Exploring the Structure, Dynamics, and Developmental Trajectory of Person Models

Minjae Kim

A dissertation

submitted to the Faculty of

the department of Psychology and Neuroscience

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Boston College
Morrissey College of Arts and Sciences
Graduate School

July 2023

© Copyright 2023 Minjae Kim

Examining the Structure, Dynamics, and Developmental Trajectory of Person Models

Minjae Kim

Advisor: Dr. Liane Young

Effective social interaction requires reasoning about people as generative models. In our day-to-day experience, we come across a remarkable amount of social information, often in the form of other people's behaviors. Observed behaviors are used to infer agents' unobservable mental states and traits – the latent causes that drive their behavior. These inferences are stored in person models, which allow us to interpret patterns of observed behaviors across multiple instances and contexts by attributing a common cause to those behaviors, and also allow us to predict people's future actions, so that we may navigate interactions smoothly and choose our social partners wisely. This dissertation pursued several open questions on flexible trait reasoning. In Paper 1, we found that the relative contributions of different traits to overall impressions may vary depending on what we know about a person. In Paper 2, we found increased neural activity in Theory of Mind regions following the violation of strong and positive prior impressions. In Paper 3, we found that 6-9-year-olds exhibit a negativity bias in impression updating, and older children are sensitive to the strength of behavioral evidence. Overall, we found evidence for flexible trait reasoning – both children and adults were sensitive to the strength and valence of available behavioral evidence, and to the overall inference context. These studies help shed light on how children and adults reason about person models and respond to new social information, and we suggest multiple avenues for further research in this arena.

NOTE:
Consistent with the Psychology and Neuroscience Department's "Three-Paper" option, portions of this dissertation are taken directly from three manuscripts which were published or under review during my time at Boston College (Kim, Young, & Anzellotti, 2022; Kim, Mende-Siedlecki, Anzellotti, & Young, 2021; Kim, Young, & McAuliffe, in prep).

Table of Contents

Acknowledgments

None of this work would have been possible without the incredible people I have been lucky enough to meet throughout this journey. Thank you to my dissertation committee, Liane Young, Stefano Anzellotti, Katie McAuliffe, and Peter Mende-Siedlecki – you have all modeled such a high standard for me in terms of what it means to be a great mentor and scientist, and I'd be thrilled if I can replicate a fraction of what you've done for me if I have students of my own. I am especially grateful for your unsparing kindness and empathy, which have kept me afloat through the uncertain seas of grad school. Thank you to all of my labmates, past and present, for making the lab such a joyful place to be, and for always being community oriented – I'm grateful we've been able to lean on each other. Thanks especially to Lily Tsoi, Ryan McManus, and Isaac Handley-Miner, who have been there for me through both the highs and lows of being a grad student and who've also patiently helped me when I wanted to seek advice or reassurance; and many thanks to Josh Hirschfeld-Kroen, Aditi Kodipady, Kevin Jiang, Sunny Liu, Nathan Liang, and Lizy Szanton, who make the lab run in both visible and invisible ways. Thanks to the Cooperation Lab, who taught me a ton and supported me through my first developmental study. Thank you to my RAs, Emma Alai, Arturo Balaguer, Catherine Kim, Mariel Kronitz, Mookie Manalili, Emma Sansom, Maria Noyes, Kayla Carew, Jordyn Mason, Skylar Hou, Christian Hamilton, and Stella Si – you're truly the ones making this work possible, and I learn more from you than you will know. Thank you to my collaborators, BoKyung Park, Melisa Kumar, Simon Karg, Panagiotis Mitkidis, Jordan Theriault, Tony Chen, Kevin Jiang, Justin Martin, Ryan Daley, Elizabeth Kensinger, and Sunny Liu – you are all so patient and brilliant, and make me so excited about science. Thank you to my former mentors, Dan Grodner and Gina Kuperberg, for instilling in me a love for research and providing me with the tools to navigate this career. My lifelong thanks to my partner Chris for being my biggest supporter and best friend, my person. Thank you to the rest of my family, my dog, and my friends – your continuous care and encouragement (and tail wags) over the past half a decade have gotten me to the finish line. Finally, to Liane: it has been the privilege of my life to know you and learn from you. You are so generous and brilliant and kind, and you've shown up for me in so many ways over the years – I can never thank you enough.

General Introduction

Effective social interaction requires reasoning about people as generative models. In our day-to-day experience, we come across a remarkable amount of social information, often in the form of other people's behaviors. Behaviors may be observed directly, via sensory input, or indirectly (e.g., by partaking in gossip or reading the news). Observed behaviors, in turn, are used to infer agents' unobservable mental states and traits – the latent causes that drive their behavior. These inferences can include: subjective states (e.g, she's happy); beliefs (e.g., she thinks the doll is in the basket); intentions (e.g., she intends to share with me); and traits (e.g., she's conscientious). These inferences are stored in person models, which are defined by a few key evaluative dimensions, and contain information about the causal links between mental states, traits, and actions (Anzellotti & Young, 2019; Tamir & Thornton, 2018). Importantly, person models allow us to interpret patterns of observed behaviors across multiple instances and contexts by attributing a common cause to those behaviors, and they allow us to predict people's future actions, so that we may navigate interactions smoothly and choose our social partners wisely.

Traits are a particularly important part of person models, because they vary across individuals but tend to remain relatively stable within an individual over time (Allport & Odbert, 1936). The usefulness of trait inference lies in its generalizability – if we relied solely on perception but not inference, we would only predict a narrow range of future behaviors; in contrast, a trait inference (e.g., she's untrustworthy) allows us to form expectations about people's actions and reactions in situations that may have never arisen before.

Furthermore, although we depend on traits to make useful predictions, we sometimes come across information that contradicts them, resulting in prediction error (PE – the difference between expectation and observation). Prediction errors are a signal that our beliefs about someone's traits – our person model – may need to be updated (Bach & Schenke, 2017). Belief updating is closely linked to causal attribution (Heider, 1958; Reeder & Brewer, 1979) and mentalizing or Theory of Mind (ToM; Koster-Hale & Saxe, 2013): mental state inferences (e.g., of innocent or harmful intent) tend to dominate our explanations of why someone did what they did (Malle, 2001), and these inferences can either reconcile the discrepancy between our prior beliefs and the surprising behavior, or support impression updating. Past work has established asymmetries in belief updating, specifically trait or impression updating, across relationship and group boundaries: people often resist updating positive beliefs about close or ingroup others, and resist updating negative beliefs about outgroup others. This form of belief maintenance has been characterized as reflecting motivated cognition; however, belief maintenance may also arise from a rational procedure, where stronger (more certain) prior beliefs about targets lead us to endorse non-dispositional explanations for unexpected behaviors, alleviating the need for updating (Gershman, 2019; Kim, Park, & Young, 2020).

Past work has also revealed a negativity bias in impression formation and updating in adults: we update our beliefs more when we receive negative information than when we receive positive information, particularly in the moral domain – e.g., if a good person behaves immorally, we will change our mind about them more swiftly than if a bad person behaves morally (Baumeister et al., 2001; Skowronski & Carlston, 1989).

This bias is thought to be driven by the differential diagnosticity of moral and immoral information: both good and bad people have reasons to behave morally, but a good person will rarely behave immorally. The precedence of negative information is not a constant throughout the lifetime, however: research on trait attributions during early and middle childhood has revealed a positivity bias instead, raising questions on the trajectory and function of valence asymmetries throughout development.

There are many open questions concerning our ability to reason about people as generative models. In the following chapters, I present research that uses multiple methods (behavioral, neuroscientific, developmental) to study some of these questions. They fall under three broad aims: (1) explore the structure of trait knowledge; (2) examine how we respond to trait-inconsistent information; and (3) investigate the developmental trajectory of dynamic trait inference.

Aim 1: Explore the structure of trait knowledge.

One of the key pursuits of social cognition research is to describe the organizational structure of our person models. How do our minds represent hundreds of possible mental states and traits in an efficient way? What are the dimensions that are prioritized during person perception, and does this change depending on how much we already know about a person? We start to examine these questions in Paper 1 (Kim, Young, & Anzellotti, 2022).

Past work has often used PCA to identify lower-dimensional 'maps' that we use to represent our inferences of other people. For instance, Thornton & Mitchell (2017) reduced 13 trait dimensions from extant models of person perception into 3 components, labeled sociality, valence, and power. This model has been well-validated across

paradigms, but a limitation of the PCA approach is that when a participant is asked to evaluate a person along a specific trait, that judgment does not contain information about the importance of that trait for the overall impression of that person. In addition, the structure of trait space may change depending on how much we know about a person – e.g., the traits that are most important for our impressions of famous people may be different from the ones that are important when thinking of unfamiliar people.

To gauge the relative importance of different traits for overall impressions of people, we conducted a representational similarity analysis, where we predicted pairwise holistic similarity ratings using pairwise trait distances. This allowed us to implicitly assess how perceivers prioritize multiple trait judgments to form overall representations of people. In addition, we conducted analyses on a set of famous people, and on a set of unfamiliar people that were described as performing a single behavior. We predicted that the dimensionality of trait space would be flexible, and adjust to the amount of information we have about a person.

Aim 2: Examine how we respond to trait-inconsistent information.

Past work in social neuroscience has shown that the neural regions for ToM often encode social prediction errors. These regions are recruited more for behaviors that violate (vs. confirm) prior expectations based on: past behavior, instructed trait knowledge, stereotypes, and reward feedback in economic games. In addition, computational neuroimaging work has shown that ToM regions track model-derived PE during learning about agents' trustworthiness (Behrens et al., 2008) and generosity (Hackel, Doll, & Amodio, 2015; Stanley, 2016). When close friends or ingroup others are the targets of inference, however, ToM activity is sometimes enhanced following their

bad behaviors, and at other times reduced (Kim, Park, & Young, 2020). To account for this puzzle, we have proposed that two different mechanisms may result in the maintenance of strong beliefs in light of surprising behaviors: in one case, the violation of prior beliefs elicits enhanced ToM activity, which supports the search for a coherent explanation of the unpredicted information. Alternatively, prior-inconsistent behavior may lead to reduced ToM activity, due to motivated disengagement from mentalizing, eliminating the need to reconcile the new information with prior beliefs.

As strong prior beliefs often co-occur with strong social motivation (such as the desire to maintain positive impressions of ingroup members), it is important that we examine the effect of prior knowledge on belief updating and ToM activity in isolation. In Paper 2 (Kim, Mende-Siedlecki, Anzellotti, & Young, 2021), we accomplished this by manipulating the strength and valence of participants' impressions of fictional targets. Participants learned about targets who first performed 2 or 4 positive or negative behaviors (leading to weak or strong initial impressions), then performed 2 behaviors of the opposite valence (eliciting an impression update). We hypothesized that, when targets engaged in trait-inconsistent behaviors, there would be increased ToM activity following the violation of strong (vs. weak) prior impressions, and following the violation of positive (vs. negative) prior impressions, consistent with the idea that, absent social motivation, more surprising and diagnostic social information should drive ToM activity.

Aim 3: Investigate the developmental trajectory of dynamic trait inference.

Past research on trait reasoning in children has documented a positivity bias that emerges by age 3, peaks during middle childhood, and begins to dissipate by age 10 (Boseovski, 2010). This is in contrast to the general negativity bias that characterizes

early development (Vaish, Grossmann, & Woodward, 2008), and the negativity bias (at least in the moral domain) that dominates impression formation and updating in adults (Baumeister et al., 2001).

Relatively little work has examined impression updating in children. A positivity bias in impression updating during middle childhood may be especially helpful for several reasons, including: (1) it may promote self-esteem in children who have to learn a lot of new skills, and who may be discouraged if they receive negative feedback; and (2) it may allow children to give others the benefit of the doubt if they do wrong, so that friendships may flourish despite setbacks. On the other hand, prioritizing negative information may help children keep track of potentially threatening actors, and be judicious in whom they approach or cooperate with.

In Paper 3 (Kim, Young, & McAuliffe, in prep), we examined the features that matter for children's impression updating, by manipulating the valence and strength of participants' impressions of fictional targets. In addition, we probed children's ability to use trait inferences along one dimension (niceness) to make behavioral predictions along another dimension (trustworthiness). We conducted the study in 6-9-year-olds, in order to potentially capture the positivity bias, and perhaps a transition point in the ability to differentiate between weak and strong impressions. We predicted that (1) children will update more in light of positive vs. negative behavior information, (2) older children will be more sensitive to the strength of the behavioral evidence, and (3) children will use evaluations of niceness to predict future trustworthiness.

1.0 EXPLORING THE REPRESENTATIONAL STRUCTURE OF TRAIT KNOWLEDGE USING PERCEIVED SIMILARITY JUDGMENTS

1.1 INTRODUCTION

How is our knowledge of other people organized? According to dimensional theories of social cognition, our knowledge of others' psychological characteristics, such as mental states and traits, can be represented by coordinates within a space defined by multiple evaluative dimensions (Cuddy, Fiske, & Glick, 2008; Bach & Schenke, 2017; Tamir & Thornton, 2018; Thornton et al., 2019). For instance, while faces can elicit many different trait inferences, variance in face-based trait inference is well-described by two underlying dimensions, called 'valence' (approximated by judgments of trustworthiness) and 'dominance' (approximated by judgments of social dominance; Oosterhof & Todorov, 2008). Traits form a particularly important part of person knowledge: they are inferred characteristics that vary across individuals but remain relatively stable within an individual over time (Allport & Odbert, 1936). As such, trait knowledge enables perceivers to tailor an understanding of behaviors to specific individuals, and generate predictions about possible future actions and reactions across contexts (Gerstenberg et al., 2018; Kryven et al., 2016; Wu et al., 2018; see Bach & Schenke, 2017 and Tamir & Thornton, 2018 for reviews on the use of social knowledge for prediction). For instance, the position of a face within the 'valence-dominance' space described above can be used to accurately predict threat evaluations, which have adaptive significance (Oosterhof & Todorov, 2008). Understanding the representational structure of trait knowledge, then, is key to understanding how people interpret and predict behavior.

A large body of psychological research has sought to identify the underlying dimensions that capture perceivers' trait knowledge of others. Thornton and Mitchell (2018) describe 4 such dimensional theories of person perception that have been influential in the literature: (1) the 5-factor model of personality, which consists of openness, conscientiousness, extraversion, agreeableness, and neuroticism (Goldberg, 1990; McCrae & Costa, 1987); (2) the stereotype content model, which consists of warmth and competence (Fiske et al., 2002); (3) the 2-factor model of mind perception, which consists of agency and experience (Gray et al., 2007); and (4) the 2-factor model of face perception, which consists of trustworthiness and dominance (Oosterhof & Todorov, 2008). Each of these theories was originally developed to account for specific phenomena: judgments of trait terms, intergroup affect, mind attribution, and face evaluation, respectively.

These theories have been tested in a common framework by harnessing the multidimensionality of fMRI data. Thornton and Mitchell (2018) scanned participants while they made social judgments (e.g., "loves to solve difficult problems"; "enjoys spending time in nature") about famous people that had been selected to span a variety of traits. Neural pattern responses to famous people in this task were predicted by each of the 4 aforementioned theories of person perception (Goldberg, 1990; McCrae & Costa, 1987; Fiske et al., 2002; Gray et al., 2007; Oosterhof & Todorov, 2008), and by a 3-factor synthetic model, produced by applying principal component analysis to the 4 extant theories. In addition, the 3-factor synthetic model outperformed all 4 extant theories in neural pattern reconstruction. These findings show that (1) dimensional theories of social cognition may partially describe the informational basis of mentalizing; (2) these theories

can generalize beyond their original contexts (of personality, intergroup affect, mind attribution, and face evaluation); and (3) pooling dimensions across inference contexts allows researchers to capture a greater proportion of the reliable variance in neural responses to famous people. In all, extant dimensional theories of person perception seem to be viable accounts of how perceivers represent other people during mentalizing.

Despite extensive previous research on the structure of trait knowledge, the importance of each individual trait in determining *overall impressions* of others is not as well understood. In addition, the traits that play a predominant role in determining overall impressions of famous or familiar people may be different from the traits that are fundamental for overall impressions of unfamiliar people. Previous fMRI studies have revealed that distinct brain regions are engaged in the representation of famous, familiar, and unfamiliar individuals (Gorno-Tempini & Price, 2001; Grabowski et al., 2001; Ramon & Gobbini 2018), suggesting the possibility that representations of famous people are organized differently than representations of unfamiliar people.

It is difficult to investigate the importance of different traits for overall representations of people using fMRI responses alone. Readout mechanisms are needed to convert neural representations of traits into behavioral judgments (Pagan et al., 2016; Park et al., 2014). As such, even if a dimension explains a large amount of variance in neural responses to people, it may still contribute to a lesser degree to behavioral judgments of people. Behavioral studies can therefore make unique contributions to the investigation of the structure of person representations.

Previous behavioral studies have largely relied on principal component analysis (PCA) to identify the key dimensions that capture variance in trait judgments. PCA is a

simple and elegant technique that identifies dimensions that account for most variance in a dataset, and as such effectively uncovers a "compressed" description of the dataset. It has been used successfully to identify lower-dimensional representational spaces that capture variance in perceivers' judgments of people along a set of specified traits (McCrae & Costa, 1987; Thornton & Mitchell, 2018). However, there is no guarantee that the dimensions that explain the most variance across trait judgments (and by extension, the *traits* that best approximate the content of these dimensions) *also* contribute the most to overall representations of people. When a participant is asked to evaluate a person along a specific trait (e.g., "how open to experience is this person?"), such judgments do not carry information about the importance of that particular trait for the overall representation of that person. To examine the importance of different traits for overall representations, we surveyed how traits judgments contribute to perceived similarity judgments between target people.

Using perceived similarity to characterize trait knowledge. In this study, we investigated the importance of 13 different traits in determining: (1) overall representations of famous people, and (2) overall representations of unfamiliar people who were described as performing a single behavior. Specifically, we aimed to identify the traits that contribute most to perceived similarity ratings between pairs of target people (collected by asking: "how similar are these two people?").

The perceived similarity approach has previously been used to test the 'summed state' hypothesis of person representations (Thornton et al., 2019): Thornton and colleagues showed that both perceived similarity ratings and neural pattern similarities were better predicted by a model that reflects how frequently targets experience mental

states, rather than by an optimized model of *traits*. The trait model, however, was still a robust predictor of similarity, and explained unique variance beyond the summed state model, indicating that traits still play a significant role in person representation.

In the current work, we examined the contributions of 13 traits (collated from extant theories of person perception by Thornton & Mitchell, 2018) to overall representations of people. To do this, we tested whether pairwise differences between targets along individual traits (i.e., trait distances) can predict pairwise holistic similarity ratings. For example, if inferences of *openness to experience* are important in determining overall representations of people, then the distance between two targets in terms of openness ratings should predict how (dis)similar the two targets are rated to be overall. Importantly, surveying how trait distances predict holistic similarity is a way to implicitly assess how perceivers prioritize and integrate multiple trait judgments to form overall representations. Additionally, the traits that perform best in predicting holistic similarity may not necessarily be ones that have traditionally been considered together; that is, the top-performing traits may cut across different theories that have been proposed for specific contexts of social inference.

Trait knowledge across inference context. We have discussed previous work that investigated representations of famous people (Thornton & Mitchell, 2018). Other studies have tested how we update representations of unfamiliar people, given information about their behaviors (e.g., Kim et al., 2021; Mende-Siedlecki et al., 2013). These paradigms involve different kinds of inference, and may elicit different person representations. When participants make social judgments about a famous person, they might draw on behavioral observations across different contexts. They might also have additional

knowledge about them acquired through language (e.g., by reading a newspaper article). By contrast, participants exposed to an unfamiliar person described as performing a single behavior have access to impoverished information for trait inferences, and they may represent that person differently.

In addition, the dimensionality of person representations could itself change as a function of the amount and type of evidence available. A higher-dimensional representation would require estimating a larger number of coordinates, and thus would require a correspondingly larger amount of data in order to obtain robust estimates. Considering this, the dimensionality of perceivers' representations of other people might be adaptive, adjusting optimally to the amount of information we have about a particular individual (e.g., representations of strangers may be lower-dimensional than representations of known individuals).

In order to study person representations across different inference contexts, we conducted the perceived similarity analyses on two datasets: ratings of famous people (collected by Thornton & Mitchell, 2018), and ratings of unfamiliar people who performed a single behavior. For each domain (famous people and unfamiliar people), we tested how well pairwise trait distances predict pairwise holistic similarity. We also tested whether the mappings between trait distance and holistic similarity generalized across the two domains. We found that distinct subsets of traits best predict holistic similarity between famous people vs. between unfamiliar people. However, the relationship between each trait and holistic similarity generalized to some extent from famous people to unfamiliar people, suggesting a degree of overlap in representational structures across inference contexts. As compared to trait ratings of famous people, trait ratings of

unfamiliar people were more intercorrelated, and they were largely driven by valence (positivity or negativity). However, removing the influence of valence information revealed that a reliable higher-dimensional structure is present even in first impressions.

1.2 METHODS

Open science. The data and analysis code for this project are available on the Open Science Framework (https://osf.io/kqc2h/?view_only=1e2116e04ea04b19accede3330e412ba).

Set of examined traits. Thirteen traits tested in a previous study of neural pattern activity during mentalizing (Thornton & Mitchell, 2018) were examined in the current study. Thornton and Mitchell (2018) took 11 of these from four extant theories of person knowledge and face perception: warmth and competence from the stereotype content model (Fiske et al., 2002); agency and experience from the two-factor model of mind perception (Gray et al., 2007); trustworthiness and dominance from the two-factor model of face perception (Oosterhof & Todorov, 2008); and the Big 5 personality dimensions, openness, conscientiousness, extraversion, agreeableness, and neuroticism (Goldberg, 1990; McCrae & Costa, 1987). Intelligence and attractiveness were also included for being widely discussed features in person knowledge (Thornton & Mitchell, 2018).

Trait ratings of unfamiliar people: Overview. There were two rounds of data collection for trait ratings of unfamiliar people. In the first round of data collection, participants rated a set of *nameless and faceless* target people, who were each described as performing a single behavior. While participants were instructed to give trait ratings of unfamiliar *people* based on their behaviors, participants may have instead rated the

behaviors themselves, as the targets were not highly personified. Thus, we conducted a conceptual replication study where *named and pictured* target people were described as performing a single behavior.

Following these two rounds of data collection, we assessed whether trait ratings of unfamiliar people with and without names and faces were comparable. We found that the two datasets were highly concordant (see **Results**). Therefore, for downstream data collection (of holistic similarity ratings) and analyses, we focused on unfamiliar targets *without* names and faces.

Behavior stimuli associated with unfamiliar people. Three hundred single-sentence descriptions of behaviors were taken from a previous study of neural activity during impression updating (Kim et al., 2021; stimuli adapted from Mende-Siedlecki et al., 2013). Of these, 120 behaviors were positive/moral (e.g., "spent a Saturday volunteering at a soup kitchen"), 120 were negative/immoral (e.g., "lost their temper at the barista"), and 60 were neutral/morally irrelevant (e.g., "walked down a sidewalk in town"). All behavior stimuli were pretested to verify valence (positivity or negativity) and moral relevance (Kim et al., 2021).

Trait ratings of unfamiliar people (without names and faces). Participants were recruited through Amazon Mechanical Turk (AMT) to rate a set of 60 unfamiliar people on a single trait. Five surveys were administered for each trait, to present all 300 behavior stimuli. We aimed to recruit approximately 30 participants for each of 65 surveys (13 traits * 5 surveys per trait); of the 2059 participants that were recruited in total, 74 were excluded for failing attention checks or for being non-native speakers of English,

resulting in a final sample of 1985 participants (995 female, 958 male, 6 nonbinary/other participants; age $M = 37.2$, $SD = 11.2$).

For each item, participants were asked to imagine someone who performed one behavior (e.g., "Imagine a person who spent a Saturday volunteering at a soup kitchen"). Participants then rated that person along the specified trait, on a scale from 1 to 7 (e.g., "Please rate the openness to experience of this person"). A short description of the trait was provided at the beginning of each survey (see Supplementary Materials p. 16 for full participant instructions).

Trait ratings of unfamiliar people (with names and faces). A new set of participants was recruited through AMT to rate a set of 60 unfamiliar people (30 female, 30 male) on a single trait. Five surveys were administered for each trait, to present all 300 behavior stimuli. We aimed to recruit approximately 10 participants for each of 65 surveys (13 traits * 5 surveys per trait); of 700 total participants, 46 were excluded for failing attention checks or for being non-native speakers of English, resulting in a final sample of 654 participants (298 female, 351 male, 3 nonbinary/other participants; age $M = 39.4$, $SE = 12.0$).

Each target person was given a name, and represented by a picture of an emotionally-neutral face from the Karolinska Directed Emotional Faces set (Lundqvist et al., 1998). Each person was described as performing one behavior (e.g., "Andrew spent a Saturday volunteering at a soup kitchen"). Participants were asked to rate each person on the specified trait, on a scale from 1 to 7 (e.g., "Please rate the openness to experience of this person"). A short description of the trait was provided at the beginning of each survey. Across participants, target identity was counterbalanced with behavior valence

(e.g., half of participants learned about Andrew performing a positive behavior, and half of participants learned about him performing a negative behavior).

Holistic similarity ratings for pairs of unfamiliar people. Holistic similarity ratings were collected for 900 randomly chosen pairs of unfamiliar people (out of $\binom{300}{2} = 44850$ possible pairs). As discussed above, we only collected holistic similarity ratings for unfamiliar targets *without* names and faces, because (1) the inclusion of names and faces did not impact trait ratings (see **Results**), and (2) participants may overweigh facial similarity in their holistic similarity ratings if pictures of faces are presented.

A new set of participants was recruited through AMT to rate 60 stimulus pairs. Fifteen surveys were administered to present all 900 stimulus pairs. We aimed to recruit approximately 5 participants for each survey; of 79 total participants, 4 were excluded for failing attention checks or for being non-native speakers of English, resulting in a final sample of 75 participants (38 female, 36 male, 1 nonbinary/other participants; age M = 30.1, SD = 12.8).

For each stimulus pair presented in the survey, participants were asked to imagine one person performing the first behavior, and another person performing the second behavior; then, participants rated how similar the two people are, on a scale from 0 (extremely dissimilar) to 100 (extremely similar). For example: "Imagine that one person spent a Saturday volunteering at a soup kitchen. Imagine that another person lost their temper at the barista. How similar are these two people?"

Following data collection, pairwise holistic similarity ratings were reflected, such that higher ratings indicated greater dissimilarity (distance) between the two targets.

All participants in the above studies provided informed consent and were compensated for their time; for full participant demographics please see Supplementary Materials (p. 18).

Ratings of famous people. Trait ratings and pairwise holistic similarity ratings of 60 famous people (e.g., Amelia Earhart, Bruce Lee, George W. Bush) were taken from Thornton and Mitchell (2018). Thornton and Mitchell (2018) collected ratings of the 60 targets on each of the thirteen traits from an online sample (N = 869). Each participant rated the entire set of 60 targets on a single trait. A short description of the relevant trait was provided. Participants gave their ratings on a continuous line scale from 1 to 7 with anchors appropriate to the trait. In addition, a separate set of participants gave holistic similarity ratings for every pair of targets (Thornton & Mitchell, 2018). Of the $\binom{60}{2} = 1770$ pairwise holistic similarity ratings, we randomly selected and retained 900 for further analysis in the current study, to match the number of holistic similarity ratings that were collected for unfamiliar people.

Trait distance calculation. For each stimulus pair for which we had holistic similarity ratings (900 pairs of unfamiliar people, 900 pairs of famous people), we computed 13 pairwise trait distances. Trait distance was calculated as the absolute difference between the average trait rating for one target and the average trait rating for the other target.

Predicting holistic similarity using trait distance. All analyses were conducted in R (R Core Team, 2013). For each domain (unfamiliar people and famous people), we fit 13 single-variable linear models (ordinary least-squares) to predict pairwise holistic similarity using pairwise trait distance. For example, one model predicted holistic

similarity between pairs of unfamiliar people as a function of their distance along openness. P-values for models were corrected using the Holm-Bonferroni method. For each domain, we also fit a cumulative linear model, where all 13 trait distances were used to predict pairwise holistic similarity, to explore how much of the variance in holistic similarity could be explained by extant theories of person perception. For cumulative models, partial correlations were calculated between each trait distance and holistic similarity.

For the domain of famous people, we also tested whether associations between holistic similarity and trait distance would be robust to adding biographical information as covariates. For each pair of famous people, we coded whether or not the targets shared the same gender, race, nationality, and industry (arts, athletics, business, media, politics, sciences), based on Wikipedia entries (entering NAs where information was not available). These four covariates were added to all models predicting holistic similarity based on trait distance.

Comparing within-domain and cross-domain predictive performance. For each linear model in each domain, five-fold cross-validation was used to examine within-domain predictive performance and cross-domain predictive performance. For instance, models trained on the unfamiliar people data were used to predict: (1) holistic similarity for held-out pairs of unfamiliar people (*within-domain generalization*), and (2) holistic similarity for pairs of famous people (*cross-domain generalization*).

To do this, we randomly split the rating data in each domain into five folds, iterating through each fold as the test (held-out) set. Standardization of all variables was

conducted separately for training and test sets. Linear models that were fitted to the training set in one domain were used to predict: (1) holistic similarity values in the same domain's test set, and (2) holistic similarity values in the other domain's test set. For example, one model, which regressed holistic similarity onto distance along openness, was trained on folds #1-4 of the unfamiliar people data; this model was then used to predict holistic similarity as a function of distance along openness in fold #5 of the unfamiliar people data, and in fold #5 of the famous people data.

The following measures of predictive performance were averaged across the five folds: coefficient of determination (*CoD*; calculated as 1 – Sum of Squares Error/Sum of Squares Total), root mean squared error (*RMSE*), and mean absolute error (*MAE*).

This five-fold cross-validation procedure was repeated with the cumulative models in each domain, where all 13 trait distances were used to predict holistic similarity.

These performance measures allowed us to examine: (1) the importance of different traits for explaining holistic similarity, and (2) whether there are correspondences in how traits relate to holistic similarity across inference contexts.

Correlation structures. We next examined whether the two domains – unfamiliar people and famous people – differ in terms of collinearity between trait ratings.

For the set of unfamiliar people, and for the set of famous people, we generated a correlation matrix that plotted the Pearson's correlation coefficient for all pairwise combinations of the 13 trait ratings. Then, we conducted Chi-squared tests of whether the Fisher-transformed correlation matrices were significantly different, using the cortest.mat function in R.

Reliability of correlation structures. Next, we examined the reliability of the correlation structures for the two domains, as any differences in intercorrelatedness may be due to greater noise in one dataset.

For each dataset, we randomly generated a subset of 60 stimuli (the minimum number of stimuli of any dataset), then split each subset into halves and calculated the correlation matrix for each split-half. We then computed the Kendall's tau-b coefficient between the lower triangles of the two correlation matrices, as a measure of reliability.

To test whether each observed Kendall's tau was significantly different from chance, we used permutation tests. By permuting the trait labels and recalculating Kendall's tau for each permuted dataset, we created a sampling distribution of Kendall's tau values under the null, from which a *p*-value can be derived.

To do this, after generating the random split-halves of data, we permuted the column names (trait labels) for one of the split-halves 10,000 times, then calculated the Kendall's tau between the correlation matrix for the permuted split-half and the correlation matrix for the other split-half, creating a sampling distribution of Kendall's tau values under the null (Fig. 6). Finally, we compared the observed Kendall's tau to the null distribution to produce a *p*-value. This allowed us to test how observed reliability compares to chance reliability for each dataset.

The role of valence in trait ratings of unfamiliar people. Overall, trait ratings were more intercorrelated within the unfamiliar people domain, compared to the famous people domain. To further investigate the correlation structure for trait ratings of unfamiliar people, we built 13 linear models (one for each trait) that predicted trait ratings as a function of target valence (whether the unfamiliar person performed a

positive or negative behavior). In addition, we conducted PCA on the 13 trait ratings, and examined component loadings as a function of target valence.

Correlation structures after removing valence information. It appeared that a single feature, valence, was capturing most of the variance in trait judgments for unfamiliar people. To examine whether there is a reliable structure in trait ratings of unfamiliar people even after removing valence information, we divided the trait rating data for unfamiliar people into two subsets – targets who performed positive behaviors, and targets who performed negative behaviors – then tested for reliable structure within each valence subset. As a complementary analysis, we removed the first PC from the trait rating data, then tested for remaining reliable structure. To do this, we (1) projected the trait rating data onto PC space; (2) removed the first PC by zeroing out all values; and (3) rotated the data back to their original coordinates using the transpose of the PCA rotation matrix.

Predicting holistic similarity after removing valence information. Given that valence may be driving perceptions of similarity between pairs of unfamiliar people, we tested whether pairs of unfamiliar people that are concordant in valence (i.e., both positive/negative/neutral) are associated with greater holistic similarity ratings, compared to counter-valenced pairs of unfamiliar people.

Then, to test whether trait distances can still predict holistic similarity after removing valence information, we added concordance in valence (i.e., whether two targets were of the same valence, or counter-valenced) as a covariate to each single-trait model. As a complementary analysis, we tested how well trait distances predict holistic

similarity between pairs of positive unfamiliar people (162 pairs) and between pairs of negative unfamiliar people (163 pairs).

1.3 RESULTS

Trait ratings of unfamiliar people: Comparing two datasets. We first assessed whether trait ratings of unfamiliar people with and without names and faces were comparable. We found that, for each of the 13 traits, there was a significant correlation between ratings of unnamed targets, and ratings of named targets (Fig. S1). In addition, for each of the 13 traits, there was a significant correlation between trait distances calculated for pairs of unnamed targets, and trait distances calculated for pairs of named targets (Fig. S1). Furthermore, for each dataset, we generated a correlation matrix comprised of Pearson's correlation coefficients for all pairwise combinations of the 13 trait ratings. These two correlation matrices were highly concordant with each other (Kendall's $\tau = 0.857$, $p < 0.0001$). These results suggest that further analyses conducted on these two datasets will be comparable.

Predicting holistic similarity: Within the set of unfamiliar people (without names and faces). We found that for each of the 13 traits, pairwise trait distance significantly predicted pairwise holistic similarity. For instance, if two unfamiliar people were given similar openness ratings, these targets were also perceived to be similar overall (by a separate group of participants); if two targets were given dissimilar openness ratings, they were perceived to be dissimilar overall. See Table 1 for statistics for each model, and Fig. 1a for a scatterplot of holistic similarity vs. distance along openness.

In addition, a cumulative model containing all 13 trait distances significantly predicted holistic similarity ($F(13,886) = 280.80$, $p < 0.0001$, *coefficient of determination* (*CoD*) = 0.800). See Table 2 for detailed statistics, and Fig. 1b for a scatterplot.

Predicting holistic similarity: Within the set of unfamiliar people (with names and faces). We found that for each of the 13 traits, pairwise trait distance significantly predicted pairwise holistic similarity (Table S1). A cumulative model containing all 13 trait distances significantly predicted holistic similarity as well ($F(13,886) = 267.70$, $p < 0.0001$, *CoD* = 0.792; Table S2). Thus, adding names and faces to the unfamiliar targets did not produce qualitatively different results. It is important to note, however, that we did not collect holistic similarity ratings for named targets; therefore, these models predicted holistic similarity between unnamed targets using trait distance between named targets. For this reason, in ensuing sections, we focus on discussing analyses of the unnamed target data; we note instances where these analyses were replicated on the named target data.

Predicting holistic similarity: Within the set of famous people. For each of the 13 traits, pairwise trait distance significantly predicted pairwise holistic similarity (Table 3; Fig. 1c). A cumulative model containing all 13 trait distances (Table 4; Fig. 1d) significantly predicted holistic similarity ($F(13,886) = 46.57$, $p < 0.0001$, *CoD* = 0.390).

Predicting holistic similarity: Within the set of famous people, controlling for biographical information. For each pair of famous people, we coded whether or not the targets shared the same gender, race, nationality, and industry (arts, athletics, business, media, politics, sciences). See Fig. S10 for visualizations of these pairwise biographical similarities.

We found that, for all traits other than neuroticism, pairwise trait distance significantly predicted pairwise holistic similarity, even after controlling for whether the two targets had the same gender, race, nationality, and industry (see Table S3 for statistics). Thus, associations between holistic similarity and trait distance are largely robust to controlling for biographical similarities. As biographical information does not exist for the unfamiliar targets, we focus on discussing the performance of models that do not include biographical similarities as covariates.

Figure 1. (a) Holistic similarity between pairs of unfamiliar people versus their distance along openness. Openness is being used as an illustrative example. (b) Holistic similarity between pairs of famous people versus their distance along openness. (c) Observed holistic similarity between pairs of unfamiliar people versus holistic similarity predicted by the cumulative model. (d) Observed holistic similarity between pairs of famous people versus holistic similarity predicted by the cumulative model.

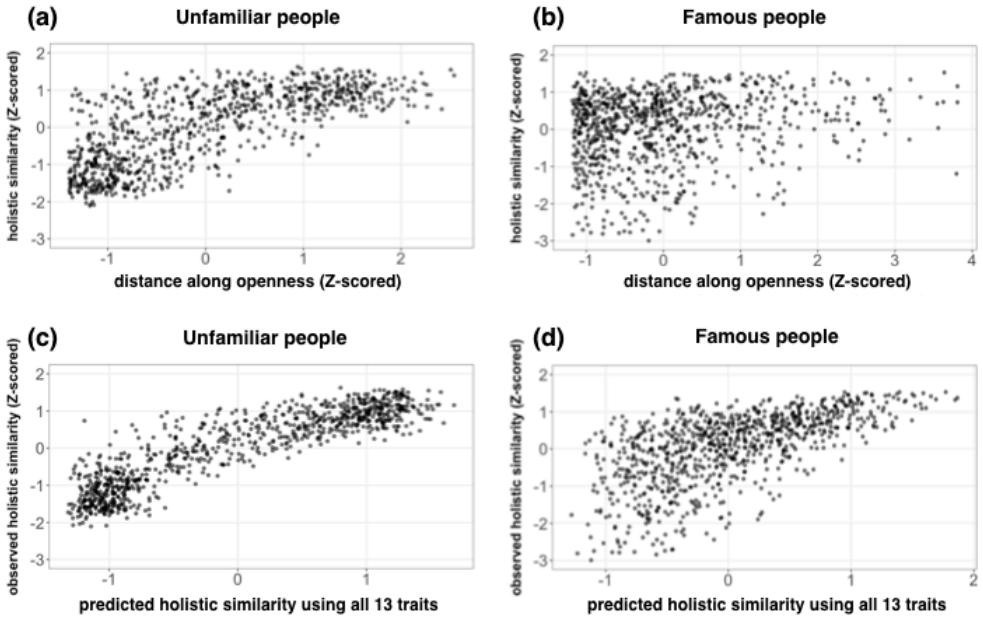


Table 1. Results from 13 linear models predicting holistic similarity between pairs of unfamiliar people, using pairwise trait distance. P-values were corrected using the Holm-Bonferroni method.

trait	theory	*b*	*SE*	*t*	*p*	adjusted *p*
openness	Big 5	0.721	0.023	31.223	1.70E-145	5.10E-145
conscientiousness	Big 5	0.871	0.016	53.230	5.42E-280	6.50E-279
extraversion	Big 5	0.078	0.033	2.354	0.019	0.019
agreeableness	Big 5	0.880	0.016	55.451	3.20E-292	4.16E-291
neuroticism	Big 5	0.842	0.018	46.714	1.41E-242	9.87E-242
dominance	face perception	0.408	0.030	13.400	1.85E-37	3.70E-37
trustworthiness	face perception	0.858	0.017	50.113	2.22E-262	2.00E-261
warmth	stereotype content model	0.866	0.017	51.981	5.37E-273	5.37E-272
competence	stereotype content model	0.844	0.018	47.218	1.50E-245	1.20E-244
agency	mind perception	0.740	0.022	32.924	1.51E-156	7.55E-156
experience	mind perception	0.735	0.023	32.475	1.24E-153	4.96E-153
intelligence	n/a	0.822	0.019	43.228	1.25E-221	7.50E-221
attractiveness	n/a	0.867	0.017	52.152	5.84E-274	6.42E-273

Table 2. Results from a cumulative linear model predicting holistic similarity between pairs of unfamiliar people, using all 13 pairwise trait distances.

variable	*b*	*SE*	*t*	*p*	partial correlation
(Intercept)	0.000	0.015	0.000	1.000	
openness	-0.009	0.033	-0.282	0.778	-0.009
conscientiousness	0.270	0.066	4.101	4.49E-05 ***	0.136
extraversion	0.052	0.016	3.301	0.001 **	0.110
agreeableness	0.235	0.099	2.361	0.018 *	0.079
neuroticism	0.091	0.046	1.979	0.048 *	0.066
dominance	0.101	0.018	5.670	1.93E-08 ***	0.187
trustworthiness	0.167	0.053	3.150	0.002 **	0.105
warmth	0.051	0.080	0.632	0.527	0.021
competence	0.129	0.061	2.111	0.035 *	0.071
agency	-0.045	0.035	-1.299	0.194	-0.044
experience	0.027	0.036	0.729	0.466	0.024
intelligence	-0.050	0.049	-1.002	0.317	-0.034
attractiveness	0.008	0.079	0.101	0.920	0.003

Table 3. Results from 13 linear models predicting holistic similarity between pairs of famous people, using pairwise trait distance. P-values were corrected using the Holm-Bonferroni method.

trait	theory	*b*	*SE*	*t*	*p*	adjusted *p*
openness	Big 5	0.202	0.033	6.180	9.70E-10	2.91E-09
conscientiousness	Big 5	0.355	0.031	11.367	4.53E-28	4.53E-27
extraversion	Big 5	0.239	0.032	7.380	3.61E-13	1.81E-12
agreeableness	Big 5	0.205	0.033	6.291	4.91E-10	1.96E-09
neuroticism	Big 5	0.124	0.033	3.759	1.82E-04	1.82E-04
dominance	face perception	0.446	0.030	14.917	4.05E-45	5.27E-44
trustworthiness	face perception	0.243	0.032	7.492	1.62E-13	1.06E-12
warmth	stereotype content model	0.154	0.033	4.671	3.46E-06	6.92E-06
competence	stereotype content model	0.335	0.031	10.642	5.47E-25	4.92E-24
agency	mind perception	0.243	0.032	7.502	1.51E-13	1.06E-12
experience	mind perception	0.392	0.031	12.751	2.39E-34	2.63E-33
intelligence	n/a	0.424	0.030	14.044	1.20E-40	1.44E-39
attractiveness	n/a	0.317	0.032	10.010	1.97E-22	1.58E-21

Table 4. Results from a cumulative linear model predicting holistic similarity between pairs of famous people, using all 13 pairwise trait distances.

variable	*b*	*SE*	*t*	*p*	partial correlation
(Intercept)	0.000	0.026	0.000	1.000	
openness	0.188	0.033	5.793	9.62E-09 ***	0.191
conscientiousness	0.145	0.047	3.083	0.002 **	0.103
extraversion	0.135	0.029	4.624	4.32E-06 ***	0.154
agreeableness	-0.105	0.080	-1.312	0.190	-0.044
neuroticism	-0.055	0.039	-1.442	0.150	-0.048
dominance	0.330	0.031	10.571	1.11E-24 ***	0.335
trustworthiness	0.100	0.057	1.747	0.081	0.059
warmth	0.029	0.058	0.491	0.624	0.016
competence	-0.329	0.068	-4.827	1.63E-06 ***	-0.160
agency	-0.188	0.043	-4.415	1.13E-05 ***	-0.147
experience	0.218	0.055	3.996	6.99E-05 ***	0.133
intelligence	0.355	0.059	6.061	2.01E-09 ***	0.200
attractiveness	0.180	0.028	6.549	9.81E-11 ***	0.215

Predictive performance: Within-domain generalization. Five-fold cross-validation was used to examine the within-domain predictive performance of models that predict holistic similarity using trait distance. Table 5 lists the cross-validated coefficient of determination (*CoD*), root mean squared error (*RMSE*), and mean absolute error (*MAE*) for each model in each domain.

We found that, for all traits other than dominance and extraversion, trait distance explained a greater proportion of variance in holistic similarity in the domain of unfamiliar people, than in the domain of famous people (see Fig. 2a-b for radar plots of performance measures). In the domain of unfamiliar people, the top-performing traits in terms of predicting holistic similarity were: agreeableness, conscientiousness, attractiveness, warmth, and trustworthiness. In the domain of famous people, the top-performing traits were: dominance, intelligence, experience, conscientiousness, and competence. In addition, the cumulative model (containing all 13 trait distances) explained a greater proportion of variance in holistic similarity in the domain of unfamiliar people, than in the domain of famous people.

These results were largely replicated when names and faces were added to the unfamiliar targets: for all traits other than dominance, trait distance explained a greater proportion of variance in holistic similarity in the domain of unfamiliar people, and the cumulative model explained a greater proportion of variance in holistic similarity in the domain of unfamiliar people (Table S5; Fig. S4).

Partial correlations in cumulative models. Fig. 3 plots, for each domain, partial correlations between each trait distance and holistic similarity, controlling for the other 12 trait distances. These partial effects within cumulative models provide one way to

evaluate the relative importance of different traits for perceived similarity. In the domain of unfamiliar people, the traits with the largest partial effects were: dominance, conscientiousness, extraversion, trustworthiness, and agreeableness. In the domain of famous people, they were: dominance, attractiveness, intelligence, openness, and extraversion.

When names and faces were added to the unfamiliar targets, three of the partial effects changed in significance (Table S2 and Fig. S2; trustworthiness became nonsignificant, while warmth and agency became significant). When biographical similarity was controlled for among famous targets, four of the partial effects changed in significance (Table S4 and Fig. S3; conscientiousness became nonsignificant, while agreeableness, neuroticism, and trustworthiness became significant).

The reported partial effects indicate the unique contributions of each trait to holistic similarity, over and above the other traits; however, these partial effects depend on the particular set of 13 traits that were tested. Thus, when evaluating the relative importance of different traits for perceived similarity, we will focus on the predictive performance of single-trait models.

Table 5. Within-domain predictive performance of models predicting pairwise holistic similarity using pairwise trait distance. Five-fold cross-validation was used to calculate performance measures. The bottom row reports performance for the cumulative model.

trait	*CoD*: unfamiliar people	*CoD*: famous people	*RMSE*: unfamiliar people	*RMSE*: famous people	*MAE*: unfamiliar people	*MAE*: famous people
openness	0.522	0.040	0.689	0.977	0.558	0.778
conscientiousness	0.759	0.122	0.488	0.934	0.385	0.734
extraversion	0.005	0.056	0.995	0.969	0.879	0.770
agreeableness	0.775	0.039	0.473	0.977	0.372	0.782
neuroticism	0.709	0.015	0.538	0.990	0.433	0.788
dominance	0.168	0.197	0.909	0.894	0.777	0.694
trustworthiness	0.737	0.057	0.511	0.968	0.409	0.767
warmth	0.751	0.022	0.497	0.986	0.392	0.786
competence	0.713	0.106	0.534	0.942	0.424	0.743
agency	0.547	0.056	0.670	0.969	0.539	0.770
experience	0.541	0.149	0.676	0.920	0.552	0.719
intelligence	0.675	0.175	0.568	0.905	0.454	0.704
attractiveness	0.752	0.098	0.496	0.947	0.393	0.745
all 13	0.800	0.390	0.446	0.778	0.352	0.602

Figure 2. Performance measures for models predicting holistic similarity, visualized as radar plots. (a) CoD values by trait, within the domain of unfamiliar people and within the domain of famous people. (b) RMSE values by trait, within the domain of unfamiliar people and within the domain of famous people. (c) Cross-domain CoD values between observed and predicted holistic similarity values. (d) Cross-domain RMSE values between observed and predicted holistic similarity values.

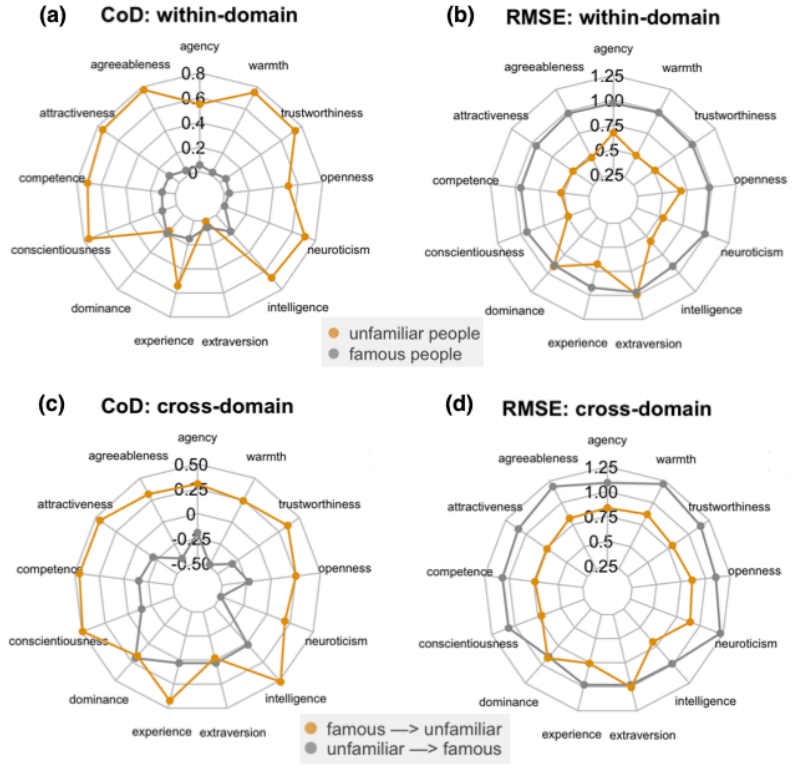
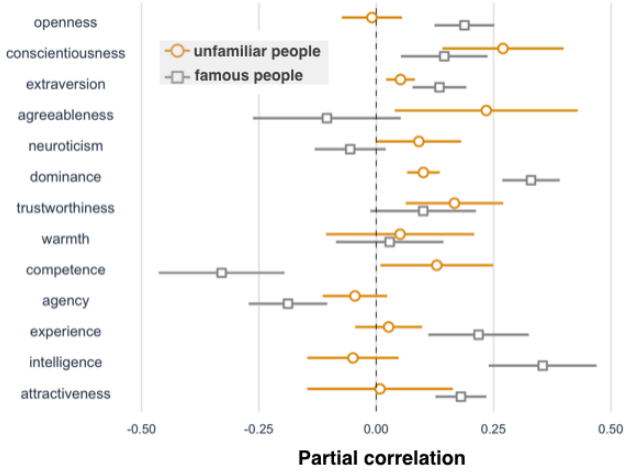


Figure 3. Partial correlations between each trait distance and holistic similarity, controlling for the other 12 trait distances, in (1) the domain of unfamiliar people, and (2) the domain of famous people.



Predictive performance: Cross-domain generalization. Five-fold cross-validation was used to examine the cross-domain predictive performance of models that predict holistic similarity using trait distance. See Table 6 for cross-validated performance measures for each model in each domain, and Fig. 4 for plots of predicted vs. observed holistic similarity.

We found that, for all traits other than dominance and extraversion, the mapping between trait distance and holistic similarity generalized better from a training set of famous people to a testing set of unfamiliar people, than from a training set of unfamiliar people to a testing set of famous people (Fig. 2c-2d; Table 6). In addition, the cumulative (13-trait) model generalized better from a training set of famous people to a testing set of unfamiliar people, than vice versa (Table 6).

The above results were largely replicated when names and faces were added to the unfamiliar targets: for all traits other than dominance, models generalized better from the famous people data to the unfamiliar people data, and the cumulative model generalized better from the famous people data to the unfamiliar people data (Table S6; Fig. S4).

Figure 4. (a) Observed holistic similarity between pairs of unfamiliar people versus holistic similarity predicted by a single-trait model trained on famous people. (b) Observed holistic similarity between pairs of famous people versus holistic similarity predicted by a single-trait model trained on unfamiliar people. (c) Observed holistic similarity between pairs of unfamiliar people versus holistic similarity predicted by a 13-trait model trained on famous people. (d) Observed holistic similarity between pairs of famous people versus holistic similarity predicted by a 13-trait model trained on unfamiliar people.

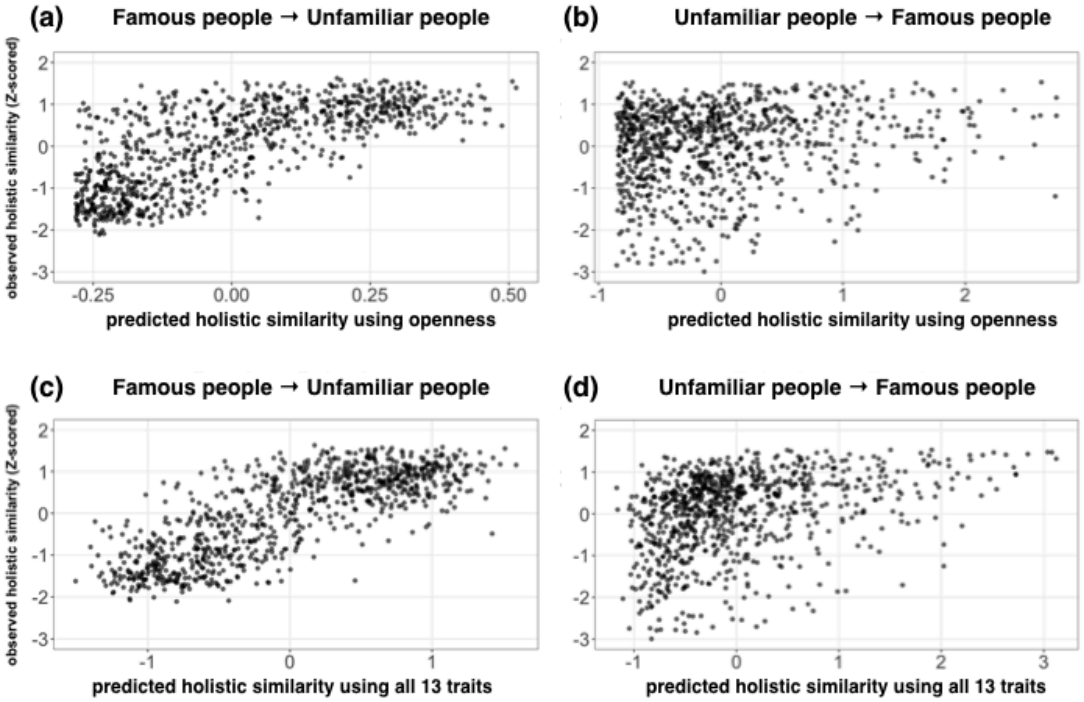


Table 6. Cross-domain predictive performance of models predicting pairwise holistic similarity using trait distance. Five-fold cross-validation was used to calculate performance measures. A negative coefficient of determination indicates poorer prediction than the mean value. The bottom row reports performance for the cumulative model.

trait	*CoD*: unfamiliar people → famous people	*CoD*: famous people → unfamiliar people	*RMSE*: unfamiliar people → famous people	*RMSE*: famous people → unfamiliar people	*MAE*: unfamiliar people → famous people	*MAE*: famous people → unfamiliar people
openness	-0.227	0.251	1.104	0.863	0.900	0.752
conscientiousness	-0.147	0.492	1.067	0.711	0.869	0.604
extraversion	0.031	-0.021	0.982	1.008	0.778	0.883
agreeableness	-0.414	0.319	1.185	0.823	0.980	0.715
neuroticism	-0.497	0.194	1.220	0.895	1.003	0.784
dominance	0.196	0.167	0.894	0.910	0.695	0.774
trustworthiness	-0.323	0.357	1.146	0.799	0.918	0.693
warmth	-0.485	0.243	1.215	0.868	0.999	0.760
competence	-0.152	0.452	1.068	0.738	0.856	0.629
agency	-0.192	0.300	1.088	0.835	0.885	0.727
experience	0.031	0.422	0.980	0.758	0.782	0.649
intelligence	0.017	0.517	0.987	0.693	0.784	0.584
attractiveness	-0.204	0.449	1.093	0.740	0.888	0.635
all 13	-0.034	0.620	1.013	0.615	0.823	0.495

Accounting for differences in generalization. The asymmetry in cross-domain generalization was pronounced. Thirteen of the models trained on the unfamiliar people data (all models except the dominance model) produced negative or near-zero coefficients of determination when predicting holistic similarity for famous people (range of *CoD*s: -0.497 to 0.031; Table 6, column 1). That is, most models exhibited poorer prediction performance than a model that just predicts the mean value for holistic similarity. In contrast, thirteen of the models trained on the famous people data (all models except the extraversion model) produced positive coefficients of determination

when predicting holistic similarity for unfamiliar people—here, model predictions performed better than a model that just predicts the mean value (range of *CoD*s: 0.167 to 0.620; Table 6, column 2). This asymmetry in prediction accuracy is consistent with the presence of an asymmetry in the dimensionality of the two representational spaces.

If we had observed that prediction accuracy was at or below chance for both directions of generalization (unfamiliar-to-famous and famous-to-unfamiliar), we would not have been able to infer that the two representational spaces have dimensions in common. However, we instead observed that nearly all models that were trained on the famous people data were able to predict holistic similarity between unfamiliar people with above-chance accuracy. For this reason, we hypothesize that the representational space of famous people includes dimensions from the representational space of unfamiliar people, *as well as* other dimensions—i.e., famous people may be represented in a higher-dimensional space than unfamiliar people. This could account for the asymmetry where generalization from famous people to unfamiliar people is more successful than generalization from unfamiliar people to famous people: inferring a higher-dimensional space from a lower-dimensional (perhaps one-dimensional) space may pose a more challenging prediction task compared to the opposite direction.

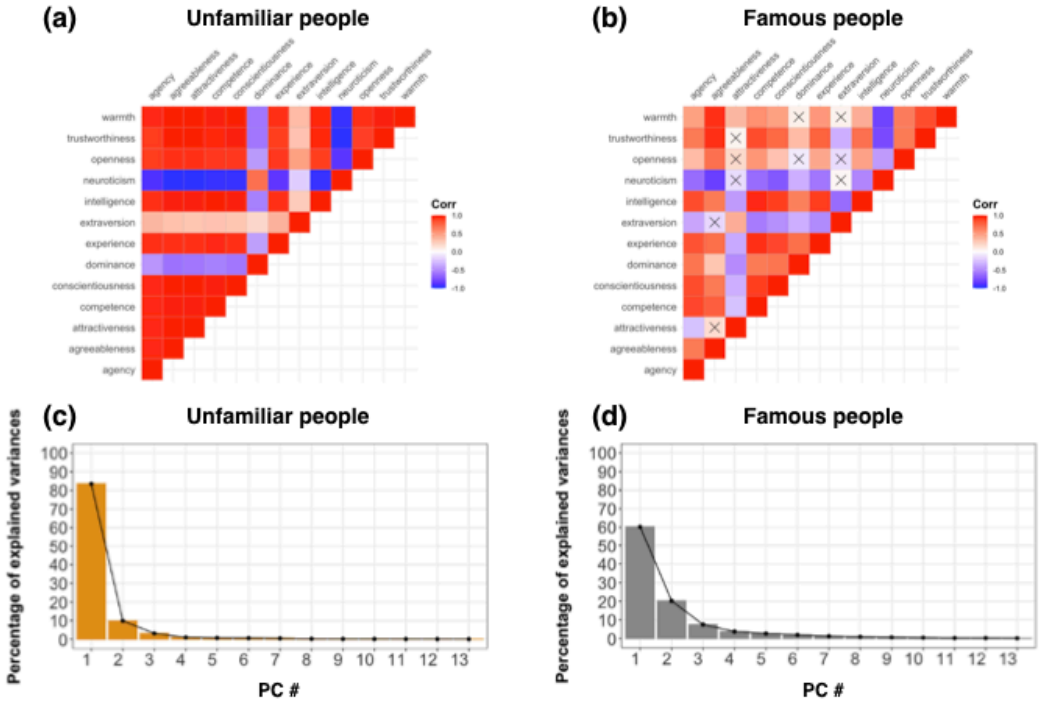
Our proposed hypothesis—that famous people are represented in a higher-dimensional space—can account for low prediction accuracy during generalization from unfamiliar people to famous people, but it does not explain why the coefficients of determination for many of these models are negative, rather than positive and near zero. To investigate this phenomenon more closely, we examined the distribution of holistic similarity ratings for each domain. We found that for the domain of unfamiliar people,

the distribution of holistic similarity ratings had a skew of -0.325 (indicating a slightly long left tail) and a kurtosis of 1.760 (indicating thinner tails and a broader peak). For the domain of famous people, the distribution of holistic similarity ratings had a skew of -0.982 (indicating a moderately long left tail) and a kurtosis of 3.544 (indicating fatter tails and a sharper peak). Thus, both distributions showed some non-normality, which may have contributed to negative coefficients of determination. However, it is important to note that non-normality should contribute to below-chance performance for both directions of generalization (unfamiliar-to-famous and famous-to-unfamiliar). Therefore, non-normality alone is not sufficient to explain (1) the marked asymmetry observed in generalization performance, or (2) the large coefficients of determination observed when generalizing from famous to unfamiliar people. Our proposed hypothesis, of a difference in representational complexity, is further explored below.

Correlation structures. We next examined the collinearity of trait ratings in each domain. As seen in Fig. 5a-b, the Pearson's correlations between trait ratings of unfamiliar people were more extreme than the correlations between trait ratings of famous people. A Chi-squared test of the two correlation matrices revealed that they significantly differ ($\chi^2(78) = 4881.09, p < 0.0001$).

PCA revealed that, in each domain, the first principal component accounted for a majority of the variance in trait ratings (see Fig. 5c-d for scree plots; see Tables S13-S14 for component loadings). The first principal component (PC1) accounted for a greater proportion of variance in the unfamiliar people domain (0.835) compared to the famous people domain (0.603).

Figure 5. Top: Pearson's correlations for all pairwise combinations of the 13 trait ratings in the domain of (a) unfamiliar people and (b) famous people. X's refer to non-significant correlations. **Bottom:** Scree plots displaying proportion of total variance explained by each principal component in the domain of (c) unfamiliar people and (d) famous people.



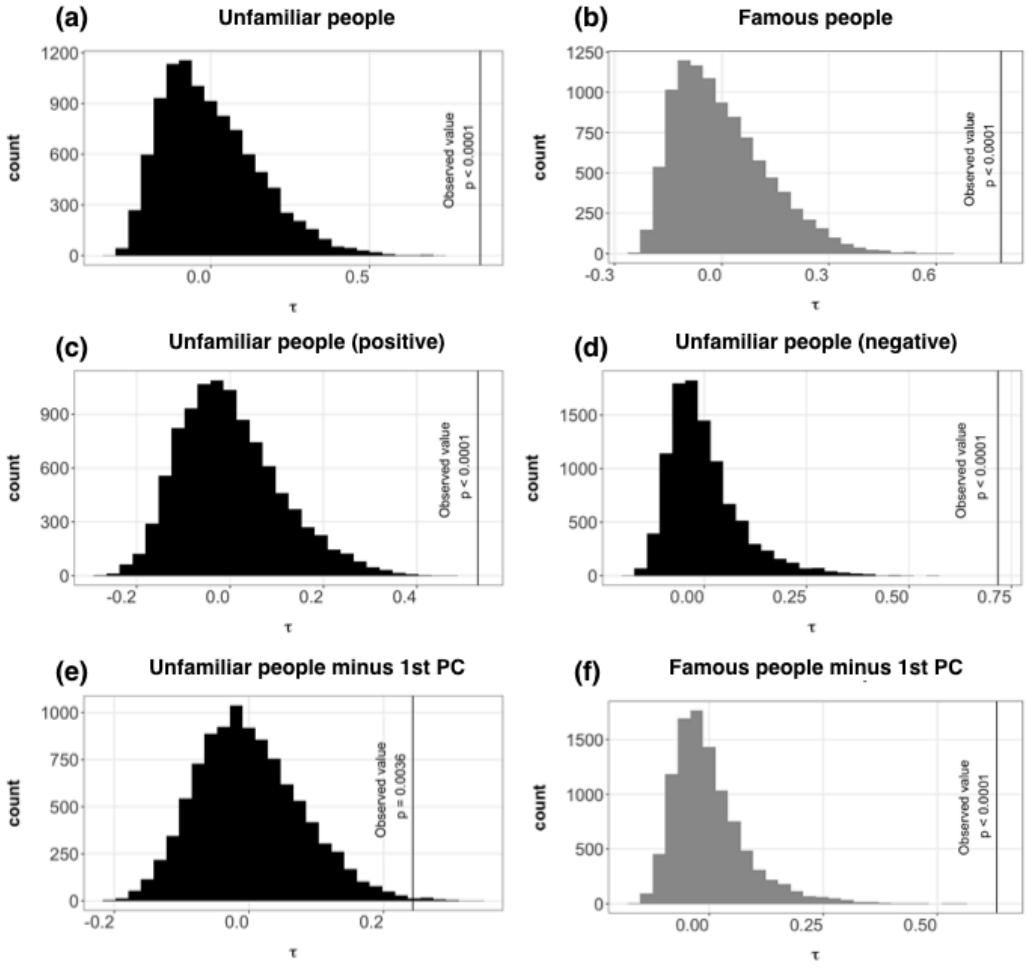
Reliability of correlation structures. Next, we examined whether the degree of intercorrelatedness between trait ratings is a robust feature of each domain, rather than being variable across different subsets of data. We found that, for the unfamiliar people domain and the famous people domain, the correlation matrices for random split-halves of data were significantly correlated (unfamiliar people: Kendall's $\tau = 0.849$, permutation $p < 0.0001$; famous people: Kendall's $\tau = 0.782$, permutation $p < 0.0001$; Fig. 6a-b). These results indicate that the degree of intercorrelatedness between traits is a robust feature in each domain. The above results were replicated when names and faces are added to the unfamiliar targets (Fig. S6; Fig. S7).

The role of valence in trait ratings of unfamiliar people. Taking a closer look at the correlation structure for trait ratings of unfamiliar people (Fig. 5a), we found that 11 of

the traits were positively correlated with each other, but largely anticorrelated with neuroticism and dominance (dominance and extraversion, however, were positively correlated). In addition, when trait ratings were regressed onto target valence (whether the target performed a positive or negative behavior), we found that negative targets (vs. positive targets) were rated significantly higher on neuroticism and dominance (neuroticism: b = -1.915, SE = 0.035, t = -54.62, $p < 0.0001$; dominance: b = -1.353, SE = 0.091, t = -14.81, $p < 0.0001$), whereas positive targets were rated significantly higher on all other traits (see Table S15 for statistics).

Furthermore, we found that the first principal component (PC1) loaded positively onto neuroticism and dominance, but negatively onto all other traits (Table S13), and PC1 scores were positive for negative targets but negative for positive targets (Fig. S5). In all, this pattern of results indicates that a single underlying feature, valence, is capturing most of the variation in trait judgments in the domain of unfamiliar people.

Figure 6. Kendall's tau distributions for permuted data in the domain of (a) unfamiliar people, (b) famous people, (c) unfamiliar people who performed positive behaviors, (d) unfamiliar people who performed negative behaviors, (e) unfamiliar people after the 1st PC was removed, (f) famous people after the 1st PC was removed.

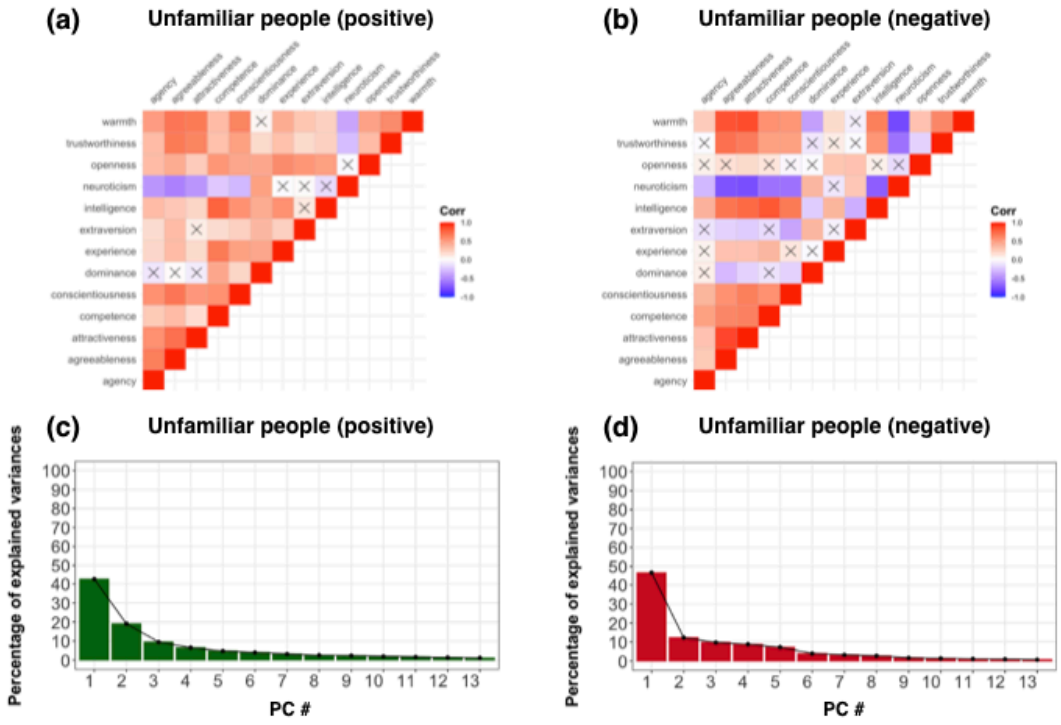


Correlation structures after removing valence information. Overall, the 13 trait ratings were more intercorrelated within the unfamiliar people domain, compared to the famous people domain, indicating that the unfamiliar targets may reside in a lower-dimensional (perhaps one-dimensional) representational space. This asymmetry might explain why it is (1) more accurate to use trait distances to predict holistic similarity between unfamiliar people than between famous people, and (2) more accurate to use models trained on famous people to predict holistic similarity between unfamiliar people

than vice versa. To test whether the representational space for unfamiliar people is one-dimensional, we examined if there is remaining reliable structure in trait ratings of unfamiliar people even after removing the feature that seems to account for most of the variance—valence.

Separate correlation matrices were generated for the subset of positive unfamiliar people, and the subset of negative unfamiliar people. We found that there was reduced intercorrelation between trait ratings of unfamiliar people of the same valence (Fig. 7a-b). The following comparisons between correlation matrices were significant: between all unfamiliar people and positive unfamiliar people ($\chi^2(78) = 9917.78, p < 0.0001$); between all unfamiliar people and negative unfamiliar people ($\chi^2(78) = 8992.06, p < 0.0001$); and between positive unfamiliar people and negative unfamiliar people ($\chi^2(78) = 640.47, p < 0.0001$). As illustrated by the scree plots (Fig. 7c-d), the first PC for each valence subset explained less than half of all variance (positive unfamiliar people: 42.6%; negative unfamiliar people: 46.5%), whereas in the set of all unfamiliar people, the first PC had explained more than 80% of all variance.

Figure 7. Top: Pearson's correlations for all pairwise combinations of the 13 trait ratings for (a) unfamiliar people who performed positive behaviors and (b) unfamiliar people who performed negative behaviors. X's refer to non-significant correlations. Bottom: Scree plots displaying proportion of total variance explained by each principal component in the domain of (c) unfamiliar people who performed positive behaviors and (d) unfamiliar people who performed negative behaviors.

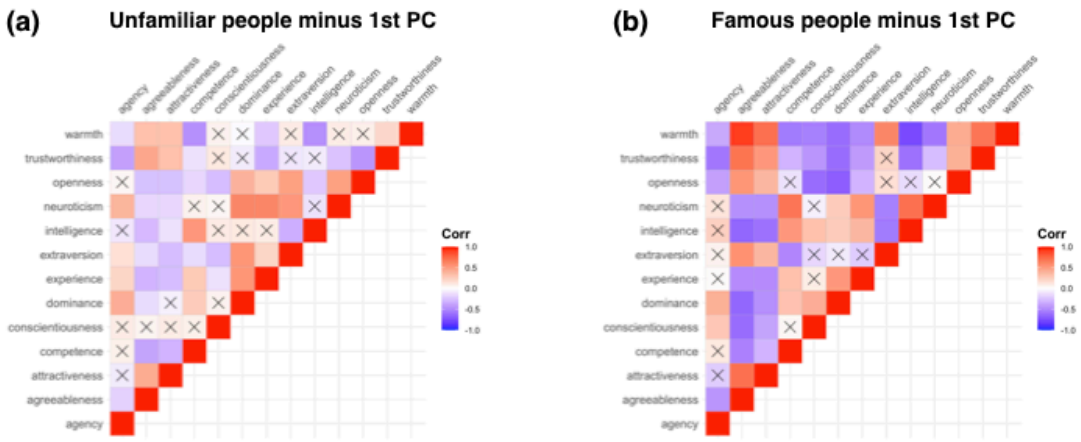


We examined the reliability of the correlation structure for each valence subset. The correlation matrices for random split-halves of data were significantly correlated for each valence subset (positive unfamiliar people: Kendall's $\tau = 0.530$, permutation $p < 0.0001$; negative unfamiliar people: Kendall's $\tau = 0.717$, permutation $p < 0.0001$; Fig. 6b-c). This indicates that the correlation structure of each valence subset is robust. These results were replicated when names and faces are added to the unfamiliar targets (Fig. S8; Fig. S7).

However, there might still be variance along the valence axis within the subset of positive unfamiliar people, and within the subset of negative unfamiliar people — some

positive unfamiliar people are more positive than others, and some negative unfamiliar people are more negative than others. Thus, as a stricter test, we removed the first PC (Tables S13-S14) from the trait rating data for each domain, then tested for remaining reliable structure (Fig. 8).

Figure 8. Pearson's correlations for all pairwise combinations of the 13 trait ratings for (a) the domain of unfamiliar people, after removing the 1st PC and (b) the domain of famous people, after removing the 1st PC.



After removing the first PC, the correlation matrix for unfamiliar people was significantly different from the original ($\chi^2(78) = 30760.81$, $p < 0.0001$), and the correlation matrix for famous people was significantly different from the original ($\chi^2(78) = 1759.05$, $p < 0.0001$). For both domains, removing the first PC still resulted in reliable correlation structures (unfamiliar people: Kendall's $\tau = 0.243$, permutation $p = 0.0036$; famous people: Kendall's $\tau = 0.630$, $p < 0.0001$, permutation $p < 0.0001$; Fig. 6e-f). Thus, for both the unfamiliar people domain and the famous people domain, the trait rating data exhibited a reliable structure even after removing the first PC. Overall, we found that a reliable higher-dimensional structure exists for both unfamiliar targets and famous targets; valence was not the *only* feature driving trait judgments.

These results were replicated when names and faces were added to the unfamiliar targets (Fig S9; Fig. S7).

Predicting holistic similarity after removing valence information. We next tested whether valence was driving perceptions of similarity between unfamiliar people. We found that pairs of unfamiliar people that performed behaviors of the same valence were rated as more similar overall, compared to pairs that performed counter-valenced behaviors ($b = -1.754, F(1,898) = 2316, p < 0.0001, R^2 = 0.721$). Again, however, valence was not the *only* feature that mattered: when concordance in valence was added as a covariate to each single-trait model, 12 of the trait distances still significantly predicted holistic similarity (Table S7).

As a complementary analysis, we tested how well trait distances predict holistic similarity within each valence subset. For pairs of positive unfamiliar people, four of the trait distances significantly predicted holistic similarity (Table S8), and the 13-trait model significantly predicted holistic similarity ($F(13,148) = 2.559, p = 0.0033, R^2 = 0.184$). For pairs of negative unfamiliar people, six of the trait distances significantly predicted holistic similarity (Table S9), and the 13-trait model significantly predicted holistic similarity ($F(13,149) = 3.105, p = 0.0004, R^2 = 0.213$). Thus, when positive and negative unfamiliar people were separated, trait distances performed worse in predicting holistic similarity; however, some traits still significantly predicted holistic similarity. The above results were replicated when names and faces were added to the unfamiliar targets (Table S10-S12).

1.4 DISCUSSION

In the current study, we examined how 13 traits from major theories of person perception (Thornton & Mitchell, 2018) contribute to overall representations of famous people and to overall representations of unfamiliar people, by probing how well each trait predicts perceived similarity between pairs of targets. This approach allowed us to examine the importance of different traits in determining perceivers' overall representations of people, and whether the representational structure depends on inference context.

Previous research on the structure of trait representations has relied on reducing the dimensionality of evaluations of targets along particular traits. However, such evaluations do not contain information about the relative importance of a particular trait for perceivers' overall representations of people; the dimensions that explain the most variance across behavioral judgments or neural responses may not necessarily contribute the most to overall representations. The current study deviates from previous research in that we used a perceived similarity approach to gauge the importance of different traits for overall representations of people. This method revealed that (1) 13 traits from extant models of person knowledge can each predict perceived similarity between pairs of targets; (2) the traits that best predict holistic similarity between unfamiliar people are different from the ones that best predict holistic similarity between famous people; and (3) trait ratings are more intercorrelated for unfamiliar people than for famous people, suggesting that the representational structure of first impressions is largely driven by one feature, valence. However, further analyses showed that for trait ratings of both unfamiliar and famous targets, a reliable structure was present even after removing the

first principal component, indicating a higher-dimensional structure for first impressions as well.

Contributions of traits to overall representations. In the domain of unfamiliar people, we found that distance along each of 13 traits, individually and together, successfully predicted pairwise holistic similarity ratings; adding names and faces to the unfamiliar targets did not qualitatively change these results. In the domain of famous people, we also found that distance along each of 13 traits, individually and together, successfully predicted pairwise holistic similarity ratings; these associations were largely robust to controlling for pairwise biographical similarities.

The significant associations between pairwise trait distance and pairwise holistic similarity indicate that the 13 tested traits contribute to perceivers' overall representations of people, both when thinking about unfamiliar people, and when thinking about famous people. Importantly, the traits did not all perform equally well in predicting holistic similarity. Differences in predictive performance allow us to make inferences about the importance of particular traits for overall representations of people.

When examining the performance of single-trait models in the domain of unfamiliar people, the top-performing traits were: agreeableness, conscientiousness, attractiveness, warmth, and trustworthiness. These traits cut across multiple theories of person perception: the Big 5 (agreeableness and conscientiousness), the stereotype content model (warmth), and the model of face perception (trustworthiness). The top-performing traits in the domain of famous people were: dominance, intelligence, experience, conscientiousness, and competence. These traits again cut across multiple theories: the model of face perception (dominance), the model of mind perception

(experience), the Big 5 (conscientiousness), and the stereotype content model (competence). Thus, the traits that best predict holistic similarity between unfamiliar people were largely different from the ones that best predict holistic similarity between famous people; in addition, within each domain, there was some conceptual overlap between the top-performing traits.

Previous research across different subfields of psychology has consistently shown that two fundamental dimensions seem to underlie social evaluations: 'communion' (captured by traits that relate to morality and sociability, such as trustworthiness and warmth) and 'agency' (captured by traits that relate to ability and assertiveness, such as competence and dominance; Abele & Wojciszke, 2014; Oliveira et al., 2018). These 'Big Two' dimensions are thought to have functional significance, as communion-related traits describe whether an individual has good or bad intentions, and whether they can garner social support for their intentions, while agency-related traits describe whether an individual can carry out their intentions, and how much power they have over others; assessments along these dimensions can carry consequences for motivations and behaviors towards individuals and groups (Abele & Wojciszke, 2014; Oliveira et al., 2018; Landy, Piazza, & Goodwin, 2016; Cuddy, Fiske, & Glick, 2008; Fiske et al., 2002).

It is of note that the traits that contribute most to holistic similarity for the studied set of famous people (dominance, intelligence, experience, conscientiousness, and competence) mostly fall under the 'agency' umbrella of traits, while the traits that contribute most to holistic similarity for the studied set of unfamiliar people (agreeableness, conscientiousness, attractiveness, warmth, and trustworthiness) mostly

fall under the 'communion' umbrella of traits. Prior work probing the relationship between valence and the Big Two has shown that communion-related traits exhibit greater overlap with valence compared to agency-related traits (Abele & Wojciszke, 2014); it may be that communion-related traits best predicted holistic similarity between unfamiliar targets in the current study, because valence was a key feature organizing trait judgments of unfamiliar targets. Below we return to the idea that members of the two studied domains are more differentiable along one umbrella dimension than the other.

Asymmetry in generalization performance. Trait distances, both together and individually, better predicted holistic similarity between unfamiliar people than between famous people. In addition, the mapping from trait distance to holistic similarity generalized better from a training set of famous people to a testing set of unfamiliar people, than vice versa. Notably, most of the models trained on the unfamiliar people data produced negative coefficients of determination when predicting holistic similarity for famous people (indicative of below-chance accuracy); in contrast, most of the models trained on the famous people data produced positive coefficients of determination when predicting holistic similarity for unfamiliar people.

As holistic similarity ratings in both domains exhibited moderate non-normality, non-normality alone is not sufficient to explain (1) the marked asymmetry observed in generalization performance, or (2) the above-chance accuracy observed when generalizing from famous to unfamiliar people. Our proposed explanation for poor generalization from the unfamiliar people domain to the famous people domain is that the unfamiliar targets are represented in a lower-dimensional space than the famous targets. In line with this, PCA of trait ratings in each domain revealed that the first PC accounts

for a greater proportion of variance in the unfamiliar people domain, compared to the famous people domain. Furthermore, we found that trait ratings of unfamiliar people were more correlated with each other than trait ratings of familiar people. These correlation structures were robust across random splits of data for both domains. This indicates that the greater intercorrelatedness between trait ratings of unfamiliar people is not just due to trait ratings of famous people being noisier; rather, the correlation structure of each domain was reliable.

Dimensionality of each domain. Given the high intercorrelatedness between trait ratings of unfamiliar people, we also tested the more specific hypothesis that one feature, valence, was driving trait ratings similarity judgments in the domain of unfamiliar people. We found that concordance in valence between pairs of unfamiliar people explained 72% of the variance in pairwise holistic similarity ratings. That is, whether two unfamiliar targets performed behaviors of the same valence successfully predicted the holistic similarity for that pair. The relative ease of classifying the unfamiliar targets as good or bad is likely why the unfamiliar people data exhibited (1) greater correlations among trait ratings, and (2) a stronger relationship between trait distance and holistic similarity — unfamiliar people of the same valence likely received more similar trait ratings and higher holistic similarity ratings than unfamiliar people of the opposite valence. This also explains why communion-related traits (agreeableness, warmth, and trustworthiness) performed the best at predicting holistic similarity between unfamiliar people.

However, several additional analyses revealed that valence is not the *only* feature that matters for representations of unfamiliar people. First, we split the unfamiliar targets into positive and negative, removing the dominant organizational feature. We found that

trait ratings were less correlated with each other in each valence subset, compared to the complete set of all unfamiliar people, but the correlation structure of each valence subset was still reliable. Second, we removed valence information in a different way, by removing the first PC from the trait ratings of unfamiliar people, and the resultant correlation structure was also reliable. Third, we found that, even when concordance in valence was added as a covariate to single-trait models predicting holistic similarity, 12 traits still significantly predicted holistic similarity. These results suggest that a higher-dimensional representational structure does exist for the unfamiliar targets, but it occurs on top of a lower-dimensional structure organized along a positive-negative axis; the content of this higher-dimensional structure is an important open question.

While the higher-dimensional structure for unfamiliar people explains less variance in holistic similarity, we found that this structure is still reliable, and it may still play a key role in social judgments and predictions, such as in everyday contexts where people's behaviors may not be as clearly valenced as the positive and negative behaviors presented in this study. We hypothesize that even when a small number of dimensions can account for most of the variance in trait ratings, a much larger number of dimensions might still be reliable and crucial for accounting for human social judgments, even if the proportion of total variance in judgments they explain is small.

Stimulus dependence. We now turn to a key observation that needs to be taken into account when interpreting the results. Differences between the two domains are likely shaped by the sets of stimuli tested (e.g., see Lin et al., 2019, for evidence that surveying a larger number of trait words than is typical yields a novel set of four dimensions that best explain trait judgments of faces). The unfamiliar people stimuli were designed to be

highly valenced (largely positive or negative; Kim et al., 2021), and the famous people stimuli were designed to be a maximally varied collection of famous people (Thornton & Mitchell, 2018). Due to these differences in stimulus selection, the observed differences in terms of which traits best predict holistic similarity should be interpreted with caution.

We note that there were some additional sources of variation among the unfamiliar people stimuli (as measured in Kim et al., 2021): the behaviors that were performed by targets varied in emotional arousal and perceived frequency. The emotional intensity of each behavior was rated on a scale from 1 to 7 (N = 30 participants/behavior), and the perceived frequency of each behavior was rated on a scale from 1 to 100 (N = 30 participants/behavior). The set of 300 behaviors displayed variance along both features (emotional arousal: $M = 3.81$, $SD = 0.94$, *range* = 1.25-6.28; perceived frequency: $M = 24.08$, $SD = 19.04$, *range* = 1-98.82). In addition, the subset of positive behaviors and the subset of negative behaviors did not significantly differ along these features ($p > 0.10$) – variation was distributed across valence, meaning that the sample of unfamiliar people in the current study was not necessarily predetermined to be one-dimensional. Thus, the unfamiliar people stimuli can provide us with limited but still useful insight into the representational structure of first impressions.

It is likely that, if a more varied set of behaviors were associated with the unfamiliar targets, the valence axis would have been less salient to perceivers, and we would have found a less robust relationship between trait judgments of unfamiliar people and holistic similarity judgments. However, it should be noted that the highly-valenced nature of the unfamiliar people stimuli in the current study made it harder, rather than easier, to identify higher-dimensional structure beyond valence. Our results show that

even in a set of stimuli that predominantly vary along the valence axis, higher-dimensional information inferred from behaviors was sufficiently strong to display reliable structure. That is, the finding that the representational structure is higher-dimensional for famous people vs. unfamiliar people is less likely to be stimulus-bound. An important direction for future work is to utilize a more varied set of unfamiliar targets, to more rigorously examine which traits are important for representations of unfamiliar people, and how representations differ across contexts.

Conclusion. In this study, we used a perceived similarity approach to gauge the importance of different traits for overall representations of famous people and overall representations of unfamiliar people for whom one behavior is known. We found that (1) 13 traits from extant models of person knowledge can predict perceived similarity between pairs of targets; (2) the traits that best predict holistic similarity partially depend on inference context (unfamiliar people vs. famous people); and (3) trait ratings are more intercorrelated for unfamiliar targets than for famous targets, but reliable higher-dimensional structure is present even for first impressions. These findings highlight a new way to probe perceivers' overall representations of people, and shed light on how trait representations are affected by inference context.

1.5 REFERENCES

Abele, A. E., & Wojciszke, B. (2014). Communal and agentic content in social cognition: A dual perspective model. In *Advances in Experimental Social Psychology* (Vol. 50, pp. 195-255). Academic Press.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological monographs*, *47*(1), i.

Bach, P., & Schenke, K. C. (2017). Predictive social perception: Towards a unifying framework from action observation to person knowledge. *Social and Personality Psychology Compass*, *11*(7), e12312.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40*, 61-149.

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.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878.

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

Goldberg, L. R. (1990). An alternative" description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216.

Gorno-Tempini, M.L., & Price, C.J. (2001). Identification of famous faces and buildings: a functional neuroimaging study of semantically unique items, *Brain*, *124*(10), 2087-2097.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148.

Grabowski, T. J., Damasio, H., Tranel, D., Ponto, L. L. B., Hichwa, R. D., & Damasio, A. R. (2001). A role for left temporal pole in the retrieval of words for unique entities. *Human Brain Mapping*, *13*(4), 199-212.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619–619.

Kim, M. J., Mende-Siedlecki, P., Anzellotti, S., & Young, L. (2021). Theory of mind following the violation of strong and weak prior beliefs. *Cerebral Cortex*, *31*(2), 884-898.

Kryven, M., Ullman, T., Cowan, W., & Tenenbaum, J. (2016). Outcome or strategy? A bayesian model of intelligence attribution. *CogSci*.

Landy, J. F., Piazza, J., & Goodwin, G. P. (2016). When it's bad to be friendly and smart: The desirability of sociability and competence depends on morality. *Personality and Social Psychology Bulletin, 42(9),* 1272-1290.

Lin, C., Keles, U., & Adolphs, R. (2019, October 2). Four dimensions characterize comprehensive trait judgments of faces. PsyArXiv. https://doi.org/10.31234/osf.io/87nex

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81.

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, *8*(6), 623–631.

Oliveira, M., Garcia-Marques, T., Garcia-Marques, L., & Dotsch, R. (2020). Good to Bad or Bad to Bad? What is the relationship between valence and the trait content of the Big Two?. *European Journal of Social Psychology, 50(2)*, 463-483.

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

Pagan, M., Simoncelli, E. P., & Rust, N. C. (2016). Neural quadratic discriminant analysis: Nonlinear decoding with v1-like computation. *Neural Computation*, *28*(11), 2291–2319.

Park, I. M., Meister, M. L., Huk, A. C., & Pillow, J. W. (2014). Encoding and decoding in parietal cortex during sensorimotor decision-making. *Nature Neuroscience*, *17*(10), 1395–1403.

R Core Team. (2013). R: A language and environment for statistical computing.

Ramon, M., & Gobbini, M. I. (2018). Familiarity matters: A review on prioritized processing of personally familiar faces. *Visual Cognition*, *26*(3), 179-195.

Ray, J. L., Mende-Siedlecki, P., Gantman, A. P., & Van Bavel, J. J. (2019). *The role of morality in social cognition*.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences*, *22*(3), 201–212.

Thornton, M. A., & Mitchell, J. P. (2018). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex*, *28*(10), 3505–3520.

Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019). The brain represents people as the mental states they habitually experience. *Nature Communications*, *10*(1), 1–10.

Wu, Y., Baker, C. L., Tenenbaum, J. B., & Schulz, L. E. (2018). Rational inference of beliefs and desires from emotional expressions. *Cognitive Science*, *42*(3), 850–884.

2.0 THEORY OF MIND FOLLOWING THE VIOLATION OF STRONG AND WEAK PRIOR BELIEFS

2.1 INTRODUCTION

It can be hard for people to change their minds about those they know well, or those who are in their groups. For instance, observers are less likely to revise their impressions when a close friend takes money from them in an economic game, compared to when a stranger does the same (Park et al. in press). In addition, observers who learn both positive information (e.g., "was awarded a research grant") and negative information (e.g., "heckled a speaker during a talk") about ingroup and outgroup members selectively downgrade their impressions of outgroup members (Hughes et al. 2017). Thus, there are differences in the magnitude of impression updating across social distance and across group membership. While such phenomena have typically been interpreted as biased or motivated, they are also compatible with rational updating over stronger (more certain) prior beliefs about close others and ingroup members (Gershman 2019; Kim et al. 2020). It can be difficult to pull apart the contributions of motivation and prior knowledge to selective belief maintenance, as they typically co-occur: we are motivated to preserve favorable impressions of groups we belong to (Van Bavel and Pereira 2018), and at the same time, ample prior experience with others can give rise to stronger beliefs that are hard to update in the face of contradictory evidence. In the current work, we ask whether differences in experimentally-induced prior beliefs, in a context absent social motivation, can lead to differences in impression updating and related neural activity. We examine changes in both rated impressions and neural activity following the violation of strong vs. weak prior beliefs, and following the violation of positive vs. negative prior beliefs.

The role of mental state inference in impression updating

A key process underlying impression updating is mentalizing, or Theory of Mind (ToM): the ability to infer, represent, and reason about others' mental states, such as beliefs, goals, and intentions. When observers generate explanations for others' behavior – *why* someone did what they did – mental state inferences tend to dominate (Malle 2001). Mentalizing can support either impression updating or impression maintenance, depending on the content of the mental state inference. For example, when we see a stranger take money from us in an economic game, we may infer that she *intended* to take the money; we may then use this inference to update our beliefs about her character. In comparison, when we see a close friend take money from us, we may infer that she did *not* intend to simply take the money; she was actually mistaken about the rules of the game, or she plans to share the spoils with us later. Such inferences allow us to maintain our positive prior beliefs about our friend's character. On the flip side, when we see an outgroup member behave prosocially, we may infer that she did so only for self-interested, reputational reasons; such inferences allow us to maintain our negative prior beliefs about the outgroup member. Inferences about transient mental states can thus be used to either support an impression update, or reconcile discrepancies between our prior impression of someone and their surprising, prior-inconsistent behavior.

The role of prior beliefs in impression updating

What are the informational factors that determine whether or not we engage in impression updating? Our prior beliefs about others can vary in both strength and valence. We tend to have stronger, and more positive, beliefs about close others compared to strangers; we also have strong beliefs based on group membership in the form of stereotypes (Fiske 1998; Dovidio et al. 2010). When we have strong prior beliefs

about someone, and they behave uncharacteristically, we may generate an explanation for their behavior based on a transient mental state, rather than update our impression of their character. Generating alternative explanations in this way can be compatible with a form of Bayesian rationality, such that the likelihood of invoking an alternative explanation depends probabilistically on the strength of the prior belief, and the likelihood of the conflicting information (Gershman 2019). Thus, when we have sufficiently strong prior beliefs about someone's character, it can be rational to generate alternative explanations for surprising behavior.

Differences in impression updating can arise from differences in belief strength, differences in belief valence, or both. For instance, one may have: (1) strong positive beliefs about a friend, and weak positive beliefs about an acquaintance; (2) strong positive beliefs about the ingroup, and strong negative beliefs about the outgroup; (3) strong positive beliefs about a friend, and weak negative beliefs about a stranger. Note that in intergroup contexts an observer can have counter-valenced but equally strong beliefs about the two groups; in these contexts, we expect that both strong positive priors about the ingroup *and* strong negative priors about the outgroup will be resistant to updating.

Strong prior beliefs about others often co-occur with social motivational factors, such as the desire to selectively maintain positive impressions of ingroup members (Van Bavel and Pereira 2018). When we continue to see close or ingroup transgressors as good or trustworthy, then, this may be because we have strong prior beliefs about their goodness, or because we are motivated to view them in a positive light, and motivated to maintain our social relationships (Park and Young 2020). Analogously, when we refuse

to improve our impressions of outgroup others, this may be because we have strong prior beliefs about their bad character, or because we are motivated to view them unfavorably. It is thus important that we examine the role of prior knowledge in belief updating in isolation. In the current work, we isolate the role of prior knowledge in belief updating by manipulating participants' beliefs about novel, fictional targets. We note that we use the term 'motivation' to refer to *social* motivation stemming from real relationships or social groups, rather than other forms of motivation. For instance, participants may have a general cognitive motivation to hold on to initial beliefs, perhaps in a heuristic manner incompatible with Bayesian reasoning; however, as all targets in the experiment were zero-acquaintance, fictional targets, we expected that participants' social motivations concerning these targets would be at floor. For this reason, we describe our paradigm as a context absent social motivation.

Mentalizing in light of strong vs. weak priors

We aimed to examine whether brain regions implicated in ToM are recruited to different degrees in light of different priors. The ToM network includes dorsomedial prefrontal cortex (DMPFC), bilateral temporo-parietal junction (TPJ), and precuneus (PC). These regions are critical for inferring moral intent and integrating mental state information with other information for moral judgment (see Young and Waytz 2013 for a review). They are also implicated in causal attributions to the self, another person, or the situation (Kestemont et al. 2015), and the formation and revision of trait inferences (Cloutier et al. 2011; Ma et al. 2012; Ferrari et al. 2016). In addition, neural activity in ToM regions is enhanced for others' behaviors that violate prior beliefs based on: past behavioral history (Mende-Siedlecki et al. 2013; Dungan et al. 2016), instructed trait

knowledge (Heil et al. 2019), and stereotypes (Cloutier et al. 2011; Li et al. 2016). ToM regions thus respond to the contradiction of prior beliefs across a variety of social contexts, and the enhanced activity may reflect an attempt to construct a mental state explanation, such as one referring to innocent intent, that reconciles the unexpected behavior with prior impressions. For example, one study found greater activity in DMPFC and bilateral TPJ when third-party observers were faced with ingroup vs. outgroup defectors in the Prisoner's Dilemma Game, and increased connectivity between DMPFC and LTPJ was associated with weaker punishment of ingroup defectors (Baumgartner et al. 2012). In this context, the more surprising event (selfish behavior from an ingroup member) was followed by greater activity in mentalizing regions and greater selective forgiveness. This increase in ToM activity may reflect the probabilistic generation of a coherent alternative explanation for the surprising event (e.g., my ingroup member did not *intend* to defect).

Past neuroimaging work has relied on participants' real-life prior beliefs about ingroup and outgroup members (Baumgartner et al. 2012), and about friends and strangers (Park et al. in press), to investigate neural differences during belief updating across relationship contexts. These contexts may have been accompanied by social motivational factors: observers may have engaged in mentalizing out of a desire to protect their beliefs about the ingroup, even though coming up with an alternative explanation was not probabilistically warranted by their prior beliefs. Therefore, it is an open question whether differences in prior strength–in a context absent social motivation–contribute to neural differences during belief updating across social distance and group membership.

The current study

The goal of the current study was to examine belief updating and mentalizing activity following the violation of strong vs. weak prior beliefs, and positive vs. negative prior beliefs, in the absence of real-life priors and motivations concerning groups and individuals. We aimed to experimentally induce prior beliefs that vary in both strength and valence, given that these features may have distinct effects on updating and mentalizing.

We adapted a paradigm developed by Mende-Siedlecki and colleagues (2012, 2013, 2016). Participants learned about fictional individuals whose behaviors were either internally consistent or internally inconsistent. The internally inconsistent targets initially performed two or four same-valenced, morally-relevant behaviors (leading to the formation of weak or strong beliefs about the agent's disposition), before performing two counter-valenced behaviors, potentially evoking an impression update. We tracked participants' impressions long the dimension of trustworthiness. We tested whether impression updating and ToM activity differ as a function of the strength of the prior (weak vs. strong) and update direction (positive-to-negative vs. negative-to-positive). We also conducted whole-brain analyses to examine overall differences in neural activity during impression updating following different types of expectation violations. Lastly, we note that, while participants may have entered the experiment with prior beliefs about the trustworthiness of people in general, we expected these real-life priors to apply equally across targets, given that they are all zero-acquaintance targets. Thus, the term "prior" in the context of this experiment will be used to refer to experimentally-induced beliefs about targets. This is in accordance with a cyclical framing of Bayesian belief updating

(Trotta 2018), where the posterior belief after a new observation (e.g., a target's first behavior) then becomes the prior belief for the next observation (e.g., a target's second behavior).

2.2 METHODS

Open science

This study was preregistered (https://aspredicted.org/blind.php?x=ti3pn4). Behavioral data, percent signal change (PSC) data, and R code are available on OSF (https://osf.io/27cjx/?view_only=df7aa52aef2048d09101df6267aca44e). Neural data are available on OpenNeuro (https://openneuro.org/datasets/ds002793).

Participants

We aimed to collect analyzable data from 28 participants (based on recent neuroimaging studies examining ToM regions, Tsoi et al. 2018, N = 25; Dungan and Young 2019, N = 26; Theriault et al. 2020, N = 25). Thirty adults from the Boston area between the ages of 18 and 35 were recruited. All participants were right-handed, native English speakers with normal or corrected-to-normal vision and no history of psychiatric disorders or learning disabilities. Participants were recruited through an online posting and given a $60 cash payment; written consent was obtained prior to participation. The study was approved by the Institutional Review Boards at Boston College and the Massachusetts Institute of Technology. Two participants were excluded for exhibiting excessive in-scanner movement, identified during spatial preprocessing (see **Neural data exclusion** below). Analyses were conducted on the remaining 28 participants (15 women; age $M = 24$, $SD = 3.92$).

Participant instructions

Participants were told that they would learn information about people, represented by pictures of faces, and that each face would be paired with a sequence of six written behavior descriptions. Participants were asked to form impressions of the people that were pictured by imagining them actually performing the behaviors. For each behavior, they were instructed to rate the target's trustworthiness, based on everything they knew about the person so far.

Sequence types

Participants learned about 50 individuals represented by male and female faces from the Karolinska Directed Emotional Faces set (Lundqvist et al. 1998). Each face was paired with a sequence of six behaviors, which was designed to be either internally inconsistent (80% of targets; 'expectation-violation sequences'), or internally consistent (20% of targets; 'control sequences').

There were four types of expectation-violation sequences, varying in prior strength and update direction: 'Strong Negative-to-Positive' (4 immoral behaviors followed by 2 moral behaviors), 'Weak Negative-to-Positive' (2 immoral followed by 2 moral then 2 neutral), 'Strong Positive-to-Negative' (4 moral, 2 immoral), and 'Weak Positive-to-Negative' (2 moral, 2 immoral, 2 neutral). Two neutral behaviors were added to the ends of Weak sequences to keep sequence length constant across sequence types. The neutral behaviors were placed at the end rather than at the beginning of Weak sequences, so that participants would not be able to detect the type of any given sequence by the very first behavior. Our aim was to minimize participants' expectations of the

upcoming sequence, and encourage participants to attend equally carefully to each sequence, regardless of sequence type.

For discussion purposes, the two behaviors immediately preceding the valence switch point will be referred to as 'pre-switch' behaviors, while the two behaviors immediately following valence switch will be referred to as 'post-switch' behaviors.

Additionally, there were two types of control sequences: 'Negative Control' (6 immoral behaviors) and 'Positive Control' (6 moral behaviors). See **Table 1** for examples of each sequence type.

Sequence type	Example
Strong Negative-to-Positive (strong neg → pos)	(B1) Megan lost her temper at the barista. (B2) Megan stood up a first date. (B3) Megan hit a car and left the scene of the accident. (B4) Megan swore at a cashier who made an error on her bill. (B5) Megan created a photo album of the family for her sister's housewarming gift. (B6) Megan purified a water source for a small village.
Weak Negative-to-Positive (weak neg → pos)	(B1) Joshua publicly mocked his sister for stuttering. (B2) Joshua got ejected from a game for getting into a fight. (B3) Joshua shared cookies from a care package with his roommates. (B4) Joshua helped an elderly woman with her groceries. (B5) Joshua called a TV station for weather information. (B6) Joshua put gas in the car.
Strong Positive-to-Negative (strong pos → neg)	(B1) Thomas had all of his wedding gifts be donations to charity. (B2) Thomas gave a stranded motorist a lift to the service station. (B3) Thomas helped his roommate prepare for a big presentation. (B4) Thomas spent a morning volunteering at a nursing home. (B5) Thomas lied to his wife about his location when visiting an ex. (B6) Thomas ordered around his housekeeper in a harsh tone of voice.
Weak Positive-to-Negative (weak pos → neg)	(B1) Emily went to her friend's teachers to get his homework when he was sick. (B2) Emily helped a neighbor fix his roof. (B3) Emily broke a valuable vase and blamed her brother. (B4) Emily smoked in a no-smoking section even though others complained. (B5) Emily mailed a letter at the post office. (B6) Emily left her shoes on the doormat.
Negative Control (neg control)	(B1) Hannah swore at her roommate for eating her leftovers. (B2) Hannah picked a fight on social media with a stranger. (B3) Hannah deliberately tripped an elderly person. (B4) Hannah charged overpriced legal fees to less educated clients. (B5) Hannah tried to steal clothes from a department store but got caught by security. (B6) Hannah deliberately excluded a friend she did not like from weekend plans.
Positive Control (pos control)	(B1) Jonathan picked up all the litter at the park. (B2) Jonathan spent a Saturday volunteering at a soup kitchen. (B3) Jonathan fixed a friend's broken laptop. (B4) Jonathan organized a free speech rally against hate in America. (B5) Jonathan helped a blind man pick out items in the grocery store. (B6) Jonathan visited a sick friend in the hospital.

Table 1. Sequence types and examples.

Stimulus balancing

Three hundred written descriptions of behaviors were used to generate 50 unique sequences of six behaviors each. A portion of the behavior descriptions were adapted from previous studies (Mende-Siedlecki et al. 2013; Mende-Siedlecki and Todorov 2016). The behavior descriptions were constructed to be relevant to morality, clearly valenced, and varying in intensity and perceived frequency. Ten sequences were generated for each of the four types of expectation-violation sequences; five sequences were generated for each of the two types of control sequences.

The expectation-violation sequences were constructed so as to control for a set of stimulus features: moral relevance, perceived frequency, emotional valence, emotional arousal, trustworthiness, and intelligence. Ratings for each feature were collected from separate groups of Amazon Mechanical Turk participants ($N \approx 30$/behavior). Two sample t-tests showed that these features did not differ significantly ($p > 0.10$) across moral and immoral behaviors, across the switch point, and across weak and strong priors within update direction.

To ensure that differences between sequence types would not be a function of specific pairings of pre-switch and post-switch behaviors, we shuffled the post-switch behaviors across participants so that they were seen following both weak and strong priors. Additionally, target name and associated target face were counterbalanced with update direction across participants, in order to control for participants' chance associations with specific names or facial features. For example, for one half of participants, Thomas appeared in the positive→negative direction, and Emily in the negative→positive direction; for the other half of participants, Thomas appeared in the

negative→positive direction, and Emily in the positive→negative direction. We expected that chance associations with specific names or facial features would not all be of the same valence, and, crucially, that experimental effects would be obscured, but not enhanced, by chance associations. These counterbalancing schemes resulted in four total stimulus lists.

Presentation

The stimuli were presented using PsychoPy v1.85.6 (Peirce 2007). Fifty total sequences were presented over ten 5.5-minute runs. Each run consisted of five sequences: one each of the expectation-violation sequences, and one control sequence. At the beginning of each sequence, the target face was presented with an introductory sentence ('This is Thomas') for 2s (**Fig. 1**). Next, the face was presented with a sequence of six written behavior descriptions for 6s each with jittered fixation (2s, 4s, or 6s, pseudorandomly assigned to keep sequence duration constant) between each face-behavior pair. On each behavior presentation, participants rated the target's trustworthiness on a scale from 1 (least trustworthy) to 7 (most trustworthy) using a button box. Participants were randomly assigned to one of four stimulus lists, and run order was randomized for each participant. Trial order within run was pseudorandomized such that, over the course of the experiment, run-initial and run-final trials were distributed evenly across sequence type.

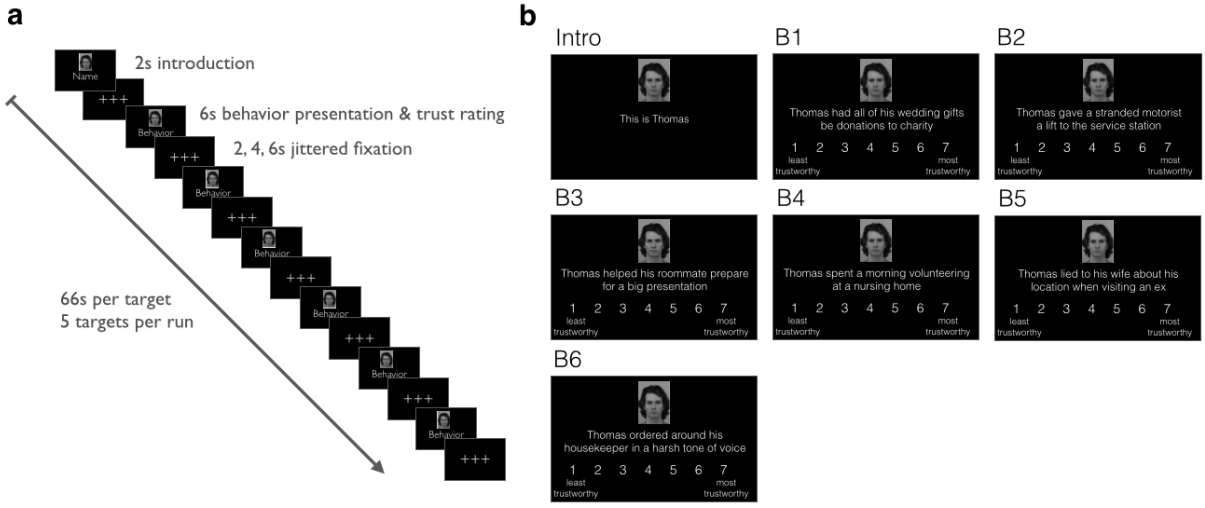


Figure 1. In-scanner stimulus presentation. (a) At the beginning of each sequence, the target face was presented with an introductory sentence ("This is Thomas") for 2 s. Next, the face was presented with a sequence of 6 behaviors, with jittered fixation (2, 4, or 6 s) between each face-behavior presentation. (b) For each behavior, participants rated the target's trustworthiness on a scale from 1 (least trustworthy) to 7 (most trustworthy).

MRI data acquisition and processing

The MRI data were collected using a 32-channel head coil in a 3T Siemens Prisma scanner at the Athinoula A. Martinos Imaging Center at the Massachusetts Institute of Technology. Functional volumes were acquired in 32 3x3x3 mm slices using a gradient-echo sequence (TR = 2s, TE = 30ms, flip angle = 90). The first 6s of each run were excluded to allow for steady state magnetization. Before the functional scans, high-resolution structural images were acquired (1mm isotropic MPRAGE, TR = 2.53s, TE = 1.69ms).

Data processing and analysis were performed using fMRIPrep (Esteban et al. (2019); see **Supplementary Materials p.1** for details), SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/), and custom software. The functional data were realigned, co-registered to the anatomical image, normalized onto a

common brain space (Montreal Neurological Institute, MNI, template), spatially smoothed using a Gaussian filter (full-width half-maximum = 8mm kernel), and high-pass filtered (128s).

Post-scan measures

Following the scanning procedure, participants completed a series of additional behavioral measures: (1) a scenario-based scale of Willingness to Forgive (DeShea 2003); (2) a measure of entity vs. incremental beliefs about morality (Chiu et al. 1997; Hughes 2015); and (3) surprisingness ratings for all behaviors in all expectation-violation sequences seen in the scanner ("Given what you know so far, how surprising is this behavior?" on a 1-7 scale).

Mixed effects models for behavioral and PSC data

Linear mixed effects models were constructed in R (RCore 2016) for all behavioral data and all percent signal change (PSC) data (package: "lme4"; Bates et al. 2014). All models initially included by-subject and by-item random intercepts. If a model failed to converge or had singular fit, we simplified the random effects structure by removing random intercepts with near-zero variance until convergence was achieved. If the random intercept for subject or item was dropped from a model, this was denoted in the results as "no by-subject intercepts" or "no by-item intercepts", respectively; if the random intercepts for both subject and item were dropped from a model, this was denoted as "no random intercepts". To obtain p-values for fixed effects, we conducted likelihood ratio tests of the full model against the model with all predictors except for the predictor of interest. We report semi-partial R-squared values (coefficients of determination; Edwards et al. 2008) as effect sizes for fixed effects.

Analyses of trustworthiness ratings

To examine behavioral impression updating, we computed an *updating measure* for each expectation-violation trial. For neg→pos targets, this was calculated as the difference between the average trustworthiness rating for the two post-switch behaviors and the average trustworthiness rating for the two pre-switch behaviors. For pos→neg targets, we multiplied this measure by -1. Here we reverse the sign of the difference (rather than taking the absolute value), to prevent overestimation of update magnitude that may occur if participants update in the unanticipated direction.

We conducted linear mixed effects analyses to test whether impression updating differed as a function of prior strength (weak or strong) and update direction (pos→neg or neg→pos). We also examined the effect of prior strength on average trust ratings for the two pre-switch behaviors and for the two post-switch behaviors.

Analyses of post-scan measures

We analyzed surprisingness ratings as a function of prior strength and update direction. We also correlated participants' scores on the Willingness to Forgive scale with behavioral impression updating, and participants' scores on the entity vs. incremental morality measure with behavioral impression updating.

Neural data exclusion

Individual functional runs were removed from further analysis if the participant exhibited >3mm movement at any point during the run, or if the average framewise displacement for the run exceeded 1mm. Participants were excluded if more than 1/3 of

collected functional runs were dropped. This resulted in the exclusion of two participants (of 30 scanned participants).

ROI analyses

A Theory of Mind localizer task (Saxe and Kanwisher 2003; Dodell-Feder et al. 2011) was used to functionally define four Regions of Interest (ROIs): dorsomedial prefrontal cortex (DMPFC, N=21), right temporoparietal junction (RTPJ, N=27), left temporoparietal junction (LTPJ, N=27), and precuneus (PC, N=27). ROIs were defined as all voxels within a 9-mm radius of the peak voxel that passed threshold in the contrast 'false belief > false photo' ($p < 0.001$, uncorrected; $k > 16$, computed via 1,000 iterations of a Monte Carlo simulation, Slotnick et al. 2003). We used the same ROI selection parameters as previous neuroimaging research examining ToM regions (Tsoi et al. 2018; Dungan and Young 2019). See **Supplementary Table S1** for peak coordinates.

As there were six sequence types (**Table 1**), and six behaviors in each sequence, the ordinal position of a behavior (1st through 6th) within a sequence type was treated as a single 'condition'. This resulted in 36 total conditions (6 ordinal positions * 6 sequence types). For example, the first behavior in Strong Negative-to-Positive sequences was treated as one condition; the second behavior in Strong Negative-to-Positive sequences was treated as a different condition. It was important to distinguish between behaviors in different ordinal positions, as we were interested in neural responses to inconsistencies that arose at different points in a sequence. In each ROI, the PSC relative to baseline was calculated for each time point for each condition, averaging across all voxels in the ROI. Baseline response, calculated separately for each run, was the average over time of the responses to fixation. PSC for each timepoint for each condition was calculated as

100[(average response for condition at time t – baseline)/baseline]. Timepoints that exhibited >1mm frame-wise displacement compared to the previous timepoint were removed prior to further analysis. PSC values were averaged across each 6-second behavior presentation (offset 4s from presentation time to adjust for hemodynamic lag) to estimate a single PSC for each condition in each ROI.

Analyses of neural activity in response to post-switch vs. pre-switch behaviors

We compared average PSC for the two pre-switch behaviors with average PSC for the two post-switch behaviors. This was done both collapsing across sequence type and within sequence type.

Analyses of neural updating

To examine neural activity associated with impression updating, we predicted PSC for the post-switch behaviors, as a function of prior strength (weak or strong) and update direction (pos→neg or neg→pos), controlling for activity for the pre-switch behaviors.

We also computed a *neural updating measure* for each expectation-violation sequence by taking the difference of the average PSC for the two post-switch behaviors and the average PSC for the two pre-switch behaviors.

Brain-behavior correlations

To explore the relationship between updating-related neural activity in each ToM ROI and behavioral impression updating, we ran linear mixed effects models predicting the magnitude of behavioral updating on each trial, with neural updating as a fixed effect.

Feature encoding models

For whole-brain analyses, we used a set of encoding models to predict activity evoked by a wide range of stimulus features that varied between behavior positions. We created parametric regressors coding for twelve behavior-wise stimulus features (see **Table 2**). For each behavior presentation, feature values were applied to all three images corresponding to the behavior. Face-only presentations were modeled separately using a condition regressor. To correct for multiple comparisons, images from group-level analyses were subjected to a voxel-wise threshold of $p < 0.001$ (uncorrected) and a cluster extent threshold ensuring $p < 0.05$ (familywise error rate-corrected; applied in SPM).

Table 2 lists the regressors that were included in each model. We were primarily interested in regions that track: *prior strength*, controlling for position of current behavior, valence of current behavior, and whether a valence change occurred (models A, D); *whether a valence change occurred*, controlling for position of current behavior, valence of current behavior, and prior strength (models B, F); and *trial-wise impression updating*, controlling for position of current behavior (models G, H). In constructing these models, each regressor was serially orthogonalized with respect to the previous regressor; the ordinal position regressor was always entered first. Rotating which regressors were added last across these different models, and examining parameter estimates of these regressors, allowed us to examine unique neural variance explained by each feature (Mumford et al. 2015).

ID	Feature description	Values
1	Ordinal position of behavior	1, 2, 3, 4, 5, 6
2	Valence of current behavior	1 for positive; −1 for negative; 0 for neutral
3	Cumulative # of consecutive positive behaviors	0, 1, 2, 3, 4, 5, 6
4	Cumulative # of consecutive negative behaviors	0, 1, 2, 3, 4, 5, 6
5	Change occurred: positive/negative previous behavior to neutral current behavior	1 for pos → neutral; −1 for neg → neutral; 0 otherwise
6	Change occurred: positive/negative previous behavior to negative/positive current behavior	1 for pos → neg; −1 for neg → pos; 0 otherwise
7	Magnitude of trustworthiness update from previous behavior	0, 1, 2, 3, 4, 5, 6
8	Positive trustworthiness update from previous behavior	0, 1, 2, 3, 4, 5, 6
9	Negative trustworthiness update from previous behavior	0, −1, −2, −3, −4, −5, −6
10	Any valence change occurred from previous behavior	1 for pos → neutral, neg → neutral, pos → neg, neg → pos; 0 otherwise
11	Any valence reversal occurred from previous behavior	1 for pos → neg, neg → pos; 0 otherwise
12	Cumulative # of consecutive valenced behaviors	1, 2, 3, 4, 5, 6
ID	**Model description**	**Regressors included**
A	Cumulative strength of positive/negative prior	1, 2, 5, 6, **3, 4**
B	Change in valence occurred from previous behavior	1, 2, 3, 4, **5, 6**
C	Valence of current behavior	1, 3, 4, 5, 6, **2**
D	Cumulative strength of prior	1, 2, 5, 6, **12**
E	Any valence change occurred from previous behavior	1, 2, 3, 4, **10**
F	Any valence reversal occurred from previous behavior	1, 2, 3, 4, **11**
G	Magnitude of behavioral updating	1, **7**, 2, 3, 4, 5, 6
H	Positive/negative behavioral updating	1, **8, 9,** 2, 3, 4, 5, 6

Notes: Twelve parametric regressors were used to describe stimulus features that varied between behavior positions (top). Regressor IDs are listed in the order in which they were added to the model; each regressor was serially orthogonalized with respect to the previous regressor (bottom). Parameter estimates were extracted for bolded regressors

Table 2. Top: behavior-wise stimulus features; bottom: regressors included in each encoding model.

2.3 RESULTS

Behavioral results

Impression updating. Participants rated trustworthiness on a scale from 1 to 7. An *updating measure* (see **Methods**) was calculated for each sequence type. The interaction between update direction and prior strength was not significant (B = -0.267, SE = 0.202, $\chi^2(1) = 1.721$, $p = 0.190$). There was a main effect of update direction (pos→neg > neg→pos, B = 1.454, SE = 0.102, $\chi^2(1) = 98.846$, $p < 0.001$, semi-partial $R^2 = 0.722$), and no main effect of prior strength (B = -0.103, SE = 0.102, $\chi^2(1) = 1.005$, $p = 0.316$, $R_{ß}^2$ = 0.013). Thus, participants updated their impressions to a greater extent following the

violation of positive priors. In other words, participants engaged in more negative updating than positive updating (**Fig. 2**).

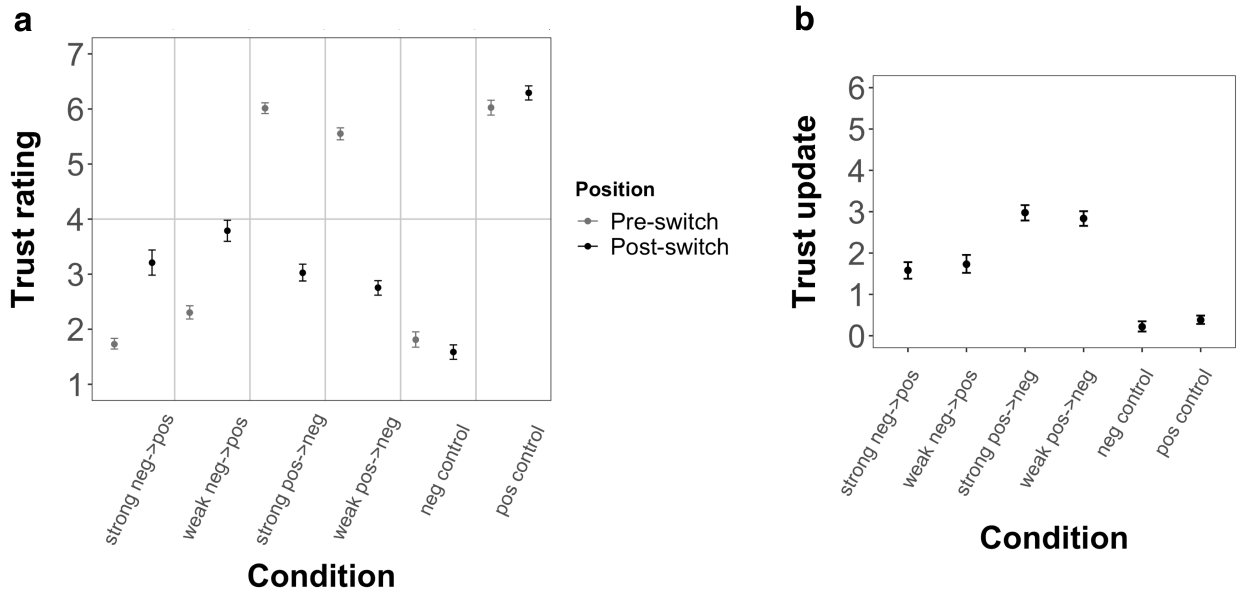


Figure 2. (a) Mean trustworthiness ratings for preswitch behaviors and postswitch behaviors, for each sequence type. Error bars indicate 95% CIs. (b) Mean magnitude of impression update, for each sequence type. For neg → pos targets, this was calculated as: (average rating for 2 postswitch behaviors) – (average rating for 2 preswitch behaviors). For pos → neg targets, this was calculated as: $-1 * [$(average rating for 2 postswitch behaviors) – (average rating for 2 preswitch behaviors)]. For control targets, this was calculated as: (average rating for last 2 behaviors) – (average rating for middle 2 behaviors). The maximum value of the impression update is 6, as the trustworthiness scale runs from 1 to 7.

Effect of prior strength on ratings. We also examined the effect of prior strength on average trust ratings elicited by the two pre-switch behaviors and by the two post-switch behaviors. The pre-switch behaviors in the Strong Positive-to-Negative condition elicited more positive trust ratings than the pre-switch behaviors in the Weak Positive-to-Negative condition (no by-item intercepts; B = -0.455, SE = 0.060, $\chi^2(1) = 61.931$, $p < 0.001$, $R_{ß}^2 = 0.119$). The pre-switch behaviors in the Strong Negative-to-Positive condition elicited more negative trust ratings than the pre-switch behaviors in the Weak

Negative-to-Positive condition (B = 0.608, SE = 0.127, $\chi^2(1) = 18.21$, $p < 0.001$, $R_ß^2$ = 0.374). In other words, impressions based on four positive behaviors more less positive than impressions based on two positive behaviors, and impressions based on four negative behaviors were more negative than impressions based on two negative behaviors (**Fig. 2**).

There was no effect of prior strength on trust ratings elicited by negative post-switch behaviors (B = -0.233, SE = 0.185, $\chi^2(1) = 1.556$, p = 0.212, $R_ß^2$ = 0.04), but post-switch behaviors in the Strong Negative-to-Positive condition elicited more negative trust ratings than post-switch behaviors in the Weak Negative-to-Positive condition (B = 0.667, SE = 0.144, $\chi^2(1) = 17.102$, $p < 0.001$, $R_ß^2$ = 0.362). That is, more negative pre-switch ratings were followed by more negative post-switch ratings (**Fig. 2**). Prior strength thus affected pre-switch impression ratings, and, to some extent, post-switch impression ratings.

Surprisingness ratings. After the scan session, participants were presented with the same expectation-violation sequences they had seen in the scanner. For each behavior in each sequence, participants rated the surprisingness of the behavior on a scale from 1 (least surprising) to 7 (most surprising). We examined average surprisingness ratings for the post-switch behaviors. The interaction between update direction and prior strength was not significant (B = -0.141, SE = 0.158, $\chi^2(1) = 0.792$, p = 0.373). There was a main effect of prior strength on post-switch surprisingness (weak < strong, B = -0.214, SE = 0.0791, $\chi^2(1) = 7.035$, p = 0.008, $R_ß^2$ = 0.089), and no main effect of update direction (B = -0.028, SE = 0.080, $\chi^2(1) = 0.127$, p = 0.722, $R_ß^2$ = 0.002). Thus, inconsistent behaviors following strong priors were rated as more surprising, compared to inconsistent behaviors

following weak priors; there was no difference in surprisingness for inconsistent behaviors following positive vs. negative priors (**Fig. 3**).

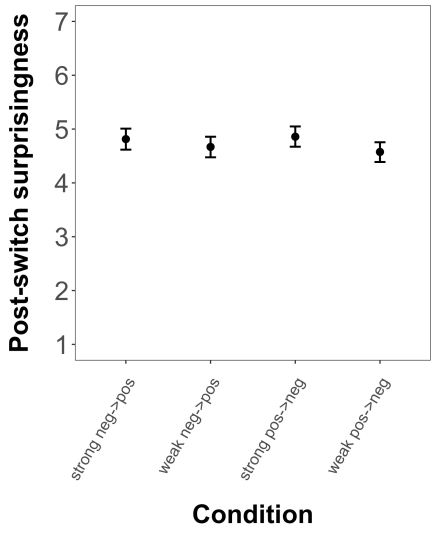


Figure 3. Mean surprisingness ratings for postswitch behaviors, for each sequence type.

Individual difference measures. Following the scan, participants completed a scale of Willingness to Forgive (DeShea 2003), and a measure of entity vs. incremental beliefs about morality (Chiu et al. 1997; Hughes 2015). Neither of these measures significantly predicted magnitude of behavioral updating (Willingness to Forgive: ß = 0.04, *SE* = 0.132, $\chi^2(1) = 0.091$, $p = 0.763$; entity vs. incremental: ß = -0.108, *SE* = 0.134, $\chi^2(1) = 0.639$, $p = 0.424$).

Neural results

Neural activity in response to post-switch vs. pre-switch behaviors. Collapsing across sequence type, all four ToM ROIs exhibited greater activity in response to post-switch behaviors than to pre-switch behaviors (DMPFC: $\chi^2(1) = 30.148$, $p < 0.001$; LTPJ: $\chi^2(1)$

= 6.353, $p = 0.012$; RTPJ: no by-item intercepts, $\chi^2(1) = 19.286$, $p < 0.001$; PC: $\chi^2(1) = 8.058$, $p = 0.005$). See **Supplementary Table S2** for analyses by sequence type.

Neural activity related to updating. We looked at PSC for the post-switch behaviors, controlling for activity during the pre-switch behaviors (**Fig. 4**). For an alternative analysis using the *neural updating measure*, see **Supplementary Materials p. 6**. In DMPFC, there was a significant main effect of update direction (pos→neg > neg→pos, $\chi^2(1) = 15.41$, $p < 0.001$), and a significant main effect of prior strength (strong > weak, $\chi^2(1) = 6.647$, $p = 0.010$). In LTPJ, there was a significant main effect of update direction (pos→neg > neg→pos, $\chi^2(1) = 14.889$, $p < 0.001$), and no main effect of prior strength ($\chi^2(1) = 0.857$, $p = 0.355$). In RTPJ, there was no main effect of update direction ($\chi^2(1) = 0.981$, $p = 0.322$), and a significant main effect of prior strength (strong > weak, $\chi^2(1) = 8.253$, $p = 0.004$). In PC, there was no main effect of update direction($\chi^2(1) = 1.164$, $p = 0.281$), and no main effect of prior strength ($\chi^2(1) = 1.815$, $p = 0.178$)

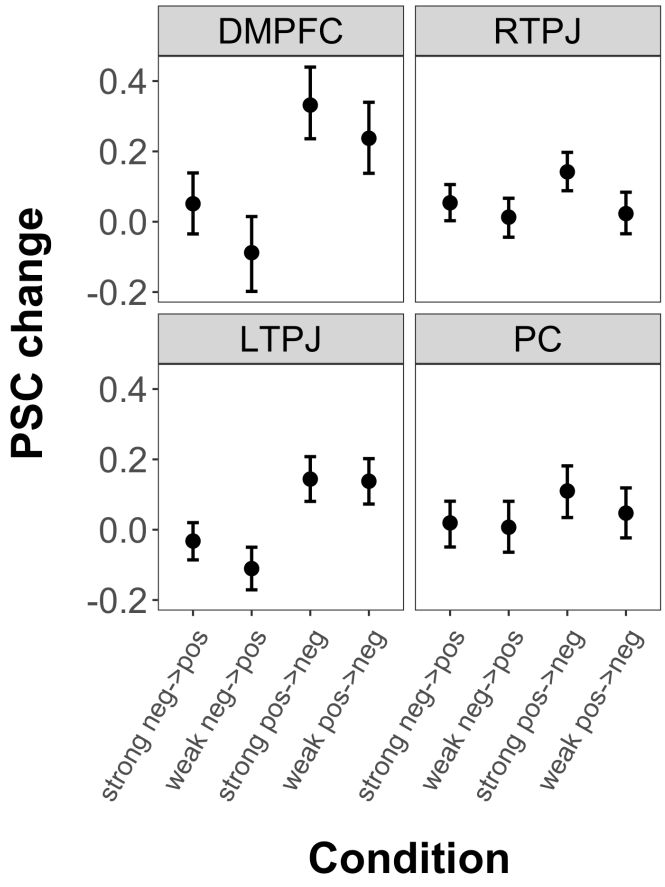


Figure 4. Mean changes in PSC from preswitch behaviors to postswitch behaviors, for each ROI and sequence type. Error bars indicate 95% CIs.

Summary of PSC analyses. We examined neural activity in response to post-switch behaviors, controlling for neural activity in response to pre-switch behaviors. This analysis revealed an effect of update direction (negative updating > positive updating) in DMPFC and LTPJ, and an effect of prior strength (strong > weak) in DMPFC and RTPJ.

Updating vs. mere valence. To test the possibility that DMPFC and LTPJ are exhibiting a mere valence effect (i.e., greater activity to negative vs. positive behaviors), rather than an updating effect, we compared the *neural updating measure* for pos→neg sequences

with an analogous *non-updating measure* for control sequences. For example, we compared the *neural updating measure* for Weak Positive-to-Negative trials with an analogous measure for Negative Control trials: average activity to the middle two behaviors minus average activity to the first two behaviors.

If activity in DMPFC and LTPJ are solely tracking valence, then we would expect to see no differences between these measures in these ROIs. However, if DMPFC and LTPJ also track updating, then we would expect to see a greater change in activity on updating trials compared to non-updating trials.

When comparing Strong Positive-to-Negative trials to Negative Control trials, we found a significant effect of updating in both DMPFC (no random intercepts; $F(1, 296) = 16.41$, $p < 0.001$) and LTPJ (no by-item intercepts; $\chi^2(1) = 10.143$, $p = 0.001$). When comparing Weak Positive-to-Negative trials to Negative Control trials, we found a significant effect of updating in DMPFC ($\chi^2(1) = 7.046$, $p = 0.008$) and LTPJ (no by-item intercepts; $\chi^2(1) = 8.681$, $p = 0.003$).

These analyses suggest that DMPFC and LTPJ are responding not just to negative valence, but also to meaningful changes in behavior. For additional analyses comparing updating measures to non-updating measures, see **Supplementary Tables S3 and S4**.

Brain-behavior analyses in ToM ROIs. Within each ROI, we examined the relationship between neural updating and behavioral impression updating. Collapsing across sequence type, no brain-behavior relationship was observed in any of the ToM ROIs (DMPFC: ß = 0.007, SE = 0.026, $\chi^2(1) = 0.076$, $p = 0.782$; LTPJ: ß = -0.009, SE = 0.021, $\chi^2(1) = 0.195$, $p = 0.659$; RTPJ: ß = -0.033, SE = 0.020, $\chi^2(1) = 2.661$, $p = 0.103$; PC: ß = -0.015, SE =

0.020, $\chi^2(1) = 0.538$, $p = 0.463$). See **Supplementary Materials p. 6** for analyses within sequence type.

Encoding model analyses. We built a set of encoding models to predict activity in voxels evoked by stimulus features that varied between behavior positions (see **Table 3** for peak coordinates). We were chiefly interested in regions that track: (i) *prior strength*, controlling for position of current behavior, valence of current behavior, and whether a valence change occurred; (ii) *whether a valence change occurred*, controlling for position of current behavior, valence of current behavior, and prior strength; and (iii) *trial-wise impression updating*, controlling for position of current behavior. For other encoding model results, see **Supplementary Table S5;** for results from condition-based GLM analyses, see **Supplementary Tables S6-8**.

(i) *Prior strength*. Activity in posterior cingulum (which overlaps with precuneus as elicited by the Theory of Mind localizer task) parametrically covaried with the cumulative number of consecutive positive or negative behaviors presented.

Activity in left middle temporal gyrus (LTPJ) and left superior frontal gyrus tracked the cumulative number of consecutive positive behaviors presented.

Activity in left posterior cingulum, left calcarine fissure, left superior frontal gyrus, left middle temporal gyrus, and left angular gyrus tracked the cumulative number of consecutive negative behaviors presented.

(ii) *Whether a valence change occurred*. Activity in right superior frontal gyrus (DMPFC), precuneus, and right inferior frontal gyrus–orbital part (VLPFC, anterior insula) tracked valence reversals (pos→neg or neg→pos).

No significant clusters responded preferentially to pos→neg changes in valence, and no significant clusters responded preferentially to neg→pos changes in valence.

(iii) *Trial-wise impression updating*. Activity in left superior temporal pole (VLPFC/IFG), right inferior frontal gyrus–orbital part (VLPFC), left SMA, and left precentral gyrus tracked the magnitude of trial-wise behavioral updating (see **Supplementary Fig. S2** for visualization). Activity in left inferior frontal gyrus–orbital part (VLPFC/IFG), left superior frontal gyrus (DMPFC), right insula, left calcarine fissure, left caudate, right superior parietal gyrus, right caudate, and left middle temporal gyrus tracked trial-wise negative behavioral updating. No significant clusters tracked trial-wise positive behavioral updating.

Region name	x	y	z	t value	# voxels	ToM	VLPFC/IFG
Model D: cumulative number of consecutive positive or negative behaviors presented							
Posterior cingulum	−3	−49	31	8.39	259	PC	
Model A: cumulative number of consecutive positive behaviors presented							
Left middle temporal gyrus	−57	−55	19	5.22	219	LTPJ	
Left superior frontal gyrus	−3	53	19	5.04	312		
Model A: cumulative number of consecutive negative behaviors presented							
Left posterior cingulum	−3	−49	28	9.07	387		
Left calcarine fissure	−9	−97	−5	7.38	168		
Left superior frontal gyrus	−6	44	49	6.53	951		
Left middle temporal gyrus	−63	−19	−14	5.64	175		
Left angular gyrus	−57	−70	34	5.05	140		
Model F: valence reversal (from pos → neg or neg → pos)							
Right superior frontal gyrus (medial)	6	53	28	6.29	219	DMPFC	
Precuneus	9	−67	49	5.04	102		
Right inferior frontal gyrus (orbital part)	36	23	−8	4.94	180		RVLPFC
Model G: magnitude of trial-wise behavioral updating							
Left superior temporal pole	−42	20	−17	5.76	553		LVLPFC
Right inferior frontal gyrus (orbital part)	45	26	−8	5.52	256		RVLPFC
Left SMA	−9	26	55	5.39	1443		
Left precentral gyrus	−24	−7	46	4.56	239		
Model H: trial-wise negative behavioral updating							
Left inferior frontal gyrus (orbital part)	−36	20	−14	7.01	1081		LVLPFC, LIFG
Left superior frontal gyrus (medial)	−9	29	55	6.33	1649	DMPFC	
Right insula	42	23	−8	5.96	315		
Left calcarine fissure	−12	−91	−2	5.85	92		
Left caudate	−15	5	13	5.70	91		
Right superior parietal gyrus	18	−64	58	5.35	112		
Right caudate	15	8	13	4.98	96		
Left middle temporal gyrus	−57	−52	7	4.41	87		

Notes: See Table 2. Coordinates are provided in MNI space. All regions survived cluster-level correction (FWE, $P < 0.05$)

Table 3. Regions that track features from encoding models.

Brain-behavior analyses in lateral prefrontal ROIs. The above whole-brain parametric analyses revealed that left IFG and VLPFC track the magnitude of impression updating. To test the robustness of these findings, we conducted exploratory ROI analyses in these regions to examine correlations between changes in PSC and changes in trustworthiness ratings. ROIs were drawn as 9mm-radius spheres centered on peak coordinates from prior work showing that left IFG and left VLPFC respond preferentially to meaningful changes in behavior (IFG: [-58, 22, 18]; VLPFC: [-48, 27, -12]; Mende-Siedlecki and Todorov 2016).

In left IFG, there was a significant relationship between neural updating and behavioral impression updating (ß = 0.048, SE = 0.021, χ^2 (1) = 5.019, p = 0.025), such that a greater change in PSC was associated with a greater change in trustworthiness ratings. We found no evidence for such a relationship in left VLPFC (ß = -0.004, SE = 0.021, χ^2 (1) = 0.026, p = 0.871).

Conceptual replication of behavioral task. We had hypothesized an effect of the strength of the prior on the magnitude of impression updating, but found no such effect in our sample of scanned participants. We thus tested this effect in a pre-registered conceptual replication on Amazon Mechanical Turk (N = 400). Across participants, all 40 expectation-violation targets from the fMRI study were presented; the behavior sequences were the same exact sequences seen in the scanner. Due to time constraints of online data collection, each participant learned about 8 targets: 2 of each sequence type (strong neg→pos, weak neg→pos, strong pos→neg, weak pos→neg). Target order was randomized for each participant. We wanted to ensure that any differences we found

either related to prior strength or update direction could not be attributed to specific pairings of pre-switch and post-switch behaviors, or to target identity. Therefore, we shuffled the post-switch behaviors across participants so that they were seen following both weak and strong priors. Additionally, we counterbalanced target name and associated target face with update direction across participants.

On each behavior presentation, participants gave two types of ratings: trustworthiness of the target on a scale from 1 (least trustworthy) to 7 (most trustworthy), and attribution of the behavior from 1 (solely due to the target's disposition) to 100 (solely due to the surrounding situation). The order in which the rating scales were presented was counterbalanced across subjects; the attribution data are not reported in this paper. To examine behavioral impression updating, we computed an *updating measure* for each trial (see **Methods**). We conducted linear mixed effects analyses to test whether impression updating differed as a function of prior strength (weak or strong) and update direction (pos→neg or neg→pos). Random intercepts for subject and item were included in the model.

We found greater impression updating following the violation of weak vs. strong priors ($B = 0.090$, $SE = 0.034$, $\chi^2(1) = 7.182$, $p = 0.007$, $R_ß^2 = 0.003$), suggesting that the prior strength manipulation in our paradigm can have an effect on the degree of belief updating. In this larger dataset we also found more updating in the positive-to-negative direction vs. the negative-to-positive direction, consistent with the in-scanner dataset (pos→neg > neg→pos, $B = 0.700$, $SE = 0.034$, $\chi^2(1) = 400.45$, $p < 0.001$, $R_ß^2 = 0.133$).

2.4 DISCUSSION

In certain social contexts, observers have strong prior impressions that can co-occur with social motivations to maintain those impressions. In the current work, we aimed to isolate the impact of experimentally-induced prior beliefs on impression updating and related neural activity. Participants learned about novel fictional individuals whose behaviors were either internally inconsistent over time or internally consistent. The inconsistent individuals performed two or four same-valence behaviors, followed by two behaviors of the opposite valence. ROI analyses of the ToM network revealed a greater change in activity in DMPFC and RTPJ following the violation of strong vs. weak priors, and a greater change in activity in DMPFC and LTPJ following the violation of positive vs. negative priors. These findings show that the ToM network is sensitive to (1) violations of strong vs. weak prior beliefs and (2) the direction of impression change. Additional analyses showed that DMPFC and LTPJ respond to meaningful changes in behavior, not just to negative valence.

Contributions of prior strength and motivation

The present study manipulated participants' priors by providing different amounts of initially positive or negative information about targets. We showed, in a context absent real-life priors and social motivations, that ToM activity is enhanced following the violation of strong vs. weak prior beliefs. This suggests that differences in neural responses to close vs. distant others, and ingroup vs. outgroup members, may be driven in part by differences in the strength of prior beliefs.

We expected that participants' social motivations would be at floor for all targets in our study, as they were zero-acquaintance, fictional targets. We thus interpret differences in ToM activity following the violation of strong vs. weak prior beliefs as

arising from our experimental manipulation, rather than differences in social motivation. Importantly, however, in real-life situations, we expect belief strength and social motivation to often co-occur and operate in parallel: people not only know more about close others and ingroup members, but also are motivated to maintain relationships with close others (Park and Young 2020) and maintain positive impressions of ingroup members (Van Bavel and Pereira 2018). The relative degrees to which prior beliefs and motivation contribute to real-life belief updating and neural activity likely depend on social goals (e.g., to affiliate with others vs. to predict others' behavior; Waytz and Young 2014), context (e.g., dyads vs. groups), and individual differences (e.g., in mentalizing ability, cognitive reflection). An important future direction is to directly compare the impact of belief strength to the impact of motivation on updating and neural activity, across a variety of paradigms. Future work can, for example, take advantage of cases where participants' prior beliefs and motivations diverge. Participants could be presented information that is consistent or inconsistent with either their beliefs or their desires (Tappin et al. 2017), enabling a comparison of neural responses to prior-inconsistent and motivation-inconsistent information.

Prior strength and updating in the current study

Prior work has found that observers typically engage in less impression updating for close and ingroup others – targets for whom they have strong (positive) priors (Hughes et al. 2017; Park et al. in press). In the current study, we did not observe an effect of the prior strength manipulation on the magnitude of impression updating. However, behavioral evidence from pre-switch and post-switch ratings suggest a difference between the strong and weak prior manipulations: 1) pre-switch ratings based

on 4 positive behaviors were more positive than pre-switch ratings based on just 2 positive behaviors; 2) pre-switch ratings based on 4 negative behaviors were more negative than pre-switch ratings based on just 2 negative behaviors; and 3) post-switch ratings following 4 negative behaviors were more negative than post-switch ratings following 2 negative behaviors. Prior strength thus had an effect on initial impressions, and, to some extent, updated impressions, in a context absent social motivation.

Why were differences in neural activation following the violation of strong vs. weak priors not accompanied by differences in the magnitude of impression updating? One possibility is that enhanced ToM activity following the violation of strong priors supported belief maintenance on some trials, and belief updating on others. Both exculpatory and condemnatory explanations of behavior involve a mental state inference: for instance, upon learning that a target took money from a tip jar, one could infer that she intended to make change for a dollar, or that she intended to steal it. Thus, the enhanced mentalizing in light of strong priors could have resulted in a prior-consistent explanation in some cases, and a prior-inconsistent explanation in others. We might expect to find a stronger relationship between mentalizing activity and belief updating when real-life priors for individuals or groups are involved: these priors may be strong enough (and/or there may be enough motivation involved) that mentalizing activity chiefly supports belief maintenance in these contexts.

In addition, we may have had insufficient power in the current study to detect a behavioral effect of prior strength on update magnitude. We conducted a conceptual replication of the behavioral task on Amazon Mechanical Turk, where we presented participants with the same stimuli presented in the scanner (N = 400). We found greater

impression updating following the violation of weak vs. strong priors, suggesting that the prior strength manipulation in our paradigm can have an effect on the degree of belief updating. We discuss below, under **Future directions**, how the strength manipulation may be made more robust.

Distinct roles for ToM regions in tracking qualities of social information

Our ROI analyses revealed that DMPFC and RTPJ are sensitive to violations of strong vs. weak prior beliefs, while DMPFC and LTPJ track the direction of impression change. In addition, surprisingness ratings indicated that, while participants were more surprised when strong priors were violated than when weak priors were violated, there was no effect of update direction on the surprisingness of inconsistent behaviors. That is, surprising negative behaviors (which led to greater updating) were not rated as more surprising than surprising positive behaviors. These results together suggest that there may be distinct roles for different ToM regions in tracking separate qualities of new social information: its surprisingness and its valence. Furthermore, the encoding model analyses revealed that precuneus tracks the strength of the prior, regardless of valence, while LTPJ tracks the strength of positive priors specifically; in addition, DMPFC tracked whether the current behavior was opposite in valence to the previous behavior. These findings suggest that different ToM regions track distinct features of new behavioral information that are dependent on the nature of previous behavioral information.

Diagnosticity of immoral behaviors

Greater belief updating and ToM activity following the receipt of new negative information vs. positive information is consistent with a diagnosticity account of

impression updating (Mende-Siedlecki et al. 2013). This account posits that immoral behaviors and highly competent behaviors elicit greater impression updates than their counterparts because they are perceived to be less frequent, and thus more informative about a person's true character. While the moral and immoral behavior stimuli in our experiment were matched on perceived frequency, we still observed both greater impression updates in the positive-to-negative direction, and greater ToM activity when positive priors were violated by negative information. In addition, as described above, post-scan surprisingness ratings indicated that there was no effect of update direction on the surprisingness of inconsistent behaviors. This raises the possibility that, at least in the context of the current experiment, factors other than perceived frequency and surprisingness contributed to the dominance of immoral behaviors for impression updating.

Why do immoral behaviors shift impressions to such a great extent? Behavioral work by Brambilla and colleagues (2019) has shown that morally-relevant behaviors in general dominate impression updating (compared to behaviors related to sociality or competence) because they are seen as containing more information about interpersonal intentions. And, in line with our findings, their mediational analyses do not support a frequency-based account of the dominance of (im)moral behaviors for updating. Relatedly, reinterpretation has been shown to play a pivotal role in reversing initial (implicit) impressions (Mann and Ferguson 2017); thus, another possibility is that immoral behaviors are more powerful because they are likelier to lead to a reinterpretation of past behaviors. That is, it is easier to generate reputation-based explanations for someone's past moral behavior (e.g., *she did that only because it would*

make her look good), rather than to conceive of prosocial explanations for past immoral behavior (Reeder and Brewer 1979). Both of the above hypotheses can also potentially account for the enhanced mentalizing activity observed when new negative information contradicts positive priors.

Relationship between ToM activity and belief maintenance

While the ToM network typically responds preferentially to unpredicted events, and, as we have shown, is sensitive to the violation of strong vs. weak prior beliefs, the relationship between ToM activity and belief updating is more complex. In some contexts, greater ToM activity facilitates belief maintenance. For instance, one study found greater activity in DMPFC and bilateral TPJ when third-party observers viewed ingroup vs. outgroup defectors in the Prisoner's Dilemma Game (Baumgartner et al. 2012). Increased connectivity between DMPFC and LTPJ was associated with weaker punishment of ingroup defectors, and disrupting RTPJ activity through transcranial magnetic stimulation reduced relative forgiveness of ingroup defectors (Baumgartner et al. 2014). Thus, ToM activity may have supported the generation of exculpatory explanations for ingroup defectors. In this context, ingroup defection may be seen as inconsistent with strong positive priors about the ingroup, while outgroup defection may be seen as consistent with strong negative priors about the outgroup. Therefore, greater ToM activity following the more surprising event (ingroup defection) vs. the less surprising event (outgroup defection) dovetails with what we find in the current study: greater ToM activity following the more surprising event (violation of strong priors) vs. the less surprising event (violation of weak priors). In the intergroup study, greater ToM activity supported belief maintenance; in our zero-acquaintance study, greater ToM

activity was not mirrored by belief maintenance (at least in a sample of 28 participants). One possibility is that social motivation to maintain positive beliefs about the ingroup (absent from the present experimental paradigm) may have played a role in the intergroup context.

In other contexts, activity in the ToM network is selectively reduced when maintaining beliefs about close or ingroup others. One study found that observers failed to downgrade their impressions of ingroup members, but not outgroup members, following negative information; furthermore, overcoming this ingroup bias (to effectively downgrade impressions) was associated with increased activity in TPJ, precuneus, LPFC, and DACC (Hughes et al. 2017). Similarly, a recent fMRI study examined impression updating for friends and strangers who gave money to, or took money from, the participant in an economic game (Park et al. in press). Reduced RTPJ activity was observed in response to friends' taking money, compared to strangers' taking money; and this neural pattern was reflected in reduced behavioral updating for friends compared to strangers. However, within the friend-taking condition, greater RTPJ activity was associated with greater negative updating, indicating greater mentalizing effort required for overcoming strong positive priors about friends. Thus, in both of these studies, disengagement of ToM regions such as RTPJ was associated with impression maintenance for ingroup members and friends, and on the flip side, recruitment of ToM regions supported belief updating, particularly negative updating. Overall, these patterns suggest that, in some intergroup contexts and social relational (friend-stranger) contexts, the passive response to prior-inconsistent information about ingroup or close others may be to disengage from mentalizing, perhaps to discount unfavorable information. These

findings stand in contrast to what we find in a zero-acquaintance context: greater mentalizing in response to information that violates strong (vs. weak) priors and positive (vs. negative) priors.

ToM activity has been found to facilitate both belief maintenance and belief updating. Our interpretation of these mixed past results, together with the findings of the current study, is that two different mechanisms can result in the maintenance of strong prior beliefs (Kim et al. 2020). In one case, the violation of strong priors is followed by enhanced ToM activity, which may reflect the generation of a coherent mentalistic explanation of the unpredicted information (e.g., my ingroup member/close friend did not *intend* to defect/make an unfair offer). The generation of alternative explanations following the violation of strong priors is compatible with a form of Bayesian rationality, where the likelihood of generating an alternative explanation depends probabilistically on the strength of the prior belief, and the likelihood of the conflicting information (Gershman 2019). Alternatively, prior-inconsistent information may be followed by reduced ToM activity, due to disengagement from mentalizing about the target, which eliminates the need to reconcile the new information with prior beliefs. Overcoming this form of passive discounting may involve the intervention of cognitive control regions, such as DACC and LPFC (Hughes and Zaki 2015; Hughes, Ambady, et al. 2017). As we have proposed, activity in ToM and control regions, then, when coupled with behavioral evidence of belief maintenance, may help distinguish between the mentalizing route to belief maintenance, which is compatible with Bayesian rationality, and the discounting route to belief maintenance, which does not account for the unexpected information.

Predictive coding in the ToM network

Greater ToM activity following the violation of strong vs. weak prior beliefs is consistent with a predictive coding view of the social brain (see Koster-Hale and Saxe 2013; Theriault et al. in press for reviews), which holds that some neural responses indicate the distance between predictions from a generative model of the world and incoming sensory information (prediction error, PE). Prior work has shown that the ToM network responds more to unpredicted vs. predicted information across a wide variety of social stimuli and task contexts, including past behavioral history (Mende-Siedlecki et al. 2013; Dungan et al. 2016), instructed trait knowledge (Heil et al. 2019), and stereotypes (Cloutier et al. 2011; Li et al. 2016). The current findings demonstrate that the ToM network is sensitive to different degrees of unpredictedness during impression updating: activity in this network was enhanced for information that violated strong prior beliefs vs. information that violated weak prior beliefs. These results are also in line with computational neuroimaging work showing that PEs generated during associative learning of social value correlate with activity in ToM regions (Behrens et al. 2008; Hackel et al. 2015).

A broader network for impression updating

Our whole-brain analyses revealed two additional regions beyond the ToM network that were consistently activated during impression updating: ventrolateral prefrontal cortex (VLPFC) and inferior frontal gyrus (IFG). In particular, VLPFC and IFG showed greater changes in activity for positive-to-negative sequences than negative-to-positive sequences. This pattern is consistent with previous work (Mende-Siedlecki and Todorov 2016) showing that these regions respond preferentially to moral-to-immoral changes in moral behavior, relative to immoral-to-moral changes in behavior,

and relative to non-meaningful changes in behavior (e.g., “Jenny went for a bike ride”; “Jenny went for a run”; “Jenny played video games”). This past study interpreted left VLPFC activity as reflecting the retrieval of stored conceptual representations, and left IFG activity as reflecting the process of resolving interference between representations (Badre et al. 2005; Badre and Wagner 2007; Satpute et al. 2014). In the current study as well, we suggest that activity in these regions is an instantiation of these more general cognitive processes, recruited to a greater degree for information that prompts an update to stored representations. In addition, our encoding model analyses revealed that bilateral VLPFC parametrically tracked the magnitude of trial-wise impression updating, and left VLPFC and left IFG parametrically tracked the magnitude of trial-wise negative impression updating. PSC analyses also revealed that greater changes in neural activity in left IFG are associated with a greater change in impression ratings from pre-switch behaviors to post-switch behaviors. Overall these results indicate that VLPFC and left IFG track changes in actual rated impressions, especially in the negative direction, during the processing of diagnostic information.

Future directions

First, in the current work, we probed the effect of different experimentally-induced priors on updating and ToM activity. It may be fruitful to manipulate whether or not updating occurs through the use of participant instructions (Trafimow and Porter 1997). That is, participants can be instructed, across blocks, to either (a) use expectation-violating information to update their prior impressions, or (b) use their prior impressions to reinterpret the expectation-violating information. This approach would allow for a direct comparison of neural activity, both in terms of magnitude of activation and patterns

of activation, for belief updating vs. belief maintenance. Recent work has shown that social information is neurally represented along a small set of representational dimensions, which in turn can facilitate the prediction of others' future mental states and actions (Tamir and Thornton 2018); analyzing patterns of brain activity elicited by our paradigm will also allow us to examine how neural representations of targets change in light of expectation violations.

Second, in the current work, we manipulated the number of same-valenced behaviors presented (2 vs. 4) before a counter-valenced behavior was presented, to induce stronger vs. weaker beliefs about the target. A limitation of this paradigm is that, in a sample of 28 participants, this manipulation was not strong enough to induce differences in update magnitude. One possibility is that there needs to be a greater difference in the number of initial behaviors (e.g., 2 vs. 6) to observe an effect on the magnitude of impression updating. An alternative way to manipulate the strength of the prior is to vary the *extremity* of behavioral information, rather than the amount of information. For example, a target who performed 2 extremely negative behaviors could be compared to a target who performed 2 mildly negative behaviors. Future work may benefit from exaggerating the diagnostic difference between strong and weak priors in this way. Yet another important future direction would be to directly compare these two implementations within the same paradigm: strong beliefs stemming from more evidence, vs. strong beliefs stemming from a more extreme piece of evidence.

Third, as we tested the impact of the strength of priors in the context of zero-acquaintance targets, the current study cannot speak to the relative importance of belief strength and motivation for real-life belief updating and neural activity. Future work

should either pit belief strength and motivation against each other, or take advantage of cases in which they diverge in participants, and then provide information that is consistent or inconsistent with beliefs or desires (Tappin et al. 2017). This unique approach would allow for the comparison of the effects of belief strength and motivation on behavioral and neural indices of updating.

Finally, another interesting area for further research is the dominance of immoral information in impression updating. Future work may explore the hypothesis that immoral behaviors are more important for updating than moral behaviors because they contain more intent information. Another possibility is that immoral behaviors are likelier to lead to a reinterpretation of past moral behaviors than vice versa. Furthermore, the boundaries of this valence effect are of interest – for example, recent work has found that, in a context where character ratings are made relative to a single type of moral behavior that evolves over time, beliefs about initially bad agents are more volatile, and thus more amenable to updating (Siegel et al. 2018).

Summary

We manipulated participants' initial beliefs about fictional targets by varying the amount of positive and negative information about targets' past behaviors. In this zero-acquaintance context, we found that activity in DMPFC and RTPJ is enhanced for information that violates strong vs. weak prior beliefs, and activity in DMPFC and LTPJ is enhanced for information that violates positive vs. negative prior beliefs. Thus, absent social motivation, differences in belief strength and belief valence can lead to differences in ToM in response to new information. These results can be compared to past work directly manipulating motivation: some studies have shown enhanced ToM for surprising

information about close others and ingroup members, while others have shown decreased ToM. We suggest that, in real-life contexts, increased mentalizing activity in light of strong positive priors may reflect the generation of alternative explanations, whereas decreased mentalizing activity may reflect motivated discounting of unfavorable information.

2.5 REFERENCES

Badre D, Poldrack RA, Paré-Blagoev EJ, Insler RZ, Wagner AD. 2005. Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. *Neuron*. 47(6):907–918.

Badre D, Wagner AD. 2007. Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia*. 45(13):2883–2901.

Bates D, Mächler M, Bolker B, Walker S. 2014. Fitting linear mixed-effects models using lme4. *arXiv*.

Baumgartner T, Götte L, Gügler R, Fehr E. 2012. The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Hum Brain Mapp*. 33(6):1452–1469.

Behrens TE, Hunt LT, Woolrich MW, Rushworth MF. 2008. Associative learning of social value. *Nature*. 456(13):245–250.

Brambilla M, Carraro L, Castelli L, Sacchi S. 2019. Changing impressions: Moral character dominates impression updating. *J Exp Soc Psychol*. 82:64–73.

Chiu C-Y, Dweck CS, Tong JY-Y, Fu H-Y. 1997. Implicit theories and conceptions of morality. *J Pers Soc Psychol*. 73(5):923–940.

Cloutier J, Gabrieli JD, O'Young D, Ambady N. 2011. An fMRI study of violations of social expectations: When people are not who we expect them to be. *Neuroimage*. 57(2):583–588.

DeShea L. 2003. A scenario-based scale of willingness to forgive. *Individ Differ Res*. 1(3):201–217.

Dodell-Feder D, Koster-Hale J, Bedny M, Saxe R. 2011. FMRI item analysis in a theory of mind task. *Neuroimage*. 55(2):705–712.

Dovidio JF, Hewstone M, Glick P, Esses VM. 2010. Prejudice, stereotyping and discrimination: Theoretical and empirical overview. The SAGE handbook of prejudice, stereotyping and discrimination. 80:3–28.

Dungan JA, Stepanovic M, Young L. 2016. Theory of mind for processing unexpected events across contexts. *Sog Cogn Affect Neurosci*. 11(8):1183–1192.

Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, Kent JD, Goncalves M, DuPre E, Snyder M, others. 2019. FMRIPrep: A robust preprocessing pipeline for functional mri. *Nat Methods*. 16(1):111–116.

Ferarri C, Lega C, Vernice M, Tamietto M, Mende-Siedlecki P, Vecchi T, Todorov A, Cattaneo Z. 2016. The dorsomedial prefrontal cortex plays a causal role in integrating social impressions from faces and verbal descriptions. *Cereb Cortex*. 26(1): 156-165.

Gershman SJ. 2019. How to never be wrong. *Psychol Bull Rev*. 26(1):13–28.

Hackel LM, Amodio DM. 2018. Computational neuroscience approaches to social cognition. *Curr Opin Psychol*. 24:92–97.

Hackel LM, Doll BB, Amodio DM. 2015. Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nat Neurosci*. 18(9):1233–1235.

Heil L, Colizoli O, Hartstra E, Kwisthout J, Pelt S van, Rooij I van, Bekkering H. 2019. Processing of prediction errors in mentalizing areas. *J Cogn Neurosci*. 31(6):900–912.

Hughes BL, Ambady N, Zaki J. 2017. Trusting outgroup, but not ingroup members, requires control: Neural and behavioral evidence. *Sog Cogn Affect Neurosci*. 12(3):372–381.

Hughes BL, Zaki J, Ambady N. 2017. Motivation alters impression formation and related neural systems. *Sog Cogn Affect Neurosci*. 12(1):49–60.

Hughes BL, Zaki J. 2015. The neuroscience of motivated cognition. *Trends Cogn Sci*. 19(2):62–64.

Hughes JS. 2015. Support for the domain specificity of implicit beliefs about persons, intelligence, and morality. *Pers Individ Differ*. 86:195–203.

Kim M, Park B, Young L. 2020. The psychology of motivated versus rational impression updating. *Trends Cogn Sci.* 24(2):101–111.

Koster-Hale J, Saxe R. 2013. Theory of mind: A neural prediction problem. *Neuron*. 7(5)9:836–848.

Li T, Cardenas-Iniguez C, Correll J, Cloutier J. 2016. The impact of motivation on race-based impression formation. *Neuroimage*. 124:1–7.

Lundqvist D, Flykt A, Öhman A. 1998. The Karolinska directed emotional faces. CD ROM from Psychology section, Department of Clinical Neuroscience, Karolinska Hospital.

Ma N, Vandekerckhove M, Baetens K, Van Overwalle F, Seurinck R, Fias W. 2012. Inconsistencies in spontaneous and intentional trait inferences. *Sog Cogn Affect Neurosci*. 7(8):937–950.

Malle BF. 2001. Folk explanations of intentional action. In: Intentions and intentionality: Foundations of social cognition. Cambridge (MA): MIT Press. p 265–286.

Mann TC, Ferguson MJ. 2017. Reversing implicit first impressions through reinterpretation after a two-day delay. *J Exp Soc Psychol*. 68:122–127.

Mende-Siedlecki P, Baron SG, Todorov A. 2013. Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *J Neurosci*. 33(50):19406–19415.

Mende-Siedlecki P, Cai Y, Todorov A. 2012. The neural dynamics of updating person impressions. *Sog Cogn Affect Neurosci*. 8(6):623–631.

Mende-Siedlecki P, Todorov A. 2016. Neural dissociations between meaningful and mere inconsistency in impression updating. *Sog Cogn Affect Neurosci*. 11(9):1489–1500.

Mumford JA, Poline J-B, Poldrack RA. 2015. Orthogonalization of regressors in fMRI models. *PloS One*. 10(4).

Muschelli J, Nebel MB, Caffo BS, Barber AD, Pekar JJ, Mostofsky SH. 2014. Reduction of motion-related artifacts in resting state fMRI using aCompCor. *Neuroimage*. 96:22–35.

Park B, Fareri D, Delgado M, Young L. Provisional acceptance. How theory-of-mind brain regions process prediction error across relationship contexts. *Soc Cogn Affect Neurosci.*

Park B, Young L. 2020. An association between biased impression updating and relationship facilitation: A behavioral and fMRI investigation. *J Exp Soc Psychol*. 87:103916.

Peirce JW. 2007. PsychoPy–psychophysics software in Python. *J Neurosci Methods*. 162:8–13.

RCore T. 2016. R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria.

Reeder GD, Brewer MB. 1979. A schematic model of dispositional attribution in interpersonal perception. *Psychol Rev*. 86(1):61–79.

Satpute AB, Badre D, Ochsner KN. 2014. Distinct regions of prefrontal cortex are associated with the controlled retrieval and selection of social information. *Cereb Cortex*. 24(5):1269–1277.

Saxe R, Kanwisher N. 2003. People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *Neuroimage*. 19(4):1835–1842.

Siegel JZ, Mathys C, Rutledge RB, Crockett MJ. Beliefs about bad people are volatile. *Nat Hum Behav.* 2(10): 750-756.

Tamir DI, Thornton MA. Modeling the predictive social mind. *Trends Cogn Sci.* 22(3): 201-212.

Theriault J, Young L, Barrett LF. In press. The sense of should: A biologically-based model of social pressure. *Phys Life Rev*.

Trafimow D, Porter PP. 1997. A comparison of updating and explanation as causes of the incongruity effect on person memory. *J Soc Psychol*. 137(4):412–420.

Van Bavel JJ, Pereira A. 2018. The partisan brain: An identity-based model of political belief. *Trends Cogn Sci*. 22(3):213–224.

Young L, Waytz A. 2013. Mind attribution is for morality. In: Understanding other minds: Perspectives from developmental social neuroscience. Oxford (UK): Oxford University Press. p 93–103.

3.0 IMPRESSION UPDATING REVEALS AN UNEXPECTED NEGATIVITY BIAS IN 6-9-YEAR-OLDS

3.1 INTRODUCTION

Flexible trait reasoning is vital for everyday social function. Our ability to make broad inferences of people's internal traits (e.g., "nice" or "mean") from a small sample of behaviors allows us to (1) predict how people will behave across a variety of future contexts, and (2) make informed decisions about who to approach, befriend, and trust. Critically, people are dynamic sources of information – their behaviors may shift in a way that prompts us to revise our prior impression of them. For instance, if Amy helps her teacher clean the classroom, her classmates may infer that she is a nice person, and predict that she will continue to be a good social partner; but if Amy then steals her friend's pencil case, her classmates should perhaps update their impression of her, and change their predictions of Amy's future behavior.

Trait reasoning in children. The ability to make flexible trait inferences from behavioral information may be particularly advantageous during middle childhood, a developmental period where social relationships are rapidly increasing in size and complexity, and children are processing large amounts of social information that can impact their interpersonal relationships and self-concepts. Past work has documented a "positivity bias" in trait understanding that emerges during early childhood and persists throughout middle childhood, in which children are relatively eager to make positive trait attributions, and relatively reluctant to make negative trait attributions, for the self and for others; Boseovski (2010) reviews the findings that support this positivity bias. For one, in a study of 5-10-year-olds' explanations for successes and failures in the social and

academic domains, Benenson & Dweck (1986) found that trait explanations for others' successes emerged by the 1st grade, but trait explanations for others' failures did not emerge until the 4th grade; there was also a decline in positive self-evaluations with age. In addition, children require different amounts of evidence to make positive vs. negative attributions: 3- to 6-year-olds only need to view 1 positive behavior in order to make the corresponding trait attribution ('nice"), but they need to view 5 negative behaviors to make "mean" attributions – i.e., they require stronger evidence for negative trait judgments (Boseovski & Lee, 2006). This asymmetry in the evidential threshold was also observed for mental-state reasoning: kindergarteners to 4th graders are reluctant to attribute intentionality to an actor who performed 3 negative behaviors (Jones & Thomson, 2001).

Children may also use positive evidence selectively for their judgments: 5- and 6-year-olds judge a character positively after hearing many positive behaviors followed by a single negative behavior, *and* after hearing many negative behaviors followed by a single positive behavior (Rholes & Ruble, 1986). Furthermore, 7- and 8-year-olds endorse the stability of positive sociomoral traits more than the stability of negative sociomoral traits (Heyman & Dweck, 1998). The positivity bias extends to the intergroup context as well: 5- and 6-year-olds make overly positive behavioral attributions to ingroup vs. outgroup members (Dunham, Baron, & Carey, 2011; Baron & Dunham, 2015).

Slightly older children display a weaker positivity bias: compared to 5- and 6-year-olds, 7- to 9-year-olds assume less stability in positive traits (Lockhart, Chang, & Story, 2002). In addition, 9- and 10-year-olds care about behavior frequency during trait

attribution – they base their judgments on whether there were more positive or negative behaviors (i.e., they do not selectively use positive information; Rholes & Ruble, 1986). Further, in contrast to 6-year-olds, both 10-year-olds and adults perceive negative information as more diagnostic about a person than positive information (Newman, 1991). In all, the positivity bias appears to emerge by age 3, peak in middle childhood, and begin to erode by age 10 (Boseovski, 2010).

Impression updating in children. While past work has established a positivity bias in children's trait reasoning, relatively little work has examined children's impression *updating* – the ability to change one's initial impression in light of new, inconsistent information. A positivity bias in impression updating may play an especially important role in establishing and maintaining friendships, by serving a protective function for pre-existing friendships that are threatened by a mistake, and for nascent relationships that get off on the wrong foot. For example, optimism may allow children to assume positive intentions in others who (appear to) do wrong, and allow them to engage in trust repair through forgiveness or rearrangement of the relationship (Boseovski, 2010; Lewicki & Brinsfield, 2017). Thus, we may expect that children will be relatively slow to revise positive impressions, and relatively quick to revise negative impressions. A key aim of the current study was to test this hypothesis by explicitly manipulating (1) the nature of children's initial impressions of targets, and (2) the nature of new information that elicits impression updating.

Impression updating in adults. A large body of past work has explored the factors that matter for impression updating in adults. Past research has shown that information related to morality has special status in impression updating: Brambilla and colleagues

found that observers revise their impressions to a greater degree after learning about behaviors related to morality, compared to behaviors related to competence or sociability (Brambilla, Carraro, Castelli, & Sacchi, 2019). A mediation analysis revealed that morally-relevant behaviors may lead to greater updating because they are perceived as more diagnostic of people's interpersonal intentions. This finding is in line with the idea that a key function of person perception is figuring out how to interact with others.

In contrast with the positivity bias in trait understanding in middle childhood, there is a negativity bias in adults, where "bad is stronger than good" (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Cone & Ferguson, 2015; Mende-Siedlecki, Baron, & Todorov, 2013; Kim, Mende-Siedlecki, Anzellotti, & Young, 2021; Reeder & Coovert, 1986; Rozin & Royzman, 2001; Skowronski & Carlston, 1989; but this bias is reversed when thinking about the self, see Sharot & Garrett, 2016). For instance, when observers are provided with both positive and negative trait information about a target, the unfavorable traits contribute more to global impression ratings than a simple averaging model would predict (Anderson, 1965); furthermore, when asked to judge a target's moral character from separate behavioral examples, participants revise their impressions more when *moral* examples are followed by *immoral* examples, compared to the opposite (Kim, Mende-Siedlecki, Anzellotti, & Young, 2021; Mende-Siedlecki, Baron, & Todorov, 2013; Reeder & Coovert, 1986). This negativity bias has been linked to a cue-diagnosticity mechanism: negative behaviors may drive impressions because they are perceived as more reflective of true character (i.e., both good and bad people have reasons to behave morally, but good people will rarely behave immorally; Skowronski & Carlston, 1987; Mende-Siedlecki, Baron, & Todorov, 2013; Cone &

Ferguson, 2015).

Negativity bias in early development. Valence asymmetries in social reasoning may undergo two reversals throughout development: the positivity bias in trait reasoning in early and middle childhood seems to be preceded by a general negativity bias in early development. Vaish, Grossmann, & Woodward (2008) review the evidence for this negativity bias. Twelve-month-olds will touch a novel toy less if their mother conveys negative emotions towards it using tactile cues (Hertenstein & Campos, 2001), and 14-month-olds will touch a toy less if an adult treats it with disgust, but infants' approach behavior is less affected by positive cues (Hertenstein & Campos, 2004). Such findings suggest an adaptive function for an early negativity bias (we are likelier to survive if we quickly learn which stimuli to avoid; Vaish, Grossmann, & Woodward, 2008). In addition, 1.5- to 2.5-year-old girls talk more about negative memories with their mothers (Miller & Sperry, 1988), and preschoolers talk to their mothers more often about causes of unpleasant emotions (Dunn & Brown, 1993). Negative events may be discussed more because negative emotions are more disruptive and have to be better regulated, or because negative events induce more motivation for resolution (Lagattuta & Wellman, 2002; Fivush et al., 2003; Vaish, Grossmann, & Woodward, 2008).

The potential trajectory from a negativity bias in infancy, to a positivity bias in childhood, and back to a negativity bias in adulthood raises questions on the causes of the positivity bias. Boseovski (2010) summarizes accounts of adaptive immaturity and socialization: young children may need a positive outlook (especially regarding their own traits) in order to persevere in learning a large number of new skills, and parents may promote this optimism by attributing positive behaviors to dispositional rather than

situational factors. It may be that, as children get older, it becomes more important to view the self and others in a more realistic light, and to be more vigilant about potentially threatening actors.

Open questions. While past work has established a positivity bias in children's trait reasoning, relatively little work has examined children's impression *updating* by explicitly manipulating the nature of the initial impression and the nature of trait-inconsistent information. There remain many open questions on flexible trait reasoning during middle childhood. For one, when children are asked to keep track of their impression of a target across multiple behavioral examples, will they exhibit a positivity bias, where they revise their impression more when a previously mean person does something nice (vs. when a previously nice person does something mean)? In addition, past work suggests that the strength of behavioral evidence may matter for impression updating. Will children exhibit a sensitivity to the cumulative strength of the behavioral evidence, such that impressions based on more behavioral examples get updated less than impressions based on fewer behavioral examples? Will this sensitivity depend on the valence of the behaviors? Furthermore, an important part of flexible trait reasoning is the appreciation of the predictive value of traits. When children are provided with mixed evidence (i.e., both positive and negative behaviors), how will they make predictions about targets' future behaviors? Exploring these questions is important for understanding how children respond to different kinds of inconsistencies in other people's behaviors, how they integrate social information over time, and how they engage in behavior-to-behavior prediction.

Current research. In the present study, we investigated 6- to 9-year-olds' flexible

trait reasoning by examining impression updating as a function of (1) the direction of change of a target's behavior, and (2) the strength of the initial behavioral evidence.

In our task, participants learned about fictional child targets who each performed a sequence of six behaviors. Behaviors were depicted in storybook-style illustrations using Vyond software, and described aloud to participants. Following the presentation of each behavior, participants evaluated the target's niceness on a 9-point scale (1 = "super mean", 9 = "super nice"). After learning about each target's sequence of behaviors, participants responded to two measures of predicted trustworthiness (described below). All stimuli and measures were embedded in a Qualtrics survey that was experimenter-controlled and shown to participants via screensharing (Sheskin et al., 2020).

Impression updating was examined for six target conditions. Some of the targets' behaviors changed over time in an attributionally meaningful fashion (*updating conditions*); other targets' behaviors did not change meaningfully over time (*control conditions*). For updating conditions, we varied both the direction of behavior change (*positive-to-negative* vs. *negative-to-positive*), and the strength of the initial behavioral evidence (*weak* vs *strong*). These targets were initially described as performing either 2 or 4 behaviors of the same valence (corresponding to weak or strong initial impressions); they were then described as performing two behaviors of the opposite valence (prompting an impression update). Control targets were described as performing either 6 positive behaviors or 6 negative behaviors. The magnitude of impression updating was compared across target conditions.

Age range. We recruited 6- to 9-year-olds in order to capture both a potential positivity bias, and a potential transition point in sensitivity to informational features. It

may be that the older children in this age range not only differentiate between positive-to-negative and negative-to-positive shifts in people's dispositions, but also differentiate between more vs. less extreme changes, because they are starting to care about features other than behavior valence (such as behavior frequency; Rholes & Ruble, 1986). Any such age-related effects may point to aspects of dynamic social evaluations that are bounded by cognitive function. We hypothesized that (1) children will update negative impressions more than positive impressions when provided with countervailing evidence, and (2) as children get older, they will become more sensitive to the strength of the behavioral evidence, such that they engage in more updating for impressions based on 2 (vs. 4) behavioral examples.

Behavioral predictions. In addition to probing children's impression updating, we also examined their ability to use trait inferences to make behavioral predictions. To do this, we asked hypothetical questions of the form, "Do you trust this kid to look after your belongings?", and "Will this kid share the spoils of your initial investment with you?", which assess a type of trust where one makes themselves vulnerable to the actions of another person (Rousseau et al., 1998). Previous studies on trust in children have often focused on trust in adult testimony (e.g. Chen & Harris, 2012; Koenig & Stevens, 2014), but findings from these studies may be more indicative of children's beliefs about who will have the best information; other work has examined trust behavior in one-shot economic games (e.g. Sutter & Kocher, 2007), but in such contexts there is no possibility of learning about one's partner through repeated exposure. We thus sought to collect preliminary data on how children combine multiple behavioral examples to make predictions about targets' trustworthiness in hypothetical scenarios. It may be the case

that children can systematically revise impressions over time, but fail to integrate this information into behavioral predictions, especially when the behavioral evidence is mixed, and when prediction involves generalization across traits (niceness → trustworthiness).

Social agents continuously give us new information to evaluate, and sometimes prompt us to revise what we think of them. Examining children's real-time impression updating will inform our understanding of the flexibility and limitations of children's implicit trait theories, and also shed light on the potential interpersonal functions of valence asymmetries in social information processing.

3.2 METHODS

Participants

All participants (N = 162) were from the United States and were tested online in moderated Zoom sessions between October 7th 2020 and February 15th 2022.[1] A target sample size of N = 160 was determined by a power simulation (*simr* package; Green & MacLeod, 2016) which revealed that we would have 90% power to detect an interaction effect size of 0.2 between: strength of the initial impression, valence of the initial impression, and participant age in months. Participants were recruited via a lab database and social media advertisements. While we initially recruited 172 participants, 9 participants were excluded from data analysis for meeting our preregistered exclusion

[1] Prior to testing participants over Zoom, we had tested 58 participants in-person between July 2019 and February 2020; when in-person studies were paused in March 2020, we preregistered that we would run an additional 58 participants online, then test for an effect of study modality. There was a significant effect of study modality on the main dependent variable, so we discarded the in-person data (as preregistered), then collected data from 104 additional online participants (see **Supp. Mat. p. 4** for details on this procedure).

criteria, including: experimenter/technical error (3), session interruption (1), parental intervention (2), and parental report of atypical development (3); 1 additional participant was excluded for not being able to consistently read the scale labels.

Participants were recruited into two age groups: 6-7-year-olds (N = 80, M = 7.00, SD = 0.56, range = 6.00-7.92, 53.8% females) and 8-9-year-olds (N = 82, M = 8.89, SD = 0.59, range = 8.00-9.83, 51.2% females). We aimed to recruit approximately 40 participants per age year, and aimed to recruit no more than 70% of each age year from a single gender (see **Table S1** for breakdown of participants by age, gender, and condition group). Participants were majority white (White = 53.1%, Asian = 11.1%, Hispanic = 6.2%, Black = 3.7%, Native American = 0.1%, Biracial = 19.1%, Other = 3.7%, unreported = 2.5%).

Participants were given a $5 Amazon gift card for their participation. The study was approved by the Boston College Institutional Review Board and preregistered prior to data collection (https://aspredicted.org/1MQ_QCQ).

Materials

Piloting behavior stimuli. A candidate set of 148 one-sentence behavior stimuli was first pre-tested in a sample of adult MTurk participants to allow for optimal stimulus selection. This pilot study allowed us to select 10 positive behaviors and 10 negative behaviors for use in the experiment that were balanced on: moral relevance, perceived frequency, absolute value trustworthiness, absolute value emotional valence, and arousal. We also selected 2 neutral behaviors that were rated low on: moral relevance, absolute value trustworthiness, and absolute value emotional valence. We then tested these stimuli

in a sample of in-person child participants, to verify that the selected behaviors elicited appropriate niceness ratings (high for positive behaviors, low for negative behaviors, moderate for neutral behaviors). See **Supplementary Materials p. 1** for detailed methods for pilot studies, and **Table S2** for full stimulus text and feature ratings for each stimulus.

Conditions. Each target character was paired with a sequence of six behaviors, designed to be internally consistent or inconsistent in terms of implied niceness. There were four updating conditions, created by crossing the valence and strength of the initial impression (**Table 1**): *Weak Positive-to-Negative* (2 positive behaviors followed by 2 negative then 2 neutral), *Strong Positive-to-Negative* (4 positive, 2 negative), *Weak Negative-to-Positive* (2 negative, 2 positive, 2 neutral), and *Strong Negative-to-Positive* (4 negative, 2 positive). Two neutral behaviors were added to the ends of *Weak* sequences to keep sequence length constant across conditions. For discussion purposes, the two behaviors immediately preceding the valence switch point will be referred to as "pre-switch" behaviors, while the two behaviors immediately following the valence switch will be referred to as "post-switch" behaviors. Additionally, there were two control conditions (**Table 1**): *Positive Control* (6 positive behaviors) and *Negative Control* (6 negative behaviors).

	Strength	Valence	Behavior #1	Behavior #2	Behavior #3	Behavior #4	Behavior #5	Behavior #6
Condition Group 1	Weak (2 initial behaviors)	Positive-to-Negative	**Amy helped a player on the opposite team who fell during soccer**	**Amy helped her sister look for her missing teddy bear**	**Amy pushed the younger kids out of the way to get to the slide**	**Amy laughed at a kid who fell from the monkey bars**	Amy got into bed	Amy drank a glass of water
	Strong (4 initial behaviors)	Positive-to-Negative	Becca brought treats for the class pet	Becca shared her dessert with others at lunch	**Becca lent her basketball to the kid next door**	**Becca gave half of her candy bar to someone at the park**	**Becca smashed a friend's lego tower**	**Becca pushed a classmate on the playground**
	Control	Negative	Carla stole her friend's pencil case	Carla stole a friend's cookie when they weren't looking	Carla threw food at a classmate during lunch time	Carla snatched her sister's toy out of her hand	Carla refused to share ice cream with her siblings	Carla laughed at another kid for not knowing how to throw a baseball
Condition Group 2	Weak (2 initial behaviors)	Negative-to-Positive	**Alex pushed a classmate on the playground**	**Alex smashed a friend's lego tower**	**Alex helped his brother look for his missing teddy bear**	**Alex helped a player on the opposite team who fell during soccer**	Alex got into bed	Alex drank a glass of water
	Strong (4 initial behaviors)	Negative-to-Positive	Ben stole his friend's pencil case	Ben stole a friend's cookie when they weren't looking	**Ben laughed at a kid who fell from the monkey bars**	**Ben pushed the younger kids out of the way to get to the slide**	**Ben gave half of his candy bar to someone at the park**	**Ben lent his basketball to the kid next door**
	Control	Positive	Chris shared his dessert with others at lunch	Chris brought treats for the class pet	Chris helped his teacher clean the classroom	Chris congratulated his brother for wining a soccer trophy	Chris gave a dollar to a friend who wanted a soda	Chris told someone their painting was nice

Table 1. Target conditions and example behavior sequences for each condition. The pre-switch behaviors in each sequence are bolded; the post-switch behaviors in each sequence are bolded and highlighted. Half of participants learned about: one Weak Positive-to-Negative target, one Strong Positive-to-Negative target, and one Positive Control target in a randomized order (Condition Group 1). The other half of participants learned about: one Weak Negative-to-Positive target, one Strong Negative-to-Positive target, and one Negative Control target in a randomized order (Condition Group 2). Two out of eight stimulus lists are presented here; see **Table S3** for description of all lists and how they were counterbalanced.

Design. Each participant learned about three targets (due to time constraints). Half of participants learned about: one Weak Positive-to-Negative target, one Strong Positive-to-Negative target, and one Negative Control target in a randomized order (Condition Group 1; **Table 1**). The other half of participants learned about: one Weak Negative-to-Positive target, one Strong Negative-to-Positive target, and one positive Control target in a randomized order (Condition Group 2; **Table 1**). Thus, when considering the updating conditions, the valence of the initial impression was a between-subjects factor, and the strength of the initial impression was a within-subjects factor.

Eight stimulus lists were created to allow for counterbalancing of behaviors and target names across conditions; participants were assigned to one of these lists in a pseudorandom manner. We employed a number of counterbalancing schemes to reduce the possibility that any differences in update magnitude across conditions will be due to properties of specific behaviors or participants' chance associations with names; see **Table S3** for detailed description of counterbalancing schemes.

Procedure

The study session was conducted remotely over Zoom and took no more than 30 minutes. Children provided verbal consent to participate and to be video recorded, and confirmed that they could track the cursor on the experimenter's shared screen.

Scale training. At the beginning of the study, participants were trained on a 9-point Likert-type scale, called a "Smiley-meter", comprised of cartoon faces whose expressions ranged from very angry to very happy (Davies & Brember, 1994; **Figure 1**). Participants were told that they will hear some stories about different kids, and answer the question "How nice is this kid?" using the Smiley-meter. Participants were told what

each scale point means, from right-to-left: *super nice, very nice, nice, kinda nice, just ok, kinda mean, mean, very mean, super mean*. Each scale point was consistently labeled with a letter below the cartoon face (*D, H, B, F, I, E, G, A, C*) to facilitate verbal response selection.

Participants practiced using the Smiley-meter on one positive behavior ("[Ellen/Kevin] shared [her/his] lunch with a friend who didn't bring lunch"; accepted responses: *kinda nice, nice, very nice, super nice*), one negative behavior ("[Megan/Joshua] stole someone's snack on the bus"; accepted responses: *kinda mean, mean, very mean, super mean*), and one neutral behavior ("[Hannah/Justin] took a bath"; accepted response: *just ok*). Participants had three attempts to answer each practice trial correctly; if they answered incorrectly, they were reminded what each scale point means, then asked again. If they answered incorrectly on their third attempt, they were told the correct answer. The number of attempts required to pass each practice trial was recorded (1: "spontaneously correct"; 2: "correct with explanation"; 3: "experience choosing answer"). Participants were not excluded on the basis of their performance on practice trials.

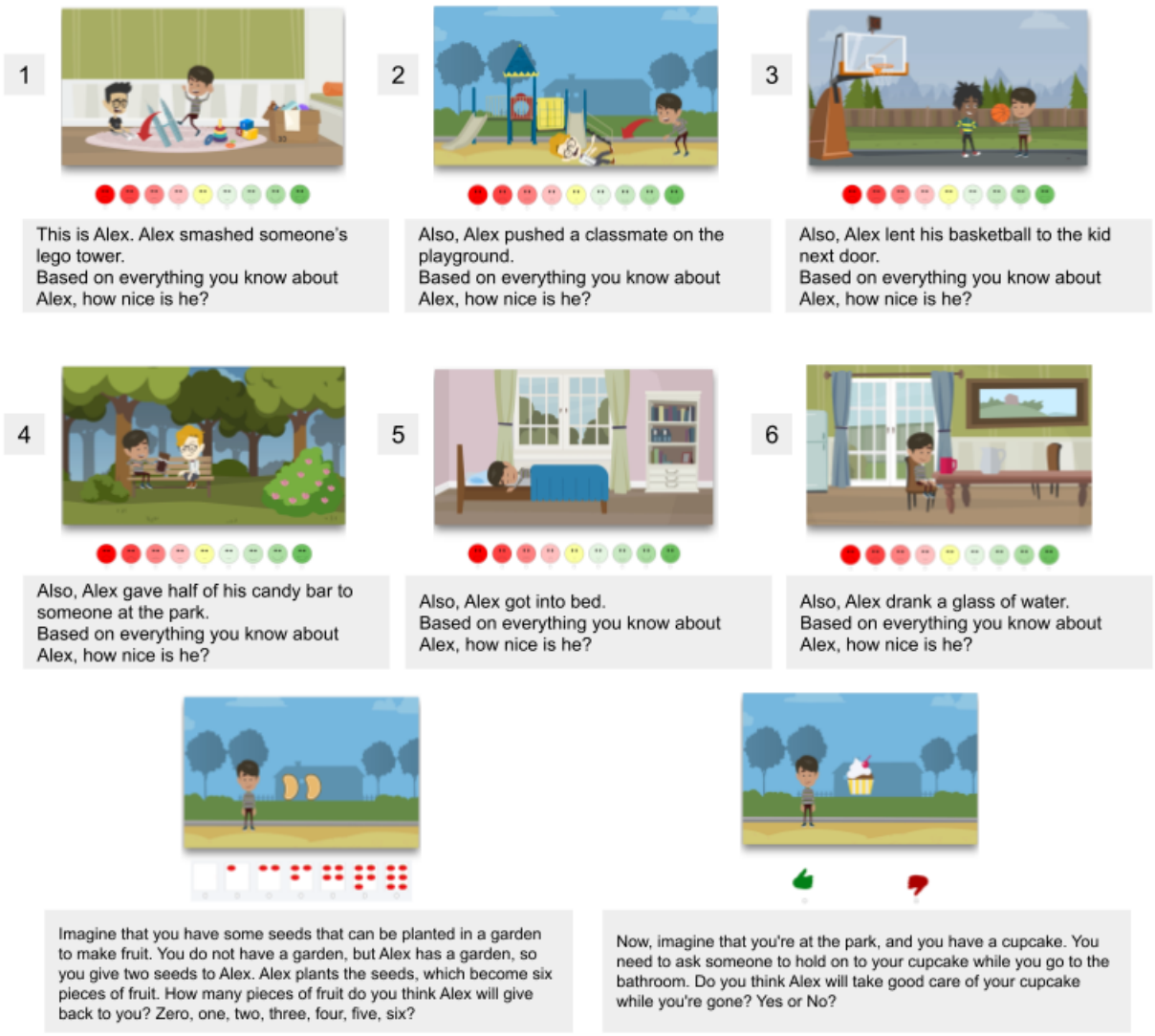


Figure 1. Example behavior sequence and experimenter script for a *Weak Negative-to-Positive* target. Vignette illustrations and measures were embedded in a Qualtrics survey which was controlled by the experimenter.

Participants were first introduced to the target character (e.g., "This is Alex"; the experimenter pointed to the target using the cursor). Next, participants heard about six successive behaviors performed by the target, presented on separate pages. Following each behavior, participants were asked, "Based on everything you know about [Target], how nice is [he/she]?". Participants reported their answer by saying the letter label underneath the desired scale point.

Afterwards, participants answered two questions about the target's predicted trustworthiness, presented on separate pages. First, participants predicted the number of fruit (0-6) the target would return to them in a modified Trust Game: "Imagine that you have some seeds that can be planted in a garden to make fruit. You do not have a garden, but [Target] has a garden, so you give two seeds to [Target]. [Target] plants the seeds, which become six pieces of fruit. How many pieces of fruit do you think [Target] will give back to you? Zero, one, two, three, four, five, six?".

Second, participants predicted whether the target would take good care of a possession of theirs: "Imagine that you're at the park, and you have a [cupcake/bike/balloon]. You need to ask someone to hold on to your [object] while you [go to the bathroom/go in the fountain/play on the slide]. Do you think [Target] will take good care of your [object] while you're gone? Yes or no?"

Main task. Participants learned about three target characters, who each performed a sequence of six behaviors (presented on separate pages; **Figure 1**). At the beginning of each sequence, participants were introduced to the target (e.g., "This is Alex"; the experimenter pointed to the target using the cursor). Then, participants heard the first behavior (e.g., "Alex smashed someone's lego tower") and were asked, "Based on everything you know about [Target], how nice is [he/she]?". Participants reported their answer by saying the letter label beneath the desired scale point. The next five behaviors proceeded the same way, with one change: each behavior description was preceded by "also" (e.g., "Also, Alex pushed a classmate on the playground").

After evaluating all six behaviors, participants answered two questions about the target's predicted trustworthiness, displayed on separate pages (**Figure 1**). First, participants answered a continuous measure that was presented within a modified Trust Game. Participants were asked to imagine that they had given two seeds to the target, who then planted them in a garden; the seeds became six pieces of fruit, which the target could choose to give back. Participants responded by saying the number of fruit (0-6) they expect to get back from the target. Second, participants answered a binary measure of predicted trustworthiness. Participants were asked to imagine that they were at the park, and they had to ask someone to hold on to their [cupcake/bike/balloon] while they went away. The dependent variable was whether the target would take good care of the object; participants responded by saying "yes" or "no". Across stimulus lists, each object type was paired with both initially positive and initially negative targets.

Analysis

Our main question of interest is whether children update their impressions to different degrees depending on the strength of the initial impression and the direction of behavior change, and whether these effects interact with age. For all interaction analyses, age in months will be treated as a continuous variable.

Condition	Rating #1	Rating #2	Rating #3	Rating #4	Rating #5	Rating #6	Update magnitude
Weak Positive-to-Negative	**7**	**8**	**2**	**1**	3	3	-1 * (1.5 - 7.5) = 6
Strong Positive-to-Negative	8	7	**8**	**8**	**2**	**1**	-1 * (1.5 - 8.0) = 6.5
Weak Negative-to-Positive	**1**	**1**	**4**	**5**	3	3	4.5 - 1.0 = 3.5
Strong Negative-to-Positive	2	1	**2**	**1**	**4**	**4**	4.0 - 1.5 = 2.5
Positive Control	8	7	8	8	8	8	8 - 7.5 = 0.5 *and* 8 - 8 = 0
Negative Control	2	1	2	1	1	1	1.5 - 1.5 = 0 *and* 1 - 1.5 = -0.5

Table 2. Example calculations for update magnitude. For control conditions, two types of update magnitude were calculated (updating between the first two behaviors and the middle two behaviors, and updating between the middle two behaviors and the last two behaviors).

Calculating update magnitude. For Negative-to-Positive targets, update magnitude was calculated as the difference between the average of the 2 post-switch ratings, and the average of the 2 pre-switch ratings (averaging was expected to guard against item-specific effects). For Positive-to-Negative targets, update magnitude was calculated as the sign-reversed difference between the average of the 2 post-switch ratings, and the average of the 2 pre-switch ratings (**Table 2**). Here we reverse the sign of

the difference, rather than taking the absolute value, to prevent overestimation of update magnitude that may occur if participants update in the unanticipated direction.

Modeling approach. Mixed effects analyses were performed using the *lme4* package in R (Bates et al., 2014; R Core Team, 2013). All models included by-participant random intercepts; stimulus list was included as a covariate in all models. P-values for fixed effects were obtained via Sattherthwaite's degrees of freedom method; contrasts were tested simultaneously using the *multcomp* package and were Tukey-corrected (Hothorn et al., 2016). model

Analyses of updating conditions. A linear mixed-effects model was fit to predict update magnitude (as calculated above), using as predictors: Strength (*weak: 0.5, strong: -0.5*), Valence (*positive-to-negative: 0.5*, *negative-to-positive: -0.5*), Age (in months, z-scored), and all interactions. In addition, pre-switch niceness ratings and post-switch niceness ratings were separately modeled using the same predictors.

We ran a Strength * Valence * Age linear mixed-effects model on the continuous measure of predicted trustworthiness (how many pieces of fruit the target would return). Similarly, we ran a Strength * Valence * Age logistic mixed-effects model on the binary measure of predicted trustworthiness (whether the target would take good care of an object; *yes: 1, no: 0*).

For the above models, if the 3-way interaction between Strength, Valence, and Age was significant, we split the participants into two age groups using median age (96.28 months); then, within the Young age group and within the Old age group, we modeled the dependent variable as a function of Strength, Valence, and their interaction.

To examine any associations between niceness ratings and predicted trustworthiness, we separately modeled the continuous measure of trustworthiness and the binary measure of trustworthiness as a function of final niceness ratings, and as a function of pre-switch niceness ratings.

Analyses of control conditions. Four measures of updating were calculated for control conditions: updating between the first two behaviors and the middle two behaviors (signed and unsigned), and updating between the middle two behaviors and the last two behaviors (signed and unsigned).

For each control condition, we compared signed update magnitudes against 0 (one-tailed t-tests), to examine whether impressions of control targets significantly improve or worsen over time. In addition, for each control condition, we compared absolute value update magnitudes to analogous measures of update magnitude for updating conditions. For each comparison, update magnitude was modeled as a function of a 2-level condition factor (*updating: 0.5, control: -0.5*). This allowed us to compare magnitudes of updating between consistent targets and inconsistent targets.

We ran a Valence * Age linear mixed-effects model on the continuous measure of trustworthiness, and a Valence * Age logistic mixed-effects model on the binary measure of trustworthiness. As above, we separately modeled the two trustworthiness measures as a function of final niceness ratings, and as a function of pre-switch niceness ratings.

3.3 RESULTS

Magnitude of impression updating: Updating conditions. We asked whether children update impressions to different degrees depending on the valence and strength of

the initial impression. Participants learned about targets who first performed 2 or 4 positive or negative behaviors (leading to weak or strong initial impressions of niceness), and then performed 2 behaviors of the opposite valence (eliciting an impression update). We modeled update magnitude as a function of Valence, Strength, and participant age (in months).[2] **Figure 3** plots how niceness ratings changed over the course of 6 behaviors.

We observed a significant 3-way interaction between Valence, Strength, and Age (*Estimate* = -0.65, *SE* = 0.25, *t*(158) = -2.60, p = 0.010; **Figure 5e**), and a 2-way interaction between Valence and Strength (*Estimate* = -0.83, *SE* = 0.25, *t*(158) = -3.36, *p* = 0.001). Contrary to predictions, we observed a negativity bias: the magnitude of impression updating was greater for Positive-to-Negative targets (*M* = 5.59, *SE* = 0.22) compared to Negative-to-Positive targets (*M* = 4.22, *SE* = 0.22) (*Estimate* = 1.37, *SE* = 0.31, *t*(155) = 4.41, *p* < 0.0001). There was also a main effect of Age, such that older children updated less than younger children (*Estimate* = -0.38, *SE* = 0.16, *t*(155) = -2.41, *p* = 0.017).

To follow up the significant 3-way interaction, we used median age (96.28 months) to split participants into two age groups: Young (approx. ages 6-7) and Old (approx. ages 8-9). When examining impression updating in the younger age group (**Figure 5d**), we observed a negativity bias: there was greater updating for Positive-to-Negative targets (*M* = 5.00, *SE* = 0.31) compared to Negative-to-Positive targets (*M* = 4.02, *SE* = 0.31) (*Estimate* = 0.97, *SE* = 0.44, *t*(76) = 2.22, *p* = 0.029; **Figure 5d**). There was no main effect of Strength, and no 2-way interaction between Valence and Strength.

[2] We also tested a model that included all possible 2-, 3-, and 4-way interactions with stimulus list; none of these interactions were significant, supporting our decision to include stimulus list as a covariate in all models but exclude it from interactions.

Thus, younger participants appeared to be sensitive to the direction of behavioral change in a categorical fashion, but not to the frequency of positive and negative behaviors.

In the older age group (**Figure 5d**), we again observed a a negativity bias when comparing updating for Positive-to-Negative targets (M = 5.00, SE = 0.31) and Negative-to-Positive targets (M = 4.02, SE = 0.31) (*Estimate* = 0.97, SE = 0.44, $t(76) = 2.22$, p = 0.029). This was qualified by a 2-way interaction between Valence and Strength (*Estimate* = -1.59, SE = 0.36, $t(79) = -4.38$, $p < 0.0001$).

Follow-up comparisons revealed *reduced* updating for Strong Negative-to-Positive targets (M = 3.64, SE = 0.34) compared to Weak Negative-to-Positive targets (M = 4.41, SE = 0.34) (*Estimate* = 0.78, SE = 0.26, $z = 3.00$, $p = 0.005$), but *greater* updating for Strong Positive-to-Negative targets (M = 5.41, SE = 0.33) compared to Weak Negative-to-Positive targets (M = 4.59 SE = 0.33) (*Estimate* = -0.82, SE = 0.26, $z = -3.20$, $p = 0.003$; **Figure 5d**). That is, older participants updated *less* when very bad (vs. slightly bad) targets changed their behavior, but they updated *more* when very good (vs. slightly good) targets changed their behavior – the amount of initial negative information and the amount of initial positive information had opposite effects on impression updating. In line with this, the comparison between Strong Positive-to-Negative targets and Strong Negative-to-Positive targets revealed the greatest difference in update magnitude (*Estimate* = 1.79, SE = 0.34, $t(203) = 5.34$, $p < 0.0001$).

In all, these findings suggest that (1) negative information can lead to durable initial impressions, especially if 4 behavioral examples are provided, and (2) negative information is effective at reversing initially positive impressions, especially if the positive impression was based on 4 behavioral examples. This sensitivity to dosage was

not observed in younger participants' impression updates.

Magnitude of impression updating: Control conditions. In addition to four updating conditions, there were two control conditions, intended to elicit negligible impression updating: Positive Control (targets who performed 6 consecutive positive behaviors) and Negative Control (targets who performed 6 consecutive negative behaviors). **Figure 3** plots how niceness ratings changed over the course of 6 behaviors.

For control conditions, we examined "updating" measures designed to be analogous to measures for the updating conditions. First, we calculated the raw change in niceness ratings between the first two behaviors and the middle two behaviors. Impressions of Positive Control targets did not change (M = 0.16, SE = 0.09, $t(81)$ = 1.56, p = 0.122), but impressions of Negative Control targets worsened significantly (M = -0.46, SE = 0.09, $t(79)$ = -5.53, $p < 0.0001$; **Figure 3**). Second, we calculated the raw change in niceness ratings between the middle two behaviors and the last two behaviors. Impressions of Positive Control targets did not change (M = -0.12, SE = 0.08, $t(81)$ = -1.29, p = 0.202), but impressions of Negative Control targets improved significantly (M = 0.19, SE = 0.08, $t(79)$ = 3.16, p = 0.002; **Figure 5a**). Thus, initially negative impressions appear to be more volatile than initially positive impressions, even when the target's behavior does not shift in valence.

Next, we computed the absolute value magnitude of updating between the first two behaviors and the middle two behaviors, and the absolute value magnitude of updating between the middle two behaviors and the last two behaviors. These measures were compared to analogous measures for the updating conditions. We observed greater impression updating (i.e., magnitude of change in niceness ratings between the first two

behaviors and the middle two behaviors) for Weak updating targets compared to Control targets (*Estimate* = 2.22, *SE* = 0.22, $t(479) = 9.96$, $p < 0.0001$). We also observed greater impression updating (i.e., magnitude of change in niceness ratings between the middle two behaviors and the last two behaviors) for Strong updating targets compared to Control targets (*Estimate* = 3.12, *SE* = 0.19, $t(322) = 16.55$, $p < 0.0001$; **Figure 5c**). Thus, while children did update their impressions to some degree when targets behaved consistently, they engaged in far greater impression updating when targets behaved inconsistently.

Are children integrating information over time? We conducted a series of analyses to rule out the possibility that children are simply basing their niceness ratings on the target's most recent behavior, rather than taking into account the target's behavioral history. We compared ratings that were matched on (1) the ordinal position of the rated behavior and (2) the valence of the rated behavior, but differed in terms of the valence of preceding behaviors. For instance, we compared the 4th niceness rating for targets who had performed 4 positive behaviors in a row, to the 4th niceness rating for targets who had performed 2 negative behaviors, then 2 positive behaviors. These two targets only differed in the valence of their initial behaviors; if participants are judging targets based on just their most recent behaviors, they should not differentiate between these targets. Four such comparisons revealed that children's niceness ratings reflect an integration of behavioral information over time – for instance, targets who performed four positive behaviors in a row were rated as nicer than targets who performed two negative behaviors then two positive behaviors (see **Supp. Mat. p. 4** for statistics for all comparisons). These analyses suggest that, when asked to form an impression of a target

based on everything they know about that target, children consider the cumulative amount of positive or negative behavioral information over time.

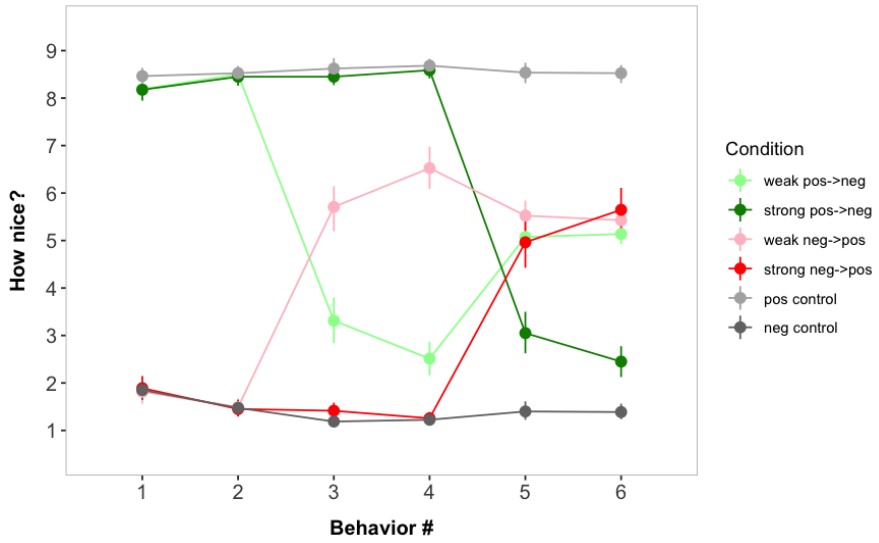
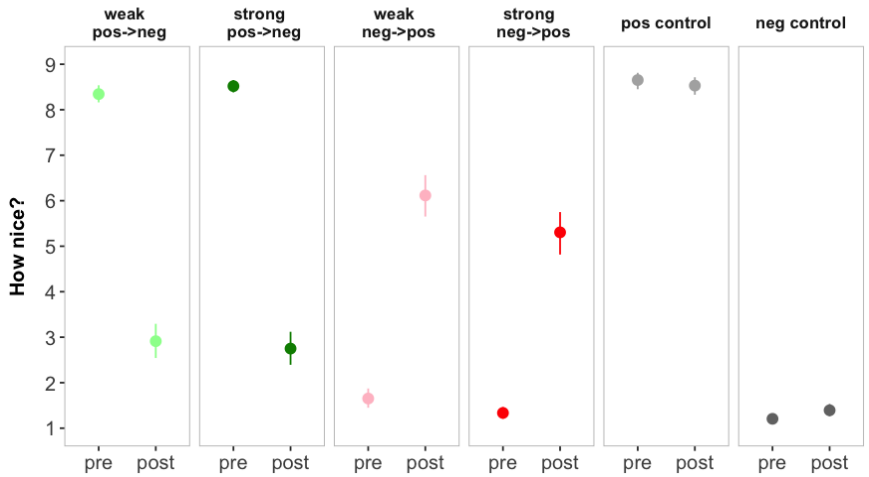


Figure 3. Average niceness ratings for each behavior in each sequence, for each target condition. Impressions of Positive Control targets did not change over time; impressions of Negative Control targets worsened between the first two behaviors and the middle two behaviors, then improved between the middle two behaviors and the last two behaviors.

(a)



(b)

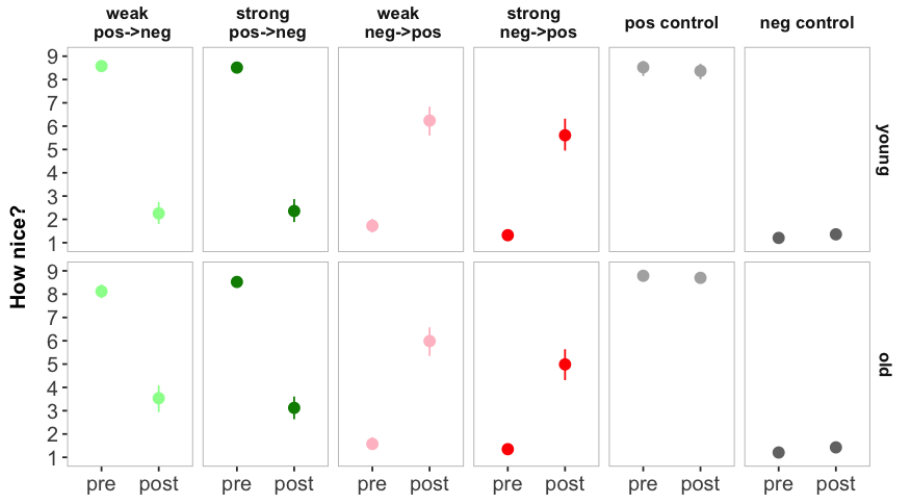


Figure 4. (a) Average niceness ratings for pre-switch behaviors and for post-switch behaviors, for each condition. For control conditions, ratings for the first middle behaviors and ratings for the last two behaviors are plotted as pre-switch and post-switch, respectively. We found that targets who performed 4 negative behaviors were perceived as less nice than targets who performed 2 negative behaviors; in addition, targets who performed 4 negative behaviors followed by 2 positive behaviors were perceived as less nice than targets who performed 2 negative behaviors followed by 2 positive behaviors. In all, the frequency of negative behaviors mattered more than the frequency of positive behaviors for children's impressions. **(b)** Average niceness ratings for pre-switch behaviors and post-switch behaviors, for each condition and age group. (See **Figure S1** for alternative graph that plots both age groups in one panel.)

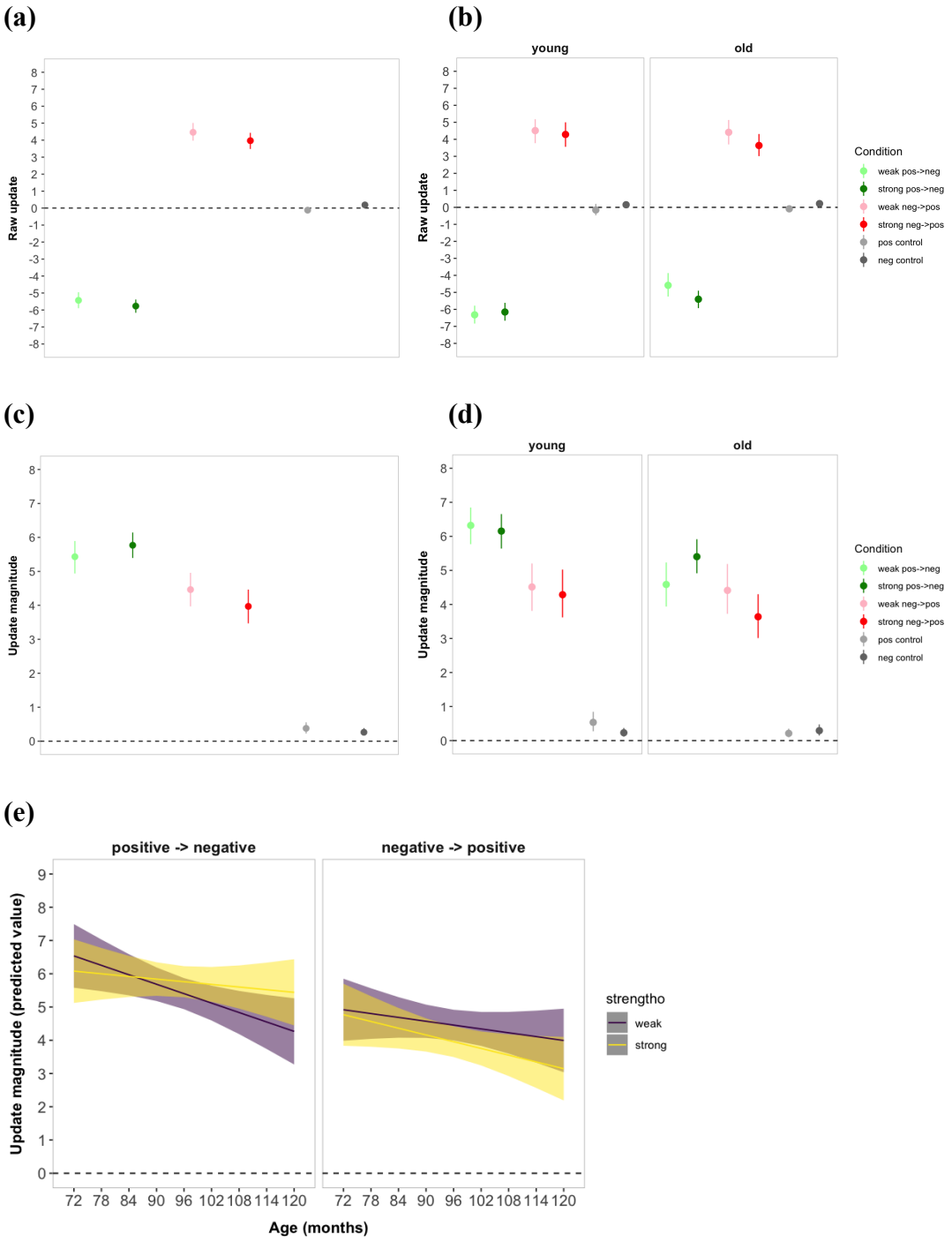


Figure 5. (**a-b)** Signed difference between post-switch and pre-switch niceness ratings, for each condition (and for each age group). **(c-d)** Magnitude of impression update for each condition (and for each age group). Within the younger age group, we observed a negativity bias. Within the older age group, we observed greater updating for Strong Positive-to-Negative vs. Weak Positive-to-Negative, and greater updating for Weak Negative-to-Positive vs. Strong-Negative-to-Positive. **(e)** Predicted effects from a model of update magnitude. We observed a significant 3-way interaction between Valence, Strength, and Age. Plot includes extrapolated values for age. (See **Figure S4** for alternative plot, faceted by strength.)

Average niceness ratings for the two pre-switch behaviors: Updating conditions. In addition to examining the magnitude of impression updating, we compared participants' initial impressions and updated impressions across conditions, to verify that our experimental design elicited appropriate responses, and to explore potential explanations for differences in updating. To gauge participants' initial impressions of targets, we averaged niceness ratings for the 2 behaviors immediately preceding the valence switch point ("pre-switch" behaviors; **Figure 4a**).

When examining pre-switch ratings for the updating conditions, we found that initially positive targets (M = 8.43, SE = 0.07) received more positive niceness ratings than initially negative targets (M = 1.49, SE = 0.07), consistent with our experimental manipulation (*Estimate* = 6.94, SE = 0.10, t(155) = 69.33, $p < 0.0001$; **Figure 4a**). Thus, we verified that our behavior stimuli elicited reasonable (valence-concordant) ratings from participants.

Next, we examined whether manipulating the strength of the initial impression (via the number of positive or negative behaviors presented consecutively) resulted in different pre-switch ratings. This analysis can help shed light on the mechanism through which behavioral frequency information impacts impression updating. We had observed greater updating for Weak Negative-to-Positive targets compared to Strong Negative-to-Positive targets, especially in older children. One possible explanation for the greater lability of weak negative impressions is that providing observers with 2 vs. 4 behavioral examples results in less extreme (less negative) impressions, which are in turn easier to update. Finding that participants gave less negative niceness ratings to Weak Negative vs. Strong Negative targets would support this explanation. Another (non-mutually

exclusive) possibility is that providing fewer behavioral examples results in more uncertain impressions, which are in turn easier to update. If we find that participants gave similar initial ratings to Weak Negative targets and Strong Negative targets, this may suggest that weak negative impressions are relatively unstable because they are less certain, not because they are less extreme.

We found that Strong Negative targets (M = 1.33, SE = 0.09) received more negative niceness ratings than Weak Negative targets (M = 1.65, SE = 0.09) (*Estimate* = 0.32, SE = 0.10, z = 3.20, p = 0.003). Thus, targets who performed four negative behaviors were perceived as less nice than those who performed two negative behaviors. On the other hand, Weak Positive targets (M = 8.34, SE = 0.09) and Strong Positive targets (M = 8.52, SE = 0.09) did not significantly differ (*Estimate* = -0.18, SE = 0.10, z = -1.76, p = 0.322l; **Figure 4a**). Thus, for initially positive targets, there was no dosage effect of behavior frequency on niceness ratings.

These comparisons were observed after following up a significant 2-way interaction between Valence and Strength (*Estimate* = -0.49, SE = 0.14, t(158) = -3.50, p < 0.001; contrasts are Tukey-corrected). There was no main effect of Strength on pre-switch niceness ratings, collapsing across Valence (*Estimate* = 0.07, SE = 0.07, t(158) = 0.99, p = 0.322). Finally, there was no main effect of Age on initial impressions (*Estimate* = -0.04, SE = 0.05, t(155) = -0.77, p = 0.445), and there were no significant interactions among Valence, Strength, and Age (see **Figure S2** for interaction plots).

In all, the amount of negative behavioral information seems to matter more than the amount of positive behavioral information for forming initial impressions. These findings can help us interpret the differences in updating we saw in older participants.

Specifically, a difference in the extremity of initial impressions may partly explain why there was more updating of weak vs. strong negative impressions; meanwhile, a difference in the uncertainty of initial impressions may partly explain why there was more updating of strong vs. weak positive impressions.

Average niceness ratings for the two post-switch behaviors: Updating conditions. After learning that a target performed 2 or 4 behaviors of the same valence, participants then learned that the target performed 2 behaviors of the opposite valence. To index participants' updated impressions of targets, we averaged niceness ratings for the 2 behaviors immediately following the valence switch point ("post-switch" behaviors; **Figure 4a**). Comparing updated impressions for Positive-to-Negative targets vs. Negative-to-Positive targets allows us to examine whether children have a more positive view of previously good actors whose behaviors change for the worse, or previously bad actors whose behaviors change for the better.

We found that, when targets' behaviors shifted in a meaningful way, Negative-to-Positive targets (M = 5.71, SE = 0.20) received more positive niceness ratings than Positive-to-Negative targets (M = 2.83, SE = 0.20) (*Estimate* = -2.88, SE = 0.29, $t(155)$ = -10.05, $p < 0.0001$; **Figure 4a**). There was a significant two-way interaction between Valence and Age (*Estimate* = 0.76, SE = 0.29, $t(155) = 2.64$, $p = 0.001$; see **Figure S3** for interaction plots). Follow-up comparisons revealed a larger effect of Valence in younger participants (*Estimate* = -3.61, SE = 0.41, $t(76)$ = -8.80, $p < 0.0001$) compared to older participants (*Estimate* = -2.15, SE = 0.40, $t(76)$ = -5.30, $p < 0.0001$; **Figure 4b**). There was no main effect of Age on updated impressions (*Estimate* = -0.01, SE = 0.14, $t(155)$ = -0.079, $p = 0.937$). In all, we found a valence asymmetry, where children had a more

positive impression of previously bad actors who got better, compared to previously good actors who got worse; this asymmetry was more pronounced in younger children.

We also examined whether post-switch ratings differed as a function of the strength of the initial impression. Comparing updated impressions for Weak vs. Strong targets is one way to test whether children are making niceness judgments that take into account targets' full behavioral history – if participants give different ratings to Weak and Strong targets when their behaviors change in valence, this would suggest that they are judging targets in light of their past history (rather than judging just their most recent behavior). We found that Weak Negative-to-Positive targets (M = 6.12, SE = 0.22) received more positive niceness ratings than Strong Negative-to-Positive targets (M = 5.31, SE = 0.22) (*Estimate* = 0.81, SE = 0.0.15, z = 5.31, $p < 0.000$). Thus, the strength of an initially negative impression affected how targets were perceived when they performed positive behaviors. On the other hand, updated impressions of Weak Positive-to-Negative targets (M = 2.91, SE = 0.22) and Strong Positive-to-Negative targets (M = 2.75, SE = 0.22) did not significantly differ (*Estimate* = 0.17, SE = 0.15, z = 1.07, p = 0.489; **Figure 4**).

These comparisons were observed after following up a significant 2-way interaction between Valence and Strength (*Estimate* = -0.64, SE = 0.22, $t(158)$ = -2.97, p = 0.003; contrasts are Tukey-corrected). In addition, there was a main effect of Strength on updated impressions: Weak targets (M = 4.52, SE = 0.15) received more positive niceness ratings than Strong targets (M = 4.03, SE = 0.15), regardless of the initial valence of the impression (*Estimate* = 0.49, SE = 0.11, $t(158)$ = 4.49, p < 0.0001). There was no interaction between Strength and Age.

Overall, these results suggest that children's niceness ratings are more sensitive to variations in the initial amount of negative (vs. positive) behavioral information, not only when forming an initial impression, but also when updating that impression in light of a meaningful change in behavior. This can be viewed as an instantiation of negativity bias, in that children are more responsive to the exact amount of negative evidence they have received.

Behavioral prediction: Modified Trust Game. After learning about each target's sequence of behaviors, participants were asked to make predictions about the target's future behavior in the context of a modified Trust Game (**Figure 1**). Participants were asked to predict how many pieces of fruit (0-6) the target will give back to them. We treated this response as a continuous measure of future trustworthiness, and examined children's ability to use trait inferences to make this form of behavioral prediction.

We first examined predicted trustworthiness for updating conditions as a function of Valence, Strength, and participant age (in months). There were no significant 3-way or 2-way interactions. We observed a main effect of Age, in which older participants expected to receive fewer pieces of fruit from the target (*Estimate* = -0.29, *SE* = 0.10, $t(155) = -2.89$, $p = 0.004$; **Figure 6a**). Follow-up tests in each age group revealed that younger children predicted they would receive significantly more than 3 pieces of fruit ($M = 3.46$, $t(161) = 3.73$, $p = 0.0003$), while older children predicted a fair split ($M = 3.04$, $t(161) = 0.37$, $p = 0.708$). This indicates a decline in optimism (or increasing alignment to the fairness norm) in predictions of trustworthiness between ages 6-7 and ages 8-9.

We also observed main effects of Valence and Strength. We found that Positive-

to-Negative targets (M = 3.47, SE = 0.14) were predicted to return more fruit than Negative-to-Positive targets (M = 3.04, SE = 0.14) (*Estimate* = 0.42, SE = 0.20, $t(155)$ = 2.11, p = 0.036) – that is, children expected that targets who were initially nice then became mean would share more of the profits of an investment, than targets who were initially mean then became nice (**Figure 6b**). In addition, Weak targets (M = 3.39, SE = 0.12) were predicted to return more fruit than Strong targets (M = 3.12, SE = 0.12) (*Estimate* = 0.27, SE = 0.13, $t(158)$ = 2.09, p = 0.038). Thus, children expected targets whose behavior shifted meaningfully after 2 (vs. 4) instances to be more trustworthy (**Figure 6b**).

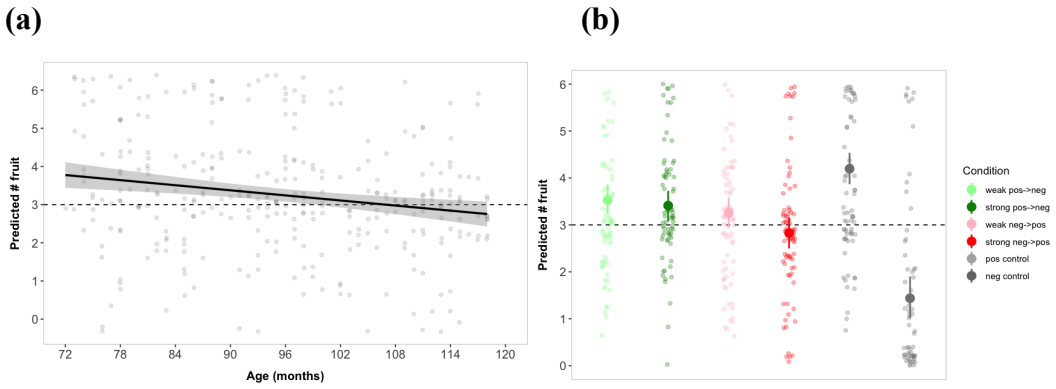


Figure 6. (a) Correlation between participant age and predictions of targets' future trustworthiness (collapsed across updating conditions). We observed a decline in optimism throughout development, such that younger children expected to receive more than a fair split of the resource, whereas older children expected to receive a fair split. **(b)** Average predictions of targets' future trustworthiness by condition. Participants expected to receive more fruit from Positive-to-Negative (vs. Negative-to-Positive) targets, and from Weak (vs. Strong) targets.

Correlations between niceness ratings and behavioral predictions. We examined how children's impressions of targets relate to predicted trustworthiness in the modified Trust Game. First, we tested the correlation between final impressions (niceness ratings

for the 6th behavior) and predicted trustworthiness, collapsing across updating conditions. We found that there was no significant relationship between final impressions and predicted number of fruit returned (*Estimate* = 0.06, *SE* = 0.04, *t*(278.91) = 1.41, *p* = 0.16; **Figure 7a**). On the other hand, for control conditions, there was a robust positive relationship between final impressions and predicted number of fruit returned (*Estimate* = 0.38, *SE* = 0.04, t(155) = 9.85, p < 0.0001; **Figure 7a**). Thus, when targets were consistently nice or mean, children based their predictions of future trustworthiness on their current impressions (we note that, for control targets, children's impressions did not change much over time, so predictions may have been based on earlier impressions).

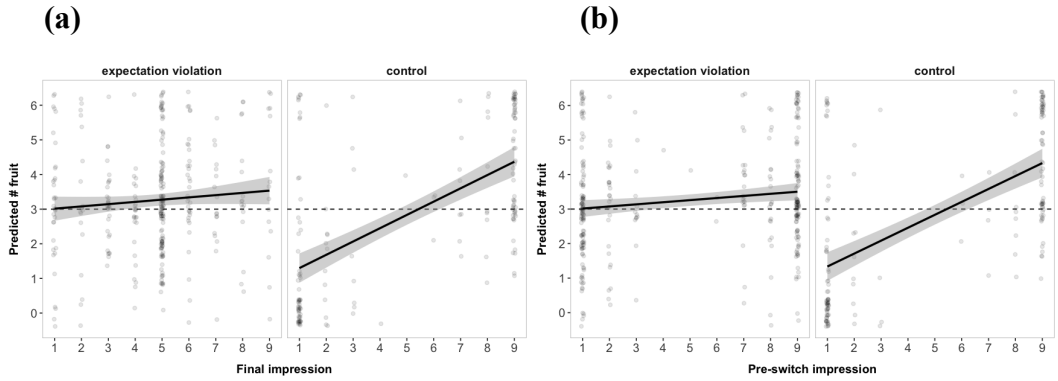


Figure 7. (a) Correlations between final impressions and predictions of targets' future trustworthiness, for updating conditions and for control conditions. When targets displayed internally consistent behaviors, children's final impressions were significantly associated with future trustworthiness. **(b)** Correlations between initial (pre-switch) impressions and predictions of targets' future trustworthiness, for updating conditions and for control conditions. Children's initial impressions were significantly associated with future trustworthiness, both when targets displayed internally inconsistent behaviors, and when targets were consistent.

Next, we considered the possibility that children engage in behavioral prediction differently when faced with mixed evidence (i.e., both positive and negative behavioral examples). Given that participants (1) gave more positive pre-switch ratings to initially positive vs. initially negative targets, and (2) gave more negative final ratings to Positive-

to-Negative targets (M = 3.79, SE = 0.13; comparison against midpoint: $t(159) = -8.19$, $p < 0.0001$) relative to Negative-to-Positive targets (M = 5.54, SE = 0.13; comparison against midpoint: $t(163) = 3.80$, $p = 0.0002$) (Estimate = -1.75, SE = 0.18, t(155) = -9.45, p < 0.0001), we hypothesized that there would be a positive relationship between pre-switch impressions and behavioral predictions, rather than between final impressions and behavioral predictions.

We observed a significant association between niceness ratings for the last pre-switch behavior and predicted trustworthiness (*Estimate* = 0.06, *SE* = 0.03, $t(167) = 2.39$, $p = 0.018$; **Figure 7b**). These results suggest that, when targets display inconsistent behaviors, children may base their predictions of future trustworthiness on *initial* impressions, rather than updated impressions.

Behavioral prediction: Trust with object. As an alternative measure of trustworthiness in a new context, we asked children to predict whether each target could be trusted to look after an object (**Figure 1**). We treated the Yes/No response as a binary measure of future trustworthiness, and examined children's ability to use trait inferences to make this type of behavioral prediction.

We first examined predicted trustworthiness for updating conditions as a function of Valence, Strength, and Age. There was a main effect of Age, in which older participants were less likely to predict that the target will take good care of their object (*OR* = 0.71, $z = -2.10$, $p = 0.04$; **Figure 8a**). Follow-up tests in each age group revealed that younger children predicted targets would take good care of their object at a rate significantly above chance (*CI of mean* = 0.53-0.73, $z = 2.59$, $p = 0.019$), while older children's predictions were at chance (*CI of mean* = 0.47-0.66, $z = 1.32$, $p = 0.337$). This

again indicates a decline in optimism in predictions of trustworthiness between ages 6-7 and ages 8-9.

We also observed a significant 2-way interaction between Valence and Strength (*OR* = 0.25, *z* = -2.55, *p* = 0.011); there were no other significant interactions. Follow-up comparisons revealed that Strong Positive-to-Negative targets (*CI of mean* = 0.63-0.85, *z* = 3.56, *p* = 0.001) were rated as more likely to take good care of the object than Strong Negative-to-Positive targets (*CI of mean* = 0.37-0.64, *z* = 0.09, *p* = 0.995) (*OR* = 0.33, *z* = -2.61, *p* = 0.045; contrasts are Tukey-corrected). Thus, children expected targets whose behavior worsened after 4 positive instances to be more trustworthy than targets whose behavior improved after 4 negative instances (**Figure 8b**).

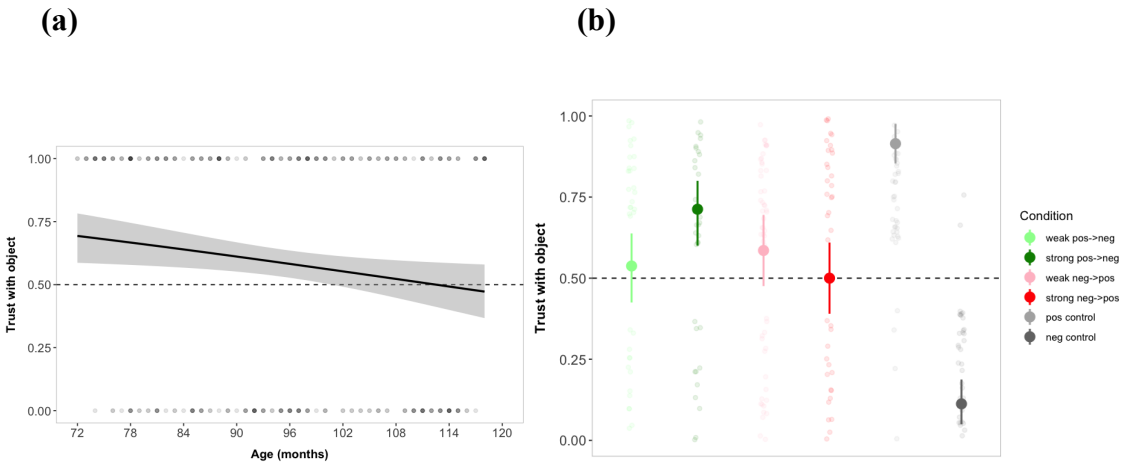


Figure 8. **(a)** Correlation between participant age and proportion of participants who predicted that targets would take good care of their belongings. We observed a decline in optimism throughout development, such that younger children were more likely to respond that targets can be trusted with their belongings, compared to older children. **(b)** Proportion of participants who predicted that targets would take good care of their belongings, for each condition. Participants expected that Strong Positive-to-Negative targets would be more likely to take good care of their belongings, compared to Strong Negative-to-Positive targets.

Correlations between niceness ratings and behavioral prediction. We then asked if participants' final impressions were associated with their predictions of whether the target would take good care of an object. For updating conditions, there was no significant relationship between niceness ratings for the 6th behavior and predicted trustworthiness ($OR = 1.05$, $z = 0.71$, $p = 0.478$; **Figure 9a**). On the other hand, for control conditions, there was a robust relationship between final impressions and predicted trustworthiness ($OR = 2.08$, $z = 7.10$, $p < 0.0001$; **Figure 9a**). Thus, when targets were consistently positive or negative, children were able to base their predictions of future trustworthiness on their current impressions (we again note that children's predictions for control targets may have been based on earlier impressions, as ratings did not change much over time).

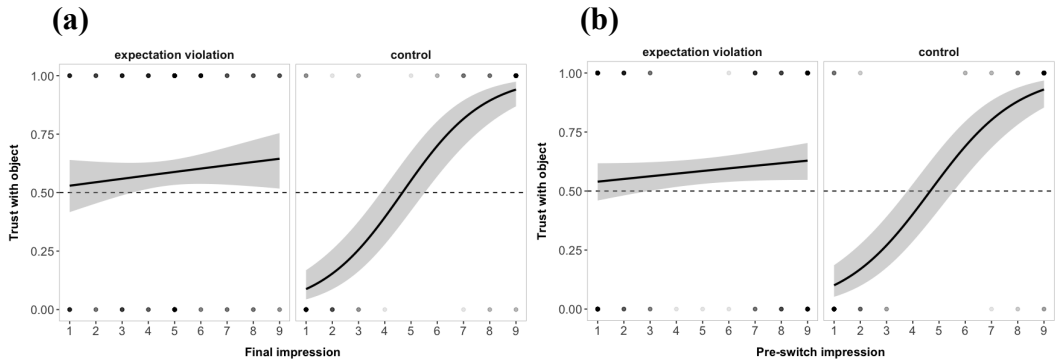


Figure 9. (a) Correlations between final impressions and proportion of participants who predicted that targets would take good care of their belongings, for expectation-violation targets and for control conditions. When targets displayed internally consistent behaviors, children's final impressions were significantly associated with future trustworthiness. **(b)** Correlations between initial (pre-switch) impressions and predictions of targets' future trustworthiness, for expectation-violation targets and for control conditions. Children's initial impressions were significantly associated with future trustworthiness when targets were consistent, but not when targets were inconsistent.

Next, we tested whether there is a relationship between initial (pre-switch) impressions and predictions of whether the target would take good care of an object. For

updating conditions, there was no significant association between niceness ratings for the last pre-switch behavior and predicted trustworthiness ($OR = 1.05$, $z = 1.31$, $p = 0.191$; **Figure 9b**). Thus, when targets displayed inconsistent behaviors, there was no relationship between children's initial impressions and their predictions of trustworthiness as indexed by a binary measure.

In all, we examined children's ability to use inferences of niceness to make predictions about two types of trustworthiness-related behaviors. We found that when targets behaved consistently, participants made predictions that aligned with their current impression of the target's niceness; in contrast, when targets behaved inconsistently, participants did not base predictions of trustworthiness on current impressions (they did, however, use initial impressions to predict trustworthiness in the Trust Game setting).

3.4 DISCUSSION

Flexible trait reasoning may be especially important during middle childhood, when children have to navigate a rapidly expanding social world, and new friendships get formed and tested. We examined impression updating in 6-9-year-olds, and systematically assessed the impact of (1) direction of behavior change, and (2) the strength of the initial impression, on the magnitude of updating.

Summary of results. We presented participants with stories of targets whose behaviors changed meaningfully over time or stayed the same. We found that children updated their impressions to a much greater extent when targets behaved inconsistently, than when they behaved consistently. When comparing different types of inconsistent targets, we found that children updated initially negative impressions more than initially

positive impressions. Older children (ages 8-9) were additionally sensitive to behavior frequency information: they updated weak negative impressions (which were based on 2 behavioral examples) more than strong negative impressions (which were based on 4 behavioral examples), and they updated strong positive impressions more than weak positive impressions.

Furthermore, for initially negative targets, behavior frequency information influenced the extremity of both initial niceness ratings and updated niceness ratings (across ages). Specifically, negative impressions based on 2 behaviors were less negative than those based on 4 behaviors, and impressions of targets who performed 2 negative then 2 positive behaviors were more positive than impressions of targets who performed 4 negative then 2 positive behaviors. In contrast, for initially positive targets, frequency information had no effect on the extremity of initial or updated niceness ratings.

We also examined children's ability to use inferences of niceness to make two types of behavioral predictions: the target's trustworthiness in a modified Trust Game, and the target's trustworthiness when asked to take care of an object. Our findings suggest that, when targets behave consistently (by performing 6 positive behaviors or 6 negative behaviors), children make predictions of future trustworthiness that align with their current impression of the target's niceness. In contrast, when targets behave inconsistently, children do not base predictions of trustworthiness on their current (updated) impression of the target; furthermore, in the context of a modified Trust Game, children appear to use their *initial* impression to predict trustworthiness.

Integrating social information over time. We found that children were quick to update their impressions when nice targets became mean, and when mean targets became

nice; overall, they updated far more for inconsistent targets compared to consistent targets. In addition, several convergent analyses showed that participants considered targets' full behavioral history when making niceness ratings (as opposed to considering just the most recent behavior). These results indicate that: (1) 6-9-year-olds can take up to 6 pieces of behavioral information into account when making trait attributions, and (2) even when a target's behaviors occur across a range of contexts (e.g., school, home, playground), and are directed towards a range of recipients (e.g., sibling, classmate, stranger), children in this age range are able to extract abstract features across these behaviors and infer an underlying character trait (niceness) that is consistent with all available evidence. This can be contrasted with younger children's trait reasoning. Boseovski and Lee (2006) probed trait attributions and behavioral predictions in 3-6-year-olds. They found that, overall, participants did not distinguish between (1) a target who performed 1 valenced behavior directed at 1 recipient, and (2) a target who performed 5 valenced behaviors directed at 5 unique recipients; however, participants did differentiate between (1) and (3): a target who performed 5 valenced behaviors directed at the same recipient. Specifically, participants were overall more likely to make appropriate trait attributions (*nice/mean*) and behavioral predictions (*share/take*) when they had access to repeated behavioral evidence. This suggests that, while preschoolers can make use of behavioral frequency information when behaviors are directed at the same recipient, they may have difficulty integrating multiple instances of one category of behavior across recipients. While this study was not designed to be an impression updating task (dependent variables were only collected once), its findings indicate that younger children's trait inferences may be constrained by a limited ability to reason about

the common dispositional cause that underlies cross-situational behavioral instances (Boseovski & Lee, 2006). On the other hand, the present research suggests that by middle childhood, children may have more advanced causal reasoning abilities in the social domain that allows them to make use of covariation information when making trait inferences across contexts, and successfully integrate social information acquired over time.

Unexpected negativity bias in middle childhood. We had predicted that, consistent with past work documenting a positivity bias in trait understanding during middle childhood, we would find greater impression updating when initially negative impressions are contradicted, compared to when initially positive impressions are contradicted. Contrary to hypotheses, we found a negativity bias: more updating from positive to negative than vice versa. We also observed another form of negativity bias: the extremity of children's niceness ratings was more sensitive to variations in the amount of negative (vs. positive) behavioral information, both for initial impressions and updated impressions.

In what contexts has the positivity bias in children's trait reasoning been observed? Past work has found that 3-6-year-olds require many behavioral examples to make negative trait attributions, and 5-6-year-olds selectively use positive evidence for their judgments, even if it's outweighed by negative evidence; in contrast, 9-10-year-olds base judgments on the predominant behavior and do not get swayed by the presence of a single positive behavior (Boseovski & Lee, 2006; Rholes & Ruble, 1986). In addition, past studies have shown that 6-year-olds view positive information as more diagnostic, and 7-8-year-olds view positive sociomoral traits as more stable (Newman, 1991;

Heyman & Dweck, 1998); other work finds that 7-9-year-olds assume less stability in positive traits, and 10-year-olds perceive negative information as more diagnostic (Lockhart, Chang, & Story, 2002; Newman, 1991). In all, positivity bias can manifest as: (1) a higher threshold for making negative judgments; (2) the overweighting of positive evidence; and (3) viewing positive information as more diagnostic.

Evidential threshold for making trait inferences. How do the findings from the current study fit into the literature? One key difference between the current study and past work is that we captured moment-to-moment changes in impressions by collecting trait ratings after each behavior presentation. This allowed us to see how children respond to each new piece of evidence as it became available. When provided with 1 initial behavior, participants were very quick to make appropriate trait attributions, following both positive behaviors (*very/super nice*) and negative behaviors (*very/super mean*; **Figure 3**). We also verified that our youngest participants, the 6-year-olds, gave highly positive initial niceness ratings for Positive targets (M = 8.35), and highly negative initial niceness ratings for Negative targets (M = 2.00). Overall, there was no valence asymmetry in the evidential threshold for making trait judgments – one behavior was sufficient for both positive and negative judgments.

In contrast to our updating paradigm, the aforementioned studies by Boseovski and Lee (2006) and Rholes and Ruble (1986) collected trait attributions just once, after presenting a set of behaviors. Furthermore, Boseovki and Lee (2006) presented sets of 6 behaviors that consisted of either: 1 positive/negative story + 5 neutral stories, or 5 positive/negative stories + 1 neutral story; in addition, the ordinal position of the odd-one-out story was counterbalanced across participants, so the trait attribution results were

averaged over different trajectories of behavior change. Thus, the evidential threshold for trait attribution that can be examined in that paradigm (whether 1 valenced story presented alongside 5 neutral stories is sufficient for trait attribution) is quite different from the one that can be examined in the current paradigm (whether 1 valenced story is sufficient for trait attribution). It is plausible that if the participants in the previous study were asked to make a trait attribution immediately after hearing 1 negative story, they would have been willing to make the 'mean' attribution (as the 6-year-olds in our study were). Nevertheless, it is notable that 1 negative story, when presented amidst 5 neutral stories, was not particularly salient for 3-6-year-olds; this is counter to the overall pattern we observed in our 6-9-year-olds, which is that negative information is more salient than positive information. This suggests that the positivity bias may not be as present throughout middle childhood as previously thought.

Overweighting and diagnosticity of negative evidence. In the current study, participants overweighted negative evidence – if they received negative evidence first, they were more resistant to updating their resultant negative impression; and if they received negative evidence second, they updated their initially positive impression to a greater degree. This is the same direction of valence asymmetry found in impression updating in adults (Kim, Mende-Siedlecki, Anzellotti, & Young, 2021; Mende-Siedlecki, Baron, & Todorov, 2013; Reeder & Coovert, 1986). The negativity bias in adults is thought to be driven by diagnosticity: negative behaviors (at least in the moral domain) are relatively infrequent, and are thus stronger indicators of true character (Skowronski & Carlston, 1987). If a particular category of social information is more diagnostic, it makes sense that we would (1) learn a lot (update more) after receiving that information, and (2)

assume stability in the traits we infer from that information. As discussed above, past work finds that 6-year-olds view positive information as more diagnostic, while 10-year-olds view negative information as more diagnostic; the findings for 7-9-year-olds are mixed (Newman, 1991; Heyman & Dweck, 1998; Lockhart, Chang, & Story, 2002). The age range in the current study may capture a transition point in children's lay theories on the diagnosticity of valenced information and the stability of valenced traits. Further work is required for a clearer understanding of this potential shift.

One possibility is that the participants in our study actually do view positive information as more diagnostic, and endorse entity theories of positive traits, but did not update their beliefs accordingly. An interesting future direction would be to explicitly probe children's beliefs about the stability of positive and negative traits, and test children's willingness to update initially positive and negative trait inferences, in the same paradigm. This would allow us to test whether children's metatheories on traits align with their actual trait inference, and if this alignment changes with age.

A related possibility is that participants in the current study were making act-based judgments, rather than person-based judgments (Uhlmann et al., 2013). Perhaps they found it odd that they were repeatedly being asked about an attribute (niceness) that they typically would not evaluate so frequently in so short a period of time, and thought it more sensible to evaluate the actions themselves. If we were to elicit both act-based judgments and person-based judgments in the same paradigm, we may find that children exhibit a negativity bias in act-based judgments, but a positivity bias in person-based judgments. Relatedly, children may have a notion of a "true self" that is distinct from the self in general (Strohminger, Knobe, & Newman, 2017). In adults, the true self, or who

someone really is “deep down”, is thought to be moral and good; this positivity bias is actor-observer invariant, and supercedes the negativity bias in impression formation. If children are asked to update impressions of a target’s “true self”, they may be more likely to prioritize positive information; this would be in line with the (mixed) evidence that 7-9-year-olds endorse the stability of positive traits.

Heightened sensitivity to the amount of negative evidence. In the current study, children’s niceness ratings reflected greater sensitivity to the exact number of negative behaviors previously performed by the target. We view this as an instantiation of negativity bias, in the sense that negative information appears to command more detailed processing. Interestingly, the differentiation between 2 vs. 4 pieces of initial negative information in the current study suggests that negative impressions can be *less stable* than positive impressions, if all available information is of the same valence; this can be contrasted with the reduced impact of new positive information on initially negative impressions (relative to the opposite). There seems to be a duality where bad beliefs are initially volatile and can be negatively updated (Siegel, Mathys, Rutledge, & Crockett, 2018), but are also less likely to be positively updated. Keeping close track of exactly how bad someone is may be more important for calibrating social interactions than tracking exactly how good they are (Baumeister, et al., 2001). One way to study this function would be to relate impression updating to social decision-making (e.g., partner choice, investment in an economic game) – we may observe a finer-grained relationship between the amount of negative behavioral evidence and social decisions, than between positive behavioral evidence and decisions.

Updating in older children. We found that older children (ages 8-9) in the current

study differentiated between the Weak Negative and Strong Negative targets not only in terms of initial niceness ratings, but also in terms of the magnitude of updating – they updated more for Weak Negative-to-Positive targets. In addition, older children updated more for Strong Positive-to-Negative targets than Weak Positive-to-Negative targets. Younger children (ages 6-7), on the other hand, did not update differently for Weak vs. Strong targets.

Broadly, the tendency to adjust one's beliefs to different degrees depending on the relationship between new information and old information is in line with a predictive framework for social cognition. In this framework, observers predict agents' future actions using prior knowledge (person models), and when unexpected actions occur, prediction errors prompt revisions of person models in a way that would minimize future errors (Bach & Schenke, 2017; Tamir & Thornton, 2018). One possibility is that older children are better able to keep track of the uncertainty of their beliefs; it may be that they are uncertain about positive impressions based on 4 behaviors, and that this uncertainty manifests in more rapid belief updating when new negative information is observed. Future work may collect confidence or uncertainty ratings to test whether representations of uncertainty can account for age-related changes in impression updating for weak and strong beliefs. In addition, it may be informative to study children's causal attributions in an impression updating task – perhaps older children engage in fine-grained adjustment of beliefs that reflect their estimate of the likelihood that the target's disposition (vs. the situation) was responsible for the action (Kim, Theriault, Hirschfeld-Kroen, & Young, 2022).

Predictions of future trustworthiness. When targets behaved consistently, children

made predictions of targets' future trustworthiness that were consistent with their current impressions (which were also their initial impressions). On the other hand, when targets behaved inconsistently, there was no association between current impressions and predictions of trustworthiness. Instead, participants based their predictions on their initial impressions (at least in the modified Trust Game).

There were several potential impediments to making behavior-to-behavior predictions for inconsistent targets. For one, final niceness ratings of three out of four target conditions ended up around the midpoint (Weak Positive-to-Negative: M = 5.14; Weak Negative-to-Positive: M = 5.43; Strong Negative-to-Positive = 5.65); the Strong Positive-to-Negative target condition ended up at a significantly lower final niceness rating compared to all other conditions (M = 2.45; t_{min} = 10.86, $p < 0.0001$; **Figure 3**). Thus, there was limited variability in final impressions of inconsistent targets; we cannot rule out the possibility that, had final impressions of inconsistent targets been more differentiated, we would have found that final impressions predict trustworthiness. That is, the current results speak less to the predictive value of initial impressions (for Trust Game predictions), and more to the consequences of ambiguous impressions. That said, it is surprising that, even though Strong Positive-to-Negative targets received the most negative final ratings, they were rated as most likely to take good care of an entrusted object.

One possibility is that, if targets have a history of behaving unpredictably, and children are asked to predict their future behavior, they revert to their initial impression of the target (a primacy effect). Future work may examine the consequences of ambiguous impressions more closely. For instance, we may compare a target who

performs 2 positive behaviors then 2 negative behaviors, to a target who alternates between positive and negative behaviors (pos-neg-pos-neg or neg-pos-neg-pos). If there is a true primacy effect, then children should base predictions on the first behavior, regardless of the trajectory of behavior change.

Second, there was a disconnect between what participants learned and what they were asked to predict. They first had to infer niceness from targets' behaviors; then, they had to translate niceness into trustworthiness, and use trustworthiness to predict trustworthiness-related behavior. While adults have been shown to entrust more money in agents whom they subjectively view as 'nice' (Siegel et al., 2018), children may not necessarily link niceness with trustworthiness. In addition, there was a disconnect between the stimuli that were presented and the ratings that were collected: most of the behaviors were morally relevant (e.g., helped a player fell, lent their basketball to someone, pushed a classmate, refused to share ice cream), but we asked about niceness, which may fall under the 'warmth' umbrella of traits (Baharloo, Fei, & Bian, 2022). A growing body of work suggests that morality has special status in social cognition, and may comprise a third primary dimension of person perception, in addition to warmth/sociability and competence (Brambilla & Leach, 2014; Ray, Mende-Siedlecki, Gantman, & Van Bavel, 2021). Children may have been better able to make behavioral predictions if (1) we had them track a morally-relevant trait, then asked them to predict trustworthiness-related behaviors, or (2) we had them track niceness, then asked them to predict niceness-related behaviors. Exploring the dimensionality of person perception and the generalizability of trait inferences throughout development is a rich arena for further research.

Implications. The current research investigated real-time impression updating in 6-9-year-olds. We found a negativity bias, where positive-to-negative updating was stronger than negative-to-positive updating. Negative information was also more salient overall – children were more responsive to variations in the amount of negative vs. positive information, and older children updated weak negative impressions more than strong negative impressions. These results were unexpected, given past documentation of a positivity bias in trait reasoning during middle childhood; in fact, the current findings are largely consistent with adult behavior in similar impression updating tasks. This suggests that the positivity bias may not be as pervasive during this developmental period as previously thought. More studies are needed to reconcile differences with past results: one possibility is that children view positive information as more diagnostic and positive traits as more stable, but do not follow these principles during impression updating; another is that children were making act-based judgments rather than person-based judgments.

This work adds to our understanding of children's flexible trait understanding in several ways. For one, we found that 6-9-year-olds are able to engage in abstract casual reasoning by integrating behavioral evidence across time, contexts, and recipients; this complements the finding that 3-6-year-olds have difficulty inferring traits if the behavioral instances are varied. We speculate that further maturation in causal reasoning in the social domain between these age ranges supports flexible trait inference. In addition, we found that children's negative impressions are easily updated in the face of negative information, but not in the face of positive information; holding onto detailed negative impressions may have an adaptive function. Furthermore, we found initial

evidence suggesting that children have difficulty translating inferences of niceness into inferences of trustworthiness; more work is needed to verify whether this is a general finding, and to further elucidate how children use person models to make predictions.

It is likely the case that different valence asymmetries will characterize different forms of trait reasoning in children. We found that when 6-9-year-olds engage in behavior-based impression updating, negative information tends to be more salient and durable. We hope future research will expand on this work, so that we may better understand when one valence is more powerful than the other, and the potential consequences of any asymmetries for social decision-making and well-being.

3.5 REFERENCES

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. Journal of Experimental Psychology, 70(4), 394.

Bach, P., & Schenke, K. C. (2017). Predictive social perception: Towards a unifying framework from action observation to person knowledge. *Social and Personality Psychology Compass*, *11*(7), e12312.

Baharloo, R., Fei, X., & Bian, L. (2022). The development of racial stereotypes about warmth and competence.

Baron, A. S., & Dunham, Y. (2015). Representing 'us' and 'them': Building blocks of intergroup cognition. *Journal of Cognition and Development*, *16*(5), 780-801.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323-370.

Benenson, J. F., & Dweck, C. S. (1986). The development of trait explanations and self-evaluations in the academic and social domains. *Child Development*, 1179-1187.

Boseovski, J. J. (2010). Evidence for "rose-colored glasses": An examination of the positivity bias in young children's personality judgments. *Child Development Perspectives*, *4*(3), 212-218.

Boseovski, J. J., Chiu, K., & Marcovitch, S. (2013). Integration of behavioral frequency and intention information in young children's trait attributions. *Social Development*, *22*(1), 38-57.

Boseovski, J. J., & Lee, K. (2006). Children's use of frequency information for trait categorization and behavioral prediction. *Developmental Psychology, 42*(3), 500.

Brambilla, M., & Leach, C. W. (2014). On the importance of being moral: The distinctive role of morality in social judgment. Social Cognition, 32(4), 397-408.

Brambilla, M., Carraro, L., Castelli, L., & Sacchi, S. (2019). Changing impressions: Moral character dominates impression updating. *Journal of Experimental Social Psychology, 82,* 64-73.

Chan, C. C., & Tardif, T. (2013). Knowing better: The role of prior knowledge and culture in trust in testimony. *Developmental Psychology, 49*(3), 591.

Chen, E. E., Corriveau, K. H., & Harris, P. L. (2013). Children trust a consensus composed of outgroup members—but do not retain that trust. *Child Development, 84*(1), 269-282.

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108*(1), 37.

Davies, J., & Brember, I. (1994). The reliability and validity of the 'Smiley' scale. *British Educational Research Journal, 20*(4), 447-454.

Dunham, Y., Baron, A. S., & Carey, S. (2011). Consequences of "minimal" group affiliations in children. *Child Development, 82*(3), 793-811.

Dunn, J., & Brown, J. R. (1993). Early conversations about causality: Content, pragmatics and developmental change. *British Journal of Developmental Psychology, 11,* 107–123.

Fivush, R. (1991). Gender and emotion in mother-child conversations about the past. *Journal of Narrative and Life History, 1,* 325–341.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493-498.

Hertenstein, M. J., & Campos, J. J. (2001). Emotion regulation via maternal touch. *Infancy, 2,* 549–566.

Hertenstein, M. J., & Campos, J. J. (2004). The retention effects of an adult's emotional displays on infant behavior. *Child Development, 75,* 595– 613.

Heyman, G. D., & Dweck, C. S. (1998). Children's thinking about traits: Implications for judgments of the self and others. *Child Development, 69(2),* 391-403.

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., Scheibe, S., & Hothorn, M. T. (2016). Package 'multcomp'. Simultaneous inference in general parametric models. *Project for Statistical Computing,* Vienna, Austria.

Hughes, B. L., Ambady, N., & Zaki, J. (2017). Trusting outgroup, but not ingroup

members, requires control: neural and behavioral evidence. S*ocial Cognitive and Affective Neuroscience, 12*(3), 372-381.

Jaswal, V. K., Croft, A. C., Setia, A. R., & Cole, C. A. (2010). Young children have a specific, highly robust bias to trust testimony. *Psychological Science, 21*(10), 1541-1547.

Jones, E. F., & Thomson, N. R. (2001). Action perception and outcome valence: Effects on children's inferences of intentionality and moral and liking judgments. *Journal of Genetic Psychology, 162*, 154–166.

Kim, M. J., Mende-Siedlecki, P., Anzellotti, S., & Young, L. (2021). Theory of mind following the violation of strong and weak prior beliefs. *Cerebral Cortex, 31*(2), 884-898.

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

Kim, M. J., Theriault, J., Hirschfeld-Kroen, J., & Young, L. (2022). Reframing of moral dilemmas reveals an unexpected "positivity bias" in updating and attributions. *Journal of Experimental Social Psychology, 101,* 104310.

Kim, M., Young, L., & Anzellotti, S. (2022). Exploring the Representational Structure of Trait Knowledge Using Perceived Similarity Judgments. *Social Cognition, 40(6),* 549-579.

Koenig, M., & Stephens, E. (2014). Characterizing children's responsiveness to cues of speaker trustworthiness: Two proposals. In *Trust and Skepticism* (pp. 13-27). Psychology Press.

Lagattuta, K. H., & Wellman, H. M. (2001). Thinking about the past: Early knowledge about links between prior experience, thinking, and emotion. *Child Psychology, 72,* 82–102.

Lewicki, R. J., & Brinsfield, C. (2017). Trust repair. *Annual Review of Organizational Psychology and Organizational Behavior, 4,* 287-313.

Lockhart, K. L., Chang, B., & Story, T. (2002). Young children's beliefs about the stability of traits: Protective optimism?. *Child Development, 73*(5), 1408-1430.

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *Journal of Neuroscience, 33*(50), 19406-19415.

Miller, P. J., & Sperry, L. L. (1988). Early talk about the past: The origins of conversational stories of personal experience. *Journal of Child Language, 15,* 293–315.

Newman, L. S. (1991). Why are traits inferred spontaneously? A developmental approach. *Social Cognition, 9*(3), 221-253.

Ray, J. L., Mende-Siedlecki, P., Gantman, A., & Van Bavel, J. J. (2021). The role of morality in social cognition. In *The Neural Basis of Mentalizing* (pp. 555-566). Cham: Springer International Publishing.

Reeder, G. D., & Coovert, M. D. (1986). Revising an impression of morality. *Social Cognition, 4*(1), 1-17.

Rholes, W. S., & Ruble, D. N. (1986). Children's impressions of other persons: The effects of temporal separation of behavioral information. *Child Development,* 872-878.

Rousseau, D.M., Sitkin, S.B., Burt, R.S., & Camerer, C. (1998). Not so different after all: a cross-discipline view of trust. *Academy of Management Review, 23,* 393–404.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296-320.

Sharot, T., & Garrett, N. (2016). Forming beliefs: Why valence matters. *Trends in Cognitive Sciences, 20*(1), 25-33.

Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., ... & Schulz, L. (2020). Online developmental science to foster innovation, access, and impact. *Trends in Cognitive Sciences, 24*(9), 675-678.

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour, 2*(10), 750-756.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*(1), 131.

Stephens, E., & Koenig, M. (2014). Characterizing children's responsiveness to cues of speaker trustworthiness: two proposals. In *Trust and Skepticism* (pp. 21-35). Psychology Press.

Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological

concept distinct from the self. *Perspectives on Psychological Science, 12*(4), 551-560.

Sutter, M., & Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Games and Economic Behavior, 59*(2), 364-382.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences, 22*(3), 201-212.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. Cognition, 126(2), 326-334.

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: the negativity bias in social-emotional development. *Psychological Bulletin, 134*(3), 383.

General Discussion

At the heart of social cognition is reasoning about agents as generative models. Per our conceptualization of this framework: we infer latent mental states and traits as the causes behind people's behaviors, and store this knowledge in person models, the organization of which may be context-dependent. We in turn use person models to make predictions about people's future behaviors. If we encounter unpredicted social information, the size of the resultant prediction error may be reflected in neural activity in ToM regions, and mentalizing may support the maintenance of strong prior beliefs (which we tend to have for close others and group members). In addition, negative social information is privileged during impression updating in adults, perhaps because it is more diagnostic; children, on the other hand, tend to be biased towards positive information during trait reasoning, perhaps because optimism serves a protective function.

We pursued several open questions within this framework across three studies. In Paper 1, we probed the structure of trait inference across contexts by conducting a representational similarity analysis on judgments of famous and unfamiliar people. We found that the relative contributions of different traits to overall impressions may vary depending on what we know about a person. We also found that, despite valence being a defining feature of the trait space for unfamiliar people, we were still able to uncover a higher-dimensional representational space on beyond this first component. In Paper 2, we examined the neural correlates of dynamic trait inference for unfamiliar people. We manipulated the strength of observers' impressions of fictional targets as trustworthy or untrustworthy by varying the amount of available behavioral evidence. We found that, when targets engaged in trait-inconsistent behaviors, there was increased ToM activity

following the violation of strong and positive prior impressions, consistent with the hypothesis that unpredicted/diagnostic information enhances ToM activity. In Paper 3, we studied dynamic trait inference in 6-9-year-olds. We manipulated the strength of children's impressions of fictional targets as nice or mean by varying the amount of available behavioral evidence. Surprisingly, children exhibited a negativity bias: they updated more in the positive-to-negative direction. Older children updated more for weak vs. strong negative impressions, and for strong vs. weak positive impressions. Finally, when targets behaved inconsistently, children were largely unable to translate niceness inferences into trustworthiness predictions. Below we discuss some open questions, common themes, and potential directions for future research.

Impact of inference context on trait space. Paper 1 highlighted a new way to investigate perceivers' overall impressions of people. In addition to probing how trait representations differ for famous and unfamiliar people, we can also ask how trait representations evolve for the same type of target as evidence accumulates. For instance, we can make use of the behavior stimuli from Paper 2, and present participants with targets who are paired with 2 behaviors or 6 behaviors; we can then compare which traits best predict pairwise similarity ratings for the 2-behavior targets, and for the 6-behavior targets. We may find that, as observers accumulate more behavioral evidence about an unfamiliar person, their trait representations begin to look more like those for famous people. In addition, it may be interesting to apply this bottom-up method to examine the traits that are important for different forms of social decision-making, especially in scenarios where our own preferences are not clear to us – e.g., people making hiring

decisions may wish to know which attributes they are implicitly using to evaluate and rank candidates.

Encoding of expectedness information in ToM regions. Paper 2 revealed that ToM regions are recruited more when strong and positive impressions are contradicted by new information, in line with a predictive coding view of social brain regions. In addition to examining overall response magnitudes for more unexpected and less unexpected information, we can also ask whether ToM regions carry information about unexpectedness, in the form of distributed patterns of neural activity. To do this, we can compare multivoxel pattern responses for behaviors that are matched on (1) ordinal position (within the sequence of 6 behaviors) and (2) valence, but differ in terms of the valence of preceding behaviors. For example, we can label the 3rd behavior in a Weak Positive-to-Negative sequence as 'unexpected', and the 3rd behavior in a Strong Negative-to-Positive sequence as 'expected', then test for above-chance classification (the former is unexpected, in that the negative behavior follows 2 positive behaviors; the latter is expected, in that the negative behavior follows 2 negative behaviors). As a stricter test, we can combine, across conditions, positive and negative unexpected behaviors, and positive and negative expected behaviors; accurate classification in this case would indicate that an abstract, potentially high-level feature of social information – unexpectedness independent of valence – is represented in brain regions that process social information. We can further test whether greater neural pattern discrimination for unexpected vs. expected information supports differential impression updating, by testing within-participant brain-behavior associations.

Translating between inferences of sociability and morality. Papers 2-3 employed largely the same paradigm, but with a couple of notable differences: (1) Paper 2 asked participants to rate targets' trustworthiness, while Paper 3 asked participants to rate targets' niceness (even though most behaviors concerned morality); and (2) Paper 3 asked participants to translate their niceness inferences into predictions of trustworthiness, while Paper 2 did not include a separate prediction task. As such, potential age-related effects are currently confounded by paradigm effects. It would be ideal for future work to study how both adults and children reason about both niceness and trustworthiness, and how they make predictions from one to the other. Such work would help examine the burgeoning hypothesis that morality should be considered as the third primary dimension of person perception, rather than being lumped in with warmth/sociality in traditional two-dimensional models (comprised of warmth-like traits and competence-like traits; Brambilla & Leach, 2014; Ray, Mende-Siedlecki, Gantman, & Van Bavel, 2021). We hypothesize that impression updating for niceness would proceed differently from impression updating for trustworthiness, especially since the behavior stimuli are mostly relevant to morality. In addition, there may be an asymmetry in generalizability, such that cross-context prediction is easier in one particular direction than the other. Furthermore, the behavior stimuli in Paper 3 give us an opportunity to apply the method used in Paper 1: future work can use those stimuli to collect pairwise holistic similarity ratings and multiple trait ratings in children, and ask which traits are most important for children's impressions of unfamiliar people.

Observations across studies. In the preceding chapters, we started to address several open questions on our ability to reason about people as generative models. In

Paper 1, we explored the structure of trait knowledge; in Paper 2, we examined how observers respond to trait-inconsistent information; and in Paper 3, we investigated the developmental trajectory of dynamic trait inference. We observed some common themes emerge across these studies. For one, they demonstrate a robust sensitivity to the amount of evidence available when reasoning about people's traits. It appears that this capacity emerges during middle childhood (Paper 3), is attuned to both stark (Paper 1) and subtle (Papers 2-3) differences in the amount of evidence, and arises in low-motivation social contexts (non-close/fictional others; Papers 1-3). These findings indicate that trait reasoning can be rationally responsive to information, and suggest that differences in trait reasoning (e.g., for close/ingroup vs. distant/outgroup others) can be partially understood in terms of differences in information. It will be interesting for future work to explore connections between this sensitivity and real-life social outcomes, such as the number and quality of friendships. It may be that people who rely more on accumulated evidence will make better decisions (e.g., whether to re-engage with someone following bad behavior); on the other hand, it may be that people who rely more on social motivations (e.g., the desire to always view friends favorably) will be preferred by social partners.

Furthermore, these studies provide us with implicit indices of flexible trait reasoning: Paper 1 infers the structure of trait space from associations between similarity ratings and trait ratings; Paper 2 examines neural responses to information that is more vs. less unexpected given priors; and Paper 3 looks at coherence between trait ratings and predictions of future behavior. These methods allow us to covertly assess: which types of social information observers find surprising, how they prioritize and integrate information and inferences to form overall impressions of people, and how they cash out

on impressions to predict future behavior and plan responses accordingly. In addition, these findings raise questions on the penetrability of different external interventions on these processes: e.g., what is the best way to get observers to disregard or reframe deeply surprising information, and would it alter implicit impressions in the long term? How do different social goals and possible future interactions shape the structure of trait space? There are many interesting possibilities for further research on the lability and context-dependence of trait reasoning.

Conclusion. Across three studies, we examined the structure, dynamics, and developmental trajectory of person models. In all, we found evidence for flexible trait reasoning – both children and adults were sensitive to the amount of available behavioral evidence, and to the overall inference context. In addition, the unexpectedness of new evidence was represented in univariate differences in ToM activity. Furthermore, children as young as 6 seemed adept at pulling out a common dispositional cause across separate behavioral instances, and 6-9-year-olds exhibited an adult-like negativity bias in impression updating. The current studies help shed light on how children and adults reason about person models and respond to new social information, and we suggest multiple avenues for further research in this arena.

References

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. Psychological monographs, 47(1), i.

Anzellotti, S., & Young, L. L. (2020). The acquisition of person knowledge. Annual Review of Psychology, 71, 613-634.

Bach, P., & Schenke, K. C. (2017). Predictive social perception: Towards a unifying framework from action observation to person knowledge. Social and Personality Psychology Compass, 11(7), e12312.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. Review of general psychology, 5(4), 323-370.

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. Nature, 456(7219), 245-249.

Boseovski, J. J. (2010). Evidence for "rose-colored glasses": An examination of the positivity bias in young children's personality judgments. Child Development Perspectives, 4(3), 212-218.

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.

Heider, F. (1982). The psychology of interpersonal relations. Psychology Press.

Kim, M. J., Mende-Siedlecki, P., Anzellotti, S., & Young, L. (2021). Theory of mind following the violation of strong and weak prior beliefs. Cerebral Cortex, 31(2), 884-898.

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

Kim, M., Young, L., & Anzellotti, S. (2022). Exploring the Representational Structure of Trait Knowledge Using Perceived Similarity Judgments. Social Cognition, 40(6), 549-579.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: a neural prediction problem. Neuron, 79(5), 836-848.

Malle BF. 2001. Folk explanations of intentional action. In: Intentions and intentionality: Foundations of social cognition. Cambridge (MA): MIT Press, pp. 265–286.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. Psychological Review, 86(1), 61.

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. Nature human behaviour, 2(10), 750-756.

Stanley, D. A. (2016). Getting to know you: general and specific neural computations for learning about people. Social Cognitive and Affective Neuroscience, 11(4), 525-536.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. Trends in cognitive sciences, 22(3), 201-212.

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: the negativity bias in social-emotional development. Psychological bulletin, 134(3), 383.