicit judgments of familiar individuals is related to social network measures such as the number of mutual friends between individuals (Peer, Hayman, Tamir & Arzy, 2021; Hyon, Kleinbaym, & Parkinson, 2020; Parkinson, Kleinbaum, & Wheatley, 2018, 2017). Our results build on these findings by showing that subjective judgments of face similarity incorporate learned social information, and the spatial organization of face similarity spaces by social traits predicts activity pattern similarity in the left IPL. This shows how the neural (IPL) and behavioral coding schemes for representing familiar others can be structured by social information and can be revealed by distance-based measures.

That social and behavioral similarity were not related to neural similarity in the same regions might be surprising, as behavioral similarity was correlated with social similarity at a group-level (Chapter 2). One likely reason for this is that some subjects incorporated the reward values (or both values) into their behavioral similarity spaces, so the spatial distances between faces were not strictly determined by social values at an

[3] Stolier and Freeman (2016) claim that this region is in the right fusiform gyrus, however, the group result map and MNI-coordinates provided in the paper do not indicate the fusiform gyrus and instead lie more superior and anterior.

individual-level (whereas the social similarity was based on similarity rankings when participants were explicitly instructed to use the given value). Neural analyses were conducted at the subject-level, such that a participant's activity patterns were compared to their specific behavioral similarity spaces. In this way, although both regions are anatomically situated between areas that process sensory input and those involved in decision-making, activity patterns in the IPL were related to value information relevant to subjective judgements of similarity, while those in the left vmATL were only modulated by social information.

Reward value similarity was related to a region in the left posterior parietal cortex, primarily situated in the posterior intraparietal sulcus (IPS). This result replicates prior work showing the involvement of the parietal cortex in processing task-reward associations (Wisniewski et al., 2015; Kahnt et al., 2014), and underscores the representation of reward values independently of a task requiring those rewards. There are several potential explanations for why reward information was not found in the left vmATL. The vmATL could incorporate person knowledge in a selective manner, such that socially relevant properties are encoded in the same region as other knowledge about a person (e.g., face identity, name) but socially irrelevant information such as point magnitudes specific to an experimental task are not (but see Rice et al., 2018; Eifuku et al., 2010). Additionally, there is evidence for category- and modality-based subdivisions of the ATL (Skipper et al., 2011; Hung et al., 2020; Persichetti et al., 2021), and task-dependent involvement of the ATL during object-based size judgments (ATL is required when making conceptual judgments of size, but not perceptual; Chiou & Lambon Ralph, 2016). Consequently, reward value information may be represented in the ATL when relevant to the task at-hand and potentially in an ATL area other than the vmATL face patch.

In conclusion, the present study shows how personality traits associated with face identities can modulate neural representations of others in face-selective parts of the left vmATL, and these changes are related to expectations of social behavior of others beyond the learning task. Together with former studies, our findings suggest that the vmATL has a task-independent role in integrating person-specific information across face processing (fusiform gyrus), social knowledge (superior, lateral ATL), and action and decision-making (orbitofrontal, prefrontal) systems. This speaks against the claim that social information associated with faces biases responses in earlier face-processing regions in the fusiform gyrus via top-down feedback interactions, at least when social characteristics are not directly cued by visual facial features.

5. Supplemental Materials

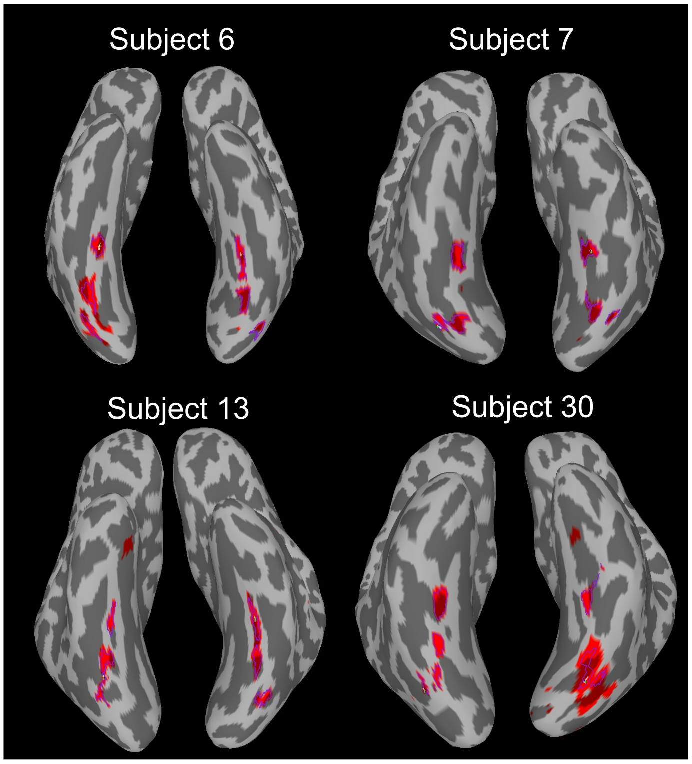


Figure 3.S7

ROI localization for three representative subjects. Univariate results of the face-localizer data displayed in red, ROIs outlined in purple. Thresholds determined within-subject for approximately equivalent ROI sizes across subjects.

Hemisphere	*Region*	*Number of surface nodes (SEM)*
LH	OFA	49 (3)
	posFFA	59 (2)
	antFFA	58 (3)
RH	OFA	52 (3)
	posFFA	68 (5)
	antFFA	58 (3)

Table 3.S6

Number of surface nodes across subjects for each manually drawn ROI (based on independent localizer data). Results shown for the occipital face area (OFA), posterior fusiform face area (posFFA) and anterior fusiform face area (antFFA) in the left and right hemispheres.

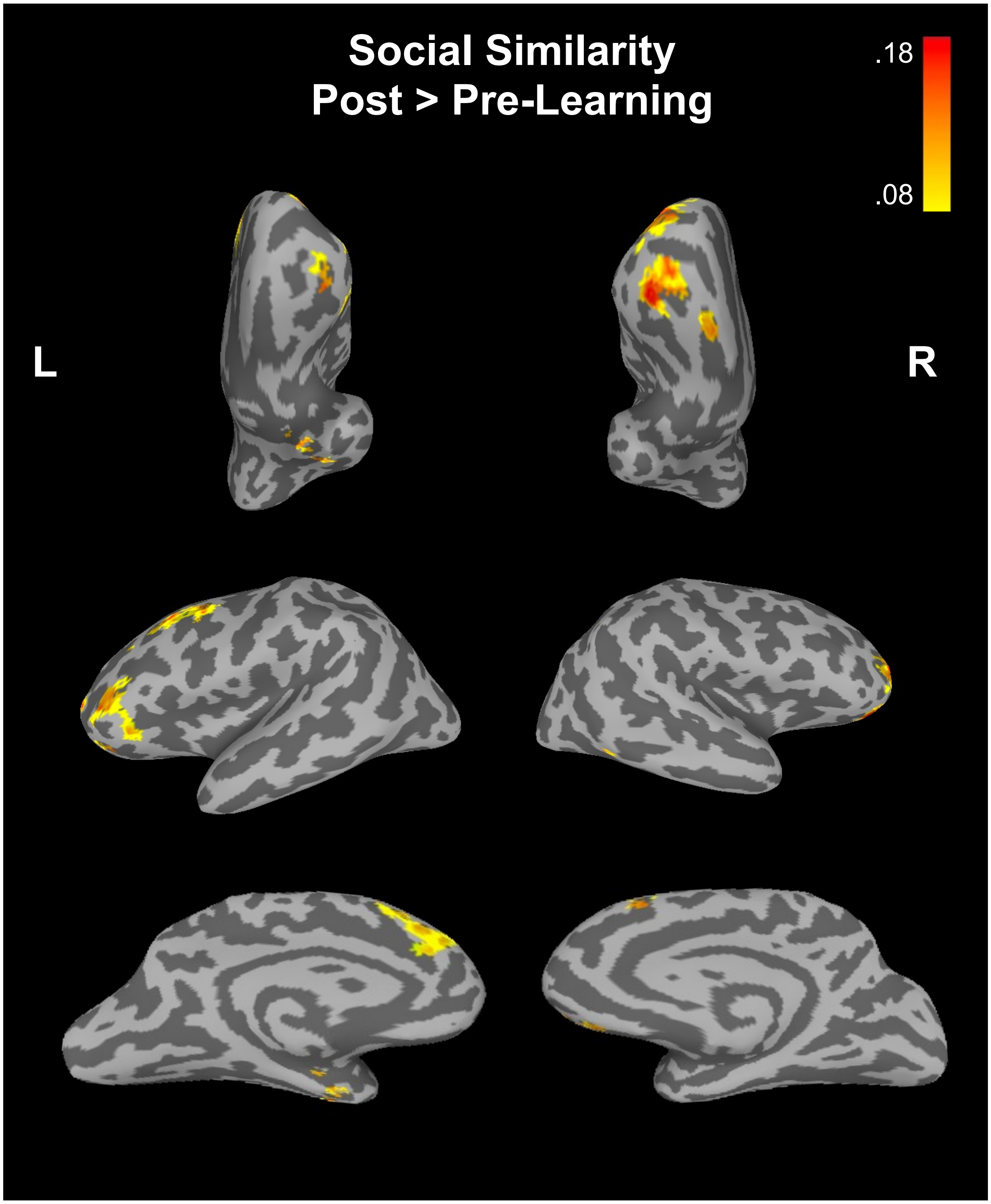


Figure 3.S8

Social similarity searchlight results. All depicted regions had significant positive correlations after, compared to before, learning ($N = 28$).

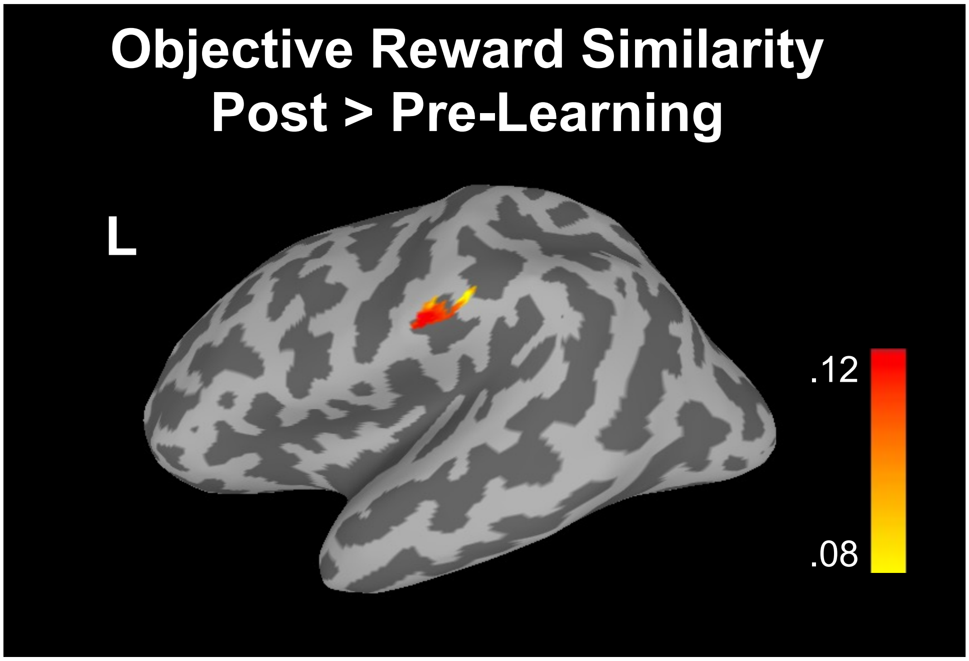


Figure 3.S9

Objective reward similarity searchlight results (N = 30).

Peak location	*Surface area (mm)*	*Post-Pre M (SEM)*	*Post-Pre Max*	*p-value M*
L. Supramarginal Gyrus	92	.11 (.001)	.12	.038)

Table 3.S7

Objective reward cluster results (N = 30).

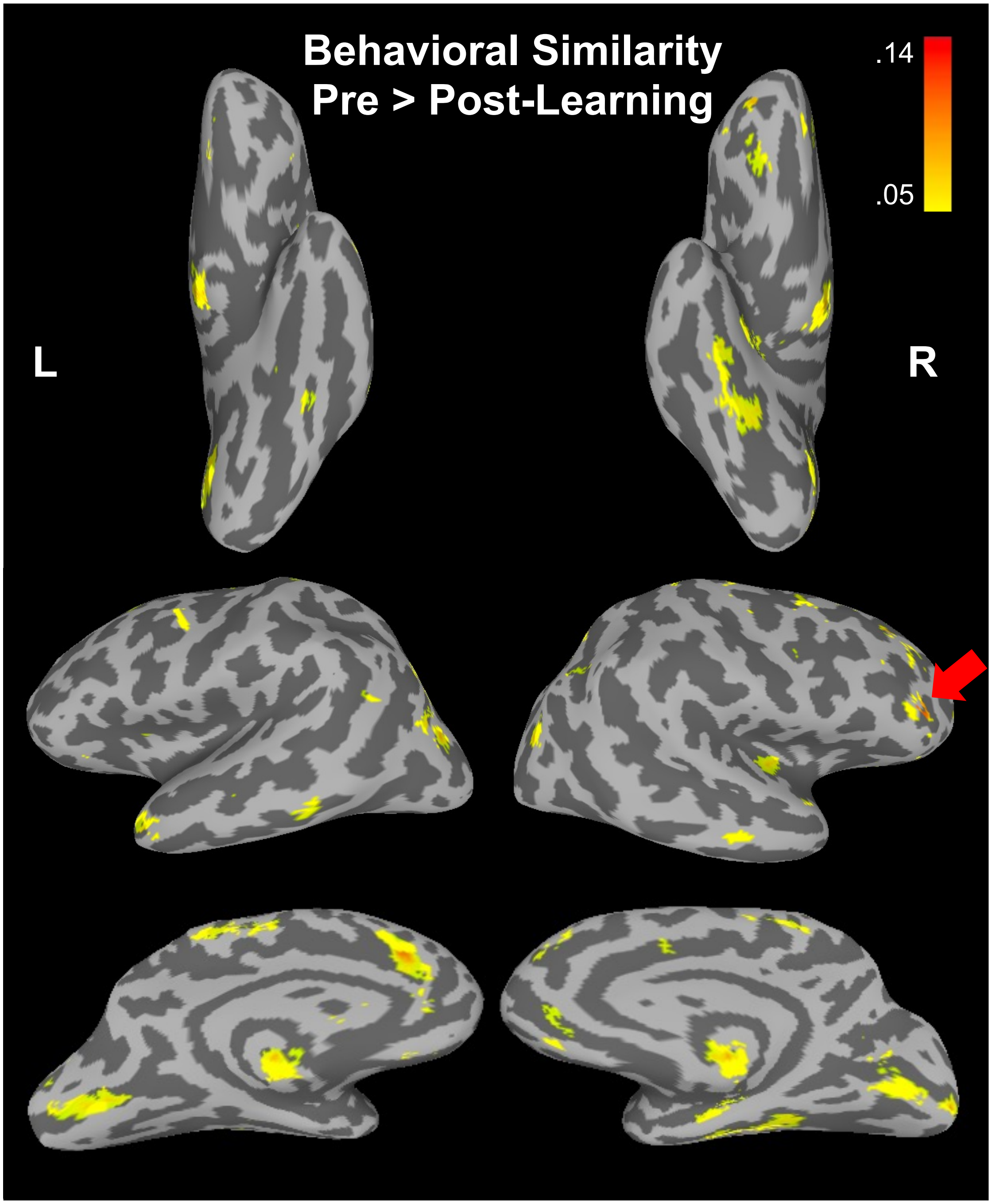


Figure 3.S10

Behavioral similarity searchlight results for the pre > post-learning comparison. No cluster reached significance; there was one cluster that was marginally significant in the right inferior frontal gyrus (red arrow)

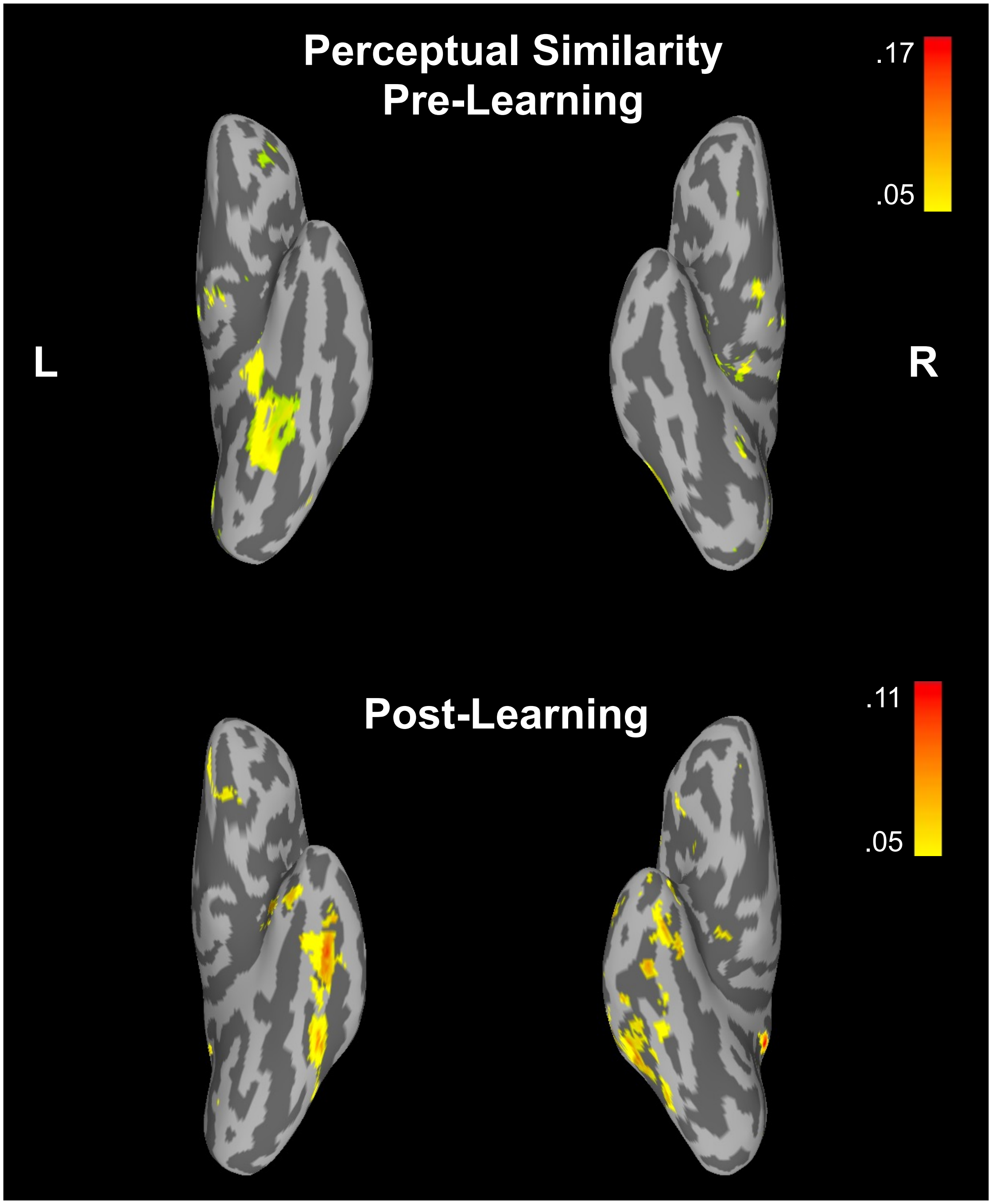


Figure 3.S11

Perceptual similarity searchlight results within pre- (top) and post-learning (bottom) scan sessions. No cluster reached significance.

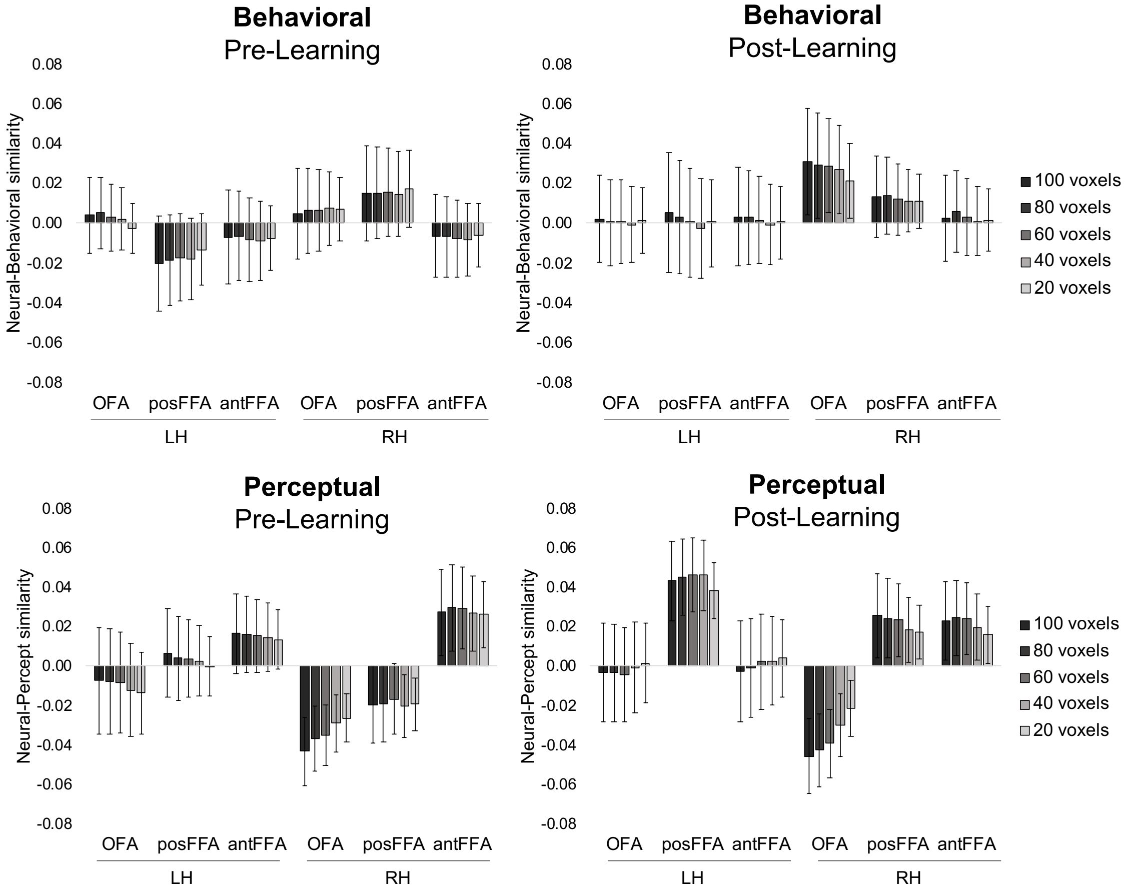


Figure 3.S12

ROI similarity analysis results comparing neural with behavioral (upper) and perceptual (lower) similarity, within pre- (left) and post-learning (right) scan sessions (*N* = 30).

Hemisphere	***Region***	***Voxel Count***	***Paired t-test***
LH	posFFA	100	*t*(29) = 1.37, *p* = .182
		80	*t*(29) = 1.62, *p* = .116
		60	*t*(29) = 1.78, *p* = .086
		40	*t*(29) = 1.93, *p* = .063
		20	*t*(29) = 1.95, *p* = .061
RH	posFFA	100	*t*(29) = 1.7, *p* = .1
		80	*t*(29) = 1.7, *p* = .1
		60	*t*(29) = 1.67, *p* = .1
		40	*t*(29) = 1.81, *p* = .081
		20	***t*(29) = 2.09, *p* = .046**

Table 3.S8

ROI-based perceptual similarity results comparing pre- and post-learning sessions (*N* = 30). All other comparisons for perceptual, social, reward, and behavioral similarity across ROIs did not reach significance.

IV. INFLUENCES OF VALUE LEARNING ON PERCEPTUAL DISCRIMINATION

1. Introduction

In Chapters 2 and 3 it was found that learning associations between values and faces modulated behavioral and neural similarity measures, such that the social value of generosity had an influence on both explicit similarity judgements and the similarity of activity patterns in face-selective brain regions during face perception. A question that follows from these findings is whether behavioral face perception is also affected by the learned value information. In the present study, we examine whether modulations of behavioral and neural face spaces by learned values influences perceptual processing when values are irrelevant for task performance. Moreover, only face stimuli were employed previously, and to address the open question of whether effects are domain-specific (i.e., are exclusive to faces due to social traits being characteristic of animate beings) we also tested associations between values and inanimate objects (flowers).

Our paradigm is based on existing literature on categorical perception. In these studies participants typically learn to categorize novel stimuli based on arbitrary category labels and the categories utilized are correlated with perceptual features that vary across stimuli. Studies have found that after learning, members of different categories are better discriminated from one another compared to members of the same category, suggesting that between-category items become further apart in representational space (e.g., Goldstone, 1994c; Folstein, Gauthier, & Palmeri, 2012; Folstein, Palmeri, & Gauthier, 2013; Goldstone, Kersten, & Carvalho, 2017). Importantly, participants become more sensitive to the perceptual information that distinguishes between-category members because it is relevant for the explicit categorization judgments, not necessarily because they are unable to learn the within-category structure (Levering & Kurtz, 2014). Here, we test whether learning to categorize faces based on social trait information results in categorical perception of faces, particularly when the traits are not determined by first impressions but instead through experience. If learned social categories modulate the structure of perceptual representations, then biases due to these categories should be found even in a perceptual task that does not require explicit social judgments, and for traits that are learned through experience.

Crucially, stimulus sets were constructed using a morphing procedure, such that all pairs of stimuli shared an equal amount of perceptual variance, and value categories were correlated with a particular visual dimension. To examine whether learned values modulated perceptual representations, we measured performance on a visual discrimination task that was unrelated to the learned values, once before and once after value learning. In this task, people sequentially viewed pairs of stimuli and reported whether they were the "same" or "different". We compared changes in discrimination of pairs as a function of their associated values. If value learning modulates perceptual spaces, then stimuli that have different values should be better discriminated after, compared to before, learning. In other words, these items will become less similar to one another and farther apart in space, and thus more distinct along the given value-relevant dimension. Notably, this task does not explicitly require any value information, thus any systematic influence of the learned values on visual discrimination performance cannot be due to task demands.

We conducted three experiments to assess whether and how learning social traits may influence perception. In Experiment 1, one group learned about four faces, and another group learned about four flowers. Changes in perceptual discrimination according to learned social and reward values were examined. Additionally, we investigated individual differences in learning behavior and changes in discrimination, as well as long-term value-related changes about one-month after learning. In Experiment 2, only reward values were learned, in order to rule out the possibilities of within-category modulations or training effects on changes in discrimination. In Experiment 3, we aimed to replicate Experiment 1 using a within-subject design, and also examine explicit social preferences for the stimuli in the context of the learning task and on a future cooperative task. This served as a measure of how the learned information generalizes to propensities in other types of social behavior.

2. Experiment 1

We first tested our main question of interest – whether learning social traits modulates perceptual discrimination performance, and whether any effects are specific to face stimuli. In Experiment 1, participants completed a training task in which they learned social and reward values associated with four stimuli (faces or flowers). Perceptual discrimination performance was examined before and after learning, in order to test whether the learned value categories influenced discrimination performance (categorical perception effect). Additionally, individual differences in the use of each value type during learning and resulting changes in perceptual discrimination performance were compared. Long-term changes in perceptual discrimination performance due to value learning were also examined to establish whether any modulations of representational spaces due to the learned values persisted over time.

2.1 Methods

2.1.1 Participants

Participants were recruited from the University of Pennsylvania to complete the study online for course credit (20-30 mins). 32 participants were included in the face condition (18-22, 9 males), and 32 were in the flowers condition (ages 18-22, 6 males; see *Supp. Methods* for eligibility criteria and sample size estimation). The University of Pennsylvania IRB approved all consent procedures.

2.1.2 Materials

All stimuli were generated by a morphing procedure. Four face images were selected based on previous studies on category learning (Goldstone & Steyvers, 2001; Goldstone, Lippa, & Shiffrin, 2001). Original face images were black-and-white photographs of bald, male heads (Kayser, 1997), which were normed in a study by Goldstone, Lippa, and Shiffrin (2001) and selected so that their average subjective similarity ratings were within 20% of one another (based on ratings of a set of 62 images). These faces were matched in their race, gender, and approximate age, in order to minimize the effects of these factors on task performance. Four flower images were selected from a database of 45 flower images from Hula & Flegr (2016) such that they had approximately equal beauty, complexity, and prototypicality ratings, and were front facing with five petals each. Prior to morphing, the flower images were made black-and-white.

For each condition (faces or flowers), the original images were morphed such that four unique stimuli were generated for each condition, and each stimulus shared a portion of its identity with two other stimuli (Morpheus Photo Morpher; Fig. 4.1, left). Consequently, an equal amount of perceptual information is shared for within-category and between-category pairs. Each stimulus contained equal percentages of two original images (faces: 37.5%, flowers: 25%) and of a fifth image (faces: 25%, flowers: 50%; percentages determined in pilot testing). All stimuli were 60 mm x 54 mm with a black background.

2.1.3 Value learning procedure

In order to have participants learn two value types, we used a modified version of a task from a previous study (Hackel, Doll, & Amodio, 2015). Participants were instructed that the four stimuli were allocated different amounts of points (point pools) and on a given trial they would donate a proportion of those points to the participant. They were told to maximize the points they received from the stimuli by learning which stimuli had larger average point pools than others, and which stimuli gave higher proportions of their points on average. Reward value is operationalized as the average magnitude of donated points (15 or 75 points), and social value is the proportion of the point pool shared on average (20 or 80%). Values were orthogonally assigned to the four stimuli, such that each stimulus shared its reward value with one stimulus and its social value with another. There were two stimulus groupings such that the same-reward/different-social value pairs in one grouping were the different-reward/same-social value pairs in the other grouping (and vice versa; Fig. 4.S1). In each condition, there were 16 participants per grouping. Results were combined across groupings for all analyses, ensuring that any resulting categorical perception effect was not due to the specific stimulus-reward value mappings.

Each participant completed 120 trials of value learning (Fig. 4.1, right). On a trial, a fixation cross was presented (1 sec), followed by two stimuli presented side-by-side (maximum 5 secs). Participants were instructed to choose either the left (F-key) or right (J-key) stimulus using their keyboard, and to be as fast and accurate as possible when responding (see *Supp. Methods* for full instructions). Once a participant made a choice, they were presented with the point value of their choice for that trial, displayed under that stimulus. If a participant did not make a choice in the allotted time, no feedback was presented.

After a participant made a choice, they were shown two pieces of information below that image: (1) how many points that stimulus shared with them on that trial (reward value), labeled "Shared", and (2) how many total points that stimulus had to share on that trial (point pool), labeled "Out of". For each trial, noise was added to the reward and social values by sampling from a normal distribution centered on the value (reward *SD* = 5, social *SD* = 0.05) and rounding up to the nearest integer. Point pools were calculated by dividing the reward values by the social values for the given trial and rounding up to the nearest integer.

Each pair of stimuli was presented 10 times (6 pairs total). Trials were divided into 10 blocks of 12 trials. Each pair was presented twice within a block, and the trial order within a block was randomized. Each participant received the same trial order and trial values.

2.1.4 Perceptual discrimination procedure

Participants completed 72 trials of a same-different discrimination task (Fig. 4.1, right), once before and once after the value learning task. In this task, they viewed pairs of stimuli, presented sequentially one after another, and reported whether the two stimuli were the same or different using the F-key and J-key on their keyboard. On a trial, a fixation cross was presented (1 sec), followed by a stimulus (1 sec), then a black-and-white noise mask (0.5 sec), and last a second stimulus with the prompt "Same (F-key) or

Different (J-key)?" above the stimulus, which were displayed until a response was made (max 5 secs). Participants were instructed to respond as fast and accurately as possible.

Across the 72 trials, each same-pair (4 total) was presented 6 times, and each different-pair (4 total) was presented 12 times. Different-pairs consisted of stimuli that shared part of their identity with each other (24 within-category pair trials, 24 between-category pair trials). Trials were divided into 6 blocks of 12 trials, and within each block there were 4 same trials (one per pair) and 8 different trials (two per pair, once for each possible presentation order within a pair). The order of trials was randomized within a block, and the same trial order was presented for all participants.

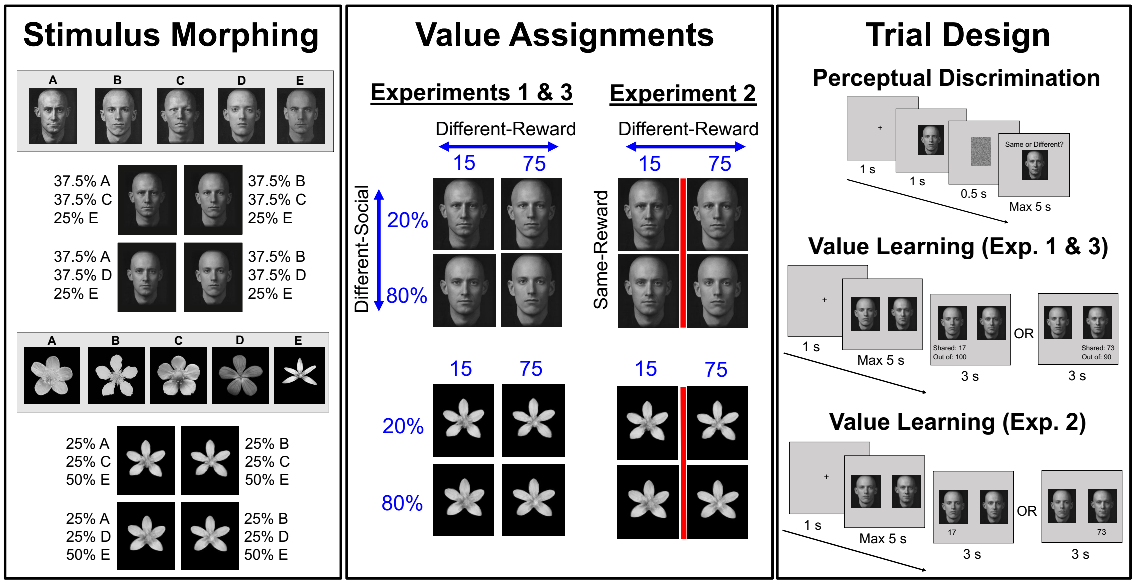


Figure 4.13

Experimental design. (Left) Morphing procedure used to generate stimuli. (Center) Example of social and reward value assignments. In each experiment, stimuli in each column were assigned the same reward value (15 or 75 points). In Experiments 1 and 3, stimuli were assigned one social (20 or 80%) and one reward (15 or 75 points) value indicated by the corresponding rows and columns. The same value mappings apply to both faces and flowers. (Right) Sample trial events for perceptual discrimination and value learning tasks. Feedback during value learning was presented for the given chosen stimulus (left/right).

2.2 Results

2.2.1 Value learning performance

Learning accuracy was calculated for each value-type separately in the second half of learning (20 trials per value). For a given value, a correct choice is choosing the higher-value stimulus, when presented with two stimuli that have different average values of that type (e.g., one high- and one low-social value face) but share the same average value of the other type (e.g., equal reward values). There was no difference between social and reward value accuracy in the faces (social: *M* = 0.8, *SEM* = 0.03, reward: *M* = 0.8, *SEM* = 0.03; paired *t*(31) = 0.05, *p* = .959) or flowers (social: *M* = 0.72, *SEM* = 0.04, reward: *M* = 0.72, *SEM* = 0.05; *t*(31) = 0.01, *p* = .990) conditions (Fig. 4.2, left). Accuracy was significantly above chance performance (50%) for both value types and conditions (Table S2). A mixed ANOVA revealed a significant effect of condition (faces vs. flowers: *F*(1, 62) = 8.49, *p* = .005), but no effect of value type (reward vs. social: *F*(1, 62) = 0.001, *p* = .972) or interaction (value x condition: *F*(1, 62) < 0.001, *p* = .990). T-tests showed no difference between conditions for social (*t*(62) = 1.61, *p* = .113) or reward (*t*(62) = 1.34, *p* = .187) values. This indicates that on average, people learned both social and reward values to the same degree. While overall learning performance was higher in the faces condition, accuracy did not differ between conditions when comparing each value type separately.

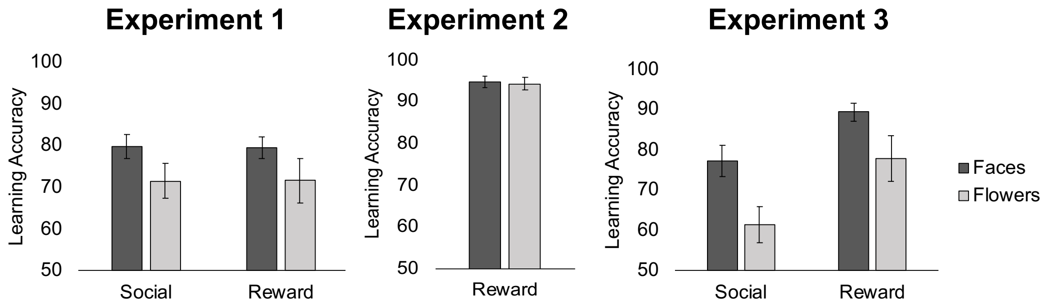


Figure 4.2

Value learning accuracy (% correct) across participants (*N* = 32 per faces/flowers condition). Note that Experiments 1 and 2 compared faces and flowers between-subjects (separate groups), while Experiment 3 was a within-subject design. Error bars represent +/- *SEM*.

2.2.2 Perceptual discrimination changes due to learning

Perceptual discrimination was measured using *d'*, separately for pre- and post-learning tasks and value types (Fig. 4.S2). Because the value types were orthogonally assigned to the stimuli, we were able to examine changes in *d'* performance for pairs with different social values but the same reward value (social-relevant), and different-reward pairs that shared the same social value (reward-relevant). There were no differences in pre-learning *d'* due to condition (faces/flowers) or value (social/reward) types (mixed ANOVA: condition, *F*(1, 62) = 0.7, *p* = .406; value, *F*(1, 62) = 3.26, *p* = .076; interaction, *F*(1, 62) = 0.2, *p* = .659).

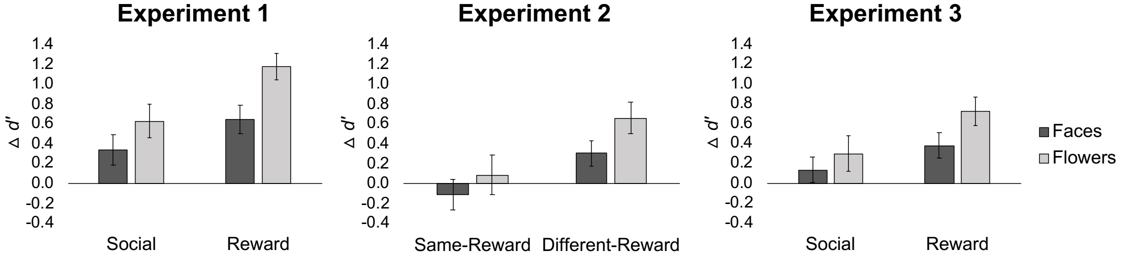


Figure 4.3

Change in discrimination performance due to value learning. Experiment 1 shows discrimination changes for social-relevant (different-social/same-reward, "Social") and reward-relevant (different-reward/same-social, "Reward") pairs of stimuli. Experiment 2 shows change in discrimination based on learned reward value categories. In Experiment 3, only reward values influenced discrimination performance (but see *Exp. 3 Discussion*). *N* = 32 per condition. Error bars represent +/- *SEM*.

To calculate a measure of the change in discrimination due to value learning for each participant, their pre-learning *d'* was subtracted from their post-learning *d'*, separately for social- and reward-relevant pairs ($\Delta d' = d'_{\text{post}} - d'_{\text{pre}}$). In both faces and flowers conditions, discrimination performance was better for social- and reward-relevant pairs after learning, indicated by positive Δ *d'* across subjects (faces: social *M* = 0.34, *SEM* = 0.15, reward *M* = 0.64, *SEM* = 0.14; flowers: social *M* = 0.63, *SEM* = 0.17, reward *M* = 1.17, *SEM* = 0.13; Fig. 4.3, left; Table 4.S3). A mixed ANOVA showed a significant within-subjects effect of value-type ($F(1, 62) = 14.07$, $p < .0001$) and between-subjects effect of condition ($F(1, 62) = 5.6$, $p = .021$), but no interaction between value-type and condition ($F(1, 62) = 1.11$, $p = .297$). T-tests showed higher Δ *d'* for the reward-relevant dimension in the flowers condition compared to faces ($t(62) = 2.77$, $p = .007$), but no difference between conditions for the social-relevant dimension ($t(62) = 1.3$, $p = .198$). Paired t-tests indicated higher Δ *d'* for reward-relevant pairs compared to social-relevant pairs in the flowers condition ($t(31) = 3.02$, $p = .005$) and in the faces condition ($t(31) = 2.23$, $p = .033$). These results indicate that learned social and reward values influenced discrimination performance for both faces and flowers. Discrimination performance was better for both value and stimulus types after learning, indicated by positive Δ *d'* in all cases, and there was a greater change for reward-relevant discriminations than social-relevant discriminations within both stimulus conditions.

2.2.3 Individual Differences

We next examined whether there were individual differences in the categorical perception effect as a function of learning behavior. Indeed, there were individual differences in learning accuracy for each value-type in both conditions (Fig. 4.S3). To quantify these differences, we calculated a measure of learning sensitivity for each participant by subtracting their reward accuracy from their social accuracy ($\textit{learning sensitivity} = \textit{accuracy}_{\text{social}} - \textit{accuracy}_{\text{reward}}$). If participants had a positive sensitivity value, they were more accurate at choosing stimuli based on the learned social values than reward values. If sensitivity was negative, they were more accurate at choosing based on reward values than social values. If a participant had sensitivity around zero, they were about equally accurate for both value types.

In order to compare an individual's learning sensitivity to changes in their discrimination performance, we also computed a measure of *d'* sensitivity. We subtracted each participant's Δ *d'* for reward-relevant pairs from their Δ *d'* for social-relevant pairs ($d'\ \textit{sensitivity} = \Delta\, d'_{\text{social}} - \Delta\, d'_{\text{reward}}$). In this case, a positive *d'* sensitivity corresponds to a greater change in perceptual discrimination for pairs along the social-relevant dimension compared to the reward-relevant dimension, while a negative *d'* sensitivity is a greater change for reward-relevant compared to social-relevant pairs. A *d'* sensitivity around zero indicates little or no difference in change along both dimensions.

Pearson correlations of the learning sensitivity and *d'* sensitivity measures revealed a significant positive correlation in the faces condition (*r*(30) = .41, *p* = .021; Fig. 4.4, left) but not the flowers condition (*r*(30) = .08, *p* = .672; Fig. 4.4, right). This shows how for faces, the difference between an individual's accuracy for each value type during category learning is related to the magnitude of change along each value-relevant dimension in the perceptual discrimination task; a relationship that was not found for flowers. Although the *d'* sensitivity measure collapses across both value types and does not differentiate between the directionality of changes for each value dimension separately, additional analyses confirmed that while all participants got better at discriminating reward-relevant face pairs, only participants with higher social value accuracy than reward value accuracy got better at discriminating social-relevant pairs (see *Supp. Results,* section 7.1.1; Fig. 4.S4).

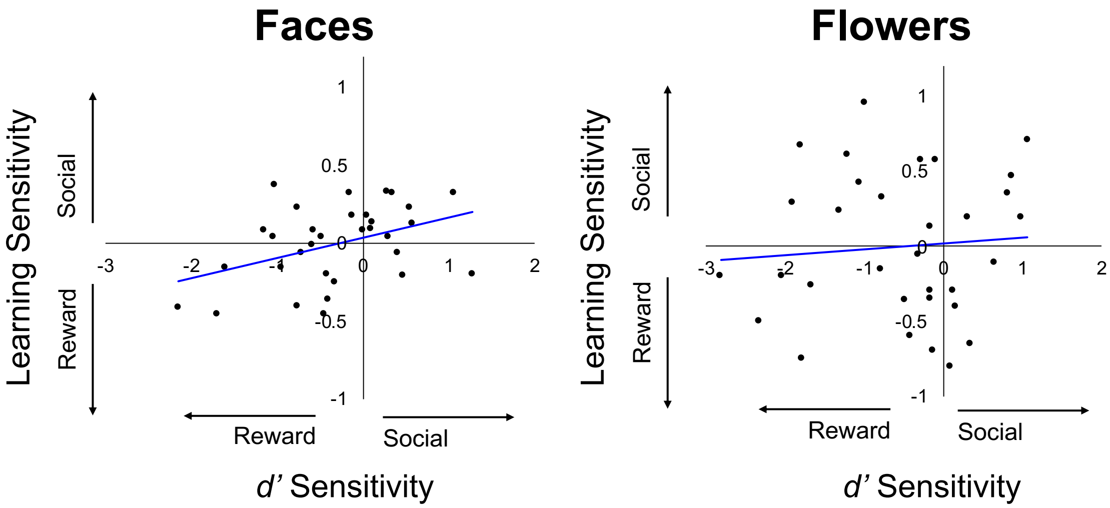


Figure 4.4

Individual differences in choice behavior during value learning and discrimination performance in Experiment 1 (*N* = 32 within condition). A significant correlation was found in the faces condition, but not the flowers condition. Each point represents one participant; blue line indicates linear trend.

2.2.4 Long-term perceptual changes

Last, we examined whether changes in perceptual discrimination performance due to value learning had long-term effects. Approximately one month following their initial session, we invited participants in both conditions to return for an additional online session of the perceptual discrimination task (72 trials) for additional course credit or \$5; a subset of the original sample completed this task (faces: *N* = 16, average 34 days delay; flowers: *N* = 10, average 32 days delay). For this group, their first session Δ *d'* results followed a similar pattern as the group-level results (Fig. 4.5); however, the effect of value was marginally significant and the effect of condition was not significant. Additionally, t-tests showed no differences between value types within or between conditions (Table 4.S4).

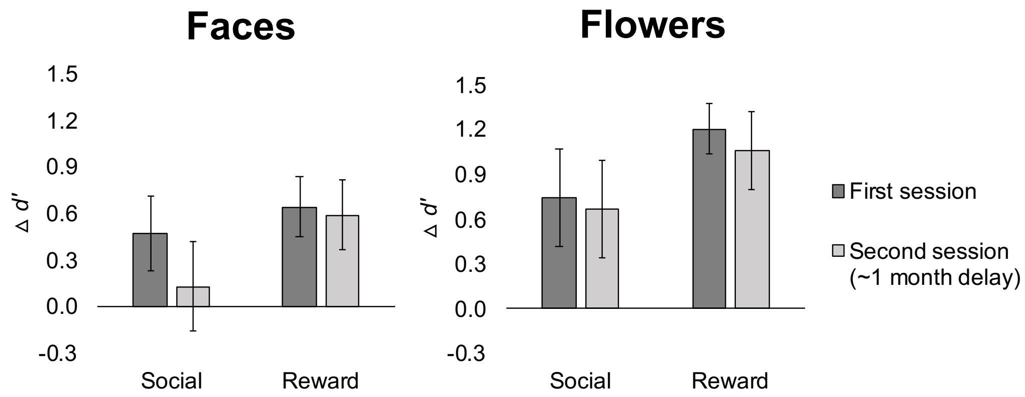


Figure 4.5
Change in discrimination performance for Experiment 1 participants who returned after a one-month delay (faces: *N* = 16; flowers: *N* = 10). Error bars represent +/- *SEM*.

Just as before, each participant's initial pre-learning *d'* was subtracted from their *d'* measured after one-month (Δ $d' = d'_{\text{delay-post}} - d'_{\text{pre}}$) for social- and reward-relevant pairs separately (Fig. 4.5). When examining Δ *d'* for faces after the delay, the positive Δ *d'* for reward pairs was comparable to the initial session (t-test against 0: $t(15) = 2.6$, $p = .02$); however, the positive social Δ *d'* effect was attenuated ($t(15) = 0.47$, $p = .645$; Table

S5). For the flowers condition, Δ *d'* after the delay was comparable to initial Δ *d'* for reward (*t*(9) = 4.08, *p* = .003) but slightly weaker for social (*t*(9) = 2.06, *p* = .069) pairs (Table S5). Together, this shows how for both stimulus types the change in discrimination performance for reward-relevant pairs persisted after the delay, but for social-relevant face pairs the effect was diminished.

2.3 Discussion

Overall, we found that at a group-level, learned social and reward values influenced discrimination performance for both face and flower stimuli. Reward value categories were behaviorally relevant to the task of maximizing one's accrued points, thus the learned reward values influenced perceptual discriminations for both types of stimuli. On the other hand, using the social values was not necessary to accomplish this goal, although participants were told to learn them. Even though irrelevant to point maximization, participants nonetheless learned the social values, and social values influenced discrimination performance of both faces and flowers; however, unlike reward values, this effect did not persist after a long-term delay.

When examining learning and discrimination performance at an individual-level, differences between faces and flowers conditions emerged. Specifically, for faces there was a relationship between an individual's performance during value learning and changes in perceptual discrimination as a function of value type. The extent to which participants were more accurate at choosing faces based on their social values compared to reward values during learning determined how much better they got at discriminating faces that differed based on their social values. This relationship was not found in the flower condition. This suggests that changes in face space were more sensitive to learning performance than modulations of flower space.

3. Experiment 2

In Experiment 2, we ruled out several alternative explanations of the main results in Experiment 1. First, the increase in discrimination performance for both value types could be due to a training effect, in which performance improved over time for all items with more practice. Instead of modulations relating to learned value categories, better discrimination for all pairs could simply be due to being able to better perceive the differences between the images. Second, the change in discriminations could be due to a single value type. For instance, the different-social pairs are also same-reward pairs, thus the changes could be due to both between- (different-reward) and within-category (same-reward) modulations for reward values. Prior studies on categorical perception typically have not found changes in discrimination for within-category comparisons; however, our learning paradigm differs significantly from previous work. To address these concerns, we sought to replicate the categorical perception effect using one value type. Participants completed perceptual discrimination and value learning tasks, but the learning task was modified such that participants only learned reward values and not social values. If changes in discrimination are not due to training, and are not influenced by within-category structure, then there should be no changes in discrimination of stimuli pairs that share the same reward value.

3.1 Methods

3.1.1 Participants

Participants were recruited from the University of Pennsylvania and completed the experiment online in about 20-30 minutes for course credit (see *Supp. Methods* for exclusion criteria). The final sample included 32 participants in the faces condition (ages 18-22, 16 males), and 32 in the flowers condition (ages 18-33, 18 males). The University of Pennsylvania IRB approved all consent procedures.

3.1.2 Value learning procedure

The stimuli and learning trials were the same as in Experiment 1, with the exception that feedback on each trial only included reward values (i.e., there was no point pool; Fig. 4.1, right). Prior to starting the task, participants were told that their goal was to maximize their total accrued points by learning which stimuli had high, and which had low, average reward values. Two stimuli were assigned a low average point value (15 points) and two stimuli had a high average point value (75 points). For each trial, noise was added to these values by randomly sampling from a normal distribution centered on the given value (*SD* = 5) and rounding up to the nearest integer. Two groupings were used to assign stimuli to value categories (Fig. 4.S1), such that the same-reward/within-category pairs for one grouping were the different-reward/between-category pairs for the other grouping (and vice versa). Within the face and flower conditions, there were 16 participants per grouping. Results were combined across groupings for all analyses.

Across 72 total trials, each pair of stimuli was presented 12 times (6 pairs total). Trials were divided into 6 blocks of 12 trials, so that each pair was presented twice within a block and each stimulus was presented on each side of the screen at an equal rate. The order of trials within a block was randomized, and each participant viewed the same trial order.

3.2 Results

3.2.1 Value learning performance

Accuracy was defined as choosing the stimulus with a higher average reward value when presented with one low- and one high-value stimulus. In the second half of learning (25 total trials), accuracy across participants was significantly above chance-performance (50%) in both the faces (*M* = 95%, *SEM* = 1%; $t(31) = 32.18$, $p < .0001$) and flowers (*M* = 94%, *SEM* = 2%; $t(31) = 30.05$, $p < .0001$) conditions (Fig. 4.2, center). Accuracy was equivalent in the flowers condition compared to faces (one-way ANOVA: $F(1, 62) = 0.04$, $p = .841$). This confirmed that participants successfully learned the reward values associated with the stimuli in both conditions.

3.2.2 Perceptual discrimination changes

Discrimination performance was measured using *d'* (Fig. 4.S5). As in Experiment 1, the change in *d'* was calculated for each participant by subtracting their pre-learning *d'* from their post-learning *d'*, separately for same- and different-reward pairs of stimuli (Δ *d'* = *d'*$_{\text{post}}$ – *d'*$_{\text{pre}}$). Note that the previous reward-relevant dimension is still relevant for reward value categorization (different-reward pairs), but the social-relevant dimension is now solely a within-category (same-reward) dimension. In both conditions, discrimination performance for different-reward pairs improved, indicated by positive Δ *d'* across subjects (faces: *M* = 0.31, *SEM* = 0.12; flowers: *M* = 0.66, *SEM* = 0.16), but did not change for same-reward pairs (faces: *M* = -0.12, *SEM* = 0.15; flowers: *M* = 0.09, *SEM* = 0.2; Fig. 4.3, center; Table 4.S6). A mixed ANOVA confirmed a significant within-subject effect of category membership (same- vs. different-reward; $F(1, 62) = 16.97$, $p < .0001$), but no between-subjects effect of condition ($F(1, 62) = 2$, $p = .162$) and no interaction of category and condition ($F(1, 62) = 0.39$, $p = .536$). Paired t-tests showed significantly higher Δ *d'* for different-reward compared to same-reward pairs in both conditions (faces: $t(31) = 2.49$, $p = .018$; flowers: $t(31) = 3.33$, $p = .002$). There was no difference between conditions for either category comparison (same-reward: $t(62) = 0.8$, $p = .425$; different-reward: $t(62) = 1.71$, $p = .093$).

3.3 Discussion

For both faces and flowers, learned reward values influenced changes in discrimination, such that items with different reward values were better discriminated after learning, and there was no difference in performance for items with the same reward value. This confirms that although our paradigm differs from category learning studies in which participants learn to map category labels to stimuli, learned reward values produce the same categorical perception effect, regardless of the type of stimulus (faces/flowers). Learned values are used to group stimuli into categories, and this results in changes in perceptual discrimination performance depending on whether the stimuli being discriminated belong to the same or different reward value category. Notably, these results establish that there are no changes in discrimination within value categories for either faces or flowers. Thus, any modulation of the discrimination of these pairs by learned information in Experiment 1 cannot be due to a training effect (overall performance increases due to practice on the task) or within-category structure.

4. Experiment 3

In Experiment 3 we sought to replicate the main results of Experiment 1 using a within-subject design. We had a single group of participants learn social and reward values associated with faces and flowers in one online session, counterbalancing the order of stimulus conditions across participants. Additionally, we examined how learned values generalized to self-reported propensities in social behaviors.

4.1 Methods

4.1.1 Participants

Participants were recruited from the University of Pennsylvania to complete the study online for course credit or $10 (40-60 mins). 32 participants were included (ages 18-29, 12 males; see *Supp. Methods* for eligibility criteria). The University of Pennsylvania IRB approved all consent procedures.

4.1.2 Procedure

The stimuli, values, and procedure were the same as in Experiment 1. Participants performed pre-learning discrimination, value learning, and post-learning discrimination tasks for one stimulus, and then repeated this for the other stimulus. The order of stimulus type was counterbalanced across participants ($N = 16$ per order). The order of stimulus grouping was also counterbalanced ($N = 16$ per grouping; $N = 8$ per order and grouping).

Afterwards, they completed preference ratings for each stimulus. They were asked to rate on a scale of 1-7 how much they preferred playing with each stimulus during the value learning task, and then how much they preferred to be paired with each stimulus in a non-economic, cooperative puzzle-solving task if they were invited back for a future study (based on Hackel, Doll, & Amodio, 2015). Surveys were conducted via Qualtrics. One subject did not complete these surveys; thus, all reported analyses are based on 31 participants.

4.2 Results

4.2.1 Value learning performance

Learning accuracy was calculated for each stimulus and value separately, following the same approach as Experiment 1 (Fig. 4.S6). A repeated-measures ANOVA showed a significant effect of condition ($F(1, 31) = 19.11$, $p < .001$), a significant effect of value ($F(1, 31) = 5.71$, $p = .023$), but no significant interaction ($F(1, 31) = 0.23$, $p = .634$). Unlike Experiment 1, paired t-tests showed significantly higher social ($t(31) = 2.95$, $p = .006$) and reward ($t(31) = 2.11$, $p = .043$) value accuracy in the faces compared to the flowers condition. Additionally, in the faces condition there was higher accuracy for reward values ($M = 0.89$, $SEM = 0.02$) compared to social values ($M = 0.77$, $SEM = 0.04$; $t(31) = 2.42$, $p = .022$; Fig. 4.2, right), while there was no such difference in Experiment 1. This comparison was marginally significant in the flowers condition (reward: $M = 0.78$, $SEM = 0.06$; social: $M = 0.61$, $SEM = 0.05$; $t(31) = 1.78$, $p = .085$). Accuracy was significantly above chance performance (50%) for both value types and conditions (Table S7). Mixed ANOVAs showed no effects of stimulus order in either condition (Table S8).

4.2.2 Perceptual discrimination changes

As before, the change in discrimination performance was calculated for each subject and value type separately ($\Delta d' = d'_{\text{post}} - d'_{\text{pre}}$). In both conditions, discrimination performance was better for reward-, but not social-, relevant pairs after learning (Fig. 4.3, right; faces: social $M = 0.13$, $SEM = 0.13$, reward $M = 0.37$, $SEM = 0.13$; flowers: social $M = 0.3$, $SEM = 0.17$, reward $M = 0.72$, $SEM = 0.14$; Table S9). A repeated-measures ANOVA showed a significant effect of value ($F(1, 31) = 8.35$, $p = .007$), with no effect of condition ($F(1, 31) = 2.2$, $p = .149$) or interaction ($F(1, 31) = 0.41$, $p = .528$). Paired t-tests indicated a significant difference between $\Delta d'$ for reward-relevant pairs compared to social-relevant pairs in the flowers condition ($t(31) = 2.11$, $p = .043$), and no difference in the faces condition ($t(31) = 1.49$, $p = .147$). There were no differences between conditions for either value type (social: *t*(31) = 0.69 *p* = .493; reward: *t*(31) = 1.65, *p* = .109).

At the same time, there was an effect of stimulus task order on changes in face discrimination performance. A mixed ANOVA showed a significant between-subject effect of order (*F*(1, 30) = 6.29, *p* = .018) on Δ *d'*, with no within-subject effect of value (*F*(1, 30) = 2.25, *p* = .144) or interaction (*F*(1, 30) = 1.58, *p* = .219). When participants performed faces first (Fig. 4.6, left), Δ *d'* was significantly above zero for reward face pairs (*M* = 0.7, *SEM* = 0.14; *t*(15) = 5.02, *p* < .0001) and marginally so for social pairs (*M* = 0.26, *SEM* = 0.15; *t*(15) = 1.72, *p* = .105). Surprisingly, those that performed flowers first had no change in face discriminations for either value type (social: *M* = 0.01, *SEM* = 0.20, *t*(15) = 0.03, *p* = .975; reward: *M* = 0.05, *SEM* = 0.18, *t*(15) = 0.25, *p* = .806; Fig. 4.6, right). Thus, a similar trend as Experiment 1 was found for those that completed the faces condition first, while there were no changes in discrimination in the group that had flowers first. There was no effect of task order on Δ *d'* for flowers (Tables S10-S11).

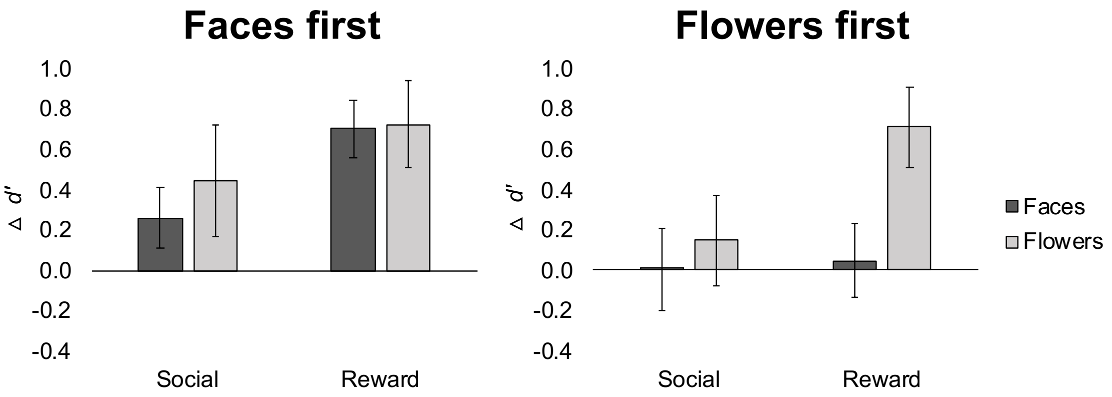


Figure 4.6

Change in discrimination performance in Experiment 3 separated by stimulus condition order (*N* = 16 per order). There was an effect of order in the faces, but not the flowers, condition. Error bars represent +/- *SEM*.

4.2.3 Social preference ratings

In order to examine the influence of learned values on preferences, participants rated each stimulus based on how much they preferred playing with them during the value learning task (current), and how much they preferred to be paired with that stimulus on a future, non-economic cooperative puzzle solving task if they were invited back for another study (future). Preference ratings across subjects were examined for each stimulus and corresponding value assignment (high/low and social/reward; Table S12).

Overall, participants preferred high value stimuli over low value stimuli (Fig. 4.7, upper). A three-way repeated-measures ANOVA on preference ratings for the faces condition showed significant effects of value level (high/low: $F(1, 30) = 25.53$, $p < .0001$) and type (social/reward: $F(1, 30) = 53.56$, $p < .0001$), but no effect of time (current/future; $F(1, 30) = 2.35$, $p = .136$) or significant interaction (Table S13). Within current and future ratings, paired t-tests showed significant differences for each value comparison except between low-social/high-reward and high-social/low reward faces (current: $t(30) = -1.71$, $p = .098$; future $t(30) = 0.58$, $p = .568$; Table S14). This shows that on average, participants used both social and reward value information to guide their face preference ratings.

Paired t-tests indicated significantly lower ratings for low-social/low-reward faces in the future, compared to current, context ($t(30) = -2.06$, $p = .048$), and a marginal difference for low-social/high-reward faces ($t(30) = 1.75$, $p = .09$), with no difference between current and future ratings for high-social/low reward ($t(30) = -0.09$, $p = .928$) or high-social/high-reward faces ($t(30) = 0.44$, $p = .662$). This indicates that when considering a future social context, preferences for low-social value faces decreased across subjects, which suggests that participants tend to avoid faces that were not generous to them during learning (low social value).

For the flowers condition, there were effects of time ($F(1, 30) = 4.3$, $p = .047$), value type ($F(1, 30) = 25.5$, $p < .0001$), and value level ($F(1, 30) = 6.54$, $p = .016$) on preference ratings, but no interactions (Table S13). T-tests indicated all pairwise comparisons were significant except between low-social/low-reward and high-social/low-reward (current: $t(30) = 1.73$, $p = .094$; future: $t(30) = 1.48$, $p = .149$), and between low-social/high-reward and high-social/high-reward (current: $t(30) = 1.39$, $p = .174$; future: $t(30) = -1$, $p = .327$; Table S14). This suggests that social values did not influence preference ratings in the flowers condition. There were no significant differences between current and future ratings for any flower comparison, although the low-social/low-reward comparison reached marginal significance (Table S15).

In order to examine whether preference ratings differed as a function of value type on an individual basis, we calculated a preference sensitivity score for each participant as follows:

$$\textit{Preference sensitivity} = (\textit{mean(high-social)} - \textit{mean(low-social)}) - (\textit{mean(high-reward)} - \textit{mean(low-reward)})$$

In this way, a positive score indicates a participant who was more sensitive to social values in their preference ratings than reward values, while a negative score corresponds to reward value sensitivity. Sensitivity was not correlated between faces and flowers conditions (current: $r_s(29) = .29$, $p =.112$; future: $r_s(29) = -.08$, $p =.663$), indicating that participants' sensitivities in their preference ratings for one stimulus did not generalize to the other stimulus type.

Comparing preference sensitivity with value learning sensitivity revealed significant positive correlations with the current ratings for both faces ($r_s(29) =.16$, $p = .01$) and flowers ($r_s(29) = .7$, $p < .0001$; Fig. 4.7, lower). For the future ratings, there was a significant correlation in the flowers ($r_s(29) = .51$, $p = .004$), but not faces ($r_s(29) =.17$, $p = .352$), condition. This shows that an individual's performance during value learning was related to their preferences during the task for faces and flowers, and for future interactions with the flowers. Specifically, the difference in accuracy for one value compared to the other predicted which values a participant was more sensitive to in their preference ratings. For example, participants who are better at choosing high-social value over low-social value faces during learning preferred high-social value over low-social values faces more than their preference between high-reward and low-reward value faces.

Preference sensitivity was not related to *d'* sensitivity ($\Delta d'_{\text{social}} - \Delta d'_{\text{reward}}$) for either faces (current: $r_s(29) = .17$, $p = .351$; future: $r_s(29) = -.11$, $p = .541$) or flowers (current: $r_s(29) = .16$, $p = .383$; future: $r_s(29) = -.02$, $p = .907$). This may be due to the lack of change in discrimination performance due to social value. Perhaps if social values influenced Δ *d'*, as in Experiment 1, the *d'* sensitivity measure would be related to individual differences in participants preference ratings depending on value type.

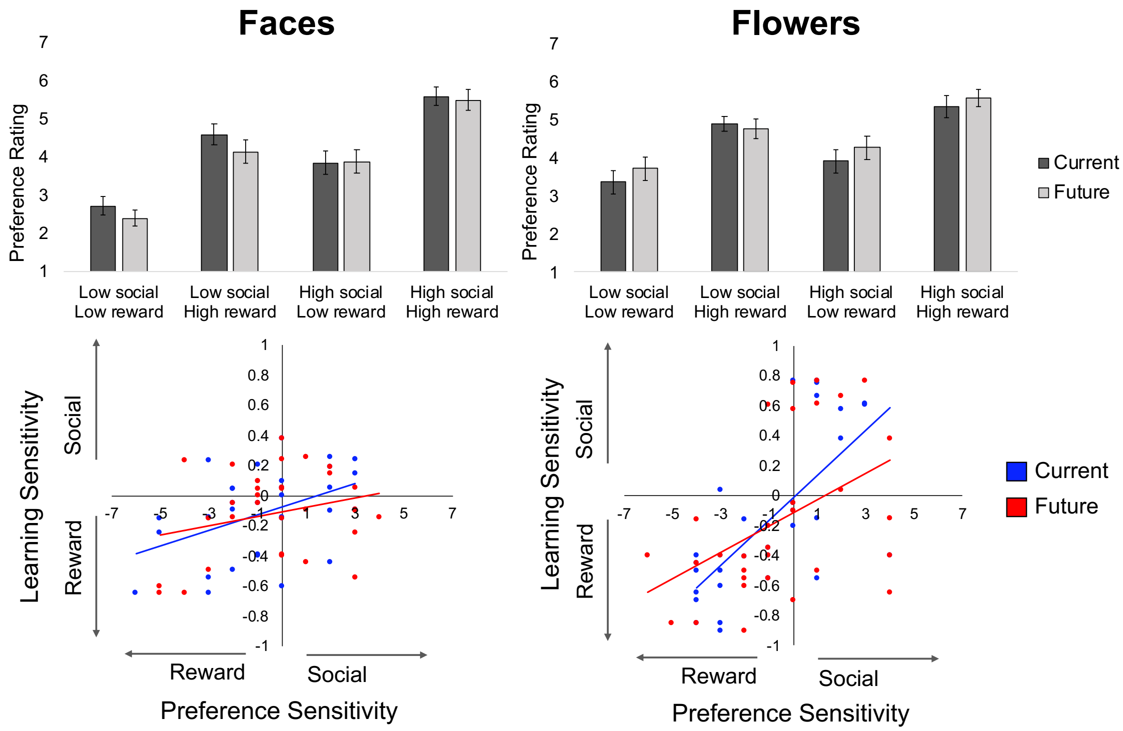


Figure 4.7
Sensitivity to social and reward values based on preference ratings (preference sensitivity) in the context of the learning task (current) and a future cooperative task (future), compared to value sensitivity during the learning task (*N* = 31). Each point represents one participant. Solid line indicates linear trend.

4.3 Discussion

Together, these results indicate that while there was an increase in discrimination performance relevant to learned reward values, the effect of learned social values on changes in discrimination performance found in Experiment 1 did not replicate at the group-level. There are several differences between the experiments which may explain this. First, there was a difference in learning accuracy for social and reward values in the faces condition which was not found in Experiment 1. Specifically, participants had higher accuracy for reward values than social values. This may have caused their discrimination performance to increase for reward-relevant, and not social-relevant discriminations. In other words, because participants were better at learning reward values compared to social values, they became better at discriminating reward pairs and not social pairs. Evidence to support this comes from the finding in Experiment 1 that learning accuracy was related to the magnitude of change in social value-based face discriminations. People who had higher reward value accuracy than social accuracy had less change in discrimination performance for social-relevant pairs.

Additionally, there was an effect of stimulus task order on changes in discrimination performance for faces. Participants who completed the faces condition before the flowers condition had increased discrimination performance for both reward and social pairs compared to those who completed flowers and then faces. Unexpectedly, participants who completed flowers first had no change in face discrimination performance for either reward or social values, even though there was no difference between the two orders in learning accuracy or overall accuracy on the discrimination task. Thus, the influence of task order on changes in discrimination performance may account for the difference in results between Experiment 1 and 3.

Finally, there was a difference in pre-learning discrimination performance (*d'*) between social and reward value stimuli that was not found in Experiment 1 (repeated-measures ANOVA: value, $F(1, 31) = 6.91$, $p = .013$; condition, $F(1, 31) = 1.49$, $p = .231$; interaction, $F(1, 31) = 0.62$, $p = .439$; paired t-test reward vs. social: faces, $t(31) = 1.79$, $p = .083$; flowers, $t(31) = 0.49$, $p = .631$; Fig. 4.S7). At baseline, participants were better at discriminating reward face pairs compared to social pairs in Experiment 3 but not Experiment 1. Because of this it may have been harder to learn the values associated social pairs (exemplified by significantly lower learning accuracy for social value compared to reward value), and consequently there were no robust changes in face discrimination performance due to the social values.

Apart from the main examination of discrimination performance, the additional preference ratings allowed us to examine how people used the learned value information in making preference judgments of each stimulus. In particular, we were interested in whether and how the learned value information may generalize to expected behavior in a separate social context. We found that across subjects, social and reward values were factored into preferences for faces in the value learning task. People preferred faces with high values the most, and low values the least, with no difference in preference for faces with one high and one low value. On the other hand, social values did not influence preferences for flowers during learning. Low reward stimuli were least preferred, and high reward stimuli were most preferred, regardless of their associated social values. This pattern was also found in future ratings of the flowers.

At an individual-level, while preferences of flowers were driven by their associated reward values in both current and future ratings, the influence of value type on preferences for faces depended on the task context. Future preferences of faces were more influenced by low social values compared to preferences for those stimuli during learning. Interestingly, ratings of low social value faces were lower in a future context, suggesting an avoidance of faces that were least generous in the learning task. Similarly, individual differences in value accuracy during learning were related to sensitivity to each value type in the ratings of flowers regardless of task context, while this relationship was only found for current, and not future, face preferences. Individual differences in discrimination changes as a function of value type were not related to value sensitivity in the preference ratings of either faces or flowers. This is potentially due to the lack of influence of social values on discrimination changes.

5. General Discussion

In the present study, we examined whether learning to associate social trait information with faces influenced perception. We employed a paradigm in which social categories were not initially derived from physical face appearance, but instead were learned through an interactive task. Moreover, we examined effects of learning on performance in a perceptual discrimination task that did not explicitly require any of the learned information.

Experiment 1 revealed that learned social categories systematically influenced the discrimination of faces as well as flowers. Importantly, this finding could not be explained by differences in learning accuracy or pre-learning discrimination performance. Individual differences in learning were related to the magnitude of this effect for faces but not for flowers, such that the better people were at learning the social categories compared to reward categories, the greater the change in their face discrimination performance. A one-month follow-up discrimination test indicated that the effect of social values on face perception had attenuated over time. Experiment 2 confirmed that the main results in Experiment 1 were not due to a training effect, or within-category modulations based on the reward values. Experiment 3 intended to replicate the results of Experiment 1 using a within-subjects instead of a between-subjects design; however, unlike in Experiment 1, there was no influence of social categories on face or flower discrimination performance. Several differences between the experiments could account for this, including lower learning accuracy for social values compared to reward values in Experiment 3, an effect of task order on changes in discrimination, and significant differences in pre-learning discrimination performance that were not found in Experiment 1. Aside from this, participants' explicit ratings revealed that they used the social and reward categories to guide their preferences for interacting with the faces; whereas only reward categories influenced preferences of flowers. There was also a tendency to avoid low generosity faces in the context of being paired with those players in a future, cooperative task, with lower preference ratings compared to the context of the current study. Together, these findings indicate that social category knowledge can bias face perception, suggestive of a re-organization of perceptual space such that this information becomes incorporated into the representational structure. Notably, these biases were sensitive to learning performance in the faces condition, indicating that this effect was only found for participants who attended to and learned the associated social categories.

Our results show how perceptual face spaces can by dynamically modulated by learning social traits, and such modulations are directly related to an individual's learning behavior. Moreover, the findings support the notion that social categories can automatically influence face perception, even when people are not actively categorizing the faces, though such effects are not persistent over a long-term delay. The current study suggests that learned social information can influence perceptual face representations, supporting the notion that both abstract and perceptual information can be linked in a common underlying representational space, particularly when people learn associations between physical resemblances and category structure.

Overall changes in discrimination performance were comparable between faces and flowers. Although social traits are characteristic of animate items, people learned the social values associated with flowers, and these associations influenced subsequent perceptual performance (except for Experiment 3 participants who completed the flowers condition before faces). We are certainly not the first to show that people can attribute social properties to objects (for review, see Scholl & Tremoulet, 2000); the novelty of these results instead is that learned social associations can influence perception of those objects. These findings suggest that categorical perception effects are a domain-general perceptual learning mechanism. Regardless of the type of stimulus, perceptual spaces are warped by learned categorical groups that are relevant for behavior.

Notably, however, there was no relationship between a participant's learning behavior and discrimination changes for flowers as a function of value type. This suggests that while social values influenced perceptual flower spaces, changes in these spaces were not sensitive to differences in learning accuracy between social and reward values. In other words, both social and reward categories became incorporated in the organization of flower space, but the structure of flower space was not modulated in a manner that was sensitive to learning behavior, as was found for face space. Additionally, participants did not use the social values in their preference ratings of flowers, showing that while there was an overall influence of social traits on perception, the traits did not affect these explicit social judgments. Thus, while there were influences of learned values in both faces and flowers conditions, there were key differences in how learning was related to perceptual space modulations as well as social preference ratings.

An important difference between the design of the present study and former work on social vision is that previous studies have focused on first impressions of a face. Visual appearance is especially relevant when characterizing unfamiliar faces in order to classify them as belonging to meaningful social groups based on the information available to the perceiver (physical features). Here, we have people learn associations between physical face features and social trait categories, and in this way the faces become familiarized. A central goal of our study was to examine how social categorizations can influence face perception, but how familiar faces are recognized and categorized may be different from the mechanisms involved in perceiving and categorizing unfamiliar faces.

One limitation of this study is that social traits in the real-world do not correlate with facial features like they do in the present study. We do not interpret our results as evidence that people who are similarly generous will appear physically similar to a perceiver. Instead, we designed the study in this way so that we could disentangle perceptual similarity and category similarity. This paradigm allowed the examination of whether social category learning could warp perceptual spaces. We show that when people learn relationships between facial features and social categories, there are related changes in how well those features can be perceptually discriminated. This provides evidence that social concepts can influence perceptual descriptions and resulting discrimination performance, as has been shown in the categorical perception literature. Another limitation is that the study focuses on the single social trait of generosity. While this allowed a careful examination of how category learning can modulate face space, impressions are often based on multiple competing and often overlapping social categories (Freeman et al., 2012; Stolier et al., 2018). Additional studies are needed to better establish the conditions under which conceptual knowledge may penetrate perceptual processes particularly in the case of social information associated with individual people.

In sum, the present findings show how learning social trait categories can implicitly modulate face perception. The re-organization of perceptual face spaces by such categories is directly tied to an individual's learning behavior, and this learned information generalizes to social expectations in a future context.

6. Supplemental Methods

6.1 Sample size estimation

An initial power analysis (alpha = .01, power = 95%) testing for a difference between two independent means (based on the categorical perception effect in Folstein, Gauthier, & Palmeri, 2012) indicated a sample size of 16 per condition (faces/flowers), or 32 total participants. This estimated *N* also allowed us to properly counterbalance the assignment of condition (2) and stimulus groupings (2). Thus, in each experiment we had 32 participants per condition.

6.2 Exclusion criteria

6.2.1 Experiment 1. 52 participants were recruited to participate in the faces condition, and 49 were recruited to the flowers condition. 14 in the faces condition, and 11 in the flowers condition were excluded for having a category learning accuracy lower than 70% for both value types. An additional 6 in the faces condition and 6 in the flowers condition were excluded for having less than 70% accuracy on either the pre- or post-learning discrimination tasks.

6.2.2 Experiment 2. 47 participants were recruited to participate in the faces condition, and 55 were recruited to participate in the flowers condition. 9 participants in the faces condition, and 8 in the flowers condition, were excluded for not reaching an accuracy of 70% on the category learning task for both value types. Additionally, 6 participants in the faces condition, and 15 in the flowers condition were excluded for having less than 70% accuracy on either pre- or post-learning discrimination tasks.

6.2.3 Experiment 3. 50 participants were recruited to participate. 14 were excluded for having a category learning accuracy lower than 70% for both value types, and 4 were excluded for having discrimination performance below 60% pre- or post-learning, for either the faces or flowers condition.

6.2 Experiment distribution

All experiments were developed using PsychoPy3 and conducted online using the Pavlovia.org platform.

6.3 Instructions

6.3.1 Same-different discrimination (All experiments)

In this task you will see two [faces/flowers], presented one after another at the center of the screen. Your task is to report whether the two [faces/flowers] were the same [faces/flowers], or different [faces/flowers]. To respond "same", press the F-key. To respond "different", press the J-key. You will have a maximum of 5 seconds to respond. Please try to be as fast and accurate as possible.

6.3.2 Category learning

6.3.2.1 Experiments 1 & 3: Faces

Now you will play a game where you learn about 4 people. Each person made a series of choices about how to divide a pool of points between themselves and a future person (i.e. you).

For each decision, we made a different pool of points available to each person. Then they chose how many points to keep for themselves, and how many to donate to you. On average, some people had larger point pools than others to work with (e.g. 10-400 points). Also, some people tended to donate more of their points than others on average (e.g. 20-80% of their point pools).

Your job is to maximize the points you receive from the players. How much you get depends on the points available to the person, and how much they chose to share with you. Remember, you will have to learn which people had more points than others to give, and which people gave more than others on average, in order to maximize your received points.

On each round, you will see two people on the screen, your task is to choose who to play with. Press the F-key to choose the person on the LEFT. Press the J-key to choose the person on the RIGHT. You will only have 5 seconds to respond, you must respond in that time for your answer to be recorded.

After you make a choice, you will see two pieces of information: (A) how many points they gave you on that round (labeled "Shared") (B) the total pool of points they had on that round (labeled "Out of"). The points they gave you are then added to your total number of points. Your goal is to maximize the total points you receive, by learning who had more points than others, and who donated more points than others.

6.3.2.2 Experiments 1 & 3: Flowers

Now you will play a game where you learn about point allocations associated with 4 flowers.

Each flower has an average point pool, either a small or large number of points (e.g. 10-400 points), and each flower donates either a small or large proportion of their point pool to you on average (e.g. 20-80%). For example, some flowers have large point pools, but only donate small proportions of the points, others have small point pools and donate large proportions.

Your job is to maximize the points you receive from the flowers. How much you get depends on the points available to the flower, and how much they share with you. Remember, you will have to learn which flowers have more points than others to give, and which flowers have more than others on average, in order to maximize your received points.

On each round, you will see two flowers on the screen, your task is to choose one to play with. Press the F-key to choose the flower on the LEFT. Press the J-key to choose the flower on the RIGHT. You will only have 5 seconds to respond, you must respond in that time for your answer to be recorded.

After you make a choice, you will see two pieces of information:(A) how many points it gave you on that round (labeled "Shared"). (B) the total pool of points it had on that round (labeled "Out of"). The points it gave you are then added to your total number of points. Your goal is to maximize the total points you receive, by learning which flowers have more points than others, and which donate more points than others.

6.3.2.3 Experiment 2

Now you will play a game where you learn about 4 [faces/flowers].

You will see two [faces/flowers] at a time and you can choose to play with one of the two. Each [person/flower] will give different amounts of points to you. Some [people/flowers] will always give you a lot of points (75 on average), and some will always give you a small amount of points (15 on average). Your job is to maximize the total points you're receiving.

For each choice, you can play with the [face/flower] on the left, or on the right. To select the [face/flower] on the LEFT, press the F-key. To select the [face/flower] on the RIGHT, press the J-key. You will have a maximum of 5 seconds to respond and potentially gain points.

After making a selection, the amounts for both [faces/flowers] are shown under their pictures. You will only receive the point amount for the [face/flower] you chose.

6.4 Stimuli

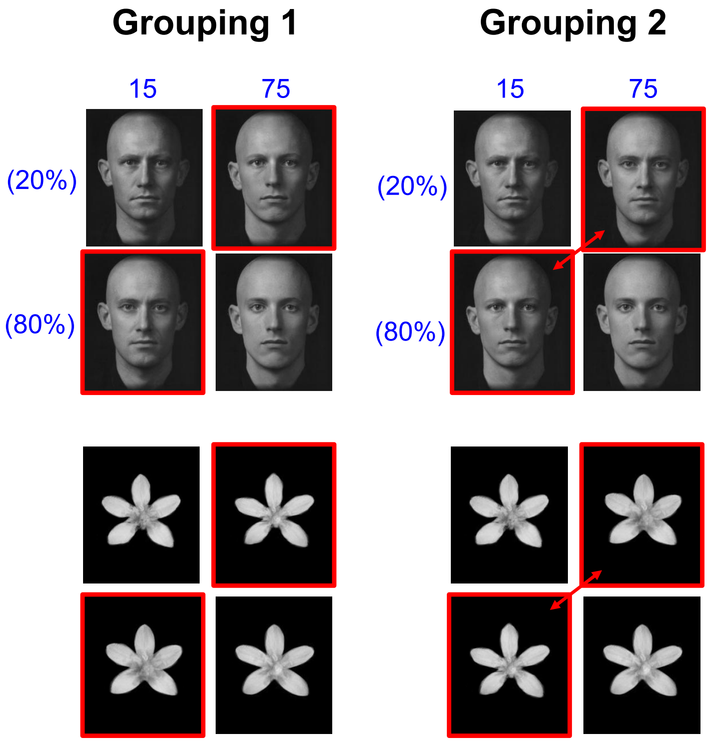


Figure 4.S1

Two groupings for each condition depending on assignment of values to stimuli in all experiments (social and reward values in Experiments 1 & 3, reward values in Experiment 2). Half of the participants received one grouping and half received the other (*N*=16 for each, within a condition). All results were combined across these groupings.

6.5 Data cleaning

For all tasks, trials in which no response was made were removed from all analyses. The percentage of trials with no response was low across subjects and experiments (Table S1).

6.6 Data analysis

6.6.1 Discrimination performance. To account for extreme hit rate and false alarm values when calculating *d'*, rates of 0 were replaced with $0.5 \div n$ and rates of 1 replaced with (*n* – 0.5) ÷ *n* where *n* is the number of signal or noise trials (Macmillan & Kaplan, 1985).

		Perceptual discrimination (pre-learning)	**Category Learning**	**Perceptual discrimination (post-learning)**
Experiment 1	*Faces*	0.04 (0.04)	0.6 (0.2)	0.04 (0.04)
	Flowers	0.2 (0.1)	0.4 (0.1)	0.04 (0.04)
Experiment 2	*Faces*	1 (0.7)	0.6 (0.3)	0.2 (0.2)
	Flowers	0.1 (0.1)	0.4 (0.2)	0
Experiment 3	*Faces*	0.07 (.04)	0.7 (0.3)	0.09 (0.06)
	Flowers	0.6 (0.5)	0.6 (0.1)	0.3 (0.1)

Table 4.S9

Average percent of trials with no response across subjects for each task and condition separately, *SEM* in parentheses.

References

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, 98(1), 185-199.

7. Supplemental Results

7.1 Experiment 1

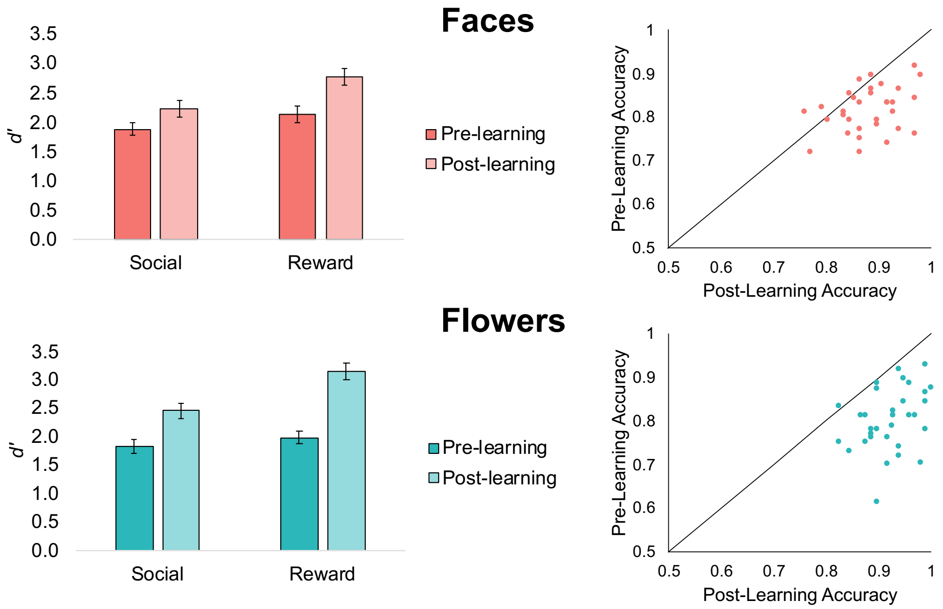


Figure 4.S2

Discrimination performance (*d'*; left) and accuracy across all trials (right; each point represents one participant) for the same-different discrimination task in Experiment 1. Error bars (left) represent +/- SEM. Solid line (right) indicates no change in pre- and post-learning accuracy. *N* = 32 per condition.

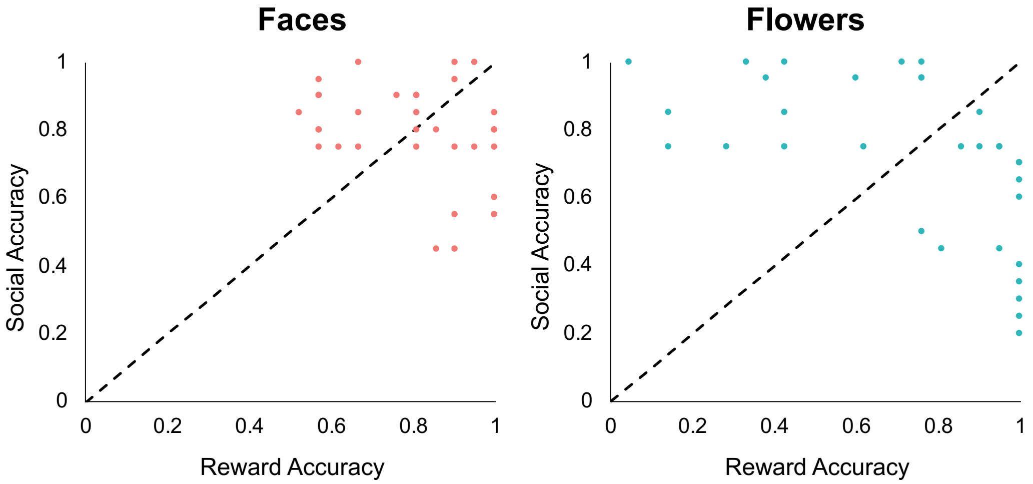


Figure 4.S3

Social and reward value accuracy during category learning in Experiment 1. Plots show individual performance for each value type, each point represents one participant, dashed line indicates equal accuracy for both values. *N* = 32 per condition.

7.1.1 Additional individual differences analysis

As the *d'* sensitivity measure collapses across both value types, it does not differentiate between the directionality of changes for each value dimension separately. For example, a positive *d'* sensitivity score could result from a greater increase in social discrimination performance than reward discriminations, or a greater decrease in reward discrimination performance compared to social discriminations. We determined whether this effect was driven by differences in Δ *d'* for social discriminations in the faces condition.

Participants were grouped according to learning sensitivity; if they had a positive sensitivity they were assigned to the Social group, if negative they were assigned to the Reward group (Social *N* = 18, Reward *N* = 14). Examining Δ *d'* by group confirmed that there was a significant increase in discrimination for social-relevant pairs due to value learning in the Social group (*M* = 0.45, *SEM* = 0.2; t-test against 0: *t*(17) = 2.32, *p* = .033) and not the Reward group (*M* = 0.19, *SEM* = 0.23; *t*(13) = 0.84, *p* = .415; Fig. S4). Reward-relevant discriminations were better after learning in both groups (Social: *M* = 0.56, *SEM* = 0.14, *t*(17) = 4.15, *p* = .001; Reward: *M* = 0.75, *SEM* = 0.27; *t*(13) = 2.78, *p* = .016). A mixed ANOVA showed a significant effect of value type (*F*(1, 30) = 6.19, *p* = .019) and a marginally significant interaction between value type and group (*F*(1, 30) = 2.88, *p* = .1), with no effect of group (*F*(1, 30) = 0.02, *p* = .893). These results suggest that while all participants got better at discriminating reward-relevant face pairs, only participants with higher social value accuracy than reward value accuracy got better at discriminating social-relevant pairs. In other words, the relationship between learning sensitivity and *d'* sensitivity is primarily accounted for by individual differences in perceptual changes associated with learned social values.

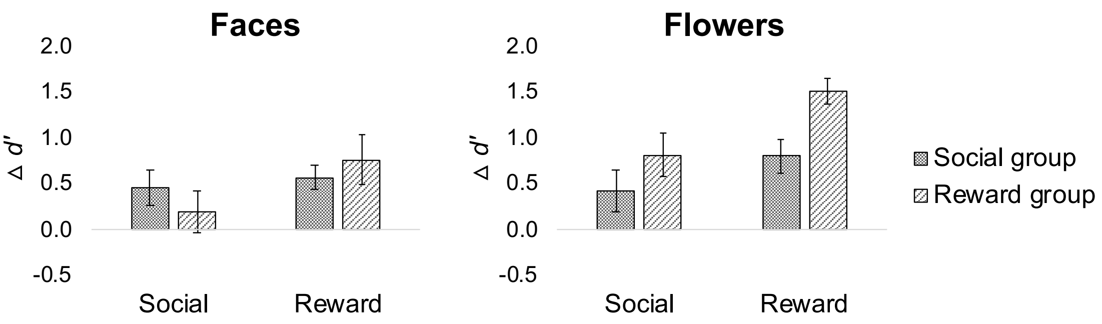


Figure 4.S4

Change in discrimination performance for Social and Reward groups in Experiment 2. Participants were grouped based on whether they had higher social or reward value accuracy on the value learning task (Social $N = 18$, Reward, $N = 14$). A difference in performance changes between groups was found only for social value-relevant discriminations. Error bars represent +/- *SEM*.

7.2 Experiment 2

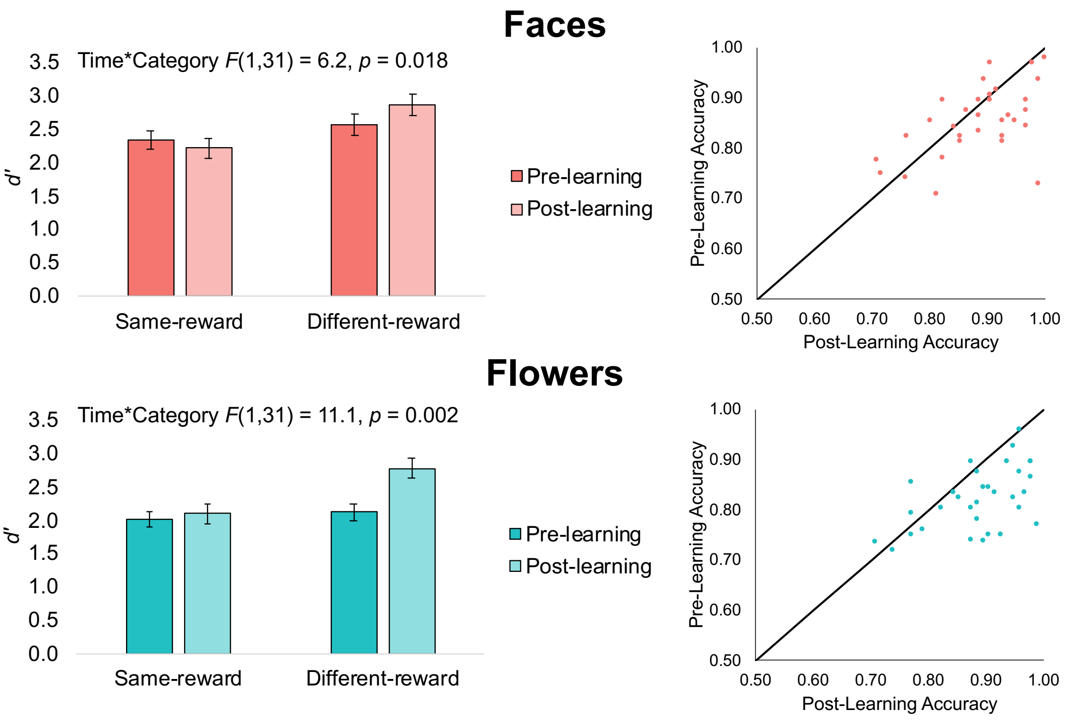


Figure 4.S5

Discrimination performance (*d'*; left) and accuracy across all trials (right; each point represents one participant) for the same-different discrimination task in Experiment 2. Error bars (left) represent +/- *SEM*. Solid line (right) indicates no change in pre- and post-learning accuracy. $N = 32$ per condition.

7.3 Experiment 3

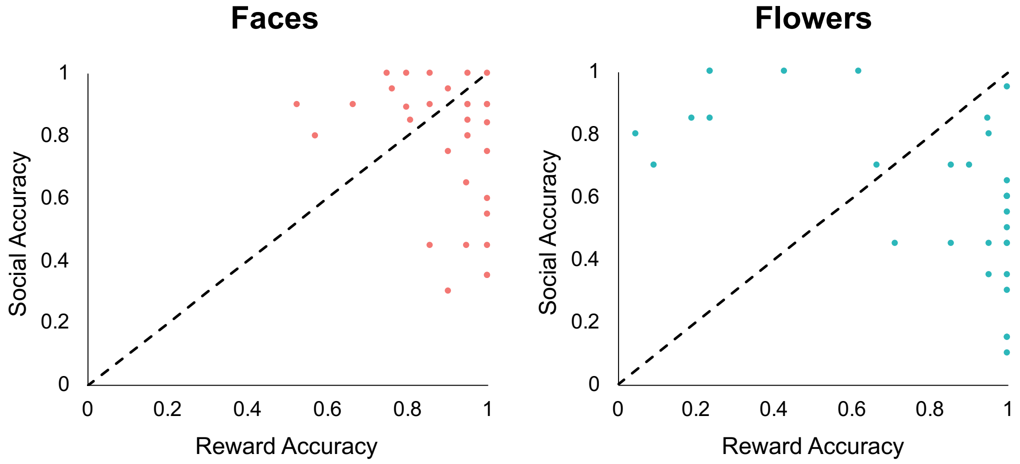


Figure 4.S6

Social and reward value accuracy during value learning in Experiment 3. Plots show individual performance for each value type, each point represents one participant, dashed line indicates equal accuracy for both values. *N* = 32.

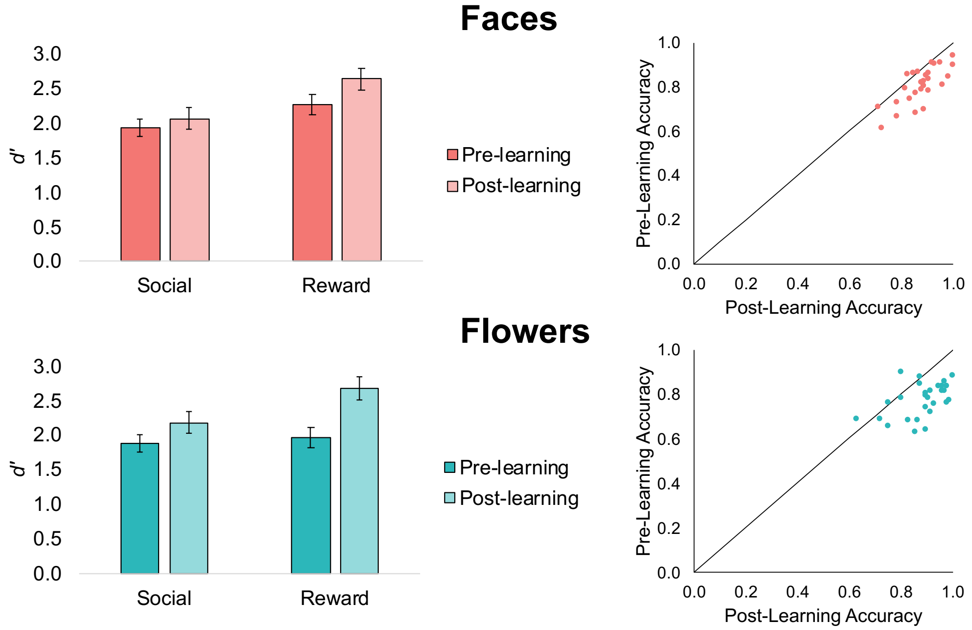


Figure 4.S7

Discrimination performance (*d'*; left) and accuracy across all trials (right; each point represents one participant) for the same-different discrimination task in Experiment 3. Error bars (left) represent +/- *SEM*. Solid line (right) indicates no change in pre- and post-learning accuracy. *N* = 32.

7.4 Additional statistics

Significant effects displayed in bold.

7.4.1 Experiment 1

Condition	Value	
Faces	Social	**_t_(31) = 11.52, _p_ < .0001**
	Reward	**_t_(31) = 10.32, _p_ < .0001**
Flowers	Social	**_t_(31) = 5.09, _p_ < .0001**
	Reward	**_t_(31) = 4.06, _p_ < .0001**

Table 4.S2

T-tests comparing category learning accuracy for each value type against chance performance (50%).

Condition	Value	
Faces	Social	**_t_(31) = 2.29, _p_ = .029**
	Reward	**_t_(31) = 4.63, _p_ < .0001**
Flowers	Social	**_t_(31) = 3.8, _p_ = .001**
	Reward	**_t_(31) = 9.06, _p_ < .0001**

Table 4.S3

T-tests comparing Δ d' against 0 (no change).

Test	Condition	Measure	
	Faces	Social	*M* = 0.47, *SEM* = 0.24
		Reward	*M* = 0.64, *SEM* = 0.19
	Flowers	Social	*M* = 0.75, *SEM* = 0.32
		Reward	*M* = 1.21, *SEM* = 0.17
T-test vs. 0	*Faces*	Social	*t*(15) = 1.98, *p* = .066
		Reward	***t*(15) = 3.36, *p* = .004**
	Flowers	Social	***t*(9) = 2.3, *p* = .047**
		Reward	***t*(9) = 6.97, *p* < .0001**
Paired t-test	*Faces*	Social vs. Reward	*t*(15) = 0.78, *p* = .437
	Flowers	Social vs. Reward	*t*(9) = 1.96, *p* = .081
T-test	*Faces vs. Flowers*	Social	*t*(24) = 0.68, *p* = .501
	Faces vs. Flowers	Reward	*t*(24) = 2.02, *p* = .054
Mixed ANOVA		Value	*F*(1, 24) = 3.74, *p* = .065
		Condition	*F*(1, 24) = 1.93, *p* = .178
		Value*Condition	*F*(1, 24) = 0.81, *p* = .378

Table 4.S4

Session 1 results group that returned after ~1 month for a second session.

Test	Condition	Measure	
	Faces	Social	*M* = 0.13, *SEM* = 0.28
		Reward	*M* = 0.59, *SEM* = 0.23
	Flowers	Social	*M* = 0.67, *SEM* = 0.33
		Reward	*M* = 1.06, *SEM* = 0.26
Repeated-Measures ANOVA	*Faces*	Value	*F*(1, 15) = 1.9, *p* = .188
		Session	*F*(1, 15) = 1.8, *p* = .199
		Value*Session	*F*(1, 15) = 1.09, *p* = .312
Paired T-Test	*Session1 vs. Session2*	Social	*t*(15) = 1.61, *p* = .129
	Session1 vs. Session2	Reward	*t*(15) = 0.27, *p* = .794
Repeated-Measures ANOVA	*Flowers*	Value	***F*(1, 9) = 7.6, *p* = .022**
		Session	*F*(1, 9) = 0.35, *p* = .569
		Value*Session	*F*(1, 9) = 0.06, *p* = .806
Paired T-Test	*Session1 vs. Session2*	Social	*t*(9) = 0.26, *p* = .801
	Session1 vs. Session2	Reward	*t*(9) = 0.76, *p* = .47

Table 4.S5

Session 2 results for group that returned after ~1 month for a second session.

7.4.2 Experiment 2

Condition	Value	
Faces	Same-reward	*t*(31) = -0.75, *p* = .458
	Different-reward	***t*(31) = 2.36, *p* = .025**
Flowers	Same-reward	*t*(31) = 0.43, *p* = .668
	Different-reward	***t*(31) = 4.12, *p* < .0001**

Table 4.S6

T-tests comparing ∆ d’ against 0 (no change).

7.4.3 Experiment 3

Condition	**Value**	
Faces	Reward	***t*(31) = 17.53, *p* < .0001**
	Social	***t*(31) = 7.08, *p* < .0001**
Flowers	Reward	***t*(31) = 4.86, *p* < .0001**
	Social	***t*(31) = 2.5, *p* = .018**

Table 4.S7

T-tests comparing learning accuracy against chance performance (50%).

Condition	**Effect**	
Faces	Value	***F*(1, 30) = 5.81, *p* = .022**
	Order	*F*(1, 30) = 0.39, *p* = .539
	Value*Order	*F*(1, 30) = 0.79, *p* = .38
Flowers	Value	*F*(1, 30) = 3.09, *p* = .089
	Order	*F*(1, 30) = 0.03, *p* = .865
	Value*Order	*F*(1, 30) = 0.21, *p* = .652

Table 4.S8

Mixed ANOVAs for learning accuracy (order (2) x value type (2)) within conditions.

Condition	**Value**	
Faces	Social	*t*(31) = 1.05, *p* = .302
	Reward	***t*(31) = 2.93, *p* = .006**
Flowers	Social	*t*(31) = 1.67, *p* = .105
	Reward	***t*(31) = 4.98, *p* < .0001**

Table 4.S9

T-tests comparing Δ d’ against 0 (no change).

Condition	**Effect**	
Flowers	Value	***F*(1, 30) = 4.36, *p* = .045**
	Order	*F*(1, 30) = 0.36, *p* = .555
	Value*Order	*F*(1, 30) = 0.48, *p* = .496

Table 4.S10

Mixed ANOVA for Δ *d'* (order (2) x value type (2)).

Condition	**Order**	**Value**	
Flowers	Faces first	Social	*t*(30) = 1.59, *p* = .132
		Reward	***t*(30) = 3.4, *p* = .004**
	Flowers first	Social	*t*(30) = 0.67, *p* = .514
		Reward	***t*(30) = 3.54, *p* = .003**

Table 4.S11

T-tests comparing Δ d' against 0 (faces first vs. flowers first).

Condition	Time	Value assignment	*M (SEM)*
Faces	Current	Low-social/Low-reward	2.71 (0.25)
		Low-social/High-reward	4.58 (0.27)
		High-social/Low-reward	3.84 (0.32)
		High-social/High-reward	5.58 (0.25)
	Future	Low-social/Low-reward	2.39 (0.22)
		Low-social/High-reward	4.13 (0.31)
		High-social/Low-reward	3.87 (0.31)
		High-social/High-reward	5.48 (0.27)
Flowers	Current	Low-social/Low-reward	3.36 (0.31)
		Low-social/High-reward	4.87 (0.2)
		High-social/Low-reward	3.9 (0.31)
		High-social/High-reward	5.32 (0.29)
	Future	Low-social/Low-reward	3.71 (0.31)
		Low-social/High-reward	4.74 (0.25)
		High-social/Low-reward	4.26 (0.31)
		High-social/High-reward	5.55 (0.22)

Table 4.S12

Preference ratings for each stimulus in Experiment 3 across subjects ($N = 31$). Ratings were made once in the context of the value learning task (Current), and once for preferences for a cooperative task in a hypothetical future study (Future).

Condition	Effect	
Faces	Value level * Value type	$F(1, 30) = 0.17, p = .684$
	Value level * Time	$F(1, 30) = 2.92, p = .098$
	Value type * Time	$F(1, 30) = 0.17, p = .682$
	Value level * Value type * Time	$F(1, 30) = 0, p = 1$
Flowers	Value level * Value type	$F(1, 30) = 0.07, p = .796$
	Value level * Time	$F(1, 30) = 0.44, p = .511$
	Value type * Time	$F(1, 30) = 0.47, p = .499$
	Value level * Value type * Time	$F(1, 30) = 0.75, p = .395$

Table 4.S13

Three-way repeated measures ANOVA of preference ratings.

Condition	Time	Value comparison		
Faces	Current	Low-S/Low-R	Low-S/High-R	***t*(30) = 5.36, *p* < .0001**
		Low-S/Low-R	High-S/Low-R	***t*(30) = 3.12, *p* = .004**
		Low-S/Low-R	High-S/High-R	***t*(30) = 8.45, *p* < .0001**
		Low-S/High-R	High-S/High-R	***t*(30) = 3.05, *p* =.005**
		High-S/Low-R	High-S/High-R	***t*(30) = 4.52, *p* < .0001**
	Future	Low-S/Low-R	Low-S/High-R	***t*(30) = 5.37, *p* < .0001**
		Low-S/Low-R	High-S/Low-R	***t*(30) = 4.78, *p* < .0001**
		Low-S/Low-R	High-S/High-R	***t*(30) = 9.6, *p* < .0001**
		Low-S/High-R	High-S/High-R	***t*(30) = 4.33, *p* < .0001**
		High-S/Low-R	High-S/High-R	***t*(30) = 5.2, *p* < .0001**
Flowers	Current	Low-S/Low-R	Low-S/High-R	***t*(30) = 3.8, *p* = .001**
		Low-S/Low-R	High-S/High-R	***t*(30) = 4.67, *p* < .0001**
		Low-S/High-R	High-S/Low-R	***t*(30) = -2.21, *p* = .035**
		High-S/Low-R	High-S/High-R	***t*(30) = 3.66, *p* = .001**
	Future	Low-S/Low-R	Low-S/High-R	***t*(30) = 2.65, *p* = .013**
		Low-S/Low-R	High-S/High-R	***t*(30) = 4.77, *p* < .0001**
		Low-S/High-R	High-S/High-R	***t*(30) = 2.88, *p* = .007**
		High-S/Low-R	High-S/High-R	***t*(30) = 3.35, *p* = .002**

Table 4.S14

Paired t-tests comparing preference ratings between stimuli (S: Social, R: Reward).

Condition	Value combination	
Flowers	Low-social/Low-reward	*t*(30) = 1.88, *p* = .07
	Low-social/High-reward	*t*(30) = -0.38, *p* = .707
	High-social/Low-reward	*t*(30) = 1.07, *p* = .295
	High-social/High-reward	*t*(30) = 0.75, *p* = .457

Table 4.S15

Paired t-tests comparing current versus future preference ratings for each stimulus.

V. GENERAL DISCUSSION

This dissertation presents three studies that explore how learning associations between perceptual and abstract information can warp psychological and neural representations. The main goals were to: (1) examine how learning face-value associations influenced behavioral and neural measures of face similarity; (2) test whether learned information generalized across task contexts and stimulus types; and (3) explore heterogeneity in learning, social preferences, and similarity metrics between individuals.

In each chapter a value learning task was used to have participants interactively learn associations between faces, social values, and reward values. Participants were under the impression that they were learning about actions that other people had taken in the same experiment. Remarkably, people learned social values even though they could ignore this information and still perform successfully on the learning task. Additionally, across cohorts there were individual differences in the extent to which participants were accurate for each value type during learning and how they weighted the values when making preferential judgments in the context of future social interactions.

Chapter 2 examined how learned values modulated behavioral similarity judgments and social preferences. Subjects sorted faces in a spatial similarity space and were free to use whatever information they thought relevant for the task. Before value learning perceptual (visual) face similarity predicted the organization of face spaces at a group-level. After learning, social, but not reward, values modulated the organization, such that faces that were more similar in their social values were closer together. In this way, the learned social trait of generosity selectively warped visual face spaces. The similarity of social preference ratings for faces was related to their social but not reward similarity. At a subject-level there was variance in how accurate participants were during learning for choosing faces based on each value type, and how much they relied on social and/or reward values in their similarity judgments as well as their social preferences. These individual differences were consistent across learning, behavioral similarity, and social preferences tasks.

Chapter 3 built on these findings by investigating whether and how learned values modulated neural responses when viewing faces. Of particular interest was whether face-selective regions along the ventral visual pathway were biased by the learned value information. Social values influenced the multi-voxel activity patterns evoked by faces in an anterior ventro-medial face region of the left temporal lobe (vmATL), and the magnitude of this relationship was related to individual learning performance as a function of value type. These results support recent proposals that this anterior face-selective area is critically involved in associating person-specific information across perception and memory and speak against theories that social information is fed-back to earlier face processing regions (at least for abstract social information associated with individuals and not perceptually cued social categories). Activity patterns in a region of the left parietal cortex, previously shown to encode distance-based properties of a person's own social network, were related to an individual's behavioral similarity space after learning, suggesting this area has a role in representing socially relevant properties of familiar others that can be revealed by spatial metrics.

Chapter 4 addressed the question of whether the behavioral and neural influences of learned values found in Chapters 2 and 3 had implicit behavioral consequences in a value-irrelevant discrimination task and probed whether effects were domain-general (specific to faces or not). Learned social and reward values influenced perceptual discrimination of both faces and flowers, and there was a relationship between individual learning and discrimination performance as a function of value type in the faces condition (but not flowers). Learned social and reward value were incorporated into social preferences to interact with the faces, but only reward values influenced preferences for the flowers. This shows initial evidence that while social values are selectively used when making explicit behavioral judgements of faces (Chapter 2), implicit discrimination judgements are influenced by both social and reward values in a domain-general manner. However, the effect of learned social values on perceptual discrimination was not found after one-month and did not replicate with a within-subject design (but see notable between-experiment differences).

Together, the results of these studies suggest that representations of individual others that are initially structured by perceptual similarity can be dynamically modulated by information that is not directly available in sensory input but instead is acquired through experience. Chapters 2 and 3 provide novel evidence that learned social traits can influence behavioral and neural responses to faces, even in the presence of additional task-relevant properties (reward values), and even when performing a value-irrelevant task (Chapter 3). This indicates that models that define face representations based on physical features, such as in social trait impressions of unfamiliar faces (Oosterhof & Todorov, 2008), are not sufficient to describe representations of familiar others with whom an observer has acquired additional knowledge. As social concepts are typically more varied and interrelated than the single trait studied in the present work (Freeman & Ambady, 2009, 2011; Stolier et al., 2018, 2020), an interesting avenue for future research is to examine how learning more complex associations, such as with multi-trait spaces, could differentially warp face spaces, and how this may depend on an individual observer's prior conceptual beliefs (Stolier et al., 2018) and social context.

That the left vmATL represents social trait information about others in a task-independent manner that is directly related to an individual's learning behavior shows for the first time that this region is recruited for face-trait associations acquired through interactive experience. These results are in line with formulations of the semantic "hub" model that posit the left ATL has a role in the storage and retrieval of semantic knowledge and integrates information across modality-specific brain regions (Lambon Ralph et al., 2010; Visser et al., 2010; Rice et al., 2018). At the same time, the connectivity of the ATL to other cortical and subcortical systems has been shown to be critical for social cognition and face processing (Tsukiura et al., 2010; Unger et al., 2016; Hampton et al., 2016), and the role of the ATL as part of a wider network that supports complex social processes remains to be fully understood (Wang et al., 2018; Wang & Olson, 2018).

The present findings have broad implications for models in perception and cognition. Prior work examining conceptual knowledge has largely focused on defining fixed spaces and their underlying dimensions. For example, social perception and social cognition research has found that the dimensions that structure social knowledge are stable across observers (e.g., warmth vs. competence, trustworthiness vs. dominance) and posits that this is due to their adaptive, ecological benefits (see Introduction). However, concepts and their use are context-dependent and recent computational models have exemplified how the structure of conceptual representations can flexibly change depending on what information is relevant for a given context (Solomon et al., 2019). Future work on social perception and person knowledge should consider not only how prior knowledge can bias current perceptions and judgments, but how it can be adaptively updated by learned associations and how it may differ for familiar individuals.

Is social information special? At first glance, it may seem that the behavioral and neural results suggest a selective influence of social value information compared to reward value. At a group-level, social values influenced explicit behavioral judgements (Chapter 2) and face-specific neural activity patterns in the visual system (Chapter 3), but reward values did not. Indeed, some have argued that social information engages specialized cognitive processes and recruits neural systems that selectively give rise to social cognition (Adolphs, 2009; Brothers, 1996). Notably, in the current set of experiments, social traits are a generalizable property that are associated with particular faces, while the point rewards are task-specific, thus it can be expected that the learned social traits would be persistent in contexts beyond the experimental task. In this way, the behavioral and neural effects specific to social, and not reward, values may be explained by their utility in predicting future behavior, and not necessarily due differences in the type of information. In other words, if participants learned to associate reward properties that are generalizable instead of task-specific, such as wealth categories (e.g., rich, poor), then we may find similar behavioral and neural effects for reward information as was found with social traits.

On the other hand, one study found that learning task-specific monetary rewards associated with novel 2D objects modulates neural activity in early visual cortex (V1; Persichetti et al., 2015). Moreover, learning to group morphed car stimuli into categories defined by their visual features has been shown to influence perceptual discrimination performance as well as responses in the lateral occipital cortex (LOC; Jiang et al., 2007) and ventral visual regions (Folstein et al., 2013). Presently, both social and reward value categories affected behavioral face discrimination performance (Chapter 3), indicating that reward value categories can implicitly influence perceptual judgements. Thus, the absence of neural modulations in visual regions as a function of learned reward values of faces may not be due to generalizability. Instead, the neural code for reward value associations in the visual system may be revealed by fMRI-adaptation methods as was found in these previous studies and not multi-voxel patterns as was tested here.

5.1 Limitations

Every study has limitations, the present ones not excluded. Arguably the most critical in terms of their generalizability is that each study uses a single social trait and faces that are matched in their perceptually-related social categories (e.g., age, race, gender). While constraining the experimental design in this way allowed us to control influences of visual and categorical information on the effects of-interest, the findings cannot be directly applied to the broader space of person knowledge that contains social categories beyond the limited set used here. Utilizing representational spaces reduced in their dimensionality allowed us to tease apart modulations based on the different types of information available to the participants, but it is important to note that this is a simplified approach to studying person knowledge that realistically is richer and more complex. As mentioned above, real-world social categories are overlapping and inter-related. Thus, learning that someone is generous will likely impact impressions of other social attributes associated with that person. Additional studies are needed to explore how such multi-dimensional spaces can be influenced by learning and under what contexts.

Another limitation is introduced by using a two-dimensional space to measure behavioral similarity and orthogonalizing the perceptual and objective values. This design was critical for measuring representational changes with visual stimuli in Chapters 2 and 3 while controlling for perceptual variance. In Chapter 2, it is impossible for subjects to use all three orthogonalized perceptual, social, and reward values together in their spatial behavioral similarity judgements. While this does not impact the finding that social values selectively modulated these behavioral spaces, we cannot rule out the possibility that a task that allowed for multiple sources of information may produce different results (e.g., a different weighting of each source of information, and different corresponding brain regions). It is worth noting that multidimensional scaling (MDS) techniques have shown that a two-dimensional configuration provides a good fit to the similarity of judgments of social traits and increasing the dimensionality does not greatly improve the fit (Rosenberg et al., 1968). However, in the present case it is nonetheless impossible to utilize the three orthogonal dimensions in a two-dimensional space. Yet another limitation of the spatial sorting method is that it assumes that the similarity judgments are symmetric (the similarity of X to Y should equal to that of Y to X; Goldstone, 1994), which to my knowledge has not been tested in research on social impressions but could potentially be violated.

5.2 Conclusion

Understanding the ways in which past experience can predict present perceptions and future actions, a hallmark of human intelligence, is a widely shared goal of psychological research. As we come to learn about our world and other people our mind and brain are tuned by learned conceptual relationships, allowing for dynamic and efficient behavior as social beings. The mechanisms that allow for this are not likely restricted to the domain of faces and person-knowledge, but instead are general processes that allow us to take the plethora of information that arrives to us, parse and make meaning of it, encode it in storage, and use it to make further abstractions, inferences, and predictions.

BIBLIOGRAPHY

Adolphs, R. (2006). How do we know the minds of others? Domain-specificity, simulation, and enactive social cognition. *Brain Research*, *1079*(1), 25-35.

Adolphs, R. (2009). The social brain: neural basis of social knowledge. *Annual Review of Psychology*, *60*, 693-716.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*(4), 268-277.

Antonakis, J., & Eubanks, D. L. (2017). Looking leadership in the face. *Current Directions in Psychological Science*, *26*(3), 270-275.

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences*, *104*(46), 17948-17953.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269-278.

Barton, J. J., Press, D. Z., Keenan, J. P., & O'Connor, M. (2002). Lesions of the fusiform face area impair perception of facial configuration in prosopagnosia. *Neurology*, *58*(1), 71-78.

Berry, D. S., & McArthur, L. Z. (1986). Perceiving character in faces: the impact of age-related craniofacial changes on social perception. *Psychological Bulletin*, *100*(1), 3-18.

Borghesani, V., Narvid, J., Battistella, G., Shwe, W., Watson, C., Binney, R. J., Sturm V, Miller, Z., Mandelli, M. L., Miller, B., & Gorno-Tempini, M. L. (2019). "Looks familiar, but I do not know who she is": The role of the anterior right temporal lobe in famous face recognition. *Cortex*, *115*, 72-85.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433-436.

Brothers, L. (1996). Brain mechanisms of social cognition. *Journal of Psychopharmacology*, *10*(1), 2-8.

Brewer, M. B. (1998). Category-based vs. person-based perception in intergroup contexts. *European Review of Social Psychology*, 9(1), 77-106.

Brooks, J. A., & Freeman, J. B. (2019). Neuroimaging of person perception: A social-visual interface. *Neuroscience Letters*, *693*, 40-43.

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology*, *102*(4), 943-958.

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., & Eickhoff, S. B. (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.

Cavanagh, P. (2011). Visual cognition. *Vision research*, *51*(13), 1538-1551.

Chiou, R., & Ralph, M. A. L. (2016). Task-related dynamic division of labor between anterior temporal and lateral occipital cortices in representing object size. *Journal of Neuroscience*, *36*(17), 4662-4668.

Collins, J. A., & Olson, I. R. (2014). Beyond the FFA: the role of the ventral anterior temporal lobes in face processing. *Neuropsychologia*, *61*, 65-79.

Coutanche, M. N., Solomon, S. H., & Thompson-Schill, S. L. (2016). A meta-analysis of fMRI decoding: Quantifying influences on human visual population codes. *Neuropsychologia*, *82*, 134-141.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, *29*(3), 162-173.

Devlin, J. T., Russell, R. P., Davis, M. H., Price, C. J., Wilson, J., Moss, H. E., ... & Tyler, L. K. (2000). Susceptibility-induced loss of signal: comparing PET and fMRI on a semantic task. *Neuroimage*, *11*(6), 589-600.

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology*, *13*(4), e1005508.

Duchaine, B., & Yovel, G. (2015). A revised neural framework for face processing. *Annual Review of Vision Science*, *1*, 393-416.

Eifuku, S., Nakata, R., Sugimori, M., Ono, T., & Tamura, R. (2010). Neural correlates of associative face memory in the anterior inferior temporal cortex of monkeys. *Journal of Neuroscience*, 30(45), 15085–15096.

Eifuku, S., De Souza, W. C., Nakata, R., Ono, T., & Tamura, R. (2011). Neural representations of personally familiar and unfamiliar faces in the anterior inferior temporal cortex of monkeys. *PLoS One*, *6*(4), e18913.

Elfgren, C., van Westen, D., Passant, U., Larsson, E. M., Mannfolk, P., & Fransson, P. (2006). fMRI activity in the medial temporal lobe during famous face processing. *Neuroimage*, *30*(2), 609-616.

Fischl, B. (2012). FreeSurfer. *Neuroimage*, *62*(2), 774-781.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology*, *23*, 1-74.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77-83.

Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2012). How category learning affects object representations: not all morphspaces stretch alike. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 807-820.

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, *23*(4), 814-823.

Fox, C. J., Moon, S. Y., Iaria, G., & Barton, J. J. (2009). The correlates of subjective perception of identity and expression in the face network: an fMRI adaptation study. *Neuroimage*, *44*(2), 569-580.

Freeman, J. B., & Ambady, N. (2009). Motions of the hand expose the partial and parallel activation of stereotypes. *Psychological science*, *20*(10), 1183-1188.

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247-279.

Freeman, J., Johnson, K., Adams Jr, R., & Ambady, N. (2012). The social-sensory interface: category interactions in person perception. *Frontiers in Integrative Neuroscience*, *6*, 81.

Freeman, J. B., & Johnson, K. L. (2016). More than meets the eye: Split-second social perception. *Trends in Cognitive Sciences*, *20*(5), 362-374.

Freeman, J. B., Stolier, R. M., & Brooks, J. A. (2020). Dynamic interactive theory as a domain-general account of social perception. In *Advances in Experimental Social Psychology* (Vol. 61, pp. 237-287). Academic Press.

Frith, C. D. (2007). The social brain?. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 671-678.

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology*, *63*, 287-313.

Gainotti, G. (2007a). Different patterns of famous people recognition disorders in patients with right and left anterior temporal lesions: a systematic review. *Neuropsychologia*, *45*(8), 1591-1607.

Gainotti, G. (2007b). Face familiarity feelings, the right temporal lobe and the possible underlying neural mechanisms. *Brain Research Reviews*, *56*(1), 214-235.

Gainotti, G., & Marra, C. (2011). Differential contribution of right and left temporo-occipital and anterior temporal lesions to face recognition disorders. *Frontiers in Human Neuroscience*, *5*, 55.

Gauthier, I., Tarr, M. J., Moylan, J., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). The fusiform "face area" is part of a network that processes faces at the individual level. *Journal of Cognitive Neuroscience*, *12*(3), 495-504.

Goesaert, E., & de Beeck, H. P. O. (2013). Representations of facial identity information in the ventral visual stream investigated with multivoxel pattern analyses. *Journal of Neuroscience*, *33*(19), 8549-8558.

Goldstone, R. L. (1994a). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*(2), 125-157.

Goldstone, R. L. (1994b). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*(4), 381-386.

Goldstone, R. L. (1994c). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178-200.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27-43.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, *130*(1), 116.

Goldstone, R. L., & Son, J. Y. (2012). Similarity. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 155–176). Oxford University Press.

Goldstone, R. L., Kersten, A., & Carvalho, P. F. (2017). Categorization and Concepts. In J. Wixted (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (Vol. 3, pp. 275-317). Wiley.

Graham, J. R., Harvey, C. R., & Puri, M. (2017). A corporate beauty contest. *Management Science*, *63*(9), 3044-3056.

Grill-Spector, K., Weiner, K. S., Kay, K., & Gomez, J. (2017). The functional neuroanatomy of human face perception. *Annual Review of Vision Science*, *3*(1), 167-196.

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife*, *7*, e32962.

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

Hampton, W. H., Unger, A., Von Der Heide, R. J., & Olson, I. R. (2016). Neural connections foster social connections: a diffusion-weighted imaging study of social networks. *Social Cognitive and Affective Neuroscience*, *11*(5), 721-727.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, *7*(1), 37-53.

Hassin, R., & Trope, Y. (2000). Facing faces: studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology*, *78*(5), 837-852.

Haxby, J. V., Ungerleider, L. G., Clark, V. P., Schouten, J. L., Hoffman, E. A., & Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. *Neuron*, *22*(1), 189-199.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223-233.

Hehman, E., Sutherland, C. A., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology*, *113*(4), 513-529.

Hehman, E., Stolier, R. M., Freeman, J. B., Flake, J. K., & Xie, S. Y. (2019). Toward a comprehensive model of face impressions: What we know, what we do not, and paths forward. *Social and Personality Psychology Compass*, *13*(2), e12431.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, *36*(5), 791-804.

Hughes, B. L., Camp, N. P., Gomez, J., Natu, V. S., Grill-Spector, K., & Eberhardt, J. L. (2019). Neural adaptation to faces reveals racial outgroup homogeneity effects in early perception. *Proceedings of the National Academy of Sciences*, *116*(29), 14532-14537.

Hůla, M., & Flegr, J. (2016). What flowers do we like? The influence of shape and color on the rating of flower beauty. *PeerJ*, *4*, e2106.

Hung, J., Wang, X., Wang, X., & Bi, Y. (2020). Functional subdivisions in the anterior temporal lobes: a large scale meta-analytic investigation. *Neuroscience & Biobehavioral Reviews*.

Hyon, R., Kleinbaum, A. M., & Parkinson, C. (2020). Social network proximity predicts similar trajectories of psychological states: evidence from multi-voxel spatiotemporal dynamics. *NeuroImage*, *216*, 116492.

Ishai, A., Haxby, J. V., & Ungerleider, L. G. (2002). Visual imagery of famous faces: effects of memory and attention revealed by fMRI. *Neuroimage*, *17*(4), 1729-1741.

Jenkins, A. C., Karashchuk, P., Zhu, L., & Hsu, M. (2018). Predicting human behavior toward members of different social groups. *Proceedings of the National Academy of Sciences*, *115*(39), 9696-9701.

Jezzard, P., & Clare, S. (1999). Sources of distortion in functional MRI data. *Human Brain Mapping*, *8*(2-3), 80-85.

Jiang, X., Bradley, E., Rini, R. A., Zeffiro, T., VanMeter, J., & Riesenhuber, M. (2007). Categorization training results in shape-and category-selective human neural plasticity. *Neuron*, *53*(6), 891-903.

Jonas, J., Jacques, C., Liu-Shuang, J., Brissart, H., Colnat-Coulbois, S., Maillard, L., & Rossion, B. (2016). A face-selective ventral occipito-temporal map of the human brain with intracerebral potentials. *Proceedings of the National Academy of Sciences*, *113*(28), E4088-E4097.

Kahnt, T., Park, S. Q., Haynes, J. D., & Tobler, P. N. (2014). Disentangling neural representations of value and salience in the human brain. *Proceedings of the National Academy of Sciences*, *111*(13), 5000-5005.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302-4311.

Kay, J., & Hanley, J. R. (1999). Person-specific knowledge and knowledge of biological categories. *Cognitive Neuropsychology*, *16*(2), 171-180.

Kayser, A. (1997). *Heads*. Abbeville Press.

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. *Cortex*, *48*(7), 805-825.

Klapwijk, A., & Van Lange, P. A. (2009). Promoting cooperation and trust in" noisy" situations: The power of generosity. *Journal of Personality and Social Psychology*, *96*(1), 83-103.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863-3868.

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences*, *104*(51), 20600-20605.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401-412.

Kubota, J. T., & Ito, T. (2017). Rapid race perception despite individuation and accuracy goals. *Social Neuroscience*, *12*(4), 468-478.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*(2), 284-308.

Lambon Ralph, M. A., Sage, K., Jones, R. W., & Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proceedings of the National Academy of Sciences*, *107*(6), 2717-2722.

Lee, D. (2008). Game theory and neural basis of social decision making. *Nature Neuroscience*, *11*(4), 404-409.

Lenz, G. S., & Lawson, C. (2011). Looking the part: Television leads less informed citizens to vote based on candidates' appearance. *American Journal of Political Science*, *55*(3), 574-589.

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition*, *43*(2), 266-282.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, 98(1), 185-199.

Maddock, R. J., Garrett, A. S., & Buonocore, M. H. (2001). Remembering familiar people: the posterior cingulate cortex and autobiographical memory retrieval. *Neuroscience*, *104*(3), 667-676.

Mahon, B. Z. (2015). The burden of embodied cognition. *Canadian Journal of Experimental Psychology*, *69*(2), 172-178.

Mason, M. F., Cloutier, J., & Macrae, C. N. (2006). On construing others: Category and stereotype activation from facial cues. *Social Cognition*, *24*(5), 540-562.

Matheson, H. E., & Barsalou, L. W. (2018). Embodiment and grounding in cognitive neuroscience. In J. T. Wixted (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (Vol. 3), John Wiley & Sons.

Mileva, M., Kramer, R. S., & Burton, A. M. (2019). Social evaluation of faces across gender and familiarity. *Perception*, *48*(6), 471-486.

Mitchell, J. P., Heatherton, T. F., & Macrae, C. N. (2002). Distinct neural systems subserve person and object knowledge. *Proceedings of the National Academy of Sciences*, *99*(23), 15238-15243.

Montepare, J. M., & Zebrowitz, L. A. (1998). Person perception comes of age: The salience and significance of age in social judgments. *Advances in Experimental Social Psychology*, *30*, 93-161.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289-316.

Nasr, S., & Tootell, R. B. (2012). Role of fusiform and anterior temporal cortical areas in facial recognition. *Neuroimage*, *63*(3), 1743-1753.

Nestor, A., Plaut, D. C., & Behrmann, M. (2011). Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proceedings of the National Academy of Sciences*, *108*(24), 9998-10003.

Olson, I. R., McCoy, D., Klobusicky, E., & Ross, L. A. (2013). Social cognition and the anterior temporal lobes: a review and theoretical framework. *Social Cognitive and Affective Neuroscience*, *8*(2), 123-133.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087-11092.

Oosterhof, N. N., Wiestler, T., Downing, P. E., & Diedrichsen, J. (2011). A comparison of volume-based and surface-based multi-voxel pattern analysis. *Neuroimage*, *56*(2), 593-600.

Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Frontiers in Neuroinformatics*, *10*, 27.

Parkinson, C., Liu, S., & Wheatley, T. (2014). A common cortical metric for spatial, temporal, and social distance. *Journal of Neuroscience*, *34*(5), 1979-1987.

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2017). Spontaneous neural encoding of social network position. *Nature Human Behaviour*, *1*(5), 1-7.

Peer, M., Hayman, M., Tamir, B., & Arzy, S. (2021). Brain coding of social network structure. *Journal of Neuroscience*, *41*(22), 4897-4909.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, *10*, 437-442.

Persichetti, A. S., Aguirre, G. K., & Thompson-Schill, S. L. (2015). Value is in the eye of the beholder: early visual cortex codes monetary value of objects during a diverted attention task. *Journal of Cognitive Neuroscience*, *27*(5), 893-901.

Persichetti, A. S., Denning, J. M., Gotts, S. J., & Martin, A. (2021). A data-driven functional mapping of the anterior temporal lobes. *Journal of Neuroscience,* JN-RM-0456-21.

Pinsk, M. A., Arcaro, M., Weiner, K. S., Kalkus, J. F., Inati, S. J., Gross, C. G., & Kastner, S. (2009). Neural representations of faces and body parts in macaque and human cortex: a comparative FMRI study. *Journal of Neurophysiology*, *101*(5), 2581-2600.

Predovan, D., Gandini, D., Montembeault, M., Rouleau, I., Bherer, L., Joubert, S., & Brambati, S. M. (2014). Loss of person-specific knowledge in Alzheimer's disease: Evidence from priming. *Neurocase*, *20*(3), 263-268.

Rajimehr, R., Young, J. C., & Tootell, R. B. (2009). An anterior temporal face patch in human cortex, predicted by macaque maps. *Proceedings of the National Academy of Sciences*, *106*(6), 1995-2000.

Rapcsak, S. Z. (2019). Face recognition. *Current Neurology and Neuroscience Reports*, *19*(7), 1-9.

Reggev, N., Brodie, K., Cikara, M., & Mitchell, J. P. (2020). Human face-selective cortex does not distinguish between members of a racial outgroup. *eNeuro*, *7*(3), ENEURO.0431-19.2020.

Rice, G. E., Caswell, H., Moore, P., Hoffman, P., & Lambon Ralph, M. A. (2018). The roles of left versus right anterior temporal lobes in semantic memory: a neuropsychological comparison of postsurgical temporal lobe epilepsy patients. *Cerebral Cortex*, *28*(4), 1487-1501.

Rice, G. E., Hoffman, P., Binney, R. J., & Lambon Ralph, M. A. (2018). Concrete versus abstract forms of social concept: an fMRI comparison of knowledge about people versus social terms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170136.

Rilling, J. K., & Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology*, *62*, 23-48.

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology*, *9*(4), 283-294.

Ross, L. A., & Olson, I. R. (2012). What's unique about unique entities? An fMRI investigation of the semantics of famous faces and landmarks. *Cerebral Cortex*, *22*(9), 2005-2015.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549-562.

Rule, N. O., & Ambady, N. (2011). Judgments of power from college yearbook photos and later career success. *Social Psychological and Personality Science*, *2*(2), 154-158.

Rule, N. O., Freeman, J. B., & Ambady, N. (2013). Culture in social neuroscience: a review. *Social Neuroscience*, *8*(1), 3-10.

Saad, Z. S., Reynolds, R. C., Argall, B., Japee, S., & Cox, R. W. (2004, April). SUMA: an interface for surface-based intra-and inter-subject analysis with AFNI. In *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*. IEEE.

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, *16*(2), 235-239.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299-309.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9-34.

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J., & Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, *147*(3), 293-327.

Schwartz, S. H. (1992). Universals in the content and structure of values: theory and empirical tests in 20 countries. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 1–65), Academic Press.

Sharon, T., Moscovitch, M., & Gilboa, A. (2011). Rapid neocortical acquisition of long-term arbitrary associations independent of the hippocampus. *Proceedings of the National Academy of Sciences*, *108*(3), 1146-1151.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390-398.

Skipper, L. M., Ross, L. A., & Olson, I. R. (2011). Sensory and semantic category subdivisions within the anterior temporal lobes. *Neuropsychologia*, *49*(12), 3419-3429.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, *23*, S208-S219.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, *44*(1), 83-98.

Solomon, S. H., Medaglia, J. D., & Thompson-Schill, S. L. (2019). Implementing a concept network model. *Behavior research methods*, *51*(4), 1717-1736.

Steiger, J. H. (1980). Testing pattern hypotheses on correlation matrices: Alternative statistics and some empirical results. *Multivariate Behavioral Research*, *15*(3), 335-352.

Stigliani, A., Weiner, K. S., & Grill-Spector, K. (2015). Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, *35*(36), 12412-12424.

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, *44*(1), 24-31.

Stoker, J. I., Garretsen, H., & Spreeuwers, L. J. (2016). The facial appearance of CEOs: Faces signal selection but not performance. *PloS one*, *11*(7), e0159950.

Stolier, R. M., & Freeman, J. B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. *Nature Neuroscience*, *19*(6), 795-797.

Stolier, R. M., & Freeman, J. B. (2017). A neural mechanism of social categorization. *Journal of Neuroscience*, *37*(23), 5711-5721.

Stolier, R. M., Hehman, E., & Freeman, J. B. (2018a). A dynamic structure of social trait space. *Trends in Cognitive Sciences*, *22*(3), 197-200.

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018b). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences*, *115*(37), 9210-9215.

Stolier, R. M., Hehman, E., & Freeman, J. B. (2020). Trait knowledge forms a common structure across social cognition. *Nature Human Behaviour*, *4*(4), 361-371.

Sutherland, C. A., Young, A. W., & Rhodes, G. (2017). Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties. *British Journal of Psychology*, *108*(2), 397-415.

Thompson, S. A., Graham, K. S., Williams, G., Patterson, K., Kapur, N., & Hodges, J. R. (2004). Dissociating person-specific from general semantic knowledge: roles of the left and right temporal lobes. *Neuropsychologia*, *42*(3), 359-370.

Thornton, M. A., & Mitchell, J. P. (2017). Consistent neural activity patterns represent personally familiar people. *Journal of Cognitive Neuroscience*, *29*(9), 1583-1594.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623-1626.

Todorov, A. (2008). Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Annals of the New York Academy of Sciences*, *1124*(1), 208-224.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27*(6), 813-833.

Todorov, A., & Oosterhof, N. N. (2011). Modeling social perception of faces. *IEEE Signal Processing Magazine*, *28*(2), 117-122.

Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological Science*, *25*(7), 1404-1417.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, 66, 519-545.

Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton University Press.

Tsao, D. Y., Moeller, S., & Freiwald, W. A. (2008). Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences*, *105*(49), 19514-19519.

Tsukiura, T., Mano, Y., Sekiguchi, A., Yomogida, Y., Hoshi, K., Kambara, T., Takeuchi, H., Sugiura, M., & Kawashima, R. (2010). Dissociable roles of the anterior temporal regions in successful encoding of memory for person identity information. *Journal of Cognitive Neuroscience*, *22*(10), 2226-2237.

Unger, A., Alm, K. H., Collins, J. A., O'Leary, J. M., & Olson, I. R. (2016). Variation in white matter connectivity predicts the ability to remember faces and discriminate their emotions. *Journal of the International Neuropsychological Society*, *22*(2), 180–190.

Van Bavel, J.J., Packer, D.J., & Cunningham, W.A. (2008). The neural substrates of in-group bias: a functional magnetic resonance imaging investigation. *Psychological Science*, *19*(11), 1131–1139.

Visconti di Oleggio Castello, M., Halchenko, Y. O., Guntupalli, J. S., Gors, J. D., & Gobbini, M. I. (2017). The neural representation of personally familiar and unfamiliar faces in the distributed system for face perception. *Scientific Reports*, *7*(1), 1-14.

Visser, M., Jefferies, E., & Lambon Ralph, M. A. (2010). Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. *Journal of Cognitive Neuroscience*, *22*(6), 1083-1094.

Von Der Heide, R. J., Skipper, L. M., & Olson, I. R. (2013). Anterior temporal face patches: a meta-analysis and empirical study. *Frontiers in Human Neuroscience*, *7*, 17.

Wang, Y., Collins, J. A., Koski, J., Nugiel, T., Metoki, A., & Olson, I. R. (2017). Dynamic neural architecture for social knowledge retrieval. *Proceedings of the National Academy of Sciences*, *114*(16), E3305-E3314.

Wang, Y., Metoki, A., Alm, K. H., & Olson, I. R. (2018). White matter pathways and social cognition. *Neuroscience & Biobehavioral Reviews*, *90*, 350-370.

Wang, Y., & Olson, I. R. (2018). The original social network: white matter and social cognition. *Trends in cognitive sciences*, *22*(6), 504-516.

Weiner, K. S., & Grill-Spector, K. (2010). Sparsely-distributed organization of face and limb activations in human ventral temporal cortex. *Neuroimage*, *52*(4), 1559-1573.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: the SHINE toolbox. *Behavior research methods*, *42*(3), 671-684.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, 17(7), 592-598.

Wisniewski, D., Reverberi, C., Momennejad, I., Kahnt, T., & Haynes, J. D. (2015). The role of the parietal cortex in the representation of task–reward associations. *Journal of Neuroscience*, *35*(36), 12355-12365.

Xiao, Y. J., Coppin, G., & Van Bavel, J. J. (2016). Perceiving the world through group-colored glasses: A perceptual model of intergroup relations. *Psychological Inquiry*, *27*(4), 255-274.

Zahn, R., Moll, J., Krueger, F., Huey, E. D., Garrido, G., & Grafman, J. (2007). Social concepts are represented in the superior anterior temporal cortex. *Proceedings of the National Academy of Sciences*, *104*(15), 6430-6435.

Zahn, R., Moll, J., Paiva, M., Garrido, G., Krueger, F., Huey, E. D., & Grafman, J. (2009). The neural basis of human social values: evidence from functional MRI. *Cerebral Ccortex*, *19*(2), 276-283.

Zahn, R., Green, S., Beaumont, H., Burns, A., Moll, J., Caine, D., ... & Ralph, M. A. L. (2017). Frontotemporal lobar degeneration and social behaviour: Dissociation between the knowledge of its consequences and its conceptual meaning. C*ortex*, *93*, 107-118.

Zebrowitz, L. A., & McDonald, S. M. (1991). The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior*, *15*(6), 603-623.

Zebrowitz, L. A. (2017). First impressions from faces. *Current Directions in Psychological Science*, *26*(3), 237-242.