Changing my mind: The socio-affective mechanisms underlying impression updating

Benjamin Michael Silver

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2025

Abstract

Changing my mind: The socio-affective mechanisms underlying impression updating

Benjamin Michael Silver

How do we change our beliefs about other people? When we learn new information about someone that contradicts what we previously believed about them, we are likely to change – or update – our beliefs about their relevant traits. But when it comes to someone we know well, or have strong feelings towards, or who is simultaneously evaluating us, the situation is more complicated. In this dissertation, I investigate the socio-affective and neural mechanisms that underly impression updating. In Chapter 1, I analyze social media posts to ask how the general public updated their perceptions of the moral character of public figures accused of sexual assault during #MeToo, and how pre-existing motivations impacted the degree of updating. I find that liking mitigates negative updating for less severe accusations, indicating forgiveness of well-liked others. In Chapter 2, I again examine the impact of motivations on impression updating, but for the competence and sociability of close others after completing a virtual escape room game together. I find that the self-enhancement bias leads perceived similarity to impact perceptions of competence, while liking impacts perceptions of sociability. Both of these chapters also investigate the durability of an update, and both demonstrate that impression updates persist beyond immediate effects. Finally, in Chapter 3, I use functional magnetic resonance imaging to study the neural mechanisms implicated after an update occurs in the context of romantic interest. I find that the mentalizing network responds to social feedback, suggesting that romantic interest updates are interdependent with a target's own evaluations.

Table of Contents

List of Figures

Acknowledgments

As if the six years and nearly 150 pages spent on topics that I find intrinsically interesting wasn't indulgent enough, I hope you'll indulge me just a bit longer as I express my gratitude to all the people who helped me get to this point.

Thank you to my advisor, Kevin Ochsner. You are the embodiment of the phrase "trust the process" – never have I worked with someone as patient and as understanding and as optimistic as you. Without you, this dissertation would be painfully boring, completely devoid of stories and phenomena and intrigue. I am the scientist I am today because of you. Thank you, thank you, thank you.

Thank you to my secondary advisor, Chris Baldassano, for somehow knowing the answer to every question anyone has ever asked. Your generosity with your time is endlessly appreciated.

Thank you to the rest of my committee – Lila Davachi, Meghan Meyer, and YC Leong – for showing such genuine interest in my work and in my development as a scientist. It's comforting to know how many people I have to turn to for help.

Thank you to all members of the SCAN Lab – past, present, and future – for making me feel like I'm a part of a community of like-minded scholars. Special shout outs to Monica, Ovidia, Jo, Eisha, and Alex for your friendship these last few years, and extra special shout out to Monica for introducing me to the wonders of Learned League.

Thank you to my cohort – Ana, Jacob, Camille, John, Anna, Margaux, and Wangjing – for helping set a department norm that it's cool to attend events. Special shout outs to Wangjing for being an amazing collaborator and to Ana for helping me become a better teacher. And thanks to my other departmental collaborators – Emma, Arlene, Danika, and Courtney – as well. We'll finish those projects eventually.

Thank you to the Public Programs team at the Zuckerman Institute, and in particular to Diana Li. My three years working with BRAINYAC provided me with some of my fondest memories from graduate school. Diana, I respect you immensely.

Thank you to my students and mentees, undergrads and precocious high schoolers alike. My students in Psychology and the Internet convinced me that all Columbia undergraduates are brilliant, and my students in BRAINYAC convinced me that all high schoolers are, too. Special shout outs to Arielle Clarke and Jana Jaran for helping me develop my mentorship skills.

Thank you to my family for a) bringing me into existence, and b) providing me with every opportunity I ever asked for. Thank you to Grandma Nadine, Poppy R, and Grandma Irma for always being proud, at even the smallest things. Thank you to Rebecca for showing me what it means to be a caring sibling. I follow your lead. And thank you to Mom and Dad for giving me my work ethic, my love of learning, and my ability to advocate for myself. I'm only here – in every sense of the word – because of you.

Lastly, thank you to Emily. You have quite literally been there for me every step of the way, and at this point you probably – against your will – know as much about my research as I do. There aren't enough pages in the world to express how grateful I am to have been able to complete this degree with you by my side. Your love allowed me to be brave, take risks, and be honest with myself. I highly recommend that everyone does a PhD while living with their best friend. You almost make me want to do it again.

Introduction

"Everything changes and nothing remains still;

and you cannot step twice into the same stream."

(Heraclitus, c. 500 BCE)

The only constant in life is change. A trite aphorism, to be sure, but it's an idea that has persisted since ancient times due to its veracity and its simplicity. As Heraclitus observes, life is like a rapidly flowing stream, where the waters before you are different from the waters that flowed past moments before. We take this to be a natural property of any body of water that runs from one place to another; so, too, should we understand our daily lives to have this natural property as well. Change is everywhere. Save for the still, secluded rooms where we write our dissertations, it would be odd to not encounter change from one day to the next. Global events, immediate surroundings, social interactions. Universal flux, indeed.

With so much change, it's a wonder that the human experience is, relatively, so stable. For this, we have our adaptability to thank. As the world around us changes, we are not stuck in silence. We can change in turn – adapt, if you will. We jump back onto the sidewalk as a car blows through a yellow light. We adopt local tipping customs when we visit another country. We become passionate about a social issue after watching an affecting documentary. Our brains and our minds are built for flexibility and adaptability – we wouldn't survive in this universe otherwise (Kashdan & Rottenberg, 2010).

Perhaps the most complex, most ever-changing domain of the human experience is our relations to other people. What makes other people unique in our story about adaptability is that other people are themselves agents, meaning they have free will to make decisions and display behaviors (Baumeister, 2008). This fact makes it nearly impossible to predict with 100%

certainty what other people will do at any given moment (Frith & Frith, 2006; Sebanz & Knoblich, 2009; Springer et al., 2012). When another person acts in a way that we fail to predict – we learn something new about them, we witness them in a new environment, we receive an indication from them about how they feel about us – we are likely to *change* our perceptions and our beliefs about who they are. Social psychology research can help us understand what this process of belief change for other people looks like. Upon first meeting someone, we instantly begin to form an impression of them based on immediately observable features such as their age, gender, and race (Freeman et al., 2010; Freeman & Ambady, 2011), as well as their facial structure (Todorov et al., 2015; Willis & Todorov, 2006). But as we spend more time with this person, we continuously learn new information about them, which may lead us to revise or update our initial impressions (Cone et al., 2017; Mende-Siedlecki, 2018; Moskowitz et al., 2022).

One of the earliest studies of impression updating demonstrated a resistance to change via a primacy effect of first-encountered traits (Asch, 1946). Since then, many impression updating studies have focused on instances where trait-impressions for a specific person conflict with a pre-existing stereotype (Crocker et al., 1983; Stangor & Ruble, 1989; V. Y. Yzerbyt et al., 1998). Other work has investigated the asymmetry in impression updating, with most studies showing negative information outweighing positive information (Baumeister et al., 2001; Crocker et al., 1984; Reeder & Coovert, 1986; Reeder & Spores, 1983), although some dimensions, such as competence, may show the reverse effect (Mende-Siedlecki, Baron, et al., 2013; Reeder et al., 1977; Wojciszke et al., 1993). Some work has also attempted to explain differences in impression updating between explicit and implicit impressions (Cone et al., 2017; Rydell & McConnell, 2006; Wyer, 2010), where explicit impressions are generally updated more readily than implicit impressions. As a result, most recent work on impression updating proposes models of impression change for implicit impressions specifically (Kurdi et al., 2022; Mann et al., 2020; Shen & Ferguson, 2021).

While person perception research has given us some insight into how beliefs about other people can change in the face of new information, much of it fails to capture the full range of situations in which impression updating can occur, leaving our understanding of this phenomenon incomplete. This dissertation seeks to fill those gaps. In particular, the roles of socio-affective motivations and prior relationships in impression updating have been left relatively unexplored, since much work on impression updating uses unfamiliar targets in hypothetical situations (Brambilla et al., 2019; Cone et al., 2017; Mende-Siedlecki, Baron, et al., 2013). But we know that relationships and pre-existing beliefs can motivate social perceptions and our susceptibility to attribution errors (Hewstone, 1990; Howard & Rothbart, 1980; Kunda, 1990; B. Park & Young, 2020; B. Schiller et al., 2014), so it's likely they would play an important role in impression updating as well.

These motivations likely stem from the fact that our own identities are implicated in our relationships; since we're motivated to view ourselves positively (Alicke & Sedikides, 2009; Sedikides & Gregg, 2008), we're motivated to maintain positive impressions of those with whom we have strong pre-existing relationships. In addition, while there is ample work in social psychology characterizing close relationships (Finkel et al., 2017; Larson et al., 2022), this literature remains largely separate from the person perception literature, making it unclear which specific factors in a relationship – how close we feel to someone, how long we've known them, how much we like them, etc. – are most likely to motivate an impression update. In Chapters 1 and 2 of this dissertation, I take a motivational approach to impression updating. I use real-life situations and personally relevant targets, and seek to characterize how the specific nature of prior relationships impacts impression updating across different well-studied dimensions of person perception.

Another gap in our understanding of impression updating pertains to the durability of an impression update. The vast majority of impression updating studies only examine immediate changes (Forscher et al., 2019), and in the few cases where impressions are re-assessed later in time, they typically don't last (Gawronski et al., 2017; Lai et al., 2016; Vuletich & Payne, 2019), perhaps because they fail to generalize to new contexts (Gawronski et al., 2018). However, the external validity of these findings is hard to assess because they also often occur in hypothetical contexts. Certainly, there are instances in daily life where impression updates stick – many of us have probably had the experience of growing apart from a long-time friend. When can we expect an update to stick vs fade? One of the few studies to find durable impression change beyond immediate effects noted that the unexpected information had to be both diagnostic and believable (Cone et al., 2021), although this study only examined durability one week later and used novel targets.

Chapters 1 and 2 of this dissertation explore the question of durability by evaluating perceptions of others at multiple timepoints. Chapter 2 compares the size of an update immediately after encountering new information to its size one week later. Chapter 1 makes a similar comparison, but looks at both day-to-day changes in the short-term (three weeks following an update) and overall changes in the long-term (one year later). Immediate changes in impressions can help us understand the immediate impact of specific types of information, but only by evaluating changes days or months later can we draw conclusions about how we incorporate this information into our long-term close relationships.

Finally, we are also limited in our understanding of the neural mechanisms that underly impression updating and changes in social evaluations because most neuroscience studies of social belief updating only investigate the moment when an update occurs. From this work, we know that the dorsomedial prefrontal cortex (dmPFC), which is often implicated in studies of person perception (Ma et al., 2014; Mitchell, Neil Macrae, et al., 2005; D. Schiller et al., 2009), demonstrates increased activity upon encountering unexpected information about another person (Cloutier et al., 2011; Mende-Siedlecki, Cai, et al., 2013; Mende-Siedlecki, 2018), and that this increase is at least partially dependent on the relevance/meaningfulness of the update (D. L. Ames & Fiske, 2013; Mende-Siedlecki & Todorov, 2016). In addition, the temporoparietal junction (TPJ) is also activated during a negative impression update, but less so for friends than for strangers, perhaps because participants updated less for friends than for strangers (M. Kim et al., 2020; B. Park & Young, 2020). Some neuroscience studies also demonstrate that social evaluations are interdependent: in the process of making social evaluations of others, the anterior cingulate cortex responds to social feedback according to its valence and its congruence with one's own evaluations (Somerville et al., 2006, 2010; van Schie et al., 2018).

However, the downstream neural impacts of an impression update are relatively unexplored, both in terms of *how* we think about someone and *how often* we think about them. In other words, we have some idea of what happens in the brain at the moment new/unexpected social information is presented, but which neural systems change as a result of that information, and what that implies for the specific psychological mechanisms involved in an impression update, is far less clear. Chapter 3 of this dissertation investigates these questions.

Overview of dissertation

This dissertation draws on literature from person perception, close relationships, and social neuroscience to demonstrate how socio-affective motivations impact the degree to which we change our beliefs about other people when faced with unexpected information about them. I argue that these motivations are frequently a result of pre-existing relationships, and that the effects of these motivations are specific to the type of belief that is updated: the psychological mechanisms that implicate one's pre-existing relationships in changing perceptions of one trait are not the same as those that are implicated in changing perceptions of a different trait. In addition, I also argue that when impression updates are meaningful – in other words, when they are a response to real-life situations for those whom we have strong feelings towards – they are also durable, and can last for weeks or even months beyond the instant that an update occurs.

Finally, I also argue that belief updates in the context of close relationships are interdependent, in that they are often updated in response to what we believe another person believes about us. This dissertation sharpens our understanding of this phenomenon through the study of neural mechanisms of belief updating, with a particular focus on the mentalizing network in the brain. I demonstrate that the structure and reactivation frequency of neural representations for other people are updated in response to social feedback in brain regions related to mentalizing and person perception.

Chapter 1: Changes in online moral discourse about public figures during #MeToo

The first chapter of this dissertation leverages the #MeToo movement, which was started by Tarana Burke and rose to prominence in 2018 after a spate of sexual assault allegations made against high-profile public figures (North et al., 2020; Tambe, 2018), as a real-world, large-scale paradigm for understanding how socio-affective motivations impact impression updating. More specifically, I measured the perceived moral character of each accused public figure, both before and after their accusation became public, and treated the accusation as an instance of encountering new information that may or may not have been congruent with how the public figure was perceived pre-allegation. I scraped over one million social media posts from the website X (formerly Twitter) about 50 different male public figures to determine how a #MeToo accusation changed collective perceptions of moral character, and how external factors, such as how well-liked the public figure was, how famous they were, and how severe their alleged actions were, impacted the degree of change that occurred. I also investigated the durability of these changes by analyzing additional social media posts from one year after each allegation occurred.

Chapter 2: What are my friends really like? How we change our perceptions of familiar others' traits and actions

The second chapter of this dissertation investigates similar questions about motivation and durability, but in regards to how beliefs change for close friends after witnessing them in an unfamiliar environment, where there is a higher chance that they will display unexpected behaviors. Specifically, I enrolled groups of 4-5 friends to complete a virtual escape room together. The virtual escape room required the group to work together to solve puzzles in a high-pressure environment. I anticipated that two types of traits were likely to be on display in this environment: competence, which is defined as someone's ability to accomplish a task and is displayed in an escape room as someone's ability to solve puzzles, and sociability, which is defined as someone's ability to win social support and is displayed as an escape room as someone's ability to collaborate with members of their team (Brambilla et al., 2021; Landy et al., 2016).

I evaluated participant perceptions of their teammates' trait-level competence and sociability before the game, immediately after the game, and one week later to investigate the durability of game-related changes in trait perceptions. I also asked how relational factors, such as liking, familiarity, and perceived similarity, impacted the degree of change, and whether or not the role of relationships in trait-perception change was specific (predicted by a single factor that was dependent on the trait type) or global (predicted by multiple factors with similar effects across trait types). Finally, I asked how one's ability to solve puzzles and collaborate with teammates during the game, as well as perceptions of these abilities, similarly impacted the degree of change in perceptions of competence and sociability.

Chapter 3: The mentalizing network updates neural representations of romantic interest in response to social feedback

The third chapter of this dissertation investigates the neural mechanisms of belief change for other people. I used romantic interest as a high-stakes, socio-affectively motivating social evaluation, and social feedback about (un-)reciprocated romantic interest as information that may elicit a belief update. Specifically, I was interested in how social feedback changes *how* we think about a potential romantic partner, as well as *how often* we think about them. I hypothesized that the mentalizing network would play a role in both of these processes, given that we consider both how we feel about someone and how we think they feel about us when evaluating romantic interest.

During an fMRI scan, participants watched dating profile videos of eight different targets, and assessed each target on romantic interest. Participants watched two videos of each target: one before receiving social feedback from the target (which was ostensibly based on a dating profile video the participant had made, but was in reality pseudorandom), and one after.

Participants also completed a resting state scan after viewing each set of eight videos. I hypothesized that neural representations of targets in brain regions related to mentalizing would change more in response to feedback that was incongruent with the participant's initial evaluation of romantic interest. I also hypothesized that targets would be reactivated more frequently in mentalizing brain regions after receiving social feedback.

Methodology

Natural language processing

Chapters 1 and 2 make use of natural language processing (NLP), or the analysis of spoken or written text to understand psychological processes (Feuerriegel et al., 2025; Jackson et al., 2022). In Chapter 1, NLP is used to quantify perceptions of moral character, while in Chapter 2, it is used to quantify team collaboration ability. The analysis of text in social psychology was perhaps best formalized by Pennebaker and colleagues with their dictionary-based Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015; Tausczik & Pennebaker, 2010), which was primarily developed to study emotion expression and cognitive styles through written text (Pennebaker et al., 2003; Pennebaker & King, 1999). Since then, NLP in psychology has expanded beyond the study of emotion and linguistic styles to make inferences about personality (Boyd & Pennebaker, 2017), well-being (Settanni & Marengo, 2015), and relations between social groups (Rathje et al., 2021). Chapter 2 uses LIWC's list of second- and first-person plural pronouns to quantify group focus of participants completing a virtual escape room, similar to previous studies of team collaboration (T. Driskell et al., 2013; Kane & Van Swol, 2023). Chapter 1 makes use of a different dictionary, the Moral Foundations Dictionary (Graham et al., 2009), which includes both positively and negatively valenced words related to morality, in order to understand how the general public perceived the moral character of public figures accused of sexual assault. Beyond its use in social and affective domains, NLP can also be used to measure the relationship between concepts based on the semantic similarity between representative texts (Evans et al., 2022; Kjell et al., 2019). This method is used in the study in Chapter 3 to analyze free-recall accuracy for dating profile videos, although is not discussed in this dissertation.

Web-scraping

If the purpose of psychology is to understand human behavior, then we need to ensure that the paradigms we use accurately capture the environments where human behavior takes place. Increasingly, human behavior takes place online, and social interactions take place on social media (Gosling & Mason, 2015; Wallace, 2015). In the past 15 years, psychology has turned to web-scraping of social media data as a valid way of understanding how human beings interact with each other (Kern et al., 2016; Kross et al., 2019; Luhmann, 2017). Social media data has been used to understand collective psychological responses to large-scale events and phenomena, from climate disasters (Sisco et al., 2017) to social movements (Ince et al., 2017; Mesquiti et al., 2025) to school shootings (B. Doré et al., 2015; Jones et al., 2016). Frequently, these studies make use of text-heavy social media websites such as X (formerly Twitter) (B. Doré et al., 2015; Simchon et al., 2020) and Reddit (Ashokkumar & Pennebaker, 2021; Mesquiti et al., 2025).

In Chapter 1 of this dissertation, I scrape social media posts on X to measure perceptions of moral character during the #MeToo movement. In addition to capturing a much larger dataset than would be possible in the lab (we analyzed over one million tweets about 50 different male public figures), social media data also allow us to capture real-time responses to real-world phenomena, rather than using hypothetical or retrospective paradigms, which we often rely on in laboratory experiments. In Chapter 1, this feature of social media data made it possible to study responses to sexual assault accusations over a very large time-period, from six months before an allegation to one year after.

Multi-level modeling

Traditionally, psychology research has averaged responses within each participant in a research study. With enough trials, the thinking goes, these averages are stable and representative of that participant. While this may be partially true, averaging obscures measures of within-participant variability, and prevents us from making estimates about the behavior of individual participants (Baayen et al., 2008; Bolger & Laurenceau, 2013; Hoffman & Rovine, 2007). Multi-level modeling, which accounts for fixed effects between-subjects *and* random effects within-subjects, can help us attain more accurate estimates of psychological effects, and allows us to leverage all of the data we collect, rather than erasing some of its richness through averaging (Gelman & Hill, 2007; Locker et al., 2007).

All three chapters in this dissertation make use of multi-level modeling. Chapter 1 treats each public figure as a grouping variable, and each social media post as a trial, so that heterogeneous effects can be estimated for each public figure. Significant heterogeneity prompted me to look for predictive factors. In Chapter 2, each participant in a group rates each teammate on competence and sociability; I treat both participant and group as grouping variables in my models to account for participant- and group-specific differences in rating tendencies. Finally, in Chapter 3, each potential romantic partner is modeled as a trial so that we can account for within-subject heterogeneity in romantic interest.

Multivariate fMRI analyses

Chapter 3 of this dissertation uses multivariate fMRI analyses to draw conclusions about patterns of representation for specific stimuli (Haxby et al., 2001; Norman et al., 2006). In particular, we make use of two multivariate methods: representational similarity analysis (RSA) and reactivation analysis. RSA uses second-order statistics to make it possible to link similarities in patterns of neural activity to similarities in a behavioral variable (Kriegeskorte, 2008; Popal et al., 2019). This method proves especially useful in social neuroscience, where stimuli often vary continuously, and where RSA has been used to identify networks that respond to facial structure (Brooks & Freeman, 2018; Stolier et al., 2018), social networks (Parkinson et al., 2017, 2018), and mental states (Tamir et al., 2016; Tamir & Thornton, 2018). In Chapter 3, I use between-subjects RSA to link neural activity to evaluations of romantic interest for potential romantic partners. In this way, I am able to detect not just which regions are more or less engaged when making this type of social evaluation, but how consistently those regions represent its structure.

In addition, I use reactivation analysis to detect how often participants think about potential romantic partners after watching their dating profile videos. Reactivation analysis is frequently used in the memory literature (Schapiro et al., 2018; Staresina et al., 2013; Tambini & Davachi, 2019; Yu et al., 2024) as evidence of consolidation of certain information after encoding. Typically, reactivation analysis is conducted by comparing the pattern of activity while viewing a stimulus to the pattern of activity during a post-viewing resting state scan. If the similarity is high enough at a particular timepoint during the resting state scan to surpass a pre-set threshold, then that timepoint is counted as an instance of reactivation. It has been shown that we more frequently reactivate more motivationally relevant stimuli, such as stimuli associated with a higher monetary reward (Gruber et al., 2016) or social stimuli (Jimenez & Meyer, 2024). In Chapter 3, I ask if certain types of social feedback from a potential romantic partner elicit larger changes in reactivation frequency for that partner.

Chapter 1: Changes in online moral discourse about public figures during #MeToo

Chapter previously published as Silver, B.M., Ochsner, K.N. (2024). Changes in Online Moral Discourse About Public Figures During #MeToo. *Affec Sci.* https://doi.org/10.1007/s42761-024-00250-4

1.1 Introduction

Activist Tarana Burke started the #MeToo movement in 2006. It went viral in 2017 and 2018 when over 250 (predominantly male) public figures were accused of committing sexual assault and/or abuse (North et al., 2020; Tambe, 2018). Many of these figures were previously revered and respected; as such, #MeToo provides a unique opportunity to study how the general public changes their discussions about public figures who are embroiled in public controversies. Here we ask whether and to what extent perceived morality of male public figures accused during #MeToo was influenced by prior familiarity with, and general liking of, the public figure, as well as the severity of the alleged actions.

At present, it is unclear how these variables may interact to cause an initial – or a lasting – change to population-level perceptions, in large part because relevant prior work has largely consisted of laboratory studies of how individuals change their beliefs about specific others. Together, these studies have shown that changes in beliefs about others happen if we receive evidence that initial attitudes or beliefs were incorrect or incompatible with subsequent behaviors the person in question demonstrates (Bhanji & Beer, 2013; Cone et al., 2021; Kovács, 2020; Mende-Siedlecki, 2018; Mende-Siedlecki & Todorov, 2016; B. Park & Young, 2020; Siegel et al., 2018).

However, for public figures – like politicians and Hollywood executives – the traditional approach of measuring shifting attitudes towards single individuals may not be the most useful level of analysis. Indeed, for public figures, it may be more important to study the ebb and flow of population-level discussions because they can determine large-scale outcomes such as who gets elected and what movies get made. While political science has long relied on public opinion polling to index population-level beliefs (e.g. Berinsky, 2017; Heath et al., 2005), here we took cues from psychological research on motivation and person perception to understand changes in public discourse surrounding figures accused during the #MeToo movement. No psychological study to date has investigated the response to or moral discourse around #MeToo accusations. Specifically, we investigated public discourse surrounding male public figures only, as the #MeToo movement was largely seen as a reckoning for powerful men, in particular (Tambe, 2018), as reflected by the fact that < 2.5% of public figures accused during #MeToo were women (North et al., 2020).

To accomplish this goal, we leveraged Twitter as a key source of data (Kachen et al., 2021; Xiong et al., 2019). For many years, natural language approaches to analyzing word usage in written texts, such as Linguistic Inquiry and Word Count (LIWC), have proved useful for understanding psychological responses to events, including emotions, beliefs, and attitudes (Mohammad, 2016; Pennebaker, 1997). Recently, these methods have been used to draw inferences about what Tweets can tell us about emotional responses to natural disasters (Sisco et al., 2017), political events (Simchon et al., 2020), violent acts that become national tragedies (B. Doré et al., 2015), and COVID-19 (Abdo et al., 2021; Metzler et al., 2023).

There are, of course, limitations to using Twitter data, including inherent difficulties in determining who/what is the subject of a tweet and understanding how to interpret the spread of a tweet (Burton et al., 2021). That said, Twitter data can provide a unique window into attitudes and beliefs on a large scale and over long periods of time. It also allows us to move beyond laboratory studies that examine impressions for novel (and often fictional) individuals about whom participants have no prior beliefs or feelings or immoral actions that are hypothetical or relatively unharmful -– which, to date, has been the norm – and ask whether changes in impressions about real-world figures endure over time.

Drawing inspiration from prior research, we sought to test three hypotheses regarding what tweets may reveal about population-level changes in moral discourse during the #MeToo movement. First, just as moral beliefs change when encountering evidence of immorality (Mende-Siedlecki, Baron, et al., 2013; B. Park & Young, 2020), we hypothesized that immoral language in tweets about public figures would increase sharply after the public figure was accused of sexual assault, as compared to baseline. Second, we hypothesized that general liking of, and familiarity with, public figures would predict the magnitude of changes in immoral language use. Lab studies have shown that we are likely to forgive close others for immoral behaviors (McCullough, 2001), which would suggest that higher liking and familiarity would lead to smaller increases in immoral language. However, harmful actions from close others can also lead to feelings of betrayal (Couch et al., 2017), which would mean that higher liking and familiarity could lead to *larger* increases in immoral language use. The question for #MeToo figures was which of these two paths public discussions would follow, and further, whether moves toward apparent forgiveness or betrayal would depend on the severity of the sexual assault allegations. Third, we sought to determine whether observed changes in immoral language use would persist over both short and longer time scales. In line with work showing that changes in beliefs and moral outrage wane over time (Crockett, 2017; Ferguson et al., 2019), we expected that immoral language use would lessen both in the short-term (the three weeks immediately after initial allegations) and in the long-term (one year later). However, given the consequential and real-world nature of the events, we anticipated that immoral language use one year later would still be higher than at baseline.

1.2 Methods

Public figure selection

A four-step procedure was used for generating a list of public figures that met specific selection criteria. First, we began with a comprehensive list of 262 individuals accused of sexual assault in the #MeToo movement as compiled by Vox.com (North, 2019). Second, within this set of 262, we focused on individuals for whom initial allegations became public during a one-year span beginning on October 5th, 2017, the day that Harvey Weinstein's allegations became public (Kantor & Twohey, 2017). That date is widely seen as launching the #MeToo era (Kachen, 2021). While the one-year cut-off is somewhat arbitrary, the vast-majority of high-profile #MeToo cases emerged during this one-year period (only 7 public figures in the Vox database have cases that emerged after the one-year cut-off), and focusing on sexual assault allegations during the #MeToo era lends a degree of consistency and shared context to the data. This cut-off led to the exclusion of 24 public figures. Third, all female figures (N = 6) were removed, as the #MeToo movement was perceived as being about powerful men, in particular (Tambe, 2018). This perception is borne out in that only 2.3% of public figures from the comprehensive Vox.com list were women. Fourth, we removed three public figures (Donald Trump, Brett Kavanaugh, and Roy Moore) whose allegations were tied to broader political events, as we predicted that discussion of these events would be present in tweets and would be confounded with the data relevant to our hypotheses. Finally, we removed one public figure (Nelly) whose name was commonly used in other contexts, which made it difficult to select tweets about sexual assault specifically. This left a list of 228 public figures.

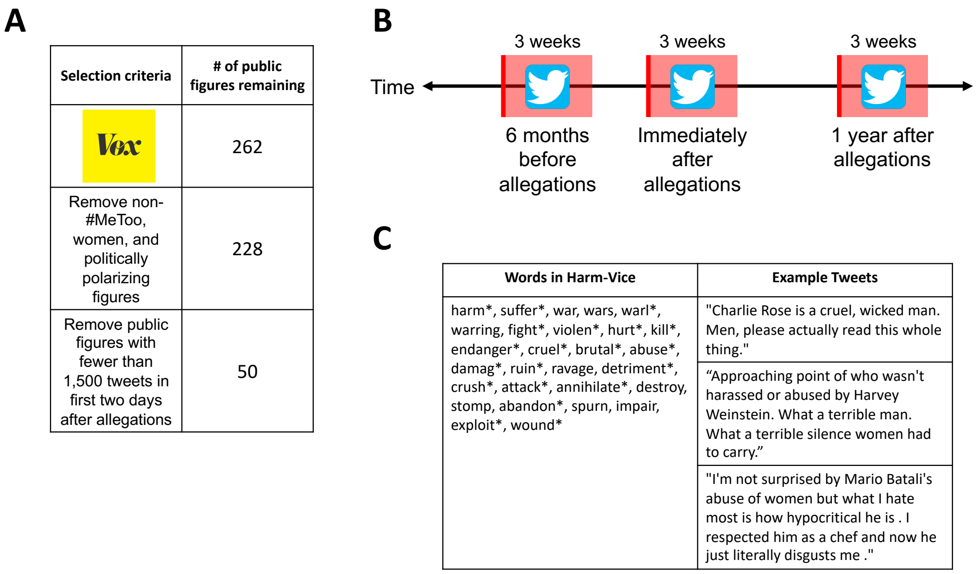
For each of these 228 individuals, an initial set of candidate tweets was selected from the first day after allegations became public. All tweets were collected using the Python package Twint (Zacharias, 2018), which scrapes tweets using Twitter's search function. Only tweets that included the full name of the public figure, or the name of the public figure without spaces, were collected to ensure that the tweet was about the public figure specifically and not the situation more broadly. Duplicate tweets were removed to reduce the influence of retweeted news articles. To ensure that we had enough tweets for each public figure to conduct robust analyses, public figures who were mentioned in fewer than 1,500 tweets on the first day were removed from the sample. Thresholding to improve the quality of Twitter data is a common practice (Murphy, 2017), although there is little agreement about what the exact threshold should be. For this study, the threshold of 1,500 tweets was chosen based on a number of factors. One was the bimodal distribution of tweet counts for the initial sample of public figures, where 1,500 tweets was a clear demarcation point. Above 1500 tweets, tweet number per public figure was distributed relatively evenly; by contrast, below 1,500 tweets, the majority of public figures had very low numbers of tweets. In addition, 1,500 tweets was a high-enough number to ensure that a) the included public figures were associated with significant and widespread discussion about their sexual assault allegations, and b) that there was enough text for each public figure's tweets to reliably analyze the data. This thresholding procedure resulted in a final list of 50 public figures *(Figure 1.1.A)*.

Tweet selection

For each of these 50 public figures, tweets mentioning that public figure's name (or their name without spaces) were collected across three time periods. To establish pre-allegation levels of moral language use in tweets about each public figure, *baseline* tweet collection was conducted for a 21-day period six months prior to the allegations. Using pre-allegation tweets as a baseline allowed us to control for the effects of the allegations and ensure that any change in immoral language was a result of the allegation and not a general feature of that public figure. To assess changes from baseline caused by allegations, *initial response* tweets were collected from the 21 days following the first public allegations about sexual assault. To investigate whether changes in perceived morality were maintained over longer time periods, tweets were collected from a 21-day period *one year after* initial public allegations (*Figure 1.1.B)*.

Data processing

Perceptions of morality in tweets were calculated using the Harm-Vice sub-list of the Moral Foundations Dictionary (MFD) (Graham et al., 2009), a natural language processing dictionary containing words related to both moral and immoral situations and characteristics. Not all words in the Harm-Vice sub-list are directly relevant to sexual assault, but because it includes words like *harm*, *abuse*, and *cruel*, it is the component of the MFD that most directly relates to the harmful and immoral behavior central to the #MeToo movement (*Figure 1.1.C*). The most widely-used software package for computing word counts, the Linguistic Inquiry and Word Count program (Pennebaker et al., 2015), calculates a score by determining the percentage of words in a tweet that are found in a particular sub-list of words. As such, tweets were concatenated by day (within each public figure's set of tweets) before being run through MFD (Tumasjan et al., 2010).



Selection criteria	# of public figures remaining
Vox	262
Remove non-#MeToo, women, and politically polarizing figures	228
Remove public figures with fewer than 1,500 tweets in first two days after allegations	50

Words in Harm-Vice	Example Tweets
harm*, suffer*, war, wars, warl*, warring, fight*, violen*, hurt*, kill*, endanger*, cruel*, brutal*, abuse*, damag*, ruin*, ravage, detriment*, crush*, attack*, annihilate*, destroy, stomp, abandon*, spurn, impair, exploit*, wound*	"Charlie Rose is a cruel, wicked man. Men, please actually read this whole thing."
	"Approaching point of who wasn't harassed or abused by Harvey Weinstein. What a terrible man. What a terrible silence women had to carry."
	"I'm not surprised by Mario Batali's abuse of women but what I hate most is how hypocritical he is . I respected him as a chef and now he just literally disgusts me ."

Figure 1.1. Tweet selection methodology.
Public figures were selected from a Vox.com database of public figures accused of sexual assault during #MeToo (A). We removed: public figures whose accusations emerged before October 5th 2017 or after October 4th 2018; the small number of female public figures included in this database (n = 6); male public figures tied to unrelated political events; and one male public figure whose name was difficult to search for in tweets. We then removed male public figures who were mentioned in fewer than 1,500 tweets the day after their allegations emerged. Tweets were collected using the Python package Twint for three weeks following allegations, for a three-week period six months before allegations, and for a three-week period one year after allegations (B). We used the Harm-Vice sub-list of the Moral Foundations Dictionary (C).

Data cleaning

For each collection period, three steps were taken to ensure that tweets accurately represented the conversation surrounding each public figure. First, all non-English tweets were removed. Second, all URLs were removed from the tweet text to reduce noise during linguistic analysis. Third, duplicate tweets were removed to reduce the influence of syndicated news articles often tweeted by bots. The resulting final tweet dataset consisted of 1,412,680 tweets, which included tweets from 6 months prior to allegations, tweets in the first three weeks after the allegations, and tweets from one year after the allegations. The median number of tweets per public figure was 10,281, with five public figures accounting for nearly half of the total number of tweets. See *Table A.1.1* for summary data showing date of first public allegations and number of tweets at each time period for each public figure included in the study.

Operationalization of factors that may motivate changes in immoral language use

We used a combination of surveys and lexical analyses to provide estimates of three factors that may affect the way in which people tweet about the #MeToo allegations levied against public figures – liking and familiarity for each figure prior to allegations being made, and the severity/harmfulness of the alleged actions. As described below, each factor was measured in multiple ways in order to make our measurements more robust, and to include both subjective and objective methods of measurement.

Liking

Pre-allegation liking was calculated in two ways: a dictionary-based approach and a machine learning approach. For the dictionary-based approach, the sentiment analysis tool AFINN (Nielsen, 2011) was used. AFINN scores each word in a text as either negative (scores ranging from -3 to -1), neutral (0), or positive (scores ranging from +1 to +3). Examples of negative words include evil (-3), awkward (-2), and demanding (-1), while examples of positive words include lenient (1), inspirational (2), and great (3). Tweets were concatenated by day and public figure, and AFINN scores were normalized based on the length of the text concatenation. For the machine learning method, a binary classification transformer model, using the uncased DistilBERT model (Sanh et al., 2020), was then trained on a dataset of 160,000 tweets from the *Sentiment140* dataset, which were each classified as either positive (+1) or negative (0), through the Simple Transformers Python package (Rajapakse, 2020). This model was then run on all of the baseline tweets, with each tweet classified as either positive or negative. Liking of the public figure was measured by averaging the transformer model score across all baseline tweets for that public figure.

Familiarity

Familiarity with the public figure was measured in four ways. The pygooglenews python package (Burgara, 2020) was used to measure trending news headline mentions of each public figure in the same 21-day period before each allegation, and number of tweets that mentioned the public figure was measured over this same 21-day period as well. These two measures were strongly correlated with each other ($p < .001$), suggesting that pygooglenews is a valid index of general news trends online.

Finally, we recruited 80 online survey participants (age range: 18-65) via Prolific to assess the prominence and power of each accused public figure. Prominence refers purely to fame – how well is the public figure known by the general population? Power refers to level of influence, which can be bestowed by money or social status. Prominence and power were separated in this survey, since some important cases during the #MeToo movement concerned public figures who were not necessarily household names before their accusations emerged, but nonetheless held significant power over others. Detailed definitions of prominence and power were shown to participants at the start of the survey. During the survey, the participants were shown the name and a photo of each public figure, and asked to retrospectively estimate, on a 0-100 scale, the prominence and power of each public figure before their sexual assault allegations emerged. (Participants were also asked to estimate current, post-allegation power and prominence, but these ratings were not used for analyses.)

Severity

Severity of the allegation was measured in two ways. First, each allegation against a public figure was summarized and anonymized, and shown to 100 participants recruited online via Prolific (age range: 18-65), who rated it on a 0-100 scale in terms of the amount of harm the event/behavior caused. Second, we created a rubric with four dimensions: number of people affected, length of time of behavior, type of behavior, and context. We scored each allegation against each of these dimensions, allowing us to calculate a summary severity score.

Prior to data analysis, the individual components for each factor were scaled around zero and averaged, so that there was one score for each factor.

Determining heterogeneity

In order to justify analyses of factors that may have motivated changes in moral discourse for public figures, we sought to demonstrate that there was indeed significant heterogeneity in the individual differences of changes across public figures. The SD of the slopes from six months prior to the first three weeks between public figures was 0.493 (95% CI: [0.394, 0.611]). Using two of the criteria laid out in Bolger et al. (2019), we determined that the heterogeneity of the slopes between time periods was significant. First, the random effect's 95% confidence interval did not surround 0 ([0.394, 0.611]), suggesting that the effect is likely not due to sampling error. Second, it is recommended that the random effect be larger than 25% of the fixed effect, and we found that in the present model it was equal to roughly 72% of the fixed effect (RE = 0.493, FE = 0.684, | RE/FE | = 72.07%). Despite this heterogeneity, 47/50 of the within-public figure slopes were positive, meaning that immoral language use increased, while the remaining three had 95% CIs surrounding 0.

1.3 Results

Did #MeToo allegations lead to a change in immoral language use?

Our first hypothesis concerned the immediate effects of sexual assault allegations on immoral language use online. We ran a multi-level Bayesian model with random intercepts and slopes, with public figure as the nesting variable (essentially, each public figure was treated as a study participant) and time period (a baseline period 6 months prior to allegations vs the first three weeks following allegations) as a random effect. Our model revealed that 0.16% of words in each day of the baseline tweets were found in the Harm-Vice list from the MFD – hereafter referred to as immoral language – with a between-public figure SD of 0.115 (95% CI: [0.066, 0.165]). In addition, there was a fixed effect of tweet time period, for which immoral language on each day were higher in initial response tweets as compared to baseline tweets (b = 0.683, SE = 0.071, 95% CI = [0.539, 0.823]), meaning that, on average, immoral language post-allegation quintupled as compared to baseline *(Figure 1.2)*.

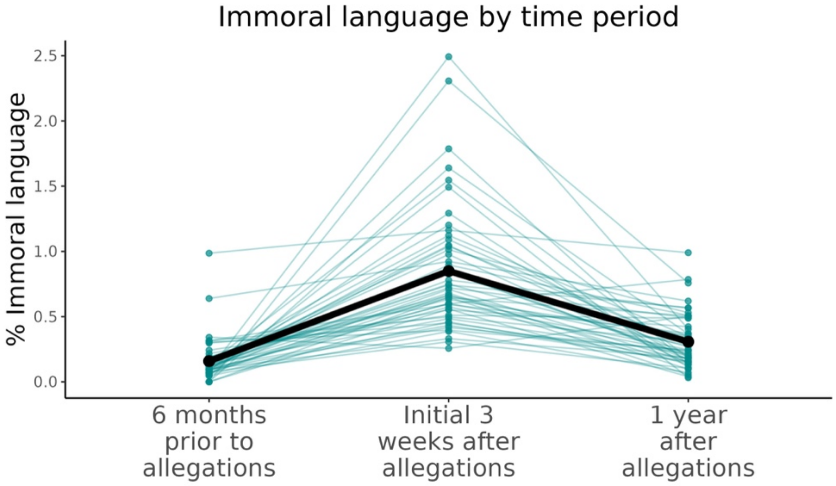


Figure 1.2. Immoral language use over time.
Average immoral language use for tweets in a three-week baseline period six months before allegations, a three-week period immediately after allegations, and a three-week follow-up period one year after allegations. Immoral language in tweets was calculated using the Moral Foundations Dictionary Care-Vice sub-list of words. Each line represents a public figure. The thick black line is the mean.

What factors predicted changes in immoral language use?

To address this question, we ran a Bayesian, multi-level model, with three potential factors – allegation severity, liking, and familiarity – as interacting predictor variables, and controlled for levels of immoral language in the baseline tweets figure (*Figure 1.3.A*). We found a significant positive effect of allegation severity (b = 0.279, SE = 0.105, 95% CI = [0.074, 0.483]), meaning that more severe actions led to tweets with more immoral language. We did not find a main effect of liking (b = -0.058, SE = 0.130 95% CI = [-0.320, 0.194]) or familiarity (b = -0.098, SE = 0.104, 95% CI = [-0.295,0.120]). However, there was an interaction between liking and allegation severity (b = 0.562, SE = 0.246, 95% CI = [0.085, 1.037]), such that at low severity levels, higher liking led to less immoral language use, while at high severity levels, higher liking did not predict a difference in immoral language use, with the interaction trending towards slightly *more* immoral language (*Figure 1.3.B*). There were no significant interactions between any of the other factors. Together, these results demonstrate that the severity of the action was most important in predicting the overall amount of immoral language following the #MeToo allegation, but that this effect differed as a function of liking for the public figure: for well-liked figures, tweets about them saw little change in immoral language use if the alleged actions were perceived to be less severe; by contrast, immoral language use increased significantly for severe allegations.

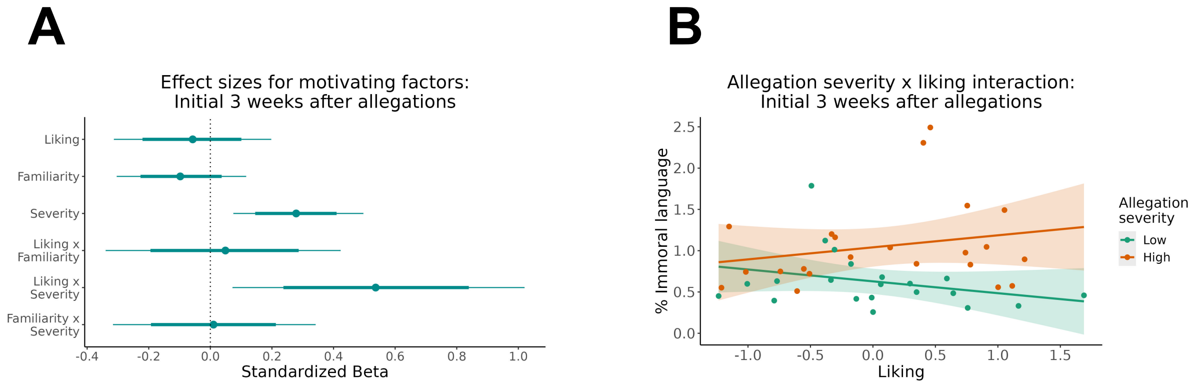


Figure 1.3. Effects of liking, familiarity, and allegation severity for initial three weeks. The effect sizes for each motivating factor's effect on overall immoral language in the initial three weeks, as defined by the MFD Care-Vice sub-list of words, are shown in (A). Thick bars are 80% credibility intervals and thin bars are 95% credibility intervals. The interaction between liking and allegation severity in the initial three weeks is shown in (B). Each point represents the average amount of immoral language in tweets about a public figure. Ribbons are 95% CIs.

What were the temporal dynamics of changes in immoral language use?

Short-term effects

To address short-term effects, we conducted an analysis of how immoral language in tweets changed in the three weeks following the allegation, on a day-to-day basis. In this model, number of days after allegations (0-20) was a predictor variable, as both a fixed and random effect. We found an effect of day (b = -0.043, SE = 0.006 95% CI: [-0.055, -0.031]), in which the percentage of words classified as immoral decreased over time in the first three weeks, but not in the tweets from six months prior (*Figure 1.4*). The decrease in immoral language appears to be exponential rather than linear, so we re-ran the model with the logarithm of Harm-Vice scores, and found a similar effect (b = -0.038, SE = 0.006, 95% CI = [-0.049, -0.027]), suggesting that the majority of the immoral language drop-off occurs early on after sexual assault allegations occur.

Next, we removed time period as an interaction term and limited our data to the first three weeks. In a model with day and all three motivating factors as interacting fixed effects and day as a random effect, we failed to find any interactions between motivating factors and day (severity: b = -0.007, SE = 0.009, 95% CI = [-0.024, 0.011]; familiarity: b = 0.005, SE = 0.009, 95% CI = [-0.013, 0.023]; liking: b = -0.013, SE = 0.011, 95% CI = [-0.035, 0.007]). This suggests that while motivating factors affect overall levels of immoral language in the three weeks following an allegation, they do not modulate the trajectory of the decrease in immoral language over time. Given that the majority of the change occurs in the first week, we ran an identical model, with only data from the first week after each allegation. However, we still failed to find any significant interactions between day and motivating factors.

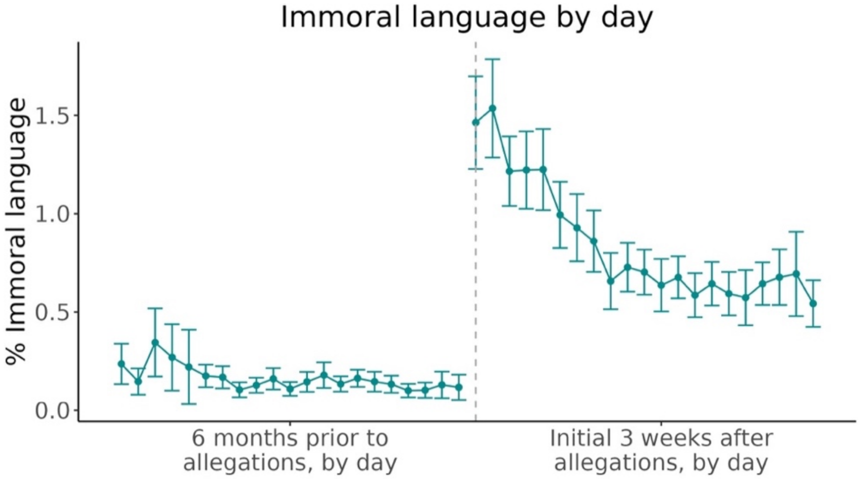


Figure 1.4. A comparison of baseline vs updating period immoral language use by day. Each dot represents a score for one day, with scores relatively constant for baseline tweets and decreasing over time for updating period tweets.

Long-term effects

To address long-term effects, for all public figures, we collected all tweets that mentioned them in a three-week period exactly one year after their allegations. We found that even one year later, the amount of immoral language was significantly higher than at baseline (b = 0.141, SE = 0.023, 95% CI = [0.097, 0.186]), but was significantly lower than in the three weeks immediately following an allegation (b = -0.541, SE = 0.070, 95% CI = [-0.680, -0.406]) (*Figure 1.2*).

We next ran an identical Bayesian multi-level model on tweets from the one year later period. While allegation severity was more predictive of immoral language in the first three weeks, in the tweets from one year later, familiarity and liking were more predictive than severity was (liking: b = -0.067, SE = 0.060, 95% CI = [-0.189, 0.053]; familiarity: b = -0.071, SE = 0.049, 95% CI = [-0.170, 0.025]; severity: b = 0.025, SE = 0.044, 95% CI = [-0.061, 0.119]), with higher levels of familiarity and liking predicting lower levels of immoral language (*Figure 1.5.A*). While the credibility intervals for effects of familiarity and liking were not completely outside zero, this result suggests that the primacy of the action in motivating changes in immoral language is a proximal effect, and that the effects of target-related motivations on moral discourse are more persistent over longer time periods. However, the interaction effect between liking and severity that was present in the initial three weeks was no longer present (*Figure 1.5.B*).

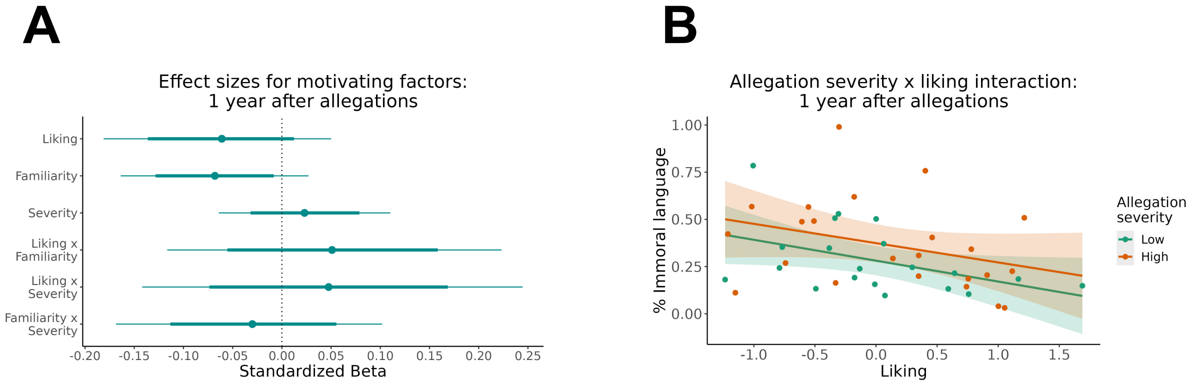


Figure 1.5. Effects of liking, familiarity, and allegation severity for tweets one year later. The effect sizes for each motivating factor's effect on overall immoral language use one year later, as defined by Care-Vice, are shown in (A). Thick bars are 80% credibility intervals and thin bars are 95% credibility intervals. The interaction between liking and allegation severity at one year later are shown in (B). Each point represents the average amount of immoral language in tweets about a public figure. Ribbons are 95% CIs.

1.4 Discussion

The #MeToo movement is thought to have sparked significant changes in the public discourse about the morality of dozens of prominent male public figures (Tambe, 2018). Here, we provide the first empirical evidence about the nature and key predictors of these changes. At multiple time points we analyzed the moral language used in tweets about male public figures accused of sexual assault to investigate the factors that predicted population-level changes in moral discourse. Three key findings were observed.

First, we found that #MeToo allegations significantly increased immoral language use in tweets about accused figures, which fits with prior lab-based findings on impression updating

(Mende-Siedlecki, Baron, et al., 2013; Reeder & Spores, 1983; Siegel et al., 2018) and responses to controversies on social media (Brady et al., 2021). Both types of research have shown that targets previously viewed as moral can be rapidly reassessed as immoral if we learn they have been accused of committing immoral acts. Among possible explanations for these findings, two are salient in the context of #MeToo: not only are immoral behaviors less common than moral ones, and therefore more diagnostic of a person's character (Fiske, 1980; Siegel et al., 2018), negative information generally is more impactful because it may signal a potential threat (Baumeister et al., 2001).

Second, the magnitude of the initial spike in immoral language depended on both the severity of alleged actions and how well-liked a public figure was before allegations emerged: liking mitigated immoral language for less severe allegations but trended towards an increase in immoral language for more severe allegations. This patterns suggests that we may collectively overlook, explain away, or forgive the immoral actions of liked individuals – so long as those actions don't seem too severe (Bradfield & Aquino, 1999; Fourie et al., 2020) – but for misdeeds of greater magnitude, then we may ignore our feelings of liking, or view them as increasingly immoral and even express moral outrage or feelings of betrayal (Couch et al., 2017; Couch & Olson, 2016). These findings also fit with lab studies showing similar effects when individuals evaluate social targets (Kihlstrom, 2013), including in the context of sexual harassment (Pryor et al., 1993).

Third, the level of immoral language in tweets one year after initial allegations was still greater than the pre-allegation baseline, but below the level seen during the three weeks immediately following the allegations. While this finding fits with studies showing that lasting changes in interpersonal beliefs happen only when an inconsistent behavior is both diagnostic and believable (Ferguson et al., 2019), it is important to note that whereas immoral language use was driven most strongly by allegation severity for the first three weeks, it was driven by liking for and familiarity with the public figures one year later. This pattern suggests that alleged actions provided an initial basis for moral discourse because they were highly accessible, concrete, and available to influence behavior (B. Doré et al., 2015; Higgins & Brendl, 1995; Rothbart et al., 1978; Schwarz et al., 1991). But after a while, the initial conversational focus faded, leaving the general public to base discourse on long-standing attitudes (such as liking and familiarity) towards a given figure. These findings fit with laboratory work showing that interpersonal beliefs can be change-resistant (Cao & Banaji, 2016), and often return to baseline quickly when changes occur (Lai et al., 2016). However, our findings go beyond this work by showing that over longer periods of time pre-existing attitudes toward people we know may be important and persistent predictors of collective judgments about them. In part, this may reflect the durability of semantic or "gist-like" representations of someone's traits, which we tend to rely on when making judgments about others (S. B. Klein et al., 1996; Sherman & Bessenoff, 1999; Wagner et al., 2019).

This leads to an important consideration – the public nature of the #MeToo movement means that there could also have been important situational and social influences on the emergence of sexual assault allegations. First, the phrase #MeToo existed for several years before the accusations against Harvey Weinstein; it is thought to have become a public and widespread movement because both the accused and the accusers had access to more public platforms (Tambe, 2018). It's precisely the public nature of this phase of the #MeToo movement that allowed us to analyze widespread conversation on social media. In this way, this method can also be considered a limitation, as it prevents us from analyzing how perceptions of moral character change for less prominent people accused of sexual assault. Second, the #MeToo movement may have involved a snowball effect: the more people who went public with allegations, the more comfortable others became with going public as well (Gallagher et al., 2019). Over time, it's possible this affected the general public's perceptions of survivors of sexual assault, which in turn could have influenced higher-level discussions each time new allegations emerged.

Similarly, the widespread coverage in conventional media channels and discussion on social media platforms could have influenced whether, when, and how a given person decided to tweet about one of the accused figures. The public nature of the #MeToo movement also may have influenced motivations to tweet, as it is known that on social media platforms people can be rewarded by their social networks for expressing moral outrage (Crockett, 2017), and that a relatively small number of users are responsible for a majority of posts (Brady et al., 2021). Many previous studies that use social media look at the spread of attitudes and information in an online context (Brady et al., 2017; Goldenberg & Gross, 2020; Schöne et al., 2021) by analyzing the salience of posts through reactions such as likes and retweets. In our study, we avoided these issues because we were not studying how information spreads, but rather, how population-level perceptions can be gleaned from aggregating social media data.

This naturally begs the question: Are these perceptions about the public figure, their alleged actions, or both? While counting immoral word usage in tweets can't differentiate these possibilities, for two reasons we think it's possible that (im)moral language in tweets may more strongly reflect beliefs/attitudes about public figures. First, beliefs about the morality of negative actions are typically stable over time (Goodwin & Darley, 2012). As such, the post-allegation increase in tweet volume and immoral language might reflect increased discourse about the qualities (e.g., the moral character) of the person implicated. Second, prior studies suggest that any immoral action is inextricably tied to the actor's perceived morality (Gantman & Van Bavel, 2015), which suggests the tweets we analyzed may reflect population-level beliefs/attitudes about the morality of accused public figures. This logic has similarly informed prior Twitter studies about reactions to public events (B. Doré et al., 2015; Metzler et al., 2023; Schöne et al., 2021; Simchon et al., 2020) and even sexual assault accusations (Maryn & Dover, 2023).

In addition to some ambiguity regarding the subject of the immoral language, there were two other limitations to our dictionary-based method of analysis. First, dictionary-based methods are not able to detect sarcasm or irony, which are popular forms of expression on social media platforms like Twitter (Sykora et al., 2020). Although there has been some work seeking to detect sarcasm using machine learning (Sarsam et al., 2020), these approaches are inexact. While there may have been some sarcastic or ironic language present in our data, the sheer quantity of our data (over one million tweets) makes it unlikely that this language significantly influenced our results. Second, the dictionary that we used, the Harm-Vice sub-list from the Moral Foundations Dictionary, may not have fully encompassed all language that is relevant to discussions of sexual assault allegations. Choosing the correct list of terms is often a topic of debate in dictionary-based research, and one that can potentially allow for large researcher degrees of freedom. We exclusively used a pre-existing list, based on a well-researched psychological construct within Moral Foundations Theory (Piazza et al., 2019), to ensure that our findings would be replicable and based on existing psychological theories.

We should also note that there are several other factors we did not test that may have impacted the magnitude of observed changes in immoral language. First, as previously stated, we excluded female public figures from analyses. The conversation about sexual assault committed by women is of a distinct nature, with additional considerations regarding power and gender (Gannon et al., 2008). Given that there were only 6 women in this dataset, we did not have enough datapoints to meaningfully compare the discourse around female public figures to the discourse around male ones. A future study may wish to systematically test these differences. Beyond excluding female public figures from our analyses, the identity of the accused, as well as the identity of the accuser, was not included in our models. Future work on this issue could examine whether population-level perceptions and discourse may have been impacted by the race of the accused, as race can play a role in perceptions of moral character (Eberhardt et al., 2006; Stanley et al., 2011). We did not test factors related to race in the current study because 43 out of 50 of our public figures were white.

Public figures accused of sexual assault also came from a wide variety of professions, from politics to Hollywood. The general public likely has different baseline assumptions about the moral character of people from different professions, perhaps because of differential perceptions of power (For example, a Hollywood executive might be deemed to have more power than a journalist.) Preliminary analyses on a subset of our data revealed that immoral language in the first three weeks increased more for figures from Hollywood than for figures from journalism/media, although this effect may be confounded with allegation severity, which was higher for Hollywood figures than for any other profession. A future analysis may wish to systematically compare population-level discussions about sexual assault across professions.

Finally, our data at the one year later timepoint may have been impacted by the fallout from the allegations: some public figures may have released genuine, well-received apologies, while others may have denied the accusations, and still others may have been cleared of wrongdoing altogether. We did not systematically test “allegation outcome” alongside our measures of liking, familiarity, and allegation severity. However, the emergence of an accusation during #MeToo typically led to more widespread media coverage than ensuing apologies and legal proceedings; thus, we believe the impacts of allegation outcome on our results are minimal. Despite these limitations, it should be noted that our aggregate results still hold. Regardless of the race and occupation of the accused or the accuser, or the outcome of the allegations, the pattern of results found in the paper still emerges in aggregate. As such, our findings may represent an average effect across race and occupation. Future work may wish to see if the present effects hold, are exacerbated, or are mitigated for specific categories of accused public figure or accuser.

In sum, the present data remind us that even people we like and are familiar with may act in ways that challenge our preconceived notions about them. Do these moments pass by without impact or influence? Are they actively explained away? Or do they profoundly change our perceptions? The current study addressed this issue in the context of the #MeToo movement, asking how society reacts when public figures that we know and like are alleged to have committed immoral acts. Changes in tweet content suggested that changes in the moral discourse about public figures did indeed occur, and that the nature and persistence of these changes was dependent on both the severity of alleged actions as well as how well-liked and well-known was a given public figure. These results highlight that collective beliefs about public figures may be constantly in flux, influenced by our prior attitudes and beliefs as well as our perceptions of their actions.

Chapter 2: What are my friends really like? How we change our perceptions of familiar others' traits and actions

Chapter previously published as Silver, B.M., Yu, W., Davachi, L., Ochsner, K.N. (in press). What are my friends really like? How we change our perceptions of familiar others' traits and actions. *Personality and Social Psychology Bulletin.* https://doi.org/10.31234/osf.io/n75xc_v2

2.1 Introduction

In everyday life, we face numerous novel situations in which we work with friends and coworkers to overcome stressful challenges and achieve common goals. An important question is what we learn from such novel situations about the character of well-known others. For example, imagine you and several of your coworkers are working on a new, unfamiliar project. Perhaps one of your coworkers, whom you previously thought to be skilled at solving problems, was not able to successfully carry out the tasks associated with this new project. Perhaps another coworker, whom you previously thought to be standoffish and isolated, took on a leadership role and successfully managed the members of your team. In each of these scenarios, the co-worker's unexpected behavior may change how we interact with and rely on them in the future. Addressing this question is clearly important, as highlighted by evolutionary theories that suggest that learning to coordinate with kin was an essential driver of the development of human social intelligence (Hayes & Sanford, 2014; Tomasello et al., 2012).

Surprisingly however, experimental behavioral work has left this question largely unexplored, as the two most relevant social psychological research literatures – person perception and close relationships – tend to operate independently and seldom focus on how trait perceptions change. On one hand, person perception research typically examines perceptions of and/or interactions with novel (or hypothetical) people (Brannon & Gawronski, 2017; Fiske, 1993). As such, this work cannot tell us how pre-existing relationships, and the factors that define them (i.e., relational factors, such as liking, familiarity, and/or perceived similarity), influence perceptions of others. Conversely, work on pre-existing relationships typically asks about relationship satisfaction (Finkel et al., 2017; Lemay & Clark, 2015) or trait perception accuracy (Biesanz et al., 2007; Kenny & Acitelli, 2001; Körner & Altmann, 2023; Wessels et al., 2020), rather than asking about how trait perceptions of close others change in light of new information. As such, the question of how we change our perceptions of a friend's traits after interacting in an unfamiliar context has received relatively little attention.

Here, we sought to address these issues by asking how our perceptions of friends may change when working with them to face unfamiliar challenges in a high-stakes environment. Specifically, we applied classic questions about person perception, which typically ask how we perceive novel or hypothetical people, to real pre-existing relationships, for which changes in trait perception are not typically examined. Although these lines of research are not commonly brought together in this way, we drew on previous person perception research to formulate the three inter-related hypotheses that we sought to address in this study.

First, we hypothesized that working with well-known others to accomplish a task would durably alter our perceptions of them. Group problem solving tasks require that someone has both the ability to accomplish the task and the ability to work well with other people (Akkerman et al., 2007; Hung, 2013). In our study, we operationalized these two abilities in terms of two well-studied dimensions of person perception: competence and sociability (Brambilla et al., 2021; Castelli et al., 2009; Landy et al., 2016). Competence broadly refers to one's ability to accomplish goals (Abele & Wojciszke, 2014) and, along with warmth, is considered one of the central dimensions of person perception in classic two-dimensional models (Fiske et al., 2007). More recently, it has been posited that the warmth dimension is an amalgamation of two other fundamental dimensions of person perception, morality and sociability, and that a morality-sociability-competence model is more accurate than a warmth-competence one (Brambilla et al., 2011, 2021; Landy et al., 2016). Sociability includes traits associated with one's ability to form relationships with others, such as extraversion and friendliness, but it also includes traits that are more relevant to working with others on a task and have some overlap with the moral dimension, such as empathy and cooperativeness (Goodwin et al., 2014; Landy et al., 2016).

While there may be other traits that also are important for determining the nature of relationships with others, we focused on traits related to competence and sociability because they are particularly relevant to a group problem-solving context. In addition, how competence and sociability are updated in response to new information, and how long lasting or durable these updates might be, is understudied compared to updating perceived morality (Brambilla et al., 2019; Mende-Siedlecki, Baron, et al., 2013; Silver & Ochsner, 2024). Some person perception work with unfamiliar or hypothetical targets suggests that interpersonal beliefs, in general, can be change-resistant (Cao & Banaji, 2016; Ferguson et al., 2019). In addition, we may be less likely to change our perceptions about well-known others' traits due to the large amount of evidence we already have about them (M. Kim et al., 2020). However, an uncommon environment that requires the use of those traits in unexpected ways may create opportunities to change our perceptions of well-known others, in both the immediate responses to the uncommon environment and several days later.

Second, we hypothesized that our relationship to a target would influence the way we make trait attributions about them (Brambilla et al., 2011, 2019; Goodwin et al., 2014; Landy et al., 2016). For people we know well, we hypothesized that at least three factors related to one's associations with, and relationship to, a target person could be important (Fiske, 1993; Kenny, 2004; W. M. Klein & Kunda, 1992; Zaki, 2014). The first is our liking of a target (Jussim et al., 1995; Leising et al., 2013; Wessels et al., 2020), which may motivate us to perceive them more favorably, thereby allowing us to maintain a view of ourselves as someone who has good judgment and likes others with positive traits. In the group problem solving example, when someone we like acts in a way that could exemplify a positive trait – such as sociability – we may be motivated to perceive them as possessing that trait more strongly than we would for someone we liked less. A second factor is familiarity (Montoya et al., 2017; Saegert et al., 1973; Zajonc, 1968), which tends to promote liking, in general (Reis et al., 2011; Zajonc, 2001). Psychology has long documented our fear of the unknown and preference for the familiar, so it's possible that we are more likely to positively assess those we know well and negatively assess those less well known. Finally, a third important factor is perceived similarity to oneself (Alves et al., 2016; D. R. Ames, 2004; Moreland & Zajonc, 1982; Mussweiler, 2003). Research suggests that we are biased to have positive views of those we are similar to (Montoya & Horton, 2013), although other work suggests we may also do the reverse, enhancing perceived similarity for those we view positively (Morry et al., 2011). In addition, work on self-enhancement suggests that we view ourselves as better and/or more important than we actually are (Beer & Hughes, 2011; Sedikides & Gregg, 2008), and it is possible that these enhancement effects might more easily extend to people we consider to be similar, rather than dissimilar, to ourselves. Indeed, prior work suggests that we often enhance similar others at the same time that we enhance ourselves (Morry, 2007; Morry et al., 2010).

Taken together, these considerations sharpened our second hypothesis: all three of these relational factors – liking, familiarity, and perceived similarity – would shape perceptions of a target's trait-level competence and sociability. However, while these three relational factors are commonly studied in relation to each other in person-perception research (Alves et al., 2016; Moreland & Zajonc, 1982; Strauss et al., 2001), how they interact to affect perceptions of friends' traits is unclear. In that context, there are two types of effects we may observe. On one hand, we may see a global effect, in which all three relational factors influence trait perceptions. This scenario would suggest that changes in perceptions of close others' traits were affected by merely the existence of a prior relationship, rather than the relationship's specific qualities. On the other hand, we may see a more selective effect, where some relational factors matter more than others. In this case, we would conclude that we value specific aspects of our relationships when re-assessing close others' traits.

Our third hypothesis posited that our perceptions of a target's competence and sociability would be related to the target's actions, as well as to our perceptions of their actions. Even when studies have examined perceptions of close others, they have rarely attempted to link perceptions of traits to perceptions of actions. To the extent that relational factors impact global trait perceptions, it's possible that these same relational factors might also impact perceptions of actions while working to achieve a common goal. For example, when working with others to solve a problem, individuals more adept at completing a task may be described as more competent, whereas people who collaborate better with others may be perceived as more sociable. In both cases, our *perceptions* of proximal behaviors – problem solving and group collaboration – may ultimately provide the impetus for updating judgments of relevant traits – competence and sociability. As such, we hypothesized that a) a target's objectively quantifiable actions during a group problem-solving task would impact perceptions of their traits, b) relational factors would bias perceptions of these actions, and c) biased perceptions of actions would bias perceptions of traits. If perceptions of in-the-moment actions and global traits are both influenced by relational factors, it is possible that there is overlap between the mechanisms that motivate trait and action perceptions. If, on the other hand, perceptions of global traits are influenced by relational factors, but perceptions of actions are not, it would suggest separate mechanisms for evaluating the actions and traits of close others.

To address these three hypotheses, we collected data from friends completing a virtual escape room game because it provided an unfamiliar and motivating environment that required people to work together to achieve a common goal. In addition, this activity allowed participants to freely interact with each other in a structured context with concrete performance metrics. Critically, the two traits of interest here – competence and sociability – are directly relevant to this type of activity: Competence is demonstrated by one's ability to find clues, solve puzzles, and ultimately "escape" a virtual room, whereas sociability is demonstrated by one's ability to coordinate with team members to solve puzzles that often require teamwork and communication.

2.2 Methods

All analysis scripts can be found on the study's github page. Model output for analyses, as well as the full surveys administered to participants, with all measures, can be found on the study's OSF page. All study procedures and data collection were performed in accordance and with the approval of the Columbia Institutional Review Board. The study was not preregistered.

Participants

142 participants completed the pre-game survey, across 30 groups of 3-5 friends (96 F, 44 M, 2 non-binary; mean age: 25.8; age range: 18-66, 32% under the age of 23). The breakdown of participant race is as follows: 0% American Indian/Alaska Native, 30% Asian, 0% Native Hawaiian/other Pacific Islander, 8% Black, 51% White, 5% other, 5% multiracial. Three groups were excluded from analyses involving the video recording due to technical errors saving

the video, leaving 128 participants for those analyses. 136 participants completed the post-game survey (122 of these participants had video data), and 129 completed the one-week-later survey. No other participants were excluded. Participants were recruited through online advertisements, email lists, and word of mouth and completed informed consent before starting the first survey. Recruitment typically began with one potential participant reaching out to the researchers to express interest in the study. Interested participants were informed of the study procedures and told to recruit four other people to participate in the study with them. Oftentimes, these were groups of friends, but sometimes certain people in the group were more familiar with each other than with others. A breakdown of relationship strength both within and between groups can be found in Supplemental Materials. Participants received $15 for participating and the costs of participating in the escape room game were covered.

Procedures

One week before the group's scheduled escape room game, each group member was sent a series of questionnaires on the Qualtrics survey platform. In addition to basic demographics, participants provided comprehensive evaluations about themselves and each group member, including their perceptions of their competence and sociability, as well as levels of familiarity, liking, and similarity (see "Definition of variables" section for how each variable was calculated). One week after receiving the questionnaires, the group participated in their virtual escape room game over Zoom. (See more information about the escape room experience in the following section.) All escape room games were recorded. Upon immediate completion of the escape room game, participants completed another series of questionnaires. They provided identical evaluations about each teammate, and also answered questions about the escape room experience. They also indicated how well they believed each teammate did in terms of solving puzzles and collaborating with teammates. An identical follow-up questionnaire was completed one week later to assess how durable changes in trait-ratings were, in line with other work that treats one week as evidence of long-term change. (Denny et al., 2015; Roediger & Karpicke, 2006; Tompary & Davachi, 2017). See *Figure 2.1* for a schematic of the study design.

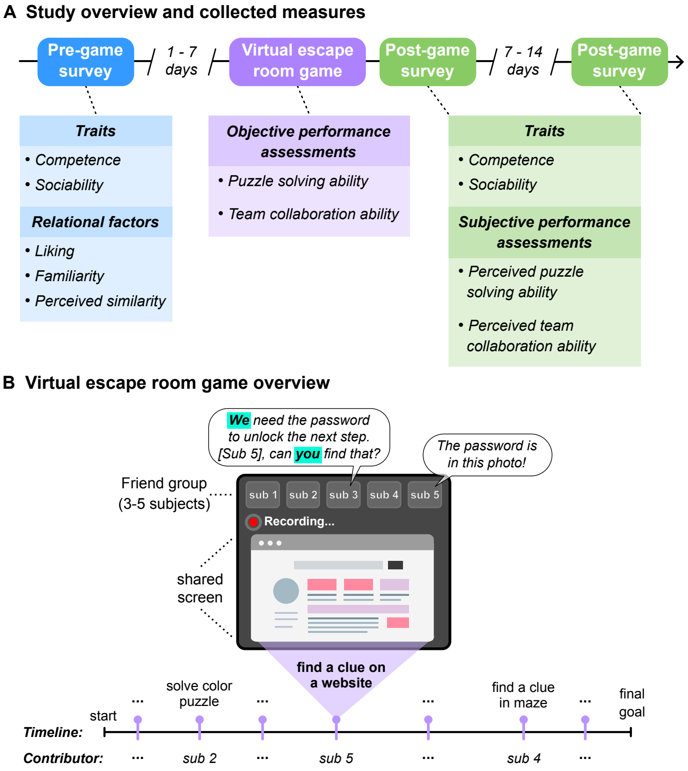


Figure 2.1. Experimental design.
A) Participants first completed a pre-game survey to assess levels of similarity, liking, and familiarity between teammates, as well as perceptions of competence and sociability. The escape room was conducted on Zoom and required participants to work together to complete a series of puzzles. Immediately after the game, participants completed a post-game survey about their game experience, as well as updated perceptions of competence and sociability. B) The Zoom recording allowed us to collect information on puzzle solving ability, and the transcript allowed us to score team collaboration.

Escape room game

The COVID-19 pandemic presented a unique opportunity to conduct our study, in which meaningful social interactions largely occurred online, where they could easily be recorded. In addition, many social activities that would typically be difficult or impossible to use as controlled experimental paradigms were translated to a more controlled, virtual space.

The virtual escape room game that we used in our study was created and administered by an escape room company called Puzzle Break LLC. The goal of the escape room, called Hackfiltration, was to solve a series of puzzles in order to hack into a company's computer system and prevent them from enacting world domination. The escape room was completed online and over Zoom. Each group completed the escape room game with the guidance of a "game manager," a Puzzle Break employee who explained the rules of the game to the participants and was available to provide hints to the group if needed. Completing the escape room required the participants to work together to find clues, solve logic puzzles, and follow a storyline across several different websites, videos, and virtual games. Typically, one group member would share their screen, and the other group members would follow along. Group members were free to speak to each other and interact as much as they wanted. Upon completion of the game, the game manager walked the group through the game solution. On average, groups took 48.9 minutes to complete the game, with completion times ranging from 26.2 minutes to 87.6 minutes. If a team was struggling to complete the game, the game manager would provide progressively helpful hints in order to move the team along and ensure that all groups completed the entire game.

Definition of variables

Our primary predictor variables were liking, familiarity, and perceived similarity for each group member, as measured before the escape room game. Liking and similarity were assessed with single questions ("How much do you like this person?" "How similar do you believe you are to this person?") on a 0-10 scale, from "Not at all" to "Extremely." In the relationship literature, familiarity is often defined in terms of both relationship longevity and interaction frequency (Berscheid et al., 1989; Fitzsimons et al., 2015; Sels et al., 2020). In the current study, we average these two dimensions together to gain a more robust measure of familiarity. Specifically, we average a question about the length of time the participant had known the target (on a 1-6 scale, from "one month" to "more than 5 years,") with a question about the frequency of interactions with them (on a 1-6 scale, from "less than several times per year" to "every day").

Trait-level competence and sociability perceptions for each group member were also measured via survey questions, both before and after the escape room game. Each dimension was the average of four component traits, indexed via descriptions of relevant behaviors, from 0 ("Strongly disagree") to 10 ("Strongly agree"). We used descriptions rather than asking about traits directly because we didn't want participants to be overtly biased against providing unfavorable ratings of their friends, especially friends with whom they were about to complete a task. Each description/behavior indexed onto a trait that previous work has shown to be related to either competence or sociability (Goodwin et al., 2014; Landy et al., 2016). For competence, the described traits were capable, effective, skillful, and talented; the questions were: 1) "[Group member] is able to succeed when faced with challenging situations." 2) "[Group member] is able to solve difficult problems." 3) "[Group member] is good at getting what they want." 4) "[Group member] is good at adapting to unfamiliar situations." For sociability, the described traits were cooperative, empathetic, humble, and kind; the questions were: 1) "[Group member] works well with other people." 2) "[Group member] is quick to understand the experiences and feelings of others." 3) "[Group member] doesn't act like they are better than other people." 4) "[Group member] is generous and considerate of others." The sociability traits were rated in previous work as being components of sociability that had some overlap with the moral dimension and were chosen for this study because they were deemed more relevant to working with others on a task than more warmth-focused sociability traits such as friendliness and extraversion. Our decision to average the four ratings came from our desire to create summary variables of competence and sociability, rather than testing effects of each specific component trait. Prior work has similarly averaged these component traits to create summary variables for competence and sociability (Goodwin et al., 2014), and our data demonstrated a high degree of reliability for both the competence trait descriptions ($\alpha = 0.88$) and the sociability trait descriptions ($\alpha = 0.90$).

We also created two different measures of performance during the escape room: puzzle solving and team collaboration. Both measures of performance were calculated using the Zoom recordings of the escape room game.

For puzzle solving performance, the escape room game was broken into 50 steps. These steps were highly specific occurrences that nearly every team needed to experience in order to complete the escape room. Each step was time-stamped, and "attributed" to a specific participant, based on which participant verbally contributed the most crucial information to complete that step, as determined by an independent coder. For example, one of the steps was figuring out the password to log into a computer system. At the exact moment that a participant indicated they knew what the password was, the time was recorded and that participant was marked as contributing to that step. If multiple participants contributed to the same step – say, by saying the answer at the same time – they split the point. Once the video was fully coded, each participant received a puzzle solving performance score, according to how many steps they contributed over the course of the entire game. We hypothesized that puzzle solving performance would be related to trait-level competence perceptions.

For team collaboration performance, we ran a linguistic analysis of the transcripts from the Zoom recordings. For each participant, we calculated a team collaboration performance score, which was the number of times they used words that focused on the group – specifically, first-person plural and second-person pronouns – over the number of total words spoken by the entire group. This approach is based on prior work using pronouns as signifiers of psychological traits and phenomena (B. P. Doré et al., 2017; Kross & Ayduk, 2011; Lyons et al., 2018; Pennebaker & Chung, 2013; Tausczik & Pennebaker, 2010), and, more specifically, studies that have used first and second person pronouns as indicators of group dynamics, such as group cohesiveness and group-focus (J. E. Driskell et al., 1999; T. Driskell et al., 2013; Gonzales et al., 2010; Kane & Van Swol, 2023; Wegner & Giuliano, 1980). Our use of pronouns to indicate team collaboration was in line with these prior uses in the group problem-solving literature. We calculated this value in Python, using words from the Linguistic Inquiry and Word Count (Pennebaker et al., 2015). First-person plural and second-person pronouns signified an attention to others and to the group as a whole. In addition, by creating a percentage that relied on both words spoken by individuals and the total number of words spoken by the group, we were able to standardize team collaboration performance scores within each group, while also taking into account how much that participant spoke. We hypothesized that team collaboration performance would be related to trait-level sociability perceptions.

Analyses

To address our three questions, we constructed a series of Bayesian multi-level models. We used Bayesian models because previous work shows that, in comparison to frequentist models, they better estimate multi-level effects models (Gelman, 2005). Each analysis consisted of two separate models: one for competence, and one for sociability. For all models, participant and team were treated as random effects, and each rating of a teammate was a repeated measure. Including team as a random effect allowed us to control for group-specific effects, which helped ensure that our results were reflective of individual perceptions. Although the participants completed the escape room in discrete groups, our questions primarily concerned evaluations and perceptions at the *interpersonal* level, rather than the group level. All models had random slopes and intercepts for the predictor variables, which were rescaled and grand mean-centered around 0.

Our first question was whether competence and sociability ratings after the game significantly differed from pre-game ratings. Specifically, our predictor variable was a Time variable consisting of pre-game ratings, post-game ratings, and one-week later ratings. Our outcome variable was the trait rating. We set up our model so that we could compare ratings at all three timepoints to each other. We also created a model with absolute change between timepoints as an outcome variable, and timepoint comparison (pre vs post or post vs one week) as the predictor variable in order to investigate rating change at the individual level.

For our second question, we asked how liking, familiarity, and perceived similarity, as measured in a pre-escape room questionnaire, interacted to affect post-game ratings of competence or sociability. Importantly, these models controlled for pre-game ratings, ensuring that any change observed from the pre- to post-game ratings was a result of the escape room game specifically. We controlled for pre-game ratings instead of using the pre-post difference as the outcome variable in order to account for potential ceiling effects in the ratings.

Our third question concerned performance during the game. We first ran a model that asked how objective puzzle solving performance (as measured from the Zoom recordings in the ways defined in the previous section) affected post-game trait-level competence perceptions. We then ran an identical model that asked how objective team collaboration performance affected post-game trait-level sociability perceptions.

We then asked whether subjective perceptions of performance were related to relevant trait perceptions. We called this subjective perception of performance a Performance Assessment Bias (PAB). PABs were calculated by subtracting the *objective* puzzle solving and team collaboration scores – as calculated in the manner described in the previous section – from *subjective* puzzle solving and team collaboration scores, as determined by participant ratings of teammates on the post-game questionnaire. Thus, we calculated one PAB for puzzle solving, and a separate PAB for team collaboration. To answer this question, we first ran models that asked how the same three relational factors predicted both PABs. We then ran a model with PAB as a predictor variable and post-game trait ratings as an outcome variable to see if biased perceptions of actions and biased perceptions of traits were related. (See *Table A.2.1* for a summary of all statistical models.)

2.3 Results

Descriptive statistics

Pre-game liking and similarity ratings were on 0-10 scales. Average pre-game liking was 8.32 (SD = 1.96) and average pre-game similarity was 5.73 (SD = 2.27). Pre-game familiarity was the product of how long the participant had known the target and how frequently they interacted, converted into 1-6 scales. Average score for time since first met was 4.36, which is in between the answers “in the past 3 years” and “in the past 5 years.” Average score for interaction frequency was 4.08, which is in between several times per month and several times per week. (A visualization of the spread of each of these scores, both within and between groups, can be found in the Supplemental Materials.) Therefore, average familiarity was 4.22 on a 6-point scale. We also calculated partial collinearity between the three relational factors, accounting for group means. The correlation coefficient values were well below values that might cause concern about collinearity: liking x similarity: 0.32; liking x familiarity: 0.09; similarity x familiarity: 0.06.

We had two independent raters code the escape room game videos to determine puzzle solving ability scores. 10% of the videos were coded by both raters. We were not able to calculate kappa values to determine agreement between the raters because our data were not binary; at every event timepoint, up to five teammates could be awarded points for contribution. Instead, we calculated how often the contributor at each event was identical between the two raters. We observed a high degree of overlap (84% of all events) between the two raters.

Question 1: Does an unfamiliar and challenging group activity lead to altered perceptions of friends' traits?

For all results, we discuss effects on competence first, followed by effects on sociability. Using Bayesian multi-level models, we examined whether post-game ratings of competence and sociability were significantly different from pre-game ratings. The average of the pre-game competence ratings was 7.58 (SD = 1.61), while for sociability ratings, it was 7.78 (SD = 1.72). When comparing ratings before the escape room game to ratings immediately after the game, we found that completing the escape room game on average led to enhanced perceptions of both competence (Post-game mean = 8.07, SD = 1.34; B = 0.48, SE = 0.09, 95% CI = [0.30, 0.67]) and sociability (Post-game mean = 8.17, SD = 1.49; B = 0.38, SE = 0.08, 95% CI = [0.22, 0.54]) (*Figure 2.2.A, 2.2.C*). In other words, given that the ratings were on a 0-10 scale, competence and sociability ratings exhibited a post-game increase of 4.8% and 3.8%, respectively. These increases were equivalent to an increase of 0.29 standard deviations for competence and 0.22 standard deviations for sociability. In addition, we found that 61% of competence ratings increased, 26% decreased, and 13% remained the same. For sociability, 53% of ratings increased, 30% decreased, and 17% remained the same. Furthermore, ratings remained on average above baseline when trait perceptions were reassessed one week later, for both competence (One week later mean = 7.79, SD = 1.39; B = 0.33, SE = 0.08, 95% CI = [0.18, 0.48]) and sociability (One week later mean = 7.94, SD = 1.58; B = 0.18, SE = 0.08, 95% CI = [0.02, 0.34]). In other words, competence ratings were 3.3% higher one week later as compared to baseline (an increase of 0.20 standard deviations), while sociability ratings were 1.8% higher (an increase of 0.10 standard deviations). A post-hoc sensitivity analysis with this model using the *pwr* package in R (Champely, 2020) revealed that our sample size provided over 80% power to detect the effect we actually observed ($f^2 = 0.16$), on the assumption that it was the true effect in the population.

Despite an average increase for both trait dimensions, there was also a high degree of heterogeneity in the amount and direction of rating change between timepoints. In order to test for lasting effects at the individual level, we calculated the absolute value of the rating change between both the pre-game ratings and the post-game ratings, and between the post-game ratings and the one-week later ratings. We found parallel results for both competence and sociability. Specifically, we found that ratings meaningfully changed (the range of our model's estimates of change did not include 0) as a result of the escape room game (Competence: B = 0.97, SE = 0.05, 95% CI = [0.87, 1.07]; Sociability: B = 0.88, SE = 0.06, 95% CI = [0.75, 1.00]). Furthermore, although there was also a rating decrease between the post-game ratings and the one-week later ratings (Competence: B = 0.64, SE = 0.05, 95% CI = [0.54, 0.74]; Sociability: B = 0.63, SE = 0.06, 95% CI = [0.52, 0.74]), this change was *smaller* than the change between the pre and post-game surveys (Competence: B = -0.32, SE = 0.06, 95% CI = [0.21, 0.43]; Sociability: B = -0.25, SE = 0.06, 95% CI = [0.13, 0.37]; *Figure 2.2.B, 2.2.D*). These results suggest that both competence and sociability ratings at the individual level were different from baseline ratings up to one week after the escape room game.

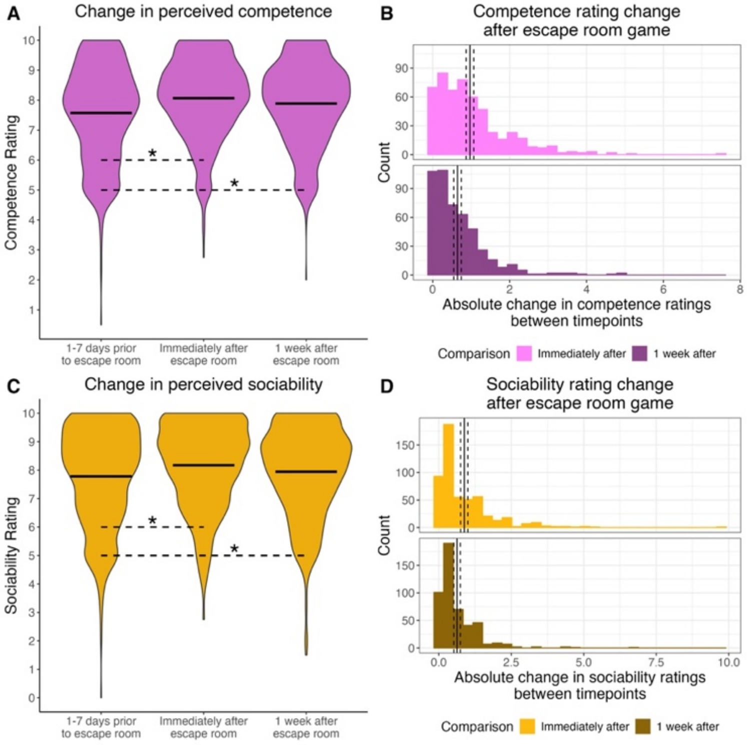


Figure 2.2. Changes in competence and sociability ratings.
A) Competence ratings on average increased after the escape room game and were still higher than baseline one week later. B) We calculated the absolute value of rating change for each individual participant-teammate dyad. These individual changes were on average larger between baseline and immediately post-game than between immediately post-game and one week later. We found similar results for sociability ratings, where average sociability ratings increased post-game and were maintained one week later (C), and individual-level absolute changes were larger post-game than one week later (D).

Question 2: How are perceptions of a friend’s traits influenced by aspects of our relationship to them (i.e., relational factors)?

Given the increases in competence and sociability, we next wanted to investigate how relational factors between the perceiver and the target prior to the game – liking, familiarity, and similarity – would affect the degree of change in competence and sociability ratings immediately after the game. All three relational factors were rescaled and mean-centered around 0 with a standard deviation of 1, and all models controlled for pre-game ratings of competence and sociability.

When examining how relational factors affected post-game competence, we found an effect of similarity (B = 0.16, SE = 0.07, 95% CI = [0.01, 0.30]), where participants rated targets that they viewed as more similar to themselves as more competent. In other words, an increase of one standard deviation of similarity led to a 1.6% increase in perceived competence. There was no effect of liking (B = 0.09, SE = 0.09, 95% CI = [-0.08, 0.27]) or familiarity (B = 0.05, SE = 0.08, 95% CI = [-0.11, 0.20]), and there were no interactions between relational factors whose estimates excluded an effect size of 0 (*Figure 2.3.A*). A post-hoc sensitivity analysis with this model with only random intercepts revealed that our sample size provided over 80% power to detect the effect of similarity that we observed ($f^2 = 0.61$), on the assumption that it was the true effect in the population.

When examining perceptions of trait-level sociability, we found that liking was the most important relational factor (B = 0.24, SE = 0.11, 95% CI = [0.03, 0.45], with more well-liked targets perceived as more globally sociable after completing the escape room game. In other words, a one-standard deviation increase in liking led to a 2.4% increase in perceived sociability. There was no effect of familiarity (B = 0.09, SE = 0.08, 95% CI = [-0.08, 0.23]), or similarity (B = -0.01, SE = 0.07, 95% CI = [-0.14, 0.14]), and no interactions with effect sizes that excluded 0 between relational factors (*Figure 2.3.B*).

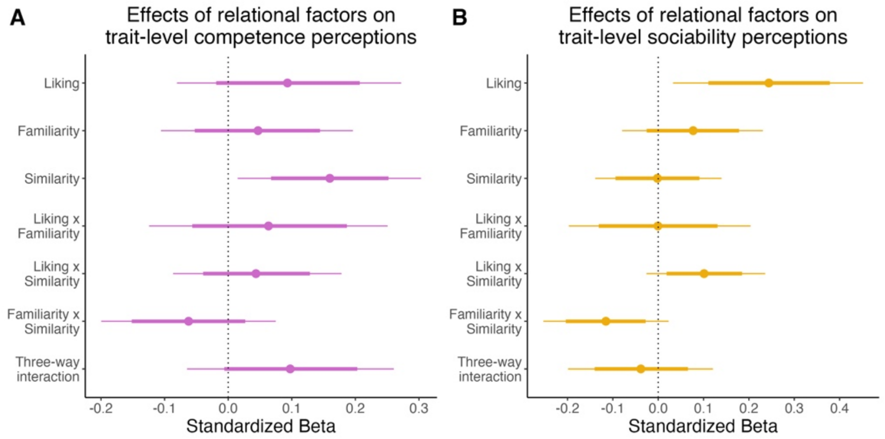


Figure 2.3. Effects of relational factors on perceptions of teammates' competence (A) and sociability (B).
X-axis is the standardized beta from a multilevel model with liking, familiarity, and similarity as predictor variables, and post-game trait ratings as the outcome variable, while controlling for pre-game trait ratings. For each variable, thick bars represent 80% credibility intervals and thin bars represent 95% credibility intervals. Competence ratings are influenced by similarity, while sociability ratings are influenced by liking.

Question 3: What is the relationship between perceptions of actions and perceptions of traits?

A friend's actions may be related to how we perceive their traits in at least two different ways. First, their *objective* actions – in this case, how they actually performed during the escape room – might have an effect on global trait perceptions. To answer this question, we ran a model with escape room game performance as a predictor variable, and post-game trait ratings as the outcome variable. Second, the *subjective perceptions* of a friend's actions might be impacted by relational factors, and in turn might predict global trait perceptions. We operationalized this bias, which we call the Performance Assessment Bias, or PAB, as the difference between the subjective rating of a friend's performance during the escape room via a questionnaire, and an objective score for their performance as determined from a Zoom recording or transcript. To answer this question, we first ran a model with our three relational factors as predictor variables, and PAB as the outcome variable. We then ran a model with PAB as the predictor variables and post-game trait ratings as the outcome variable to determine whether PAB was directly related to trait perceptions.

Effect of escape room game behavior on trait perception

When considering how perceived trait-level competence was predicted by puzzle solving performance during the escape room (specifically, the number of steps for which each participant contributed solutions), we found a main effect of puzzle solving (B = 0.14, SE = 0.03, 95% CI = [0.07, 0.21]), meaning that teammates who solved more puzzles were rated as more competent after the game. We defined team collaboration performance during the escape room game as the frequency of group-focused words (first-person plural and second-person pronouns) over total words spoken by all members of the group. When considering the effects of team collaboration performance on perceived global sociability, we found no effect of team collaboration score on trait-level sociability ratings (B = -0.05, SE = 0.04, 95% CI = [-0.12, 0.02]) (*Figure 2.4)*.

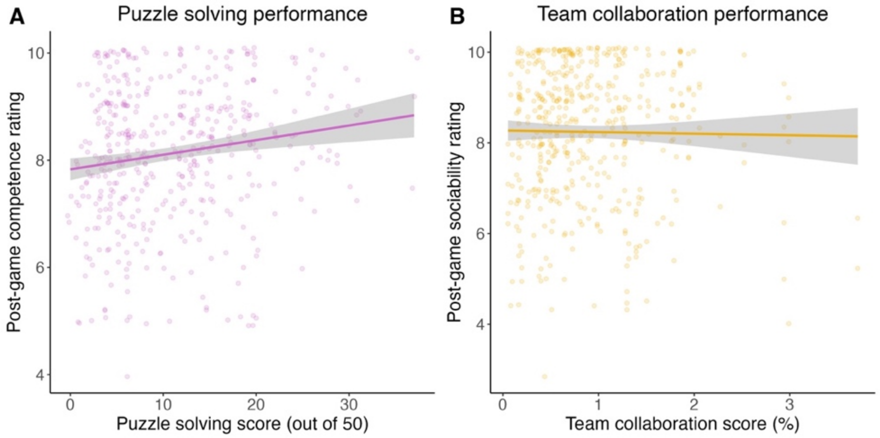


Figure 2.4. Relationship between escape room game behavior and post-game perceptions of general traits.
A) The effect of puzzle solving score, calculated from the Zoom recording as the number of contributions each participant made to solving a puzzle, on post-game competence ratings. B) The effect of team collaboration score, calculated from the transcript of the Zoom recording as the number of group-focused words each participant used relative to the total words spoken, on post-game sociability ratings. The ribbons around the regression line represent 95% credibility

intervals. Puzzle solving performance predicted competence ratings, but team collaboration performance did not predict sociability ratings.

Effect of relational factors on Performance Assessment Bias (PAB)

We next asked whether biased perceptions of the target's actions – the PAB, as defined in Methods – were similarly impacted by the same relational factors.

The participant's *objective* performance scores were rescaled to a range of 0-10 to align with the range of the *subjective* performance questions. Since the PAB was the difference between these two scores, a positive score indicated that participants reported their teammates as having performed better than they actually did, while a negative score indicated that participants reported teammates as having performed worse than they actually did. There were two PABs: One that was a measure of how well the participant did on solving puzzles, which we hypothesized would be related to trait-level competence, and one that was a measure of how well the participant collaborated with members of the team, which we hypothesized would be related to trait-level sociability.

We ran two models with the same three relational factors – liking, familiarity, and similarity – as predictor variables and each PAB as an outcome variable. We did not find that any relational factor predicted puzzle solving PAB. However, with a positive intercept of 5.07, we found that familiarity was the strongest predictor of puzzle solving PAB (B = 0.26, SE = 0.16, 95% CI = [-0.05, 0.56]), with smaller effects for liking (B = 0.20, SE = 0.17, 95% CI = [-0.14, 0.54]) and similarity (B = 0.17, SE = 0.15, 95% CI = [-0.12, 0.46]) (*Figure 2.5.A*). A post-hoc sensitivity analysis with this model with only random intercepts revealed that our sample size provided over 80% power to detect these effects (f^2 = 0.16). For team collaboration PAB, which had a positive intercept of 6.16, we found an effect of familiarity (B = 0.44, SE = 0.21, 95% CI = [0.03, 0.85]). The effects of liking (B = 0.35, SE = 0.18, 95% CI = [-0.00, 0.72]) and similarity (B = -0.20, SE = 0.16, 95% CI = [-0.51, 0.11]) were smaller (*Figure 2.5.B*).

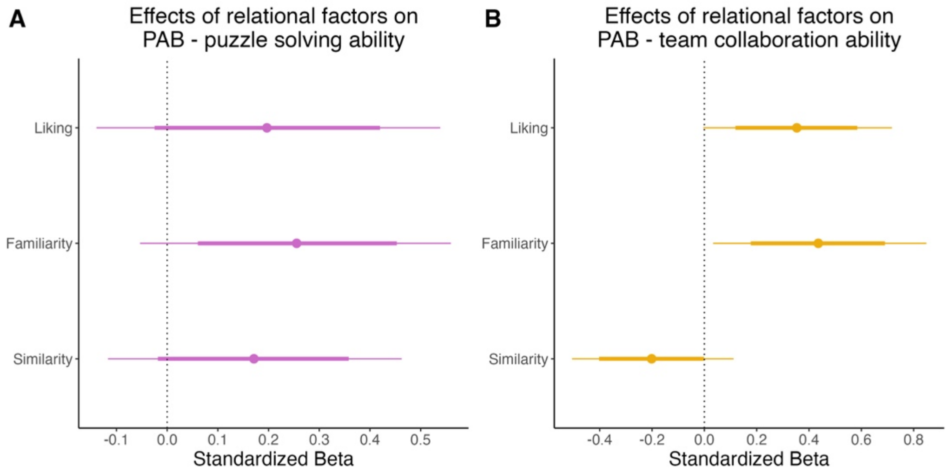


Figure 2.5. Effects of relational factors on Performance Assessment Bias (PAB) for puzzle solving (A) and team collaboration (B).
PAB is the difference between perceived performance as reported by teammate ratings and actual performance as determined by a video recording or video transcript of the escape room game. The X-axis is the standardized beta from a multilevel model with liking, familiarity, and similarity as predictor variables, and PAB as the outcome variable. For each variable, thick bars represent 80% credibility intervals and thin bars represent 95% credibility intervals. (A) For competence, task performance is defined as ability to solve puzzles. No relational factor predicted puzzle solving ability PAB. (B) For sociability, task performance is defined as ability to collaborate with one's team. Familiarity and liking are the strongest predictors of team collaboration PAB.

Effect of PAB on trait perception

We followed up these results with an additional Bayesian multi-level model that included the relevant PAB as a fixed and random effect in order to determine if PAB was related to the relevant trait perception. Since our predictor variable was a difference score, we included the sum of the difference components (subjective and objective performance) as a control variable in our model. We found that PAB was indeed related to post-game trait ratings, for both competence (B = 0.15, SE = 0.02, 95% CI = [0.11, 0.19]) and sociability (B = 0.15, SE = 0.02, 95% CI = [0.11, 0.20]).

It was also possible that the size of the bias may have been more meaningful at different performance levels. If a participant overestimated a poor-performing teammate's performance, it may have had more of an impact on post-game trait perceptions than if a participant overestimated a well-performing teammate's performance. On the flip side, overestimations of performance may have been equally meaningful across the spectrum of objective performance levels, meaning that greater overestimation always led to greater bias in perceptions of traits. To address this issue, we conducted an exploratory analysis testing whether the effect of PAB on post-game traits depended on the objective performance levels. We found that puzzle solving PABs made more of an impact on impressions for poor performers than high performers (B = -0.14, SE = 0.04, 95% CI = [-0.22, -0.05]). The interaction effect for team collaboration PABs trended in the same direction, although the range of estimates did not exclude 0 (B = -0.07, SE = 0.04, 95% CI = [-0.15, 0.01]).

We followed up these models with a model that had separate terms for the objective and subjective scores (sometimes termed a condition-based regression analysis) so that we would be able to determine if the impact of PAB was driven more strongly by one term in the difference score (Humberg et al., 2018). According to condition-based analysis, the difference score is the "true" predictor, rather than being driven by one component, if the effects for the subjective and objective scores are in opposite directions. For trait-level ratings of competence, we found a positive effect of subjective performance ratings (B = 0.29, SE = 0.03, 95% CI = [0.23, 0.34]) and no effect of objective performance score (B = -0.02, SE = 0.02, 95% CI = [-0.05, 0.02]). A follow-up analysis using the lavaan package in R (Rosseel, 2012) revealed that the effect of puzzle solving PAB on competence ratings was not significant (B = 0.07, SE = 0.05, p = 0.09). For trait-level ratings of sociability, we found a positive effect of subjective performance (B =

0.27, SE = 0.04, 95% CI = [0.20, 0.34]) and a trending negative effect of objective performance (B = -0.05, SE = 0.03, 95% CI = [-0.9, -0.01]). A follow-up analysis using the lavaan package revealed that the effect of team collaboration PAB on sociability ratings was significant (B = 0.13, SE = 0.07, p = 0.03). In sum, the condition-based regression analysis revealed that PAB predicted trait-level sociability ratings (over and above each individual component of PAB, which was a difference score), but that PAB did not predict trait-level competence ratings.

2.4 Discussion

In this study we investigated how one's prior relationship to a target – in terms of liking, familiarity, and perceived similarity – influenced perceptions of friends' traits and actions as they worked together to overcome a shared problem. Previous person perception research has not been able to adequately address this question because it has traditionally focused on novel targets who are not motivationally significant to the participant, while close relationships work that does focus on motivationally relevant friends/close others has not typically asked how trait perceptions change over time. Here we bridged these research traditions using a novel method: participants completed a virtual escape room game conducted on Zoom that allowed them to freely interact and solve puzzles for approximately one hour in a novel environment. Three key findings were obtained.

First, we found that ratings of friends' trait-level competence and sociability increased as a result of an unfamiliar experience – in this case, the escape room game – and that these changes persisted to some degree up to one week after the game. Second, we found that pre-existing relationships *selectively*, rather than globally, impacted trait perceptions. For competence, we found that higher baseline similarity led to higher post-game competence ratings, and for sociability, we found that higher baseline liking led to higher post-game sociability ratings. Third, we found that a teammate's performance during the game, as well as how one perceives that performance, also predicts post-game trait perceptions, but that relational factors do not consistently bias performance perceptions. Below, we unpack each of these findings.

Changes in perceptions of trait competence and sociability

First, the escape room game elicited updated trait ratings that – to some degree – lasted for at least a week. Belief updating work demonstrates that we typically only update our beliefs when faced with information that disconfirms those beliefs (M. Kim et al., 2020; Kube & Rozenkrantz, 2021). In daily life, we don't often witness our friends and close others in unfamiliar situations where they will act in ways that we are unable to predict. A virtual escape room presented an opportunity to study belief updating because it presented many unfamiliar elements: Most of our participants likely had not previously participated in a virtual escape room before with this specific group of friends (and even if they had, it likely wasn't a particularly common occurrence). In addition, we also thought that an escape room was an opportunity for an impression update because people's actions during the game would be perceived as *meaningful*: The time pressure might have been revealing of one's "true" traits, as time pressure is often used to reveal implicit attitudes (Karpinski & Hilton, 2001; Stepanikova, 2012), and the competitive nature might have caused participants to become more invested, as they do in minimal groups paradigms (Dunham, 2018; Otten, 2016). These aspects of the game could have created a high-stakes environment, in which people were perceived as behaving authentically and not hypothetically. Thus, the escape room offered a combination of controlled and dynamic elements that – as our data show – was able to elicit impression updates.

Second, we saw that both competence and sociability ratings increased overall as a result of the escape room game. Previous work has shown a positivity bias for competence (Reeder et al., 1977; Wojciszke et al., 1993), in that it's easier to update competence beliefs in the positive direction than the negative direction. Our study is to our knowledge the first to demonstrate a parallel effect in the sociability dimension, since sociability is often lumped in with morality in studies of impression updating. It has been hypothesized that the reason for this asymmetry is that competent actions are more "diagnostic" than incompetent ones (Mende-Siedlecki, Baron, et al., 2013).

Finally, the increases in competence and sociability ratings were maintained to some degree one week later, as evidenced by the smaller rating changes between the post-game ratings and the one-week later ratings than between the pre-game ratings and the post-game ratings. Work investigating the maintenance of impression updates is sparse and mixed, and limited by the fact that many impression updating studies examine short timescales (i.e., less than one day) and use strangers, whose behavior is hypothetical, as targets (Murphy, 2017). Some have argued that an action needs to be both diagnostic and believable to elicit an impression update that lasts beyond an immediate effect (Ferguson et al., 2019). Our participants interacted with each other in a real (not hypothetical) and high-intensity environment, which may be why we see increases in competence and sociability persist for at least one week.

For at least two reasons it is meaningful that change was maintained to some degree one week later. First, the persistence beyond immediate effects seems to preclude typical explanations for temporary fluctuations in impressions related to variability in contexts (Geukes et al., 2017) or mood inductions (Forgas & Bower, 1987). Second, many studies of impression updating fail to show any updates at all, and when they do, the update is often assessed only in the immediate-term (Cone & Ferguson, 2015; Gregg et al., 2006; Kurdi et al., 2023). The fact that we saw maintenance of change at all beyond the immediate-term is noteworthy, especially because the changes we observed were a result of a one-hour interaction, amidst relationships that on average spanned 3-5 years. This demonstrated that the escape room game was perceived as meaningful enough by our participants that it continued to play a role in trait perceptions at least one week beyond immediate assessments.

That said, it is worth acknowledging that in the current study, we did not assess impressions farther out than one week after the game, so it is possible that impressions may revert to baseline after this time point. In addition, we did not track how often escape room teammates interacted between completing the game and the one week later-assessment, so it is unclear whether additional meaningful interactions between the two timepoints might have influenced the presence or absence of effects observed at the one-week follow-up.

How relationships impact trait perceptions

For both competence and sociability, specific components of one's prior relationship with a target impacted perceptions of that target's traits after working together to solve an unfamiliar challenge. What the competence and sociability findings share is an emphasis on prior relationships, which motivate our perceptions. Even for those we know well, relative degrees of motivation continue to impact our assessments of others. This is in line with previous work that goes above and beyond investigating changes in impressions for strangers, and instead uses in-group members and close friends (Hughes et al., 2017; B. Park & Young, 2020). However, even this work is still comparing friends to strangers, or an ingroup to an outgroup. Our study demonstrated that within a close group, our beliefs are biased about some friends more than others.

Where the competence and sociability findings differed, however, was in the type of relational factor that matters. When controlling for baseline (i.e., pre-game) ratings of competence and sociability, we found that higher perceived similarity led to higher post-game competence ratings, and higher liking led to higher post-game sociability ratings.

This similarity-competence connection might have been related to the self-enhancement effect. Long documented in psychology (Alicke & Sedikides, 2009; Sedikides & Gregg, 2007, 2008), self-enhancement theories posit that we have a motivation to view ourselves positively or favorably. Given that competence is a desired quality for oneself (Anderson et al., 2012; Heck & Krueger, 2016), it makes sense that we would be motivated to perceive people who are similar to us as more competent as well, even if we might still enhance perceptions of our own competence more than we enhance perceptions of the competence of similar others (Morry, 2007; Morry et al., 2010). The liking-sociability connection, on the other hand, may have been because sociability signifies a person's ability to maintain a successful relationship. If we like someone, we'll want to maintain a successful relationship with them, which would motivate us to view them as more sociable. Sociability is also a trait more desired in others than the self (similar to morality (Wojciszke, 2005)), although there is limited work in this area (Soral & Kofta, 2020). Our results demonstrated a separability between the desired qualities in oneself and the desired qualities of others (Brambilla & Leach, 2014; Wojciszke, 2005), and extended previous work into the sociability domain.

How actions and action perceptions relate to trait perceptions

In addition to being affected by relationships, our perceptions of others' traits are likely also based on their actions, as well as perceptions of those actions. We found mixed evidence for the role of actions on trait perceptions and the role of relationships on action perceptions, both within the competence and sociability dimensions, and between them. For competence, objective puzzle-solving performance and subjective ratings of puzzle-solving performance both independently predicted post-game competence ratings, but a condition-based regression analysis revealed that the difference between objective and subjective scores – what we term the performance assessment bias (PAB) – did not. These results imply that estimates of competence were impacted by instantiations of that competence in a particular situation, and that estimates of situation-specific competence were tied to actual, objective demonstrations of it.

Conversely, we found that sociability ratings were predicted by subjective ratings of team-collaboration performance and by team collaboration PAB, but not by objective team collaboration scores, suggesting that biased perceptions of sociability are more swayed by biased perceptions of sociability-specific actions than by actual, objective demonstrations of those actions. In other words, how sociable someone actually was in a specific situation was less important for perceptions of their trait than how sociable *we perceived them to be* in that situation.

These results might be partially explained by state-trait models of person perception. State-trait models distinguish between qualities a person is deemed to possess generally and qualities they display in a particular situation (Hamaker et al., 2007; Trope, 1998). Judgments of each can converge or diverge depending on the context and the type of judgment one is being asked to make (Gilbert et al., 1988; Kruse & Degner, 2021). In the current study, participants who demonstrated state-like competence were more likely to be rated as possessing trait-like competence, but such a relationship did not exist between state- and trait-sociability.

Why the discrepancy between the competence and sociability findings? Or more specifically, why did perceptions of competence performance appear to be more strongly tied to reality than they were for sociability performance, and why did perceptions matter more than reality for sociability than for competence? Prior research has focused less on how sociability is updated, but we can put forth several potential reasons for this discrepancy. First, prior work suggests that demonstrations of competence are more accurately perceived than other dimensions of person perception, such as morality (Abele et al., 2021; V. Yzerbyt, 2018). In addition, we know that there is a bias towards updating competence impressions in the positive direction (Mende-Siedlecki, Baron, et al., 2013), but it is not clear if this same bias exists for sociability. As such, it is possible that objective demonstrations of competence led to larger revisions of trait-level competence perceptions than did occur for sociability. Finally, it’s also possible that puzzle solving was simply perceived as either more important or more variable during the escape room game than was team collaboration. If so, then we might have expected people would tether their trait-level competence perceptions to reality more than they would for trait-level sociability perceptions.

How relationships impact action perceptions

Our study demonstrated that relationships and actions both motivate trait perceptions, albeit in different ways. However, the question remains: What motivates perceptions of actions? In our study, we saw that performance was often overestimated, so we next must ask why, and what causes the amount of overestimation to vary. We believe there are several potential interpretations of our findings that can help answer this question.

First, most of our estimates for the impact of relational factors on PABs included 0; the one that didn’t (familiarity for team collaboration PAB) was a small effect. This would suggest that perceptions of actions were largely biased by different factors than the relational factors that biased traits. These might have been other relational factors (such as social closeness or trust), or external factors, such as one’s mood or overall group cohesion.

Second, for both puzzle solving PAB and team collaboration PAB, familiarity was the strongest predictor. This would suggest that while trait perceptions are motivated by specific relational factors that differ by dimension, action perceptions are largely motivated by familiarity, regardless of the action type. This might be because it’s easier to remember the contributions of teammates who are more familiar (Koriat & Levy-Sadot, 2001; Poppenk et al., 2010), or because we are more likely to make intentional attributions to teammates who are more familiar (Idson & Mischel, 2001; Malle et al., 2007; Malle & Pearce, 2001).

Finally, many of the effect sizes were relatively close in size, so we may also wish to consider overlap between the mechanisms that motivate trait updating and the mechanisms that motivate action perceptions. Under this explanation, relational factors may impact both trait perceptions and perceptions of performance, as opposed to simply altering trait-based assessments. This distinction is important because it sheds light on the mechanisms by which motivations may indirectly shape perceptions of others’ traits (Zaki, 2013). In addition, while our study wasn’t set up to conduct formal mediation analyses, partially due to the cross-sectional nature of one component of the PAB calculation and the post-game trait ratings (Maxwell & Cole, 2007), it’s possible that subjective perceptions of actions actually mediated the relationship between relational factors and trait updating. (For example, liking may have biased perceptions of collaborative behavior during the game, which would then in turn have led to altered perceptions of trait-level sociability.) Future studies should directly test whether biased perceptions of actions and biased perceptions of traits are independently affected by relational factors, or whether action perceptions mediate the impact of relational factors on trait ratings.

Limitations and future directions

The majority of the analyses in this paper concern the beliefs one person (the observer) holds about another person (the target). As dyadic interactions unfold across time, however, individuals may alternate between the target and observer roles. Future studies may wish to take this into account and ask how an observer's perceptions of a target impact the target's perceptions of the observer, and vice versa (Back & Kenny, 2010; Human et al., 2020). More broadly, the actions of any individual may be embedded within the actions of a larger group. In a complex and collaborative problem-solving environment (like an escape room game), we may wish to ask how group dynamics, such as the structure of the social network and the nature of social interactions between group members, impact group performance and group well-being. Organizational psychology has long investigated the factors that create successful groups in a workplace context (Cannon-Bowers & Bowers, 2011; Hesse et al., 2015; Mathieu et al., 2019; Neubert et al., 2015), but there has been comparatively little work in social psychology that seeks to understand how group dynamics, relationships between people, and/or feelings towards others affect a group's ability to accomplish a goal.

It's also worth noting that a virtual escape room is an uncommon environment to interact with friends, and it's possible that some of our effects are specific to this distinctive environment. Future studies may wish to utilize other types of events beyond a virtual escape room, and test other dimensions of person perception, such as morality, in order to determine how well our findings generalize across contexts. In addition, it's important to keep in mind that this study was conducted in 2021; thus, all effects should be interpreted within the context of the COVID-19 pandemic. The pandemic is partially what made this research project possible, as escape room companies deployed complex games to be completed over Zoom during this time.

However, baseline social activity was much lower than typical, which may have skewed participant responses. A virtual escape room may have been perceived as a more meaningful event than it would be outside of the pandemic, and participants may have been biased towards perceiving others as sociable after the game, given that other social activities were so much less frequent. It will be important to replicate these results outside of the context of the pandemic, when people were starved for social interaction.

Finally, we made several decisions about how to define certain variables in our study, and future work should seek to compare the impacts of these decisions on outcomes. For example, while previous work has assessed group processes by measuring pronoun use (J. E. Driskell et al., 1999; T. Driskell et al., 2013; Gonzales et al., 2010), additional features of group transcripts can also be used to predict group cohesion, such as verb tense (T. Driskell et al., 2013), language style matching (Kane & Van Swol, 2023), and bottom-up machine learning approaches (Stewart et al., 2019). In addition, future work should continue to explore the relationship between competence, sociability, and morality, and how susceptible each one is to an impression update. Our study relied on a narrow definition of sociability to ensure relevance to a group problem solving task. Future studies may wish to test the likelihood of an impression update for purer sociability traits, as has been done for morality (Mende-Siedlecki, Baron, et al., 2013). Relatedly, although we hypothesized that overall competence would be expressed via puzzle solving performance and that overall sociability would be expressed via team collaboration performance, it's possible that the component traits of our summary competence and sociability dimensions mapped onto our performance measurements to varying degrees. Future work might seek to examine these within-dimension variations, and their relations to specific actions.

Finally, we define durable or meaningful change as change that persists for one week. While one week has been used in other domains, such as memory (Meltzoff, 1988; Roediger & Karpicke, 2006; Tompary & Davachi, 2017) and emotion regulation strategies (Denny et al., 2015), to indicate long-term change, it's possible that evidence for durability would be strengthened by evaluating impressions after more than one week has passed. Future studies that wish to focus on durability of impression updates should evaluate impressions periodically, for at least six months after the update occurs.

When people work together to achieve a common goal, they draw conclusions about each other's traits based on how each person performed during their shared experience. In the present study, we showed that perceptions of a target's traits are impacted by one's prior relationship with that target *and* one's in-the-moment actions that demonstrate that trait. When groups of friends completed a virtual escape room together, prior perceived similarity and one's ability to solve puzzles both impacted perceptions of trait-level competence, while prior liking and one's *perceived* ability to collaborate with others both impacted perceptions of trait-level sociability. How we are perceived by another person is consistently impacted by our relationship to them, but the role of subjective vs objective actions is dependent on the trait being perceived. In some cases, we're not only biased in our perceptions of others' traits, but also in our subjective perceptions of their actions. It's these biased perceptions of individual actions that add up to global trait assessments inextricably tied to our relationships.

Chapter 3: The mentalizing network updates neural representations of romantic interest in response to social feedback

3.1 Introduction

Finding a romantic partner is essential for our survival as a species and is linked to increases in well-being (Proulx et al., 2007). It is also one of the most complex social tasks that we engage in. At a basic level, you're trying to learn as much as you can about this person: What are they like? Do you like them? Do you approve of their values and life goals? Most social neuroscience research has focused on these types of questions. However, you are not merely concerned with your own romantic interest and social evaluations of this other person, but you are also trying to figure out how they feel about you. Are they romantically interested in you? Do their feelings towards you match your feelings towards them? This information is gleaned from social feedback that your date provides to you, and will often lead you to update your own romantic interest. Thus, in addition to inferences about others' traits, your evaluations of romantic interest likely also require you to mentalize about another person's feelings about yourself. In the current study, our goal was to unpack the psychological and neural mechanisms involved in the formation and updating of romantic interest. To accomplish this goal, we investigated whether brain regions involved in mentalizing responded to social feedback, both in terms of *how* we think about another person as well as *how often* we think about them.

Although prior work has not addressed these specific questions directly, we can draw on two decades of social neuroscience research on person perception to develop hypotheses. A consistent network of brain regions, often referred to as the mentalizing network (Atique et al., 2011; Baetens et al., 2014; Sahi & Eisenberger, 2021), is often implicated in social cognitive processes that involve perceiving other people and making inferences about their mental states and traits. Specifically, the mentalizing network is made up of regions that include the dorsomedial prefrontal cortex (dmPFC), which shows greater activity and more distinct multivoxel representations when accessing information that pertains to other people (Denny et al., 2012; Lieberman et al., 2019; Wagner et al., 2019) and also prioritizes consolidation of social information during rest (Jimenez & Meyer, 2024; Meyer et al., 2019); the temporoparietal junction (TPJ), which has been linked to thinking about the thoughts and beliefs of other people (Saxe & Kanwisher, 2004; Van Overwalle & Baetens, 2009); as well as the ventromedial prefrontal cortex (vmPFC), which is linked to assessing value and self-perception (Hiser & Koenigs, 2018; Roy et al., 2012); the precuneus, which is linked to a wide range of social processes such as self-referential processing and attribution (Cabanis et al., 2013; Cavanna & Trimble, 2006); and the temporal pole, which helps link faces, identities, and emotional responses (Deen et al., 2024; Olson et al., 2007).

We had three primary research questions that we aimed to answer in this study, First, we wanted to extend prior work linking these regions to social evaluations, and ask whether or not they are involved in forming evaluations of romantic interest. Given that one's romantic interest towards another person is closely tied to that person's romantic interest in oneself, we hypothesized that this process would be linked to both the person perception regions and the mentalizing regions that make up the mentalizing network in the brain. Specifically, we expected that within the mentalizing network, the dmPFC and the TPJ would both play important roles in evaluating romantic interest. The dmPFC is typically linked to person knowledge, which includes a broad array of social judgments based on abstract information. For example, the dmPFC is activated when forming impressions based on written or verbal information that convey a person's traits (Ferrari et al., 2016; Ma et al., 2014; Mitchell, Banaji, et al., 2005; D.

Schiller et al., 2009). Other types of social judgments, such as selecting collaborators in a business setting or social network position, are also tracked via multivoxel patterns in the dmPFC, the TPJ, and the precuneus (S. A. Park et al., 2021; Parkinson et al., 2017).

While our first research question focused on evaluating romantic interest more generally, our second and third research questions pertained to how neural representations of others in the mentalizing network are *updated* in response to social feedback. Specifically, we asked whether social feedback changes *how* we think about someone, as well as *how often* we think about them. These questions are in contrast to past social neuroscience work, which typically studies these kinds of updates by focusing on the moment that an update occurs (Mende-Siedlecki, 2018). Here, the dmPFC and the TPJ play a particularly important role, with previous work showing increased activation in these regions in response to information about another person that was incongruent with one's initial beliefs (D. L. Ames & Fiske, 2013; Cloutier et al., 2011; Mende-Siedlecki, Cai, et al., 2013). In addition, both of these regions have also been linked to between-subject differences in motivated impression updating, where updating occurs to different degrees based on prior beliefs and relationships (M. J. Kim et al., 2021; B. Park et al., 2021; B. Park & Young, 2020).

Our second research question asked how social feedback changes *how* we think about another person. To answer this question, we had participants watch two 90-second dating profile videos for a series of potential romantic partners while in an fMRI scanner, and indicate their romantic interest in the video targets. In between the two videos, participants received social feedback from the targets, which systematically varied across targets in terms of valence and congruence with the participant's initial romantic interest. Past work shows that patterns of neural activity can reliably distinguish between the identities or mental states of others (Freeman

& Stolier, 2014; Hassabis et al., 2014; Thornton & Mitchell, 2017; Visconti di Oleggio Castello et al., 2017). In addition, our representation of a specific other is not simply based on their identity, but on our feelings towards, and our relationship with, that person. Indeed, political differences can alter the way that participants represent political and emotional stimuli in the dmPFC, TPJ, and precuneus (Jacoby et al., 2024; Leong et al., 2020; van Baar et al., 2021). We hypothesized that multivoxel neural representations in the mentalizing network would respond to unexpected information, meaning they would change more in response to incongruent feedback than congruent feedback.

Our third research question asked how social feedback changes *how often* we think about another person. To answer this question, we had participants undergo three resting state scans: once before viewing any target videos, once after viewing the first set of videos but before receiving feedback, and once after receiving feedback. We then calculated how often individual targets were reactivated in the mentalizing network during each resting state. Past work shows that patterns of activity for specific stimuli can be reliably detected during post-encoding rest, in both the hippocampus and prefrontal cortex (Schuck & Niv, 2019; Staresina et al., 2012, 2013). In addition, reactivation frequency can differ based on our priorities or our responses to the information we are encoding (Gruber et al., 2016; Jimenez & Meyer, 2024; Schapiro et al., 2018; Yu et al., 2024). Thus, we hypothesized that reactivation frequency in the mentalizing network would be dictated by what information participants found most motivating; specifically, that it would increase in response to social feedback, and would increase more in response to positive feedback.

In summary, in the current study we investigated the role of the mentalizing network in the formation and updating of romantic interest for specific other people. We asked a) whether the mentalizing network tracks romantic interest ratings, b) whether and how patterns of activity in the mentalizing network for specific other people change in response to social feedback, and c) whether and how reactivation frequencies of specific other people in the mentalizing network change in response to social feedback. To investigate these questions, participants completed an fMRI scan while watching multiple dating profile videos for a series of targets, both before and after receiving social feedback from the targets. In addition, participants completed resting state scans after watching each set of dating profile videos.

3.2 Methods

Participants

All study procedures were approved by the Columbia University Institutional Review Board. Participants were recruited via the RecruitMe website associated with the Columbia University Irving Medical Center. Participants completed a baseline survey to determine eligibility. Participants were required to be between 18 and 29 years old to ensure that a large age difference between the participant and a potential romantic partner was not a factor in the participant's romantic interest. In addition, we required participants to be currently using dating apps to ensure a) that they were actively interested in finding a romantic partner, and b) to ensure they would feel comfortable indicating romantic interest in someone they'd never met. Finally, we excluded participants who were not eligible to complete an fMRI scan.

Our final sample consisted of 30 total participants (age M = 23.50, SD = 2.67). 13 participants were men (of these, 5 requested to view dating profile videos of men and 8 requested to view dating profile videos of women) and 17 were women (14 requested to view videos of men and 3 requested to view videos of women). The breakdown of participants' race was as follows: 9 Asian, 2 Black/African American, 2 White-Hispanic or Latino, 14 White-Not Hispanic or Latino, 3 Other.

Stimuli

During the fMRI scan, participants watched 90-second videos of potential romantic partners. These videos were developed for this study with actors hired from the website Backstage. All videos, as well as meta-information about each video, can be found on the study's OSF page.

We hired 19 actors (9 men and 10 women) to make two videos each, for a total of 38 videos in the full stimulus set. Each actor was sent two lists of three prompts that are commonly found on dating profiles, such as: What do you like to do in your free time? What are your goals for the future? (A full list of prompts for each video can be found on the OSF page.) The actor was instructed to film the videos in a quiet room and to speak directly to the camera. They were told to speak about each prompt for roughly 30 seconds, and that each video should be in total between 80 and 100 seconds. They were told to speak truthfully about themselves, but to not discuss where they currently lived or their political views to ensure that those were not factors in participants' romantic interest.

After all videos were created, we recruited 50 participants via Prolific to provide ratings on the videos to ensure that the videos were perceived similarly on relevant dimensions, including physical attractiveness, perceived age, and perceived sexuality. We did not want any targets who were outliers on physical attractiveness, to ensure that that target's videos were not treated categorically differently across our participants. In addition, we wanted to ensure that the targets appeared to be in the same age range as our participants. Finally, given that in our main study, we had some participants who viewed same-gender videos and some participants who viewed other-gender videos, we wanted to ensure that the targets presented as potentially being interested in the gender of the participant, regardless of whether the participant's gender was the same as or different from the gender of the target. We excluded one male target and two female targets based on the results of the pre-test, for a total of eight male targets and eight female targets (32 videos total) used in our study.

Behavioral procedures

Before coming in for an fMRI scan, participants completed a pre-scan survey that included both demographic information as well as questionnaires designed to assess their motivations to find a partner. Specifically, participants completed the BIS/BAS (Carver & White, 1994) to assess their general response to receiving positive feedback, the Rosenberg self-esteem scale (Rosenberg, 1965) and the Rejection Sensitivity scale (Mendoza-Denton et al., 2002) to assess how they might feel after receiving negative feedback, and the UCLA Loneliness scale (Russell et al., 1978) to assess to assess their desire to find a romantic partner.

Upon arriving at the neuroimaging center for their fMRI scan, participants were given a cover story for the study. Specifically, they were told that they were going to have 10 minutes to make a 90 second dating profile video, in which they responded to three prompts, for approximately 30 seconds each; the video that they created was intended to be similar in format to the target videos that they were going to view during the scan. Participants were told that their video would be uploaded to a study database, and that upon uploading, a large network of "at-home" participants who had previously made similar videos would be pinged to view the actual participant's video and indicate their perceived romantic interest in the actual participant. The actual participant was told that they would be "matched" with the first eight at-home participants to respond, and that during the study, they would get to watch and rate their romantic interest in these at-home participants. In addition, they were told that they would get to see the at-home participants' ratings for their video, but the at-home participants would not see the actual participant's ratings of the at-home participants. Finally, participants were told that after the scan, they would get an opportunity to virtually chat with the at-home participants that they were most romantically interested in.

In reality, the "at-home" participants in the videos were merely actors and were not active participants in the study. The actual participant's video was not actually uploaded to a database, and no other participants in the study viewed their video. The romantic interest ratings that the participant saw during the scan (more on these can be found below) were pseudorandom. Finally, after the scan, participants were debriefed on the cover story (including that they would not get to chat with an at-home participant) and were asked to confirm that they understood.

Scanner procedures

Participants began the neuroimaging session with a 6.5-minute resting state scan, in which participants were told to keep their eyes open while looking at a white crosshair on a gray screen. Then, participants completed a photo-viewing scan, in which they viewed a photo of each of the at-home participants (hereafter called targets) whose videos they would be viewing during the scan. Each photo was presented for three seconds and was separated by a jittered inter-trial interval, between 1.5 and 5.5 seconds. A photo of each of the eight targets was presented four times, for a total of 32 photo presentations. The order of the photo presentations was randomized. Analyses of brain data during the photos scan are not included in this paper.

After the photo-viewing scan, participants watched the first set of dating profile videos, hereafter called the pre-feedback videos. In a single run, participants watched one video of each of the eight targets, in a random order (*Figure 3.1.A*). After each video finished playing, there was a 2-second ITI, followed by five slider questions, presented one at a time. The five questions were as follows: “How romantically compatible do you believe you are with this person?” “How similar do you believe you are to this person?” “How physically attractive do you find this person?” “How interested were you in what this person was talking about?” “How much do you like this person?” Questions were answered on a 1-9 slider. All five questions were highly correlated with each other and so were reduced to a single factor using an exploratory factor analysis. We term this factor romantic interest and use it in our behavioral analyses. Participants had eight seconds to respond to each question; if the participant had not submitted an answer by the end of the eight seconds, their mouse position along the slider was recorded. Following the slider questions, there was a 5-second ITI, and then participants viewed the next video.

Following the pre-feedback videos, participants completed another 6.5 resting state scan and another photos scan. Then, participants completed a cued-recall task while in the fMRI scanner. Participants were cued with the name and photo of a target and were told to say out loud everything they could remember from that target’s video. After participants finished speaking, they moved on to the next target. The results from the cued-recall task are not discussed in this paper.

After the cued-recall task, participants viewed a second set of dating profile videos for the same targets, hereafter called the post-feedback videos. (As a reminder, each target made two videos. The order of each target’s video presentation – whether it was in the pre-feedback videos or post-feedback videos – was counterbalanced between participants.) The post-feedback video-viewing was nearly identical to the pre-feedback video-viewing, in that participants viewed videos of each of the targets and provided the same five ratings. However, each video presentation was immediately preceded by a feedback presentation. Participants were shown a photo of the target whose video they were about to view, along with one of two feedback messages: "[The target] said that [she/he] thinks she/he is romantically compatible with you" (positive feedback) or "[The target] said they [she/he] does not think that [she/he] is romantically compatible with you" (negative feedback). The feedback was pseudorandom, in that the task was designed so that approximately half of the feedback from all targets were congruent with participants' initial compatibility ratings, and half were incongruent. (The participants' initial compatibility ratings were binarized into positive ratings (5 or above) or negative ratings (below 5).) Thus, there were four feedback conditions: Positive-congruent, positive-incongruent, negative-congruent, and negative-incongruent. The feedback was presented for 5 seconds. Following a 2-second ITI, participants viewed that target's post-feedback video.

After viewing the post-feedback videos, participants completed a third resting state scan, a third photos scan, and a second cued-recall task. The second cued-recall task was nearly identical to the first, except that participants were told to speak about the target's second video (the more recent one). Following the post-feedback cued-recall, participants left the scanner and completed their cover story debriefing in a testing room.

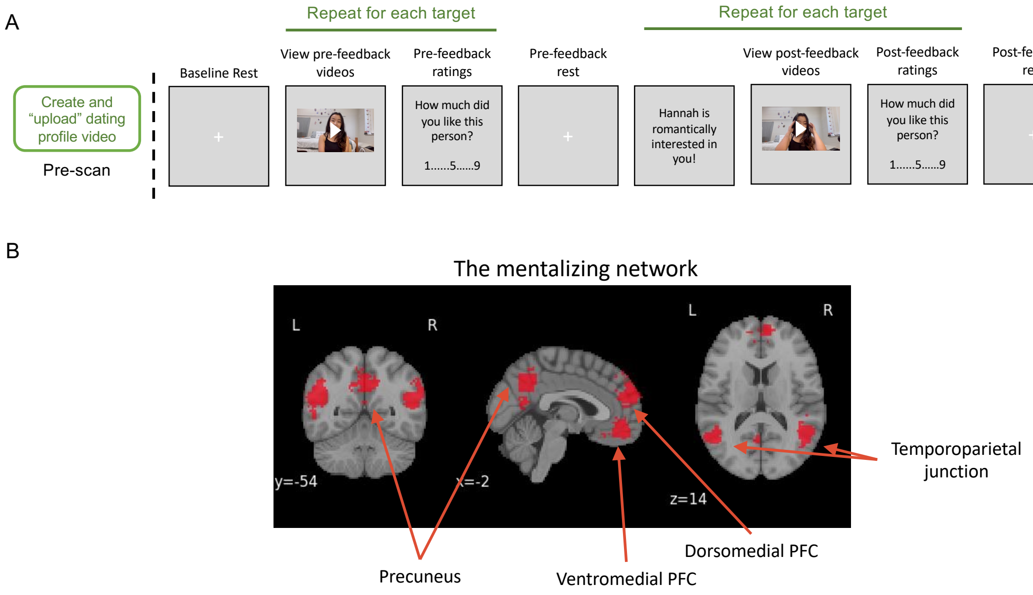


Figure 3.1. Paradigm and ROIs.
A) Participant procedures before and during the fMRI scan. Participants watched videos of eight different targets, both before and after receiving social feedback from the target. B) The brain regions that make up the mentalizing network: Bilateral temporoparietal junction, bilateral temporal pole, precuneus, dorsomedial prefrontal cortex, and ventromedial prefrontal cortex. All analyses were first conducted on the mentalizing network as a whole. Follow up analyses were conducted separately in each component region.

fMRI scanning

Imaging data were acquired on a 3T Siemens scanner with a Siemens head coil at the Zuckerman Mind Brain and Behavior Institute at Columbia University. Anatomical images were collected using a T1-weighted protocol (1 mm^3 voxels, 176 slices per slab with 1 mm thickness, TR = 2300 ms, TE = 2.98 ms). Functional images were collected with a T2*-weighted gradient-echo EPI sequence (TR = 1000 ms; TE = 30; flip angle = 62°; 2.5 mm^3 voxels), with 48 slices oriented parallel to the anterior commissure - posterior commissure (AC–PC) line. We also collected in-plane field map scans to improve co-registration between anatomical and functional images.

fMRI preprocessing

Results included in this manuscript come from preprocessing performed using *fMRIPrep* 20.2.6 (Esteban et al., 2019). See *Supplemental Methods* for a full description of fMRIPrep preprocessing steps.

Our primary ROI was the entire mentalizing network, as defined by the mentalizing network map on Neurosynth (Yarkoni et al., 2011) and resampled to our data using AFNI’s 3dresample function (Cox, 1996). The mentalizing map contained 7 clusters, which served as additional ROIs for follow-up analyses: bilateral temporoparietal junction, bilateral temporal pole, dmPFC, vmPFC, and the precuneus. Each cluster was defined using the label() function from the SciPy package (Virtanen et al., 2020). Our hippocampal ROI was created using the Freesurfer output from fMRIPrep and the hippocampus was divided into anterior and posterior subregions by dividing the hippocampal ROI into thirds along the long axis of the hippocampus. The anterior third became our anterior hippocampus ROI, while the posterior two thirds became the posterior hippocampus ROI, in line with previous work that divides the hippocampus into anterior and posterior at the uncal apex (Thorp et al., 2022).

For whole brain analyses, we used the Shen parcellation (Shen et al., 2013), which divides the brain into 268 parcels. Neural templates for targets (see below) were created for each ROI and each Shen parcel.

fMRI analyses

The majority of our neural analyses were conducted by analyzing the pattern similarity between temporally-averaged neural templates. For each participant, for each video, we extracted a TR x voxel matrix of neural activity, where each TR was a timepoint during the video and each voxel was a location within our ROI. For each of these matrices, we averaged across all TRs to create a single, one-dimensional vector, where each point in the vector represents the average activity in a particular voxel across all TRs for that video. We did this separately for every participant, so that every participant had 16 neural templates (one for each video they viewed). The majority of our analyses concerned Pearson correlations between these templates.

First, we wanted to determine if any of our ROIs were tracking romantic interest for the targets (*Figure 3.3.A*). In each of our ROIs, we ran a representational similarity analysis (RSA) for each video by correlating a neural similarity matrix (where each cell represented the Pearson correlation between two participants' neural templates) and a behavioral rating distance matrix (where each cell represented the Euclidean distance between two participants' behavioral ratings). This process generated 32 pairs of matrices (one for each video), and thus, 32 correlation values. We then ran a permutation analysis (N = 1000) by scrambling the neural similarity correspondences for each neural similarity matrix. For each permutation, we calculated the average correlation across all videos so we could determine if the true average correlation was lower than the 5th percentile of our distribution of correlations. (We expected the correlation to be negative, since we were correlating a similarity matrix and a distance matrix.) We repeated this analysis with variations of our spatial template, including spatial templates made up of just the first 10 TRs, as well as spatial templates with just the last 10 TRs, to determine if the brain's tracking of romantic interest was stronger at the start or end of the video.

Next, we wanted to determine the impact of feedback on neural representations. For each subject, for each profile, we calculated the Pearson correlation between the profile's pre-feedback template and the profile's post-feedback template (*Figure 3.4.A*). Higher correlation values indicated more similarity between the two templates and thus less change in representations between pre- and post-feedback. We split these profile similarities into two feedback groups. In one analysis, the two feedback groups were split according to feedback *valence*, i.e. positive and negative. In another analysis, the two feedback groups were split according to feedback *congruence*, i.e. congruent and incongruent. Feedback was considered congruent if the participant's initial rating was below 5 AND the target feedback was negative, OR if the participant's initial rating was 5 or above AND the target feedback was positive. Feedback was considered incongruent when the participant's initial rating did not align with the binary target feedback. In a third analysis, we split the profile similarities according to the participant's initial ratings (a low group for ratings under 5, a high group for ratings 5 or over) to determine whether initial high or low romantic interest elicited more stability in neural patterns.

Our final suite of analyses concerned reactivation of profiles during post-encoding rest (*Figure 3.5.A*). For each temporally-averaged template, we calculated the Pearson correlation between the template and the pattern of neural activity at each TR during the baseline resting state scan, before the participants had viewed any photos or videos of the targets. This gave us a baseline correlation distribution for each template. We used the correlation value at the 95th percentile of each distribution as our threshold for correlations during the pre-feedback rest and post-feedback rest. We then correlated the pre-feedback template with every TR of the pre-feedback rest. Any correlation value above the 95th percentile of the corresponding baseline distribution counted as a reactivation. We did the same procedure with the post-feedback template and the post-feedback rest. We then summed each instance of reactivation within each profile, within each rest, so that for each subject, every profile had a reactivation frequency score for each resting state scan. (The reactivation frequency score for the baseline resting state scan was always 20, since that number is equal to approximately 5% of all TRs in a single resting state scan.) With these reactivation frequency scores, we were able to ask whether reactivation frequency increased relative to baseline, as well as whether it was higher pre-feedback or post-feedback. We also examined whether reactivation frequency post-feedback was related to the type of feedback the participant received for that target.

3.3 Results

Descriptive statistics and demographics

Each participant provided five different ratings (Compatibility: “How romantically compatible do you believe you are with this person?” Similarity: “How similar do you believe you are to this person?” Attractiveness: “How physically attractive do you find this person?” Interest: “How interested were you in what this person was talking about?” Liking: “How much do you like this person?”) for each of the eight targets, both before receiving feedback and after receiving feedback. Ratings. were on a 1-9 scale. Average pre-feedback ratings for each question are as follows: Compatibility: M = 4.03, SD = 2.03; Similarity: M = 4.12, SD = 1.93; Attractiveness: M = 4.68, SD = 2.14; Interest: M = 4.97, SD = 1.94; Liking: M = 5.08, SD = 1.84. A factor analysis revealed that all five ratings had relatively high loadings onto a single factor (Compatibility: 0.92; Similarity: 0.81; Attractiveness: 0.76; Interest: 0.73; Liking: 0.86). Thus, for analyses in which we needed a single "romantic interest" score, we combined all five dimensions into a single factor using a weighted average based on the standardized loadings. In addition, this romantic interest score was binarized when determining whether target feedback was congruent or incongruent with the participant's initial romantic interest.

How do participants' romantic interest change in response to social feedback?

To determine how participants' romantic interest in a target changed in response to social feedback from that target, we first calculated a romantic interest difference score, which was the difference between post-feedback romantic interest and pre-feedback romantic interest. We next ran a model with feedback valence (positive vs negative) and feedback congruence (congruent vs incongruent) as fixed effects and participant as a random effect (since each participant had eight "trials," or targets).

We found a significant effect of feedback valence (B = 1.52, SE = 0.21, $p < 0.001$), where positive feedback led to an increase in romantic interest and negative feedback led to a decrease (*Figure 3.2.A*). There was also a significant interaction between feedback valence and feedback congruence (B = -2.04, SE = 0.38, $p < 0.001$). This interaction revealed that the majority of the romantic interest change came from incongruent feedback: There was a significant increase (M = 0.938, t(63) = 5.11, $p < 0.001$) in romantic interest for incongruent positive feedback (meaning the participant's initial romantic interest was low) and a significant decrease (M = -1.52, t(50) = -7.78, $p < 0.001$) in romantic interest for incongruent negative feedback (meaning the participant's initial romantic interest was high). When the feedback was congruent, there was very little change in romantic interest, regardless of if the feedback was positive (M = 0.38) or negative (M = -0.17).

We followed up these results with a model where the value of romantic interest change for negative feedback was multiplied by -1 so we could directly compare the magnitude of the change for positive vs negative feedback (*Figure 3.2.B*). We again found larger changes in response to incongruent feedback than congruent feedback (B = -1.02, SE = 0.19, p < 0.001), but did not see an effect of feedback valence (B = -0.11, SE = 0.20, p = 0.59) or an interaction between the two dimensions of feedback type (B = 0.65, SE = 0.45, p = 0.16). A model that directly compared the effect of feedback valence for incongruent data only also failed to find an effect (B = -0.50, SE = 0.32, p = 0.13), although the effect trended towards larger changes for negative incongruent feedback.

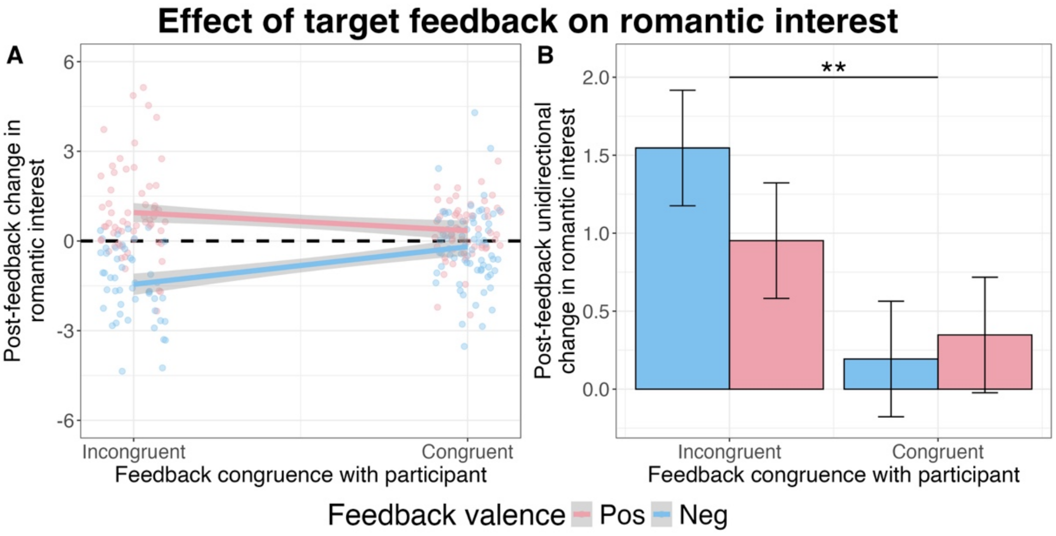


Figure 3.2. The effect of feedback on romantic interest.
The effect of feedback on romantic interest. A) Participants increased their ratings in response to incongruent positive feedback and decreased their ratings in response to incongruent negative feedback. B) Participants' ratings changed more for incongruent feedback than congruent feedback, but there was no significant difference in the size of the change based on feedback valence.

Does the mentalizing network represent romantic interest?

For each video that participants watched, we ran an RSA, whereby we correlated a rating distance matrix (distance here being the Euclidean distance between behavioral rating vectors, where each element of the vector is a rating dimension) and a neural similarity matrix (similarity here being the correlation between the pattern of average activity across all voxels in a particular ROI). If a particular brain area was tracking romantic interest, we would expect to see a negative correlation. A one-tailed t-test revealed a significant negative correlation across the entire Neurosynth mentalizing network (t = -2.21, p = 0.02) (*Figure 3.3.B*). Follow-up tests revealed a similar, but stronger, effect in the right TPJ (t = -3.32, p = 0.001).

We also tested significance via a permutation test, where we correlated the neural similarity matrix with 1,000 scrambled version of the behavioral distance matrix. We found that the true correlation value for the right TPJ (r = -0.095) was lower than the correlation value for all 1,000 permutations, and the true correlation value for the mentalizing network (r = -0.063) was lower than the correlation value for 99.9% of permutations (*Figure 3.3.C*). Thus, our permutation tests reinforced our finding that the mentalizing network as a whole, and especially the right TPJ, were tracking romantic interest ratings.

We next wanted to know if the effects we found above were driven by neural activity during a particular segment of the video. We conducted identical RSAs with two different versions of the neural template: the first 10 TRs and the last 10 TRs. We then compared the strength of the correlations calculated with the first 10 TR template to the correlations from the whole-video template and the last 10 TR template. We chose to compare these templates to determine if romantic interest was most strongly represented in the brain in the opening moments of an encounter or at the end of an encounter.

When using the first 10 TRs as a template, we again found significantly negative correlations in the entire mentalizing network (t = -3.70, p < 0.001) and the right TPJ (t = -2.80, p = 0.004). We also found a significant negative correlation in the dmPFC (t = -1.96, p = 0.03) (*Figure 3.3.B*). Permutation tests reinforced these findings, with the true correlation value lower than the correlation value for 100% of permutations, 99.9% of permutations, and 99% of permutations, respectively (*Figure 3.3.D*). When using the last 10 TRs as a template, we found no significant correlations in any of our ROIs. When comparing correlation strength across templates, we found that correlations were significantly stronger with the first 10 TR template than with the last 10 TR template in the mentalizing network (t = -2.23, p = 0.03) and the right TPJ (t = -2.26, p = 0.03). We also found that correlations were marginally stronger with the first 10 TR template than with the whole video template in the dmPFC (t = -1.96, p = 0.05) (*Figure 3.3.E*). These results suggest that the relationship between brain activity and reported romantic interest is driven by brain activity in the first 10 seconds across the whole mentalizing network, particularly in the right TPJ and to a lesser extent in the dmPFC.

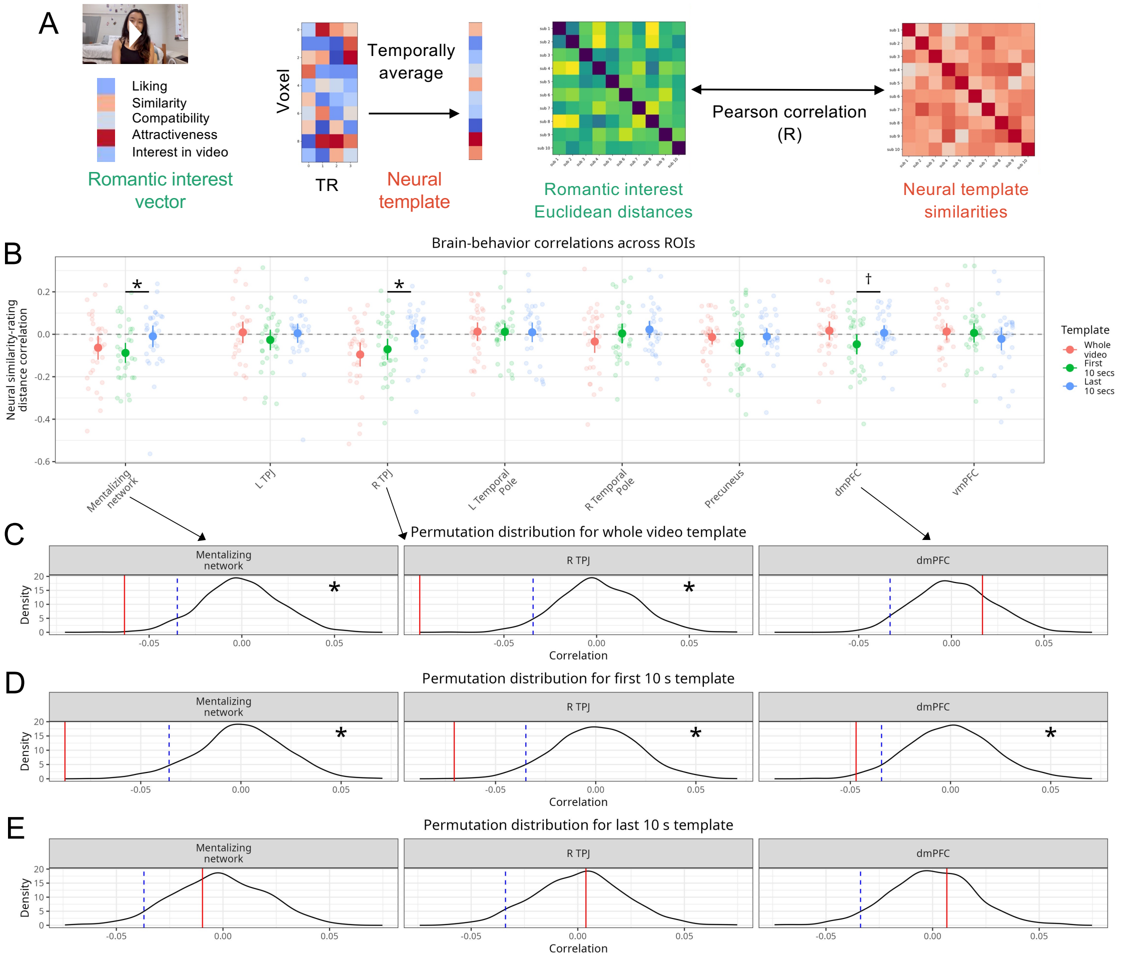


Figure 3.3. Representational Similarity Analysis.
A) To conduct an RSA, for each video we created a Euclidean distance matrix, where each cell is the distance between a vector of behavioral ratings, and a neural similarity matrix, where each cell is the Pearson correlation between neural templates. B) Comparisons of brain behavior correlations for each neural template, in each ROI. The neural template for the first 10 seconds of each video was significantly more predictive of romantic interest ratings than the neural template for the last 10 seconds of video in the mentalizing network and the right TPJ. C) A permutation analysis revealed that the brain-behavior correlation when using the whole video neural template was significantly stronger than chance in the mentalizing network and the right TPJ. The true correlation is the red line, and the 5% threshold of the distribution is the dotted blue line. D) The same as B), but with a neural template with the first 10 seconds of each video. Correlations stronger than chance were found in the mentalizing network, the right TPJ, and the dmPFC. E) No significant brain-behavior correlations were found when using the last 10 seconds of each video as a neural template.

Are neural representations updated in response to feedback?

Given that the mentalizing network, and several ROIs within that network, appeared to be tracking romantic interest, we next asked if neural representations of targets changed more in response to certain types of feedback. To answer this question, we correlated pre-feedback and post-feedback templates, for each participant and for each target. We did not find that neural representations for specific targets changed more in response to a specific feedback valence (positive vs negative) across the mentalizing network or in any of our ROIs (*Figure 3.4*). However, for feedback congruence, a one-tailed test of significance revealed that the mentalizing network showed greater similarity (in other words, less change) between the pre-feedback and post-feedback templates when the participant received congruent feedback, as compared to incongruent feedback (B = 0.03, SE = 0.01, p = 0.03). Follow-up tests revealed this effect in several of our other ROIs, including the left TPJ (B = 0.04, SE = 0.02 p = 0.04), the right TPJ (B = 0.04, SE = 0.02, p = 0.04), the precuneus (B = 0.04, SE = 0.02, p = 0.03), and the right temporal pole (B = 0.03, SE = 0.02, p = 0.04).

We followed up our mentalizing network analyses with a whole brain analysis, where we calculated similarities in each parcel of the Shen parcellation. While no effects survived FDR correction for multiple comparisons across all parcels, we found greater similarity in response to congruent feedback (as compared to incongruent feedback) in 79% of all parcels (N = 225, after excluding parcels with null data from whole-brain masking). In other words, there was a 129-parcel difference between the number of parcels that showed higher similarity for each feedback type. This difference was higher than 100% of differences that resulted from 100 permutations of randomly assigned feedback congruence (largest permutation difference = 53). This result suggests that across the brain, neural representations of other people change less in response to additional information that aligns with initial beliefs than in response to additional information that contradicts initial beliefs.

We found similar results when we compared pre-post similarity between targets for whom participants initially had low romantic interest vs targets for whom participants initially had high romantic interest. Specifically, we found greater similarity (less change) between the two templates across the whole the mentalizing network when initial romantic interest was high (B = 0.03, SE = 0.01, p = 0.03), although we did not find any parallel effects in mentalizing network ROIs. When we looked across the whole brain, we found that greater similarity for initially high romantic interest in 88% of parcels, although no effects survived multiple comparisons. The true 173-parcel difference was larger than 100% of differences that resulted from 100 permutations of randomly assigned initial romantic interest (largest permutation difference = 69). This result might suggest that being initially romantically interested in someone tags neural representations, and these tags are maintained even as the representations change in other ways. It might also be an effect of attention, whereby participants paid greater attention upon a second encounter to targets they were more romantically interested in, leading to a greater neural similarity between the two encounters.

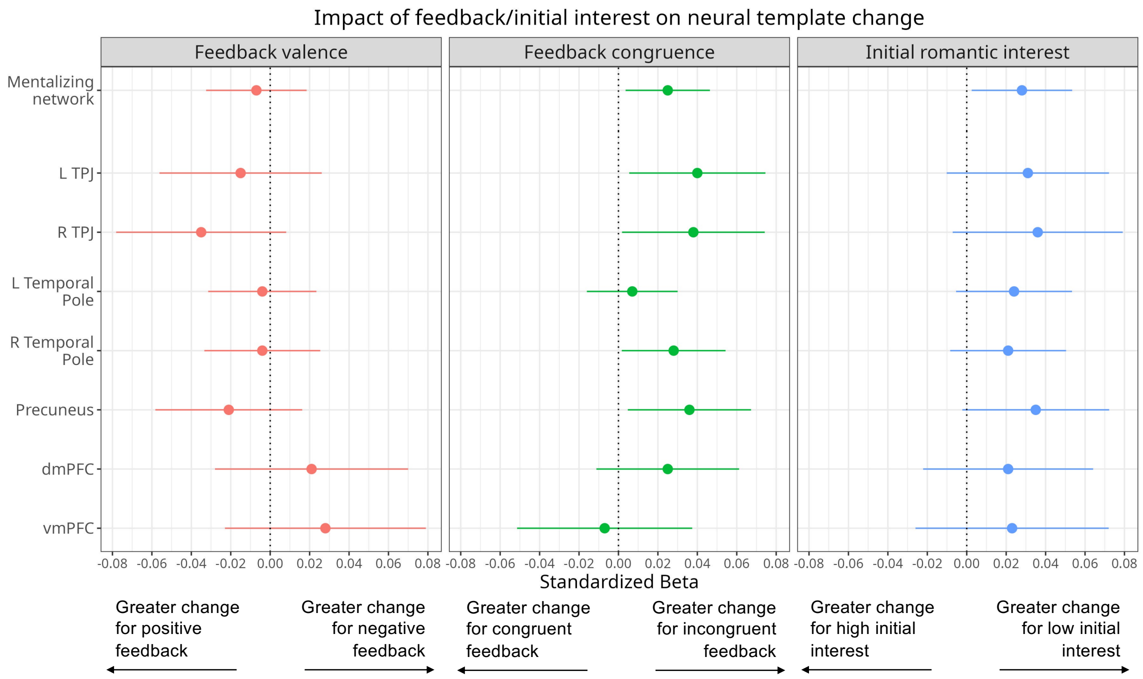


Figure 3.4. How neural templates for targets changed between pre- and post-feedback videos.

Effect sizes for each ROI for each comparison type. There were no significant differences in neural template similarity between targets who provided positive feedback and targets who provided negative feedback. One-tailed tests revealed that neural similarity was higher for targets who provided congruent feedback than those who provided incongruent feedback in the mentalizing network, the left and right TPJ, the right Temporal Pole and the precuneus. For initial romantic interest, pre-post neural similarity was higher in the mentalizing network for high initial interest targets than for low initial interest targets.

How often are potential romantic partners reactivated in the brain during a post-encounter rest?

We next looked to post-encoding rest scans to understand how frequently participants were reactivating targets. More frequent reactivation post-feedback would suggest that the participant finds the feedback meaningful and thinks more about a target in response to receiving feedback from them. In line with our hypothesis, we found a significant increase in reactivation frequency during post-feedback rest, as compared to pre-feedback rest, across the whole mentalizing network (B = 4.16, SE = 1.77, p = 0.03). Follow-up tests revealed stronger effects in the left temporal pole (B = 4.69, SE = 1.46, p = 0.003) and the dmPFC (B = 5.25, SE = 1.85, p = 0.008). We did not see a significant increase in reactivation frequency in any of our other ROIs (*Figure 3.5.B*).

In addition, none of these ROIs demonstrated a significant increase in reactivation frequency during pre-feedback rest as compared to baseline (mentalizing network: B = 1.86, SE = 1.19, p = 0.13; left temporal pole: B = 1.25, SE = 1.05, p = 0.24; dmPFC: 2.55, SE = 1.53, p = 0.11), which suggests that reactivation increased in these regions in response to receiving feedback, and not simply in response to an encounter. On the flip side, we also examined changes in reactivation frequency in the hippocampus, given extensive work that it demonstrates reactivation during post-encoding rest (Schapiro et al., 2018; Yu et al., 2024). Here, we found an increase in reactivation frequency when comparing pre-feedback rest to baseline (B = 1.88, SE = 0.68, p = 0.009), and the effect appeared to be driven by the anterior hippocampus (B = 1.58, SE = 0.58, p = 0.009) and not the posterior hippocampus (B = 0.28, SE = 0.65, p = 0.67). In addition, we did not observe a significant increase during post-feedback rest as compared to pre-feedback rest (Hippocampus: B = 2.20, SE = 1.14, p = 0.06; anterior hippocampus: B = 1.80, SE = 1.27, p = 0.17; posterior hippocampus: B = 1.23, SE = 0.64, p = 0.07). These results supported our hypothesis that targets were reactivated in the hippocampus after an initial encounter, and were reactivated in the mentalizing network – and more specifically, in the left temporal pole and the dmPFC – after a second encounter and after receiving social feedback.

How does feedback type impact reactivation frequency during post-encounter rest?

We next wanted to examine if the increase in reactivation frequency during post-feedback rest was driven by a particular kind of feedback. In a test of feedback valence (positive vs negative), we did not find a significant change in reactivation frequency based on feedback valence in the mentalizing network or in any of our ROIs (*Figure 3.5.C*). However, when we looked at reactivation frequency change as a function of feedback congruence (congruent vs incongruent), we found that incongruent feedback led to a larger increase in reactivation frequency in the anterior hippocampus (B = -3.13, SE = 1.34, p = 0.02), while congruent feedback led to a larger increase in reactivation frequency in the right TPJ (B = 4.41, SE = 1.75, p = 0.01) (*Figure 3.5.D*). The role of feedback in reactivation in these areas is interesting, given that they did not demonstrate an overall increase in reactivation post-feedback as compared to pre-feedback. These results suggest the impact of feedback on neural reactivation is dependent on both the dimension of the feedback (valence or congruence) as well as the role of the brain area where it is being reactivated.

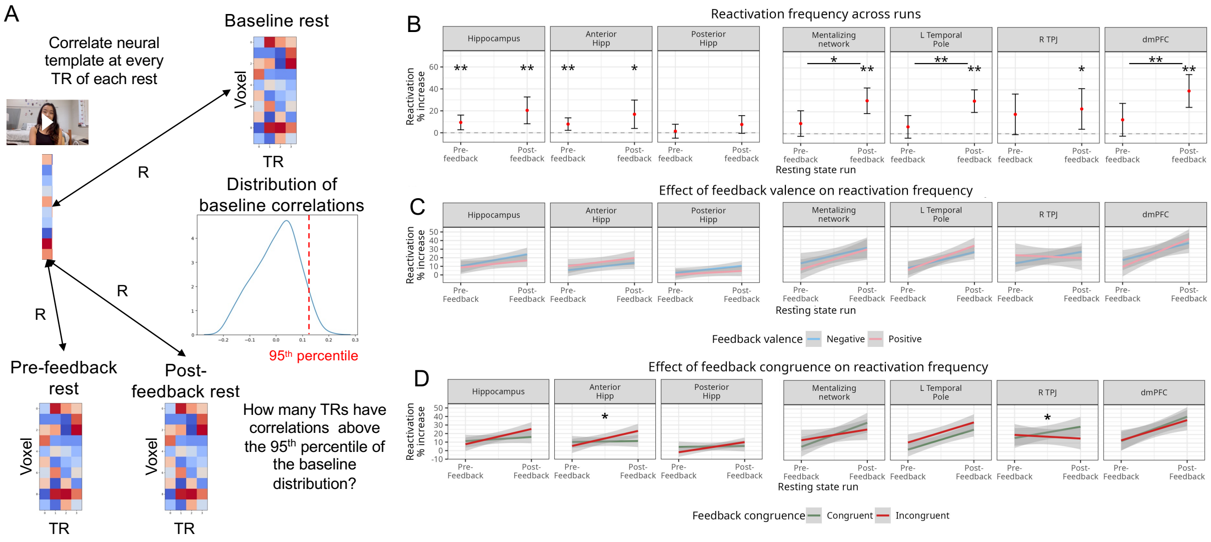


Figure 3.5. Changes in reactivation frequency during post encoding rest.
A) We calculated reactivation during post-encoding rest by establishing a reactivation threshold, which was the 95th percentile of a distribution of baseline rest-neural template correlation values. Then, we asked how many TRs during pre- and post-feedback rest had correlations that were above that threshold. B) The hippocampus, and specifically the anterior hippocampus, demonstrated an increase in reactivation frequency over baseline rest, but not between pre- and post-feedback rest. The mentalizing network, the left temporal pole, and the dmPFC did not demonstrate increases over baseline, but did show increases during post-feedback rest as compared to pre-feedback rest. C) Changes in reactivation frequency were not dependent on feedback valence. D) In the anterior hippocampus, reactivation frequency increased more

between pre- and post-feedback rest in response to receiving incongruent feedback, while in the right TPJ, reactivation frequency increased more in response to congruent feedback.

3.4 Discussion

In the current study, we set out to investigate how people represent potential romantic partners in the brain, and how these representations change, both in terms of structure and reactivation frequency, over time. In order to investigate these questions, we had participants watch pairs of 90-second dating profile videos for eight different potential romantic partners (targets) during an fMRI scan. After each video, participants provided a series of five ratings about their romantic interest in the target. For each target, participants watched two videos: one before receiving social feedback from the target, and one after. The social feedback was ostensibly based on a video the participants had made before entering the scanner, but in reality, the feedback systematically varied in terms of valence and congruence with the participant's initial romantic interest evaluations. Participants also completed a series of resting state scans, both before and after receiving feedback.

We found that romantic interest was represented in the mentalizing network, and more specifically, in the right TPJ and the dmPFC, and that neural representations were more closely tied to romantic interest ratings at the start of a video than at the end of a video. We also found that neural representations of targets who provided incongruent feedback changed more (within-subjects) than targets who provided congruent feedback in the mentalizing network, bilateral TPJ, and the precuneus. Finally, we found that the amount participants reactivated targets during post-feedback rest increased as compared to pre-feedback rest in the mentalizing network, and specifically in the dmPFC, left temporal pole. Contrary to our hypothesis, the size of the increase was not dependent on feedback valence, but was instead dependent on feedback congruence, in both the anterior hippocampus and the right TPJ. We discuss each of these findings in turn below.

The mentalizing network represents romantic interest

An RSA revealed that in the right TPJ and across the entire mentalizing network, videos with more similar romantic interest ratings between participants also exhibited more similar patterns of neural representation. This finding implies that the right TPJ – and to a lesser extent, the mentalizing network – is representing romantic interest. The TPJ is commonly associated with mentalizing (Saxe & Kanwisher, 2004; Van Overwalle & Baetens, 2009), and previous work using similar multivariate analyses has found that the TPJ tracks others' mental states (Golec-Staśkiewicz et al., 2022; Tamir et al., 2016; Thornton et al., 2019), most consistently across the dimensions of rationality, social impact, and valence. Recent work suggests the TPJ may be tracking other types of social evaluation as well, such as trait impressions (Chwe et al., 2024), social value (Morelli et al., 2018), or personal relevance (Bayer et al., 2021). To our knowledge, no other study has demonstrated that the TPJ also represents romantic interest. However, it may be the case that findings concerning different types of social evaluations in the TPJ, including ours about romantic interest, are united by an attention to the mental states of others. Indeed, the dating context is one in which we are particularly attuned to others' intentions, and our level of romantic interest is at least partially dependent on what we believe is their romantic interest towards us.

However, not all contexts where we track social evaluations or mental states are equivalent. In our study, we also found that in the TPJ, the mentalizing network, and the dmPFC, neural templates made up of data averaged over the first ten seconds of each video better predicted romantic interest than neural templates made up of data averaged over the last ten seconds of each video. Indeed, none of our ROIs demonstrated significant correlations with behavior in the last ten seconds. These results are particularly striking given that the last ten seconds of the video were closest in time to the point of providing romantic interest ratings. These results imply that in the dating context, the start of an encounter is most important in determining social evaluations.

This finding fits with previous behavioral work that shows that we form impressions on the order of hundreds of milliseconds (Todorov et al., 2015; Willis & Todorov, 2006), and that early impressions are predictive of impressions much later in time (Ambady & Rosenthal, 1992), including in romantic relationships (Baxter et al., 2022). The vast majority of social neuroscience studies that implicate the dmPFC and TPJ in social evaluations use static images as stimuli (Ferrari et al., 2016; Ma et al., 2014; D. Schiller et al., 2009), so they are not able to tell us anything about the temporal dynamics of social evaluation formation in the brain. Our results suggest that, at least in the context of meeting a potential romantic partner, our neural representations at the start of an encounter are most predictive of our romantic interest later on. Neural activity later in an encounter, on the other hand, appears to have little relationship with this type of social evaluation.

Incongruent information causes larger changes in neural representations than congruent information

Participants viewed videos of each target twice: once before receiving feedback, and once after receiving feedback. We found that, within-subject, similarity between and pre- and post-feedback neural templates in the mentalizing network, the bilateral TPJ, and the precuneus was higher when the target provided congruent feedback than when the target provided incongruent feedback. These results align with our behavioral findings, which showed that incongruent feedback led to larger changes in romantic interest ratings. While previous work implicates the TPJ in the process of impression updating (Cloutier et al., 2011) and prediction error (Dohmatob et al., 2020; Shulman et al., 2007; Vetter et al., 2011), relatively few studies have demonstrated what we found: neural representations in the TPJ and across the mentalizing network changed more as a result of unexpected information. The idea of representational change is often investigated in the context of memory updating (Wahlheim & Zacks, 2024; Zadbood et al., 2022), where representations for past events in regions in the default mode network (which has significant overlap with the mentalizing network) change upon the receipt of new information. Here, we show that representations for different stimuli that pertain to the same target change according to how well information provided by the target aligns with expectations.

We also found effects of within-subject representational change across the entire brain, with significantly more parcels than would be expected by chance showing higher similarity for targets who provide congruent feedback, and for targets for whom participants had initial high levels of romantic interest. The widespread nature of the effect of congruent feedback on similarity beyond the mentalizing network is in contrast to some prior work about memory updating (Zadbood et al., 2022); our findings suggest widespread representational changes across the brain for a target who provides unexpected information.

In addition, we saw similar whole-brain similarity for targets who participants are more romantically interested in, which may be because those targets elicit greater attention (Compton, 2003; Langeslag & van Strien, 2019; Nakamura et al., 2017). Indeed, greater attention elicits more reliable neural responses (Hasson et al., 2008; Ki et al., 2016), which may lead to higher levels of neural similarity in sensory regions across time. An alternative explanation would be that initial social evaluations are “stickier” when they are positive, or are tagged in some way that negative evaluations are not. However, given that we did not see effects of initial impression on neural similarity in social brain regions specifically, we deem this interpretation of our whole-brain findings unlikely.

Reactivation increases are dependent on different types of feedback in different brain systems

Our experiences with another person, and how well our evaluations of them align with their evaluations of ourselves, might also impact how often we think about them. To test this question, we calculated reactivation frequency of neural templates during post-encoding rest, both in terms of run (pre- vs post-feedback) and feedback type (valence and congruence). Consistent with previous work on reactivation (Gruber et al., 2016; Schapiro et al., 2018; Staresina et al., 2012; Tambini & Davachi, 2019; Yu et al., 2024), we found that reactivation frequency increased in the hippocampus, and in particular the anterior hippocampus, compared to a baseline rest before participants had seen the stimuli. We also found that in the anterior hippocampus, reactivation increased between pre- and post-feedback more for targets who provided incongruent feedback than congruent feedback. Incongruent feedback can be thought of as an expectancy violation (Somerville et al., 2006, 2010). The hippocampus is known to play a role in schema updating as a result of expectancy violations or schema-inconsistent information (Bein et al., 2014; van Kesteren et al., 2013). Most of this work demonstrates an increase in hippocampal activity upon receipt of the information or during consolidation. Our study extends these findings by showing that incongruent information leads to increased hippocampal reactivation for stimuli associated with that incongruent information.

We also saw an increase in overall reactivation of targets between pre- and post-feedback rest in the dmPFC, the left temporal pole, and across the mentalizing network. Previous work has demonstrated that these brain areas, and areas across the mentalizing network, prioritize social information during post-encoding rest (Jimenez & Meyer, 2024; Meyer et al., 2019), so it makes sense that they would demonstrate an increase in reactivation for social stimuli. Intriguingly, we only see a significant increase in these areas after viewing the post-feedback video, which likely suggests that the receipt of social feedback causes one to think more about the person who provided that feedback. In addition, we saw that reactivation frequency increased more in the right TPJ in response to congruent feedback than incongruent feedback, the reverse of the finding we saw in the anterior hippocampus. (We did not, however, see an overall increase in reactivation in the TPJ.) Our study is the first to investigate reactivation in the context of changing social stimuli, which may be why we see an effect in the TPJ when others haven't. In essence, the right TPJ is prioritizing the reactivation of social stimuli who provide congruent feedback, regardless of if the feedback is positive or negative. Previous work demonstrates that we engage in mentalizing more for those who are more similar to us (Mitchell et al., 2006; Tamir & Mitchell, 2010); perhaps participants engaged in more mentalizing during rest for targets whose feedback aligned with initial evaluations.

In the current study, we demonstrated that mentalizing plays a crucial role in forming and updating romantic interest evaluations for potential romantic partners, but that the type of updating that occurs depends on the type of feedback we receive. An RSA revealed that the mentalizing network – and specifically, the right TPJ and the dmPFC – tracked romantic interest over the course of an entire encounter and even more strongly at the start of an encounter. In addition, neural representations in several regions of the mentalizing network of specific other people changed more in response to feedback that was incongruent with one's initial evaluation than in response to congruent feedback, suggesting that how we are perceived by others impacts our perceptions of them as well. Finally, reactivation frequency increased in the mentalizing network – especially in the dmPFC – after receiving social feedback. Reactivation frequency increased more in the anterior hippocampus in response to incongruent feedback but more in the TPJ in response to congruent feedback, suggesting that others with whom we share similar social evaluations linger more in our minds. Overall, these results demonstrate the importance of mentalizing in evaluating our romantic interest in other people.

Conclusion

Heraclitus's never-ending stream tells us that change occurs, constantly. On why, or when, or how change occurs, Heraclitus does not have much to offer. This is where psychologists must step in. Our job is to characterize the flowing stream, to describe the principles by which it flows, and to predict how the streams flow in the next polis over. In my dissertation, I argue that the phenomenon of changing our beliefs about other people can only be completely understood if we account for the socio-affective motivations and pre-existing social relationships that color the way we perceive unexpected information. In addition, these same motivations also determine the persistence of a belief update over time. Finally, I argue that only through an understanding of post-belief change neural processing can we fully understand the psychological mechanisms that underly why and how we update our social evaluations of other people.

Overview of findings

In Chapter 1, we found that we were able to detect relatively large changes in perceived moral character from social media data, with language associated with immorality and harm (as defined by the Moral Foundations Dictionary (Graham et al., 2009)) becoming five times more common post-allegations, in comparison to baseline. We also found that in the first three weeks after each allegation became public, the magnitude of the increase was dependent on both the severity of the allegation and an interaction between the severity and how well-liked the public figure was pre-allegation. Specifically, for less severe allegations, being well-liked mitigated an increase in immoral language, while for more severe allegations, being well-liked had no effect. Finally, we found that immoral language was still elevated over baseline one year later. At that point, unlike immediately after the allegations, we found that person-specific factors, such as liking and familiarity, were more predictive of immoral language than were situation-specific factors, such as allegation severity.

In Chapter 2, we found similar evidence of longer-lasting impression updates. Perceptions of teammates' trait-level competence and sociability on average increased immediately after the escape room, and remained elevated over baseline one week later. In addition, the magnitude of the absolute value of change – which we calculated to account for heterogeneity in impression update direction – was larger between pre-game and post-game than between post-game and one week later, again suggesting that the impression update persisted beyond an immediate effect. We also saw further specificity in the types of relationships that mattered between different dimensions of trait impressions: Greater perceived similarity impacted ratings of others' competence while greater liking impacted ratings of others' sociability. These factors alone predicted the magnitude of each impression update. Finally, we saw that these dimensions also differed in the role that one's actions played in an impression update: An objective measure of puzzle solving performance predicted competence ratings, but biased perceptions of team collaboration performance predicted sociability ratings.

In Chapter 3, my investigation of the neural mechanisms underlying how evaluations of romantic interest are updated revealed that mentalizing regions likely play a large role. Specifically, neural representations in these regions responded most strongly to social feedback that was incongruent with one's initial evaluation, likely because our romantic interest in another person is at least partially dependent on how we believe they feel about us. Within the mentalizing network, the right TPJ and the dmPFC most strongly represented romantic interest at the beginning of an encounter. In addition, representations in the TPJ and the precuneus changed more in response to incongruent feedback, while reactivation *frequency* in the TPJ changed more in response to congruent feedback, demonstrating that how much we think about someone is not necessarily linked to how much we've changed our perceptions of them.

Discussion

The three studies in this dissertation were all designed to investigate the impression updating process. For that reason, all three studies are structurally similar. First, each of them includes an instance of receiving information that may or may not contradict previously held beliefs. #MeToo was chosen as the context for Chapter 1 because it was a real-world instance of repeated impression updating; part of the reason #MeToo made such an impact was precisely *because* the revelations were often unexpected and surprising. A virtual escape room was chosen as the context for Chapter 2 because it was a dynamic, unfamiliar environment, which presented ample opportunities to encounter unexpected information. In Chapter 3, unlike in Chapters 1 and 2, the information – in this case, social feedback about romantic interest – was systematically varied to either align or misalign with expectations.

Second, each study includes both pre and post periods so that changes in impressions can be assessed. In Chapter 1, the pre period was tweets from six months prior to an allegation; in Chapter 2, it was a survey completed up to one week before the escape room game; and in Chapter 3, it was a set of eight videos viewed before receiving feedback. A shared feature of Chapters 1 and 2 was that they both had two post periods: Impressions in Chapter 1 were assessed for three weeks after an allegation as well as one year later, while impressions in Chapter 2 were assessed immediately after the game and one week later. Having multiple post periods allowed me to make observations about the durability of an update. Chapter 3 only had one post period, which was the set of videos viewed after receiving social feedback.

An important difference between all three chapters was in the type of evaluation that was made; this distinction was particularly important between Chapters 1 and 2 because it differentiated the aspect of a relationship that most strongly motivated the impression update. Chapter 1 focused on updating of morality, while Chapter 2 directly compared updating of competence and sociability. These three traits form the basis of a common model of person perception (Brambilla et al., 2011; Goodwin et al., 2014; Landy et al., 2016), where morality and sociability are seen as two overlapping components of the more traditional warmth dimension (Fiske et al., 2007). Goodwin and colleagues have attempted to differentiate the roles of these three traits in overall person perception, and some work has investigated asymmetries in updating between warmth and competence (Mende-Siedlecki, Baron, et al., 2013; Reeder et al., 1977). However, there is comparatively much less work on the updating tendencies of sociability; I demonstrate that sociability may behave similarly to competence in an impression updating context, as both dimensions showed a positive bias.

In addition, there is also very little work on how these traits interact with pre-existing relationships. As such, Chapters 1 and 2 of my dissertation present two additional novel contributions to the field of person perception. First, I bring together work on close relationships and work on person perception to demonstrate that socio-affective motivations associated with close relationships – which are often not present upon first meeting someone – impact the impression updating process. Second, I show that the impact of pre-existing relationships is specific and dependent on the dimension that is being updated: Liking mitigated negative updating for morality and exacerbated positive updating for sociability, while perceived similarity exacerbated positive updating for competence. These are not global effects of relationships on overall social evaluations, but rather, a demonstration of the distinctive nature of different kinds of relationships and perceptions of different traits. Future work that investigates how impressions of close others change in response to unexpected information must properly characterize the relationship and consider the dimension that is being assessed.

Chapter 3, on the other hand, does not investigate impression updating for traditional person perception traits, but instead investigates how we update a different sort of social evaluation: romantic interest. There is extensive work on how relationships form and their quality changes over time (Joel et al., 2020; Larson et al., 2022; Meltzer & McNulty, 2019), but the interplay of romantic interest between two people at the very start of a romantic relationship is understudied. Indeed, neural analyses in this study reveal that interdependence – or, what we believe another person believes about us – are another important property of how we update our beliefs about others. I found that neural representations in the mentalizing network, and specifically the TPJ, responded to social feedback, both in terms of the structure of the representation as well as how frequently it was reactivated. Given that the TPJ has long been shown to play a role in taking the perspective of others (Saxe & Kanwisher, 2004; Van Overwalle & Baetens, 2009), and that beliefs about others' mental states can be decoded from multivariate patterns in the TPJ (Tamir et al., 2016; Thornton et al., 2019), the results from Chapter 3 suggest that we actively consider what others think about us when we change how we think about them.

These findings have widespread implications for the study of social belief change more generally, and it's interesting to consider the impression updating processes in Chapters 1 and 2 in light of these results. On the one hand, it's less likely that people would be as concerned about the interdependence of their evaluations because strong pre-existing relationships either already existed (Chapter 2) or the relationship was unidirectional (Chapter 1). These factors suggest that the specific context of Chapter 3 – meeting a potential romantic partner for the first time – uniquely implicates mentalizing processes, more so than would be required when updating impressions for others you already know. On the other hand, participants in Chapter 2 who were evaluating their teammates' competence and sociability were also demonstrating these traits themselves, since all teammates were equal participants in the escape room game. It's possible that people's evaluations of their friends' traits were dependent on how they believed their friends would evaluate them. This open question demonstrates the importance of neural data in Chapter 3 to helping us understand a complex psychological phenomenon.

Finally, all three chapters speak to debates about the durability of an impression update. The vast majority of impression updating studies only examine immediate changes (Forscher et al., 2019), which makes it unclear if observed changes are simply a result of in-the-moment responses to new information, or if they represent real, permanent alterations in one's perceptions of another person. While Chapter 3 also only looked at immediate changes, it goes a step further than most previous social neuroscience studies of impression updating, which only examine the moment when updating occurs (D. L. Ames & Fiske, 2013; Cloutier et al., 2011; B. Park & Young, 2020). Instead, Chapter 3 demonstrates that there are changes in the structure of neural representations in response to unexpected information. While these results can't speak to how long-lasting the changes are, they do help to disentangle debates about responses to unexpected information vs alterations in perceptions of the target as a result of the information.

On the other hand, Chapters 1 and 2, and especially Chapter 1, can speak directly to the longevity of an impression update. Chapter 2, which examined the durability of an update one week later, is in line with the timescale of previous work that has tested specifically for durability (Cone et al., 2021; Cone & Ferguson, 2015; Lai et al., 2016). Chapter 1 had both a more fine-grained examination of belief updating – I measured it every day for three weeks after an update occurred – as well as a far longer time-scale – I also measured it one year later – than other studies of impression updating, and even of belief updating in general (Costello et al., 2024; Sharot et al., 2023). Both chapters demonstrated similar results at the long-term timepoint: impressions demonstrated a slight reversion to, but were still significantly different from, baseline measurements. These results demonstrate that in real-world, non-hypothetical paradigms like the ones highlighted in this dissertation, belief updates can persist beyond immediate effects. This result is in line with previous work that notes that belief changes persist when the new information is both diagnostic and believable (Cone et al., 2021; Siegel et al., 2018). Indeed, in Chapters 1 and 2, "believability" is not much of a concern since the paradigms were real demonstrations of one's traits, highlighting the importance of ecologically valid psychology paradigms.

In terms of the slight reversion to baseline, Chapter 1 showed that this reversion began immediately, but then leveled off after about a week at a level that was similar to the value one year later. This finding implies that the competence and sociability values measured one week later in Chapter 2 might be the same if measured a year later. However, I think that this is unlikely, as a major difference between Chapters 1 and 2 is in the scale and importance of the new information that was presented, further highlighting the role that a paradigm's ecological validity can play in one's findings. The #MeToo movement was a large-scale social phenomenon that remained a prominent news item for much of the time when levels of perceived morality were measured in Chapter 1, meaning that the information remained easily accessible long after people first encountered it and therefore could more easily continue to influence perceptions (Higgins & Brendl, 1995; Rothbart et al., 1978). The friends who were assessing each other in

Chapter 2 likely continued to have many and varied interactions long after the escape room game. These additional interactions likely provided further information that one incorporated into their perceptions, reducing the importance of the information learned from the escape room game.

This final point is a neat demonstration of the quote that we began with. Change is constant, especially when it comes to perceptions of other people. We build models of the world, and revise them when we make unexpected observations. Other people are constantly surprising us, shocking us, throwing us off balance, defying expectations. The factors that impact how we respond to these shocks and defiances are the subject of this dissertation. Chapters 1 and 2 highlight the importance and specificity of pre-existing relationships and socio-affective motivations in parsing our responses to unexpected actions from other people. These two chapters also help us understand the temporal dynamics of these responses, and that in meaningful situations, they can last quite a long time. Chapter 3 elucidates the neural mechanisms that support and explain how our perceptions of others change after the moment new information is presented, both in terms of how we think about someone and how often we think about them. These results make clear that social evaluations are updated interdependently with how other people evaluate us. In sum, this dissertation demonstrates that just because change is constant does not mean it is impossible to understand. I have shown that impression updates cannot be fully understood without accounting for motivation, durability, and interdependence. These principles provide a strong foundation for future investigations of real-world impression updating.

References

Abdo, M. S., Alghonaim, A. S., & Essam, B. A. (2021). Public perception of COVID-19's global health crisis on Twitter until 14 weeks after the outbreak. *Digital Scholarship in the Humanities*, *36*(3), 509–524. https://doi.org/10.1093/llc/fqaa037

Abele, A. E., Ellemers, N., Fiske, S. T., Koch, A., & Yzerbyt, V. (2021). Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological Review*, *128*(2), 290–314. https://doi.org/10.1037/rev0000262

Abele, A. E., & Wojciszke, B. (2014). Communal and Agentic Content in Social Cognition. In *Advances in Experimental Social Psychology* (Vol. 50, pp. 195–255). Elsevier. https://doi.org/10.1016/B978-0-12-800284-1.00004-7

Akkerman, S., Van den Bossche, P., Admiraal, W., Gijselaers, W., Segers, M., Simons, R.-J., & Kirschner, P. (2007). Reconsidering group cognition: From conceptual confusion to a boundary area between cognitive and socio-cultural perspectives? *Educational Research Review*, *2*(1), 39–63. https://doi.org/10.1016/j.edurev.2007.02.001

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology*, *20*(1), 1–48. https://doi.org/10.1080/10463280802613866

Alves, H., Koch, A., & Unkelbach, C. (2016). My friends are all alike—The relation between liking and perceived similarity in person perception. *Journal of Experimental Social Psychology*, *62*, 103–117. https://doi.org/10.1016/j.jesp.2015.10.011

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256–274. https://doi.org/10.1037/0033-2909.111.2.256

Ames, D. L., & Fiske, S. T. (2013). Outcome dependency alters the neural substrates of impression formation. *NeuroImage*, *83*, 599–608. https://doi.org/10.1016/j.neuroimage.2013.07.001

Ames, D. R. (2004). Strategies for Social Inference: A Similarity Contingency Model of Projection and Stereotyping in Attribute Prevalence Estimates. *Journal of Personality and Social Psychology*, *87*(5), 573–585. https://doi.org/10.1037/0022-3514.87.5.573

Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, *103*(4), 718–735. https://doi.org/10.1037/a0029395

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, *41*(3), 258–290. https://doi.org/10.1037/h0055756

Ashokkumar, A., & Pennebaker, J. W. (2021). Social media conversations reveal large psychological shifts caused by COVID-19's onset across U.S. cities. *Science Advances*, *7*(39), eabg7843. https://doi.org/10.1126/sciadv.abg7843

Atique, B., Erb, M., Gharabaghi, A., Grodd, W., & Anders, S. (2011). Task-specific activity and connectivity within the mentalizing network during emotion and intention mentalizing. *NeuroImage*, *55*(4), 1899–1911. https://doi.org/10.1016/j.neuroimage.2010.12.036

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Back, M. D., & Kenny, D. A. (2010). The Social Relations Model: How to Understand Dyadic Processes: The Social Relations Model. *Social and Personality Psychology Compass*, *4*(10), 855–870. https://doi.org/10.1111/j.1751-9004.2010.00303.x

Baetens, K., Ma, N., Steen, J., & Van Overwalle, F. (2014). Involvement of the mentalizing network in social and non-social high construal. *Social Cognitive and Affective Neuroscience*, *9*(6), 817–824. https://doi.org/10.1093/scan/nst048

Baumeister, R. F. (2008). Free Will in Scientific Psychology. *Perspectives on Psychological Science*, *3*(1), 14–19. https://doi.org/10.1111/j.1745-6916.2008.00057.x

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is Stronger than Good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Baxter, A., Maxwell, J. A., Bales, K. L., Finkel, E. J., Impett, E. A., & Eastwick, P. W. (2022). Initial impressions of compatibility and mate value predict later dating and romantic interest. *Proceedings of the National Academy of Sciences*, *119*(45), e2206925119. https://doi.org/10.1073/pnas.2206925119

Bayer, M., Berhe, O., Dziobek, I., & Johnstone, T. (2021). Rapid Neural Representations of Personally Relevant Faces. *Cerebral Cortex*, *31*(10), 4699–4708. https://doi.org/10.1093/cercor/bhab116

Beer, J. S., & Hughes, B. L. (2011). Self-enhancement: A social neuroscience perspective. *Handbook of Self-Enhancement and Self-Protection*, 49–65.

Bein, O., Reggev, N., & Maril, A. (2014). Prior knowledge influences on hippocampus and medial prefrontal cortex interactions in subsequent memory. *Neuropsychologia*, *64*, 320–330. https://doi.org/10.1016/j.neuropsychologia.2014.09.046

Berinsky, A. J. (2017). Measuring Public Opinion with Surveys. *Annual Review of Political Science*, *20*(1), 309–329. https://doi.org/10.1146/annurev-polisci-101513-113724

Berscheid, E., Snyder, M., & Omoto, A. M. (1989). The Relationship Closeness Inventory: Assessing the closeness of interpersonal relationships. *Journal of Personality and Social Psychology*, *57*(5), 792–807. https://doi.org/10.1037/0022-3514.57.5.792

Bhanji, J. P., & Beer, J. S. (2013). Dissociable Neural Modulation Underlying Lasting First Impressions, Changing Your Mind for the Better, and Changing It for the Worse. *Journal of Neuroscience*, *33*(22), 9337–9344. https://doi.org/10.1523/JNEUROSCI.5634-12.2013

Biesanz, J. C., West, S. G., & Millevoi, A. (2007). What do you learn about someone over time? The relationship between length of acquaintance and consensus and self-other agreement in judgments of personality. *Journal of Personality and Social Psychology*, *92*(1), 119–135. https://doi.org/10.1037/0022-3514.92.1.119

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford press.

Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, *18*, 63–68. https://doi.org/10.1016/j.cobeha.2017.07.017

Bradfield, M., & Aquino, K. (1999). The Effects of Blame Attributions and Offender Likableness on Forgiveness and Revenge in the Workplace. *JOURNAL OF MANAGEMENT*, *25*(5), 25.

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), eabe5641. https://doi.org/10.1126/sciadv.abe5641

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Brambilla, M., Carraro, L., Castelli, L., & Sacchi, S. (2019). Changing impressions: Moral character dominates impression updating. *Journal of Experimental Social Psychology*, *82*, 64–73. https://doi.org/10.1016/j.jesp.2019.01.003

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. https://doi.org/10.1521/soco.2014.32.4.397

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology*, *41*(2), 135–143. https://doi.org/10.1002/ejsp.744

Brambilla, M., Sacchi, S., Rusconi, P., & Goodwin, G. P. (2021). The primacy of morality in impression development: Theory, research, and future directions. In *Advances in Experimental Social Psychology* (Vol. 64, pp. 187–262). Elsevier. https://doi.org/10.1016/bs.aesp.2021.03.001

Brannon, S. M., & Gawronski, B. (2017). A Second Chance for First Impressions? Exploring the Context-(In)Dependent Updating of Implicit Evaluations. *Social Psychological and Personality Science*, *8*(3), 275–283. https://doi.org/10.1177/1948550616673875

Brooks, J. A., & Freeman, J. B. (2018). Conceptual knowledge predicts the representational structure of facial emotion perception. *Nature Human Behaviour*, *2*(8), 581–591. https://doi.org/10.1038/s41562-018-0376-6

Burgara, A. (2020). *Pygooglenews* (Version 0.1.2) [Python]. https://github.com/kotartemiy/pygooglenews

Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*, *5*(12), 1629–1635. https://doi.org/10.1038/s41562-021-01133-5

Cabanis, M., Pyka, M., Mehl, S., Müller, B. W., Loos-Jankowiak, S., Winterer, G., Wölwer, W., Musso, F., Klingberg, S., Rapp, A. M., Langohr, K., Wiedemann, G., Herrlich, J., Walter, H., Wagner, M., Schnell, K., Vogeley, K., Kockler, H., Shah, N. J., … Kircher, T. (2013). The precuneus and the insula in self-attributional processes. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(2), 330–345. https://doi.org/10.3758/s13415-012-0143-5

Cannon-Bowers, J. A., & Bowers, C. (2011). Team development and functioning. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology, Vol 1: Building and developing the organization.* (pp. 597–650). American Psychological Association. https://doi.org/10.1037/12169-019

Cao, J., & Banaji, M. R. (2016). The base rate principle and the fairness principle in social judgment. *Proceedings of the National Academy of Sciences*, *113*(27), 7475–7480. https://doi.org/10.1073/pnas.1524268113

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology*, *67*(2), 319–333. https://doi.org/10.1037/0022-3514.67.2.319

Castelli, L., Carraro, L., Ghitti, C., & Pastore, M. (2009). The effects of perceived competence and sociability on electoral outcomes. *Journal of Experimental Social Psychology*, *45*(5), 1152–1155. https://doi.org/10.1016/j.jesp.2009.06.018

Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: A review of its functional anatomy and behavioural correlates. *Brain*, *129*(3), 564–583. https://doi.org/10.1093/brain/awl004

Champely, S. (2020). *_pwr: Basic Functions for Power Analysis_*. https://CRAN.R-project.org/package=pwr

Chwe, J. A. H., Vartiainen, H. I., & Freeman, J. B. (2024). A Multidimensional Neural Representation of Face Impressions. *The Journal of Neuroscience*, *44*(39), e0542242024. https://doi.org/10.1523/JNEUROSCI.0542-24.2024

Cloutier, J., Gabrieli, J. D. E., 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. https://doi.org/10.1016/j.neuroimage.2011.04.051

Compton, R. J. (2003). The Interface Between Emotion and Attention: A Review of Evidence from Psychology and Neuroscience. *Behavioral and Cognitive Neuroscience Reviews*, *2*(2), 115–129. https://doi.org/10.1177/1534582303002002003

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), 37–57. https://doi.org/10.1037/pspa0000014

Cone, J., Flaharty, K., & Ferguson, M. J. (2021). The Long-Term Effects of New Evidence on Implicit Impressions of Other People. *Psychological Science*, *32*(2), 173–188. https://doi.org/10.1177/0956797620963559

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing Our Implicit Minds: How, When, and Why Implicit Evaluations Can Be Rapidly Revised. In *Advances in Experimental Social Psychology* (Vol. 56, pp. 131–199). Elsevier. https://doi.org/10.1016/bs.aesp.2017.03.001

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), eadq1814. https://doi.org/10.1126/science.adq1814

Couch, L. L., Baughman, K. R., & Derow, M. R. (2017). The Aftermath of Romantic Betrayal: What's Love Got to Do with It? *Current Psychology*, *36*(3), 504–515. https://doi.org/10.1007/s12144-016-9438-y

Couch, L. L., & Olson, D. R. (2016). Loss Through Betrayal: An Analysis of Social Provision Changes and Psychological Reactions. *Journal of Loss and Trauma*, *21*(5), 372–383. https://doi.org/10.1080/15325024.2015.1108789

Cox, R. W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173. https://doi.org/10.1006/cbmr.1996.0014

Crocker, J., Fiske, S. T., & Taylor, S. E. (1984). Schematic Bases of Belief Change. In J. R. Eiser (Ed.), *Attitudinal Judgment* (pp. 197–226). Springer New York. https://doi.org/10.1007/978-1-4613-8251-5_10

Crocker, J., Hannah, D. B., & Weber, R. (1983). Person memory and causal attributions. *Journal of Personality and Social Psychology*, *44*(1), 55–66. https://doi.org/10.1037/0022-3514.44.1.55

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771. https://doi.org/10.1038/s41562-017-0213-3

Deen, B., Husain, G., & Freiwald, W. A. (2024). A familiar face and person processing area in the human temporal pole. *Proceedings of the National Academy of Sciences*, *121*(28), e2321346121. https://doi.org/10.1073/pnas.2321346121

Denny, B. T., Inhoff, M. C., Zerubavel, N., Davachi, L., & Ochsner, K. N. (2015). Getting Over It: Long-Lasting Effects of Emotion Regulation on Amygdala Response. *Psychological Science*, *26*(9), 1377–1388. https://doi.org/10.1177/0956797615578863

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A Meta-analysis of Functional Neuroimaging Studies of Self- and Other Judgments Reveals a Spatial Gradient for Mentalizing in Medial Prefrontal Cortex. *Journal of Cognitive Neuroscience*, *24*(8), 1742–1752. https://doi.org/10.1162/jocn_a_00233

Dohmatob, E., Dumas, G., & Bzdok, D. (2020). Dark control: The default mode network as a reinforcement learning agent. *Human Brain Mapping*, *41*(12), 3318–3341. https://doi.org/10.1002/hbm.25019

Doré, B., Ort, L., Braverman, O., & Ochsner, K. N. (2015). Sadness Shifts to Anxiety Over Time and Distance From the National Tragedy in Newtown, Connecticut. *Psychological Science*, *26*(4), 363–373. https://doi.org/10.1177/0956797614562218

Doré, B. P., Morris, R. R., Burr, D. A., Picard, R. W., & Ochsner, K. N. (2017). Helping Others Regulate Emotion Predicts Increased Regulation of One's Own Emotions and Decreased Symptoms of Depression. *Personality and Social Psychology Bulletin*, *43*(5), 729–739. https://doi.org/10.1177/0146167217695558

Driskell, J. E., Salas, E., & Johnston, J. (1999). Does Stress Lead to a Loss of Team Perspective? *Group Dynamics: Theory, Research, and Practice*, *3*(4), 291–302.

Driskell, T., Blickensderfer, E. L., & Salas, E. (2013). Is three a crowd? Examining rapport in investigative interviews. *Group Dynamics: Theory, Research, and Practice*, *17*(1), 1–13. https://doi.org/10.1037/a0029686

Dunham, Y. (2018). Mere Membership. *Trends in Cognitive Sciences*, *22*(9), 780–793. https://doi.org/10.1016/j.tics.2018.06.004

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking Deathworthy: Perceived Stereotypicality of Black Defendants Predicts Capital-Sentencing Outcomes. *Psychological Science*, *17*(5), 383–386. https://doi.org/10.1111/j.1467-9280.2006.01716.x

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Evans, A. M., Rosenbusch, H., & Zeelenberg, M. (2022). Using semantic similarity to understand the psychological constructs related to prosociality. *Current Opinion in Psychology*, *44*, 226–230. https://doi.org/10.1016/j.copsyc.2021.09.019

Ferguson, M. J., Mann, T. C., Cone, J., & Shen, X. (2019). When and How Implicit First Impressions Can Be Updated. *Current Directions in Psychological Science*, *28*(4), 331–336. https://doi.org/10.1177/0963721419835206

Ferrari, 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. *Cerebral Cortex*, *26*(1), 156–165. https://doi.org/10.1093/cercor/bhu186

Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C. E., Rathje, S., Hartmann, J., Mohammad, S. M., Netzer, O., Siegel, A. A., Plank, B., & Van Bavel, J. J. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, *4*(2), 96–111. https://doi.org/10.1038/s44159-024-00392-z

Finkel, E. J., Simpson, J. A., & Eastwick, P. W. (2017). The Psychology of Close Relationships: Fourteen Core Principles. *Annual Review of Psychology*, *68*(1), 383–411. https://doi.org/10.1146/annurev-psych-010416-044038

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906. https://doi.org/10.1037/0022-3514.38.6.889

Fiske, S. T. (1993). Social Cognition and Social Perception. *Annual Review of Psychology*, *44*(1), 155–194. https://doi.org/10.1146/annurev.ps.44.020193.001103

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fitzsimons, G. M., Finkel, E. J., & vanDellen, M. R. (2015). Transactive goal dynamics. *Psychological Review*, *122*(4), 648–673. https://doi.org/10.1037/a0039654

Forgas, J. P., & Bower, G. H. (1987). Mood effects on person-perception judgments. *Journal of Personality and Social Psychology*, *53*(1), 53–60. https://doi.org/10.1037/0022-3514.53.1.53

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, *117*(3), 522–559. https://doi.org/10.1037/pspa0000160

Fourie, M. M., Hortensius, R., & Decety, J. (2020). Parsing the components of forgiveness: Psychological and neural mechanisms. *Neuroscience & Biobehavioral Reviews*, *112*, 437–451. https://doi.org/10.1016/j.neubiorev.2020.02.020

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–279. https://doi.org/10.1037/a0022327

Freeman, J. B., Pauker, K., Apfelbaum, E. P., & Ambady, N. (2010). Continuous dynamics in the real-time perception of race. *Journal of Experimental Social Psychology*, *46*(1), 179–185. https://doi.org/10.1016/j.jesp.2009.10.002

Freeman, J. B., & Stolier, R. M. (2014). The medial prefrontal cortex in constructing personality models. *Trends in Cognitive Sciences*, *18*(11), 571–572. https://doi.org/10.1016/j.tics.2014.09.009

Frith, C. D., & Frith, U. (2006). How we predict what other people are going to do. *Brain Research*, *1079*(1), 36–46. https://doi.org/10.1016/j.brainres.2005.12.126

Gallagher, R. J., Stowell, E., Parker, A. G., & Foucault Welles, B. (2019). Reclaiming Stigmatized Narratives: The Networked Disclosure Landscape of #MeToo. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–30. https://doi.org/10.1145/3359198

Gannon, T. A., Rose, M. R., & Ward, T. (2008). A Descriptive Model of the Offense Process for Female Sexual Offenders. *Sexual Abuse*, *20*(3), 352–374. https://doi.org/10.1177/1079063208322495

Gantman, A. P., & Van Bavel, J. J. (2015). Moral Perception. *Trends in Cognitive Sciences*, *19*(11), 631–633. https://doi.org/10.1016/j.tics.2015.08.004

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal Stability of Implicit and Explicit Measures: A Longitudinal Analysis. *Personality and Social Psychology Bulletin*, *43*(3), 300–312. https://doi.org/10.1177/0146167216684131

Gawronski, B., Rydell, R. J., De Houwer, J., Brannon, S. M., Ye, Y., Vervliet, B., & Hu, X. (2018). Contextualized Attitude Change. In *Advances in Experimental Social Psychology* (Vol. 57, pp. 1–52). Elsevier. https://doi.org/10.1016/bs.aesp.2017.06.001

Gelman, A. (2005). Comment: Fuzzy and Bayesian p-Values and u-Values. *Statistical Science*, *20*(4). https://doi.org/10.1214/088342305000000368

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Geukes, K., Nestler, S., Hutteman, R., Küfner, A. C. P., & Back, M. D. (2017). Trait personality and state variability: Predicting individual differences in within- and cross-context fluctuations in affect, self-evaluations, and behavior in everyday life. *Journal of Research in Personality*, *69*, 124–138. https://doi.org/10.1016/j.jrp.2016.06.003

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology*, *54*(5), 733–740. https://doi.org/10.1037/0022-3514.54.5.733

Goldenberg, A., & Gross, J. J. (2020). Digital Emotion Contagion. *Trends in Cognitive Sciences*, *24*(4), 316–328. https://doi.org/10.1016/j.tics.2020.01.009

Golec-Staśkiewicz, K., Pluta, A., Wojciechowski, J., Okruszek, Ł., Haman, M., Wysocka, J., & Wolak, T. (2022). Does the TPJ fit it all? Representational similarity analysis of different forms of mentalizing. *Social Neuroscience*, *17*(5), 428–440. https://doi.org/10.1080/17470919.2022.2138536

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language Style Matching as a Predictor of Social Dynamics in Small Groups. *Communication Research*, *37*(1), 3–19. https://doi.org/10.1177/0093650209351468

Goodwin, G. P., & Darley, J. M. (2012). Why are some moral beliefs perceived to be more objective than others? *Journal of Experimental Social Psychology*, *48*(1), 250–256. https://doi.org/10.1016/j.jesp.2011.08.006

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org/10.1037/a0034726

Gosling, S. D., & Mason, W. (2015). Internet Research in Psychology. *Annual Review of Psychology*, *66*(1), 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*(1), 1–20. https://doi.org/10.1037/0022-3514.90.1.1

Gruber, M. J., Ritchey, M., Wang, S.-F., Doss, M. K., & Ranganath, C. (2016). Post-learning Hippocampal Dynamics Promote Preferential Retention of Rewarding Events. *Neuron*, *89*(5), 1110–1120. https://doi.org/10.1016/j.neuron.2016.01.017

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. M. (2007). The integrated trait–state model. *Journal of Research in Personality*, *41*(2), 295–315. https://doi.org/10.1016/j.jrp.2006.04.003

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2014). Imagine All the People: How the Brain Creates and Uses Personality Models to Predict Behavior. *Cerebral Cortex*, *24*(8), 1979–1987. https://doi.org/10.1093/cercor/bht042

Hasson, U., Furman, O., Clark, D., Dudai, Y., & Davachi, L. (2008). Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding. *Neuron*, *57*(3), 452–462. https://doi.org/10.1016/j.neuron.2007.12.009

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Hayes, S. C., & Sanford, B. T. (2014). Cooperation came first: Evolution and human cognition: COOPERATION AND COGNITION. *Journal of the Experimental Analysis of Behavior*, *101*(1), 112–129. https://doi.org/10.1002/jeab.64

Heath, A., Fisher, S., & Smith, S. (2005). The globalization of public opinion research. *Annual Review of Political Science*, *8*(1), 297–333. https://doi.org/10.1146/annurev.polisci.8.090203.103000

Heck, P. R., & Krueger, J. I. (2016). Social Perception of Self-Enhancement Bias and Error. *Social Psychology*, *47*(6), 327–339. https://doi.org/10.1027/1864-9335/a000287

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A Framework for Teachable Collaborative Problem Solving Skills. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 37–56). Springer Netherlands. https://doi.org/10.1007/978-94-017-9395-7_2

Hewstone, M. (1990). The ‘ultimate attribution error’? A review of the literature on intergroup causal attribution. *European Journal of Social Psychology*, *20*(4), 311–335. https://doi.org/10.1002/ejsp.2420200404

Higgins, E. T., & Brendl, C. M. (1995). Accessibility and Applicability: Some “Activation Rules” Influencing Judgment. *Journal of Experimental Social Psychology*, *31*(3), 218–243. https://doi.org/10.1006/jesp.1995.1011

Hiser, J., & Koenigs, M. (2018). The Multifaceted Role of the Ventromedial Prefrontal Cortex in Emotion, Decision Making, Social Cognition, and Psychopathology. *Biological Psychiatry*, *83*(8), 638–647. https://doi.org/10.1016/j.biopsych.2017.10.030

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, *39*(1), 101–117. https://doi.org/10.3758/BF03192848

Howard, J. W., & Rothbart, M. (1980). Social categorization and memory for in-group and out-group behavior. *Journal of Personality and Social Psychology*, *38*(2), 301–310. https://doi.org/10.1037/0022-3514.38.2.301

Hughes, B. L., Zaki, J., & Ambady, N. (2017). Motivation alters impression formation and related neural systems. *Social Cognitive and Affective Neuroscience*, *12*(1), 49–60. https://doi.org/10.1093/scan/nsw147

Human, L. J., Carlson, E. N., Geukes, K., Nestler, S., & Back, M. D. (2020). Do accurate personality impressions benefit early relationship development? The bidirectional associations between accuracy and liking. *Journal of Personality and Social Psychology*, *118*(1), 199–212. https://doi.org/10.1037/pspp0000214

Humberg, S., Dufner, M., Schönbrodt, F. D., Geukes, K., Hutteman, R., Van Zalk, M. H. W., Denissen, J. J. A., Nestler, S., & Back, M. D. (2018). Enhanced versus simply positive: A new condition-based regression analysis to disentangle effects of self-enhancement from effects of positivity of self-view. *Journal of Personality and Social Psychology*, *114*(2), 303–322. https://doi.org/10.1037/pspp0000134

Hung, W. (2013). Team-based complex problem solving: A collective cognition perspective. *Educational Technology Research and Development*, *61*(3), 365–384. https://doi.org/10.1007/s11423-013-9296-3

Idson, L. C., & Mischel, W. (2001). The personality of familiar and significant people: The lay perceiver as a social–cognitive theorist. *Journal of Personality and Social Psychology*, *80*(4), 585–596. https://doi.org/10.1037/0022-3514.80.4.585

Ince, J., Rojas, F., & Davis, C. A. (2017). The social media response to Black Lives Matter: How Twitter users interact with Black Lives Matter through hashtag use. *Ethnic and Racial Studies*, *40*(11), 1814–1830. https://doi.org/10.1080/01419870.2017.1334931

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). From Text to Thought: How Analyzing Language Can Advance Psychological Science. *Perspectives on Psychological Science*, *17*(3), 805–826. https://doi.org/10.1177/17456916211004899

Jacoby, N., Landau-Wells, M., Pearl, J., Paul, A., Falk, E. B., Bruneau, E. G., & Ochsner, K. N. (2024). Partisans process policy-based and identity-based messages using dissociable neural systems. *Cerebral Cortex*, *34*(9), bhae368. https://doi.org/10.1093/cercor/bhae368

Jimenez, C. A., & Meyer, M. L. (2024). The dorsomedial prefrontal cortex prioritizes social learning during rest. *Proceedings of the National Academy of Sciences*, *121*(12), e2309232121. https://doi.org/10.1073/pnas.2309232121

Joel, S., Eastwick, P. W., Allison, C. J., Arriaga, X. B., Baker, Z. G., Bar-Kalifa, E., Bergeron, S., Birnbaum, G. E., Brock, R. L., Brumbaugh, C. C., Carmichael, C. L., Chen, S., Clarke, J., Cobb, R. J., Coolsen, M. K., Davis, J., de Jong, D. C., Debrot, A., DeHaas, E. C., … Wolf, S. (2020). Machine learning uncovers the most robust self-report predictors of relationship quality across 43 longitudinal couples studies. *Proceedings of the National Academy of Sciences*, *117*(32), 19061–19071. https://doi.org/10.1073/pnas.1917036117

Jones, N. M., Wojcik, S. P., Sweeting, J., & Silver, R. C. (2016). Tweeting negative emotion: An investigation of Twitter data in the aftermath of violence on college campuses. *Psychological Methods*, *21*(4), 526–541. https://doi.org/10.1037/met0000099

Jussim, L., Nelson, T. E., Manis, M., & Soffin, S. (1995). Prejudice, Stereotypes, and Labeling Effects: Sources of Bias in Person Perception. *Journal of Personality and Social Psychology*, *68*(2), 228–246.

Kachen, A., Krishen, A. S., Petrescu, M., Gill, R. D., & Peter, P. C. (2021). #MeToo, #MeThree, #MeFour: Twitter as community building across academic and corporate institutions. *Psychology & Marketing*, *38*(3), 455–469. https://doi.org/10.1002/mar.21442

Kane, A. A., & Van Swol, L. M. (2023). Using linguistic inquiry and word count software to analyze group interaction language data. *Group Dynamics: Theory, Research, and Practice*, *27*(3), 188–201. https://doi.org/10.1037/gdn0000195

Kantor, J., & Twohey, M. (2017). Harvey Weinstein Paid Off Sexual Harassment Accusers for Decades. *The New York Times*. https://www.nytimes.com/2017/10/05/us/harvey-weinstein-harassment-allegations.html

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology*, *81*(5), 774–788. https://doi.org/10.1037/0022-3514.81.5.774

Kashdan, T. B., & Rottenberg, J. (2010). Psychological flexibility as a fundamental aspect of health. *Clinical Psychology Review*, *30*(7), 865–878. https://doi.org/10.1016/j.cpr.2010.03.001

Kenny, D. A. (2004). PERSON: A General Model of Interpersonal Perception. *Personality and Social Psychology Review*, *8*(3), 265–280. https://doi.org/10.1207/s15327957pspr0803_3

Kenny, D. A., & Acitelli, L. K. (2001). Accuracy and bias in the perception of the partner in a close relationship. *Journal of Personality and Social Psychology*, *80*(3), 439–448. https://doi.org/10.1037/0022-3514.80.3.439

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods*, *21*(4), 507–525. https://doi.org/10.1037/met0000091

Ki, J. J., Kelly, S. P., & Parra, L. C. (2016). Attention Strongly Modulates Reliability of Neural Responses to Naturalistic Narrative Stimuli. *Journal of Neuroscience*, *36*(10), 3092–3101. https://doi.org/10.1523/JNEUROSCI.2942-15.2016

Kihlstrom, J. F. (2013). The person-situation interaction. In *The Oxford handbook of social cognition.* (pp. 786–805). Oxford University 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. https://doi.org/10.1093/cercor/bhaa263

Kim, M., Park, B., & Young, L. (2020). The Psychology of Motivated versus Rational Impression Updating. *Trends in Cognitive Sciences*, *24*(2), 101–111. https://doi.org/10.1016/j.tics.2019.12.001

Kjell, O. N. E., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, *24*(1), 92–115. https://doi.org/10.1037/met0000191

Klein, S. B., Sherman, J. W., & Loftus, J. (1996). The Role of Episodic and Semantic Memory in the Development of Trait Self-Knowledge. *Social Cognition*, *14*(4), 277–291. https://doi.org/10.1521/soco.1996.14.4.277

Klein, W. M., & Kunda, Z. (1992). Motivated person perception: Constructing justifications for desired beliefs. *Journal of Experimental Social Psychology*, *28*(2), 145–168. https://doi.org/10.1016/0022-1031(92)90036-J

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 34–53. https://doi.org/10.1037/0278-7393.27.1.34

Körner, R., & Altmann, T. (2023). Personality is related to satisfaction in friendship dyads, but similarity is not: Understanding the links between the big five and friendship satisfaction using actor-partner interdependence models. *Journal of Research in Personality*, *107*, 104436. https://doi.org/10.1016/j.jrp.2023.104436

Kovács, G. (2020). Getting to Know Someone: Familiarity, Person Recognition, and Identification in the Human Brain. *Journal of Cognitive Neuroscience*, *32*(12), 2205–2225. https://doi.org/10.1162/jocn_a_01627

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. https://doi.org/10.3389/neuro.06.004.2008

Kross, E., & Ayduk, O. (2011). Making Meaning out of Negative Experiences by Self-Distancing. *Current Directions in Psychological Science*, *20*(3), 187–191. https://doi.org/10.1177/0963721411408883

Kross, E., Verduyn, P., Boyer, M., Drake, B., Gainsburg, I., Vickers, B., Ybarra, O., & Jonides, J. (2019). Does counting emotion words on online social networks provide a window into people's subjective experience of emotion? A case study on Facebook. *Emotion*, *19*(1), 97–107. https://doi.org/10.1037/emo0000416

Kruse, F., & Degner, J. (2021). Spontaneous state inferences. *Journal of Personality and Social Psychology*, *121*(4), 774–791. https://doi.org/10.1037/pspa0000232

Kube, T., & Rozenkrantz, L. (2021). When Beliefs Face Reality: An Integrative Review of Belief Updating in Mental Health and Illness. *Perspectives on Psychological Science*, *16*(2), 247–274. https://doi.org/10.1177/1745691620931496

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kurdi, B., Mann, T. C., & Ferguson, M. J. (2022). Persuading the Implicit Mind: Changing Negative Implicit Evaluations With an 8-Minute Podcast. *Social Psychological and Personality Science*, *13*(3), 688–697. https://doi.org/10.1177/19485506211037140

Kurdi, B., Morehouse, K. N., & Dunham, Y. (2023). How do explicit and implicit evaluations shift? A preregistered meta-analysis of the effects of co-occurrence and relational information. *Journal of Personality and Social Psychology*, *124*(6), 1174–1202. https://doi.org/10.1037/pspa0000329

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., … Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016. https://doi.org/10.1037/xge0000179

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. https://doi.org/10.1177/0146167216655984

Langeslag, S. J. E., & van Strien, J. W. (2019). Romantic love and attention: Early and late event-related potentials. *Biological Psychology*, *146*, 107737. https://doi.org/10.1016/j.biopsycho.2019.107737

Larson, G. M., Faure, R., Righetti, F., & Hofmann, W. (2022). How do implicit and explicit partner evaluations update in daily life? Evidence from the lab and the field. *Journal of Experimental Psychology: General*, *151*(10), 2511–2533. https://doi.org/10.1037/xge0001199

Leising, D., Ostrovski, O., & Zimmermann, J. (2013). "Are We Talking About the Same Person Here?": Interrater Agreement in Judgments of Personality Varies Dramatically With How Much the Perceivers Like the Targets. *Social Psychological and Personality Science*, *4*(4), 468–474. https://doi.org/10.1177/1948550612462414

Lemay, E. P., & Clark, M. S. (2015). Motivated cognition in relationships. *Current Opinion in Psychology*, *1*, 72–75. https://doi.org/10.1016/j.copsyc.2014.11.002

Leong, Y. C., Chen, J., Willer, R., & Zaki, J. (2020). Conservative and liberal attitudes drive polarized neural responses to political content. *Proceedings of the National Academy of Sciences*, *117*(44), 27731–27739. https://doi.org/10.1073/pnas.2008530117

Lieberman, M. D., Straccia, M. A., Meyer, M. L., Du, M., & Tan, K. M. (2019). Social, self, (situational), and affective processes in medial prefrontal cortex (MPFC): Causal, multivariate, and reverse inference evidence. *Neuroscience & Biobehavioral Reviews*, *99*, 311–328. https://doi.org/10.1016/j.neubiorev.2018.12.021

Locker, L., Hoffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods*, *39*(4), 723–730. https://doi.org/10.3758/BF03192962

Luhmann, M. (2017). Using Big Data to study subjective well-being. *Current Opinion in Behavioral Sciences*, *18*, 28–33. https://doi.org/10.1016/j.cobeha.2017.07.006

Lyons, M., Aksayli, N. D., & Brewer, G. (2018). Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior*, *87*, 207–211. https://doi.org/10.1016/j.chb.2018.05.035

Ma, N., Baetens, K., Vandekerckhove, M., Kestemont, J., Fias, W., & Van Overwalle, F. (2014). Traits are represented in the medial prefrontal cortex: An fMRI adaptation study. *Social Cognitive and Affective Neuroscience*, *9*(8), 1185–1192. https://doi.org/10.1093/scan/nst098

Malle, B. F., Knobe, J. M., & Nelson, S. E. (2007). Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology*, *93*(4), 491–514. https://doi.org/10.1037/0022-3514.93.4.491

Malle, B. F., & Pearce, G. E. (2001). Attention to behavioral events during interaction: Two actor–observer gaps and three attempts to close them. *Journal of Personality and Social Psychology*, *81*(2), 278–294. https://doi.org/10.1037/0022-3514.81.2.278

Mann, T. C., Kurdi, B., & Banaji, M. R. (2020). How effectively can implicit evaluations be updated? Using evaluative statements after aversive repeated evaluative pairings. *Journal of Experimental Psychology: General*, *149*(6), 1169–1192. https://doi.org/10.1037/xge0000701

Maryn, A. G., & Dover, T. L. (2023). Who gets canceled? Twitter responses to gender-based violence allegations. *Psychology of Violence*, *13*(2), 117–126. https://doi.org/10.1037/vio0000436

Mathieu, J. E., Gallagher, P. T., Domingo, M. A., & Klock, E. A. (2019). Embracing Complexity: Reviewing the Past Decade of Team Effectiveness Research. *Annual Review of Organizational Psychology and Organizational Behavior*, *6*(1), 17–46. https://doi.org/10.1146/annurev-orgpsych-012218-015106

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, *12*(1), 23–44. https://doi.org/10.1037/1082-989X.12.1.23

McCullough, M. E. (2001). Forgiveness: Who Does It and How Do They Do It? *Current Directions in Psychological Science*, *10*(6), 194–197. https://doi.org/10.1111/1467-8721.00147

Meltzer, A. L., & McNulty, J. K. (2019). Relationship formation and early romantic relationships. In D. Schoebi & B. Campos (Eds.), *New Directions in the Psychology of Close Relationships* (1st ed., pp. 9–27). Routledge. https://doi.org/10.4324/9781351136266-2

Meltzoff, A. N. (1988). Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. *Developmental Psychology*, *24*(4), 470–476. https://doi.org/10.1037/0012-1649.24.4.470

Mende-Siedlecki, P. (2018). Changing our minds: The neural bases of dynamic impression updating. *Current Opinion in Psychology*, *24*, 72–76. https://doi.org/10.1016/j.copsyc.2018.08.007

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. https://doi.org/10.1523/JNEUROSCI.2334-13.2013

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, *8*(6), 623–631. https://doi.org/10.1093/scan/nss040

Mende-Siedlecki, P., & Todorov, A. (2016). Neural dissociations between meaningful and mere inconsistency in impression updating. *Social Cognitive and Affective Neuroscience*, *11*(9), 1489–1500. https://doi.org/10.1093/scan/nsw058

Mendoza-Denton, R., Downey, G., Purdie, V. J., Davis, A., & Pietrzak, J. (2002). Sensitivity to status-based rejection: Implications for African American students' college experience. *Journal of Personality and Social Psychology*, *83*(4), 896–918. https://doi.org/10.1037/0022-3514.83.4.896

Mesquiti, S., Seraj, S., Weyland, A. H., Ashokkumar, A., Boyd, R. L., Mihalcea, R., & Pennebaker, J. W. (2025). Analysis of social media language reveals the psychological interaction of three successive upheavals. *Scientific Reports*, *15*(1), 5740. https://doi.org/10.1038/s41598-025-89165-z

Metzler, H., Rimé, B., Pellert, M., Niederkrotenthaler, T., Di Natale, A., & Garcia, D. (2023). Collective emotions during the COVID-19 outbreak. *Emotion*, *23*(3), 844–858. https://doi.org/10.1037/emo0001111

Meyer, M. L., Davachi, L., Ochsner, K. N., & Lieberman, M. D. (2019). Evidence That Default Network Connectivity During Rest Consolidates Social Information. *Cerebral Cortex*, *29*(5), 1910–1920. https://doi.org/10.1093/cercor/bhy071

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The Link between Social Cognition and Self-referential Thought in the Medial Prefrontal Cortex. *Journal of Cognitive Neuroscience*, *17*(8), 1306–1315. https://doi.org/10.1162/0898929055002418

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable Medial Prefrontal Contributions to Judgments of Similar and Dissimilar Others. *Neuron*, *50*(4), 655–663. https://doi.org/10.1016/j.neuron.2006.03.040

Mitchell, J. P., Neil Macrae, C., & Banaji, M. R. (2005). Forming impressions of people versus inanimate objects: Social-cognitive processing in the medial prefrontal cortex. *NeuroImage*, *26*(1), 251–257. https://doi.org/10.1016/j.neuroimage.2005.01.031

Mohammad, S. M. (2016). Sentiment Analysis. In *Emotion Measurement* (pp. 201–237). Elsevier. https://doi.org/10.1016/B978-0-08-100508-8.00009-6

Montoya, R. M., & Horton, R. S. (2013). A meta-analytic investigation of the processes underlying the similarity-attraction effect. *Journal of Social and Personal Relationships*, *30*(1), 64–94. https://doi.org/10.1177/0265407512452989

Montoya, R. M., Horton, R. S., Vevea, J. L., Citkowicz, M., & Lauber, E. A. (2017). A re-examination of the mere exposure effect: The influence of repeated exposure on recognition, familiarity, and liking. *Psychological Bulletin*, *143*(5), 459–498. https://doi.org/10.1037/bul0000085

Moreland, R. L., & Zajonc, R. B. (1982). Exposure effects in person perception: Familiarity, similarity, and attraction. *Journal of Experimental Social Psychology*, *18*(5), 395–415. https://doi.org/10.1016/0022-1031(82)90062-2

Morelli, S. A., Leong, Y. C., Carlson, R. W., Kullar, M., & Zaki, J. (2018). Neural detection of socially valued community members. *Proceedings of the National Academy of Sciences*, *115*(32), 8149–8154. https://doi.org/10.1073/pnas.1712811115

Morry, M. M. (2007). The attraction-similarity hypothesis among cross-sex friends: Relationship satisfaction, perceived similarities, and self-serving perceptions. *Journal of Social and Personal Relationships*, *24*(1), 117–138. https://doi.org/10.1177/0265407507072615

Morry, M. M., Kito, M., & Ortiz, L. (2011). The attraction–similarity model and dating couples: Projection, perceived similarity, and psychological benefits. *Personal Relationships*, *18*(1), 125–143. https://doi.org/10.1111/j.1475-6811.2010.01293.x

Morry, M. M., Reich, T., & Kito, M. (2010). How Do I See You Relative to Myself? Relationship Quality as a Predictor of Self- and Partner-Enhancement Within Cross-Sex Friendships, Dating Relationships, and Marriages. *The Journal of Social Psychology*, *150*(4), 369–392. https://doi.org/10.1080/00224540903365471

Moskowitz, G. B., Olcaysoy Okten, I., & Schneid, E. (2022). The Updating of First Impressions. In E. Balcetis & G. B. Moskowitz, *The Handbook of Impression Formation* (1st ed., pp. 348–392). Routledge. https://doi.org/10.4324/9781003045687-21

Murphy, S. C. (2017). A Hands-On Guide to Conducting Psychological Research on Twitter. *Social Psychological and Personality Science*, *8*(4), 396–412. https://doi.org/10.1177/1948550617697178

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review*, *110*(3), 472–489. https://doi.org/10.1037/0033-295X.110.3.472

Nakamura, K., Arai, S., & Kawabata, H. (2017). Prioritized Identification of Attractive and Romantic Partner Faces in Rapid Serial Visual Presentation. *Archives of Sexual Behavior*, *46*(8), 2327–2338. https://doi.org/10.1007/s10508-017-1027-0

Neubert, J. C., Mainert, J., Kretzschmar, A., & Greiff, S. (2015). The Assessment of 21st Century Skills in Industrial and Organizational Psychology: Complex and Collaborative Problem Solving. *Industrial and Organizational Psychology*, *8*(2), 238–268. https://doi.org/10.1017/iop.2015.14

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv:1103.2903 [Cs]*. http://arxiv.org/abs/1103.2903

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

North, A., Grady, C., McGann, L., & Romano, A. (2020). 262 celebrities, politicians, CEOs, and others who have been accused of sexual misconduct since April 2017. *Vox*. https://www.vox.com/a/sexual-harassment-assault-allegations-list

Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The Enigmatic temporal pole: A review of findings on social and emotional processing. *Brain*, *130*(7), 1718–1731. https://doi.org/10.1093/brain/awm052

Otten, S. (2016). The Minimal Group Paradigm and its maximal impact in research on social categorization. *Current Opinion in Psychology*, *11*, 85–89. https://doi.org/10.1016/j.copsyc.2016.06.010

Park, B., Fareri, D., Delgado, M., & Young, L. (2021). The role of right temporoparietal junction in processing social prediction error across relationship contexts. *Social Cognitive and Affective Neuroscience*, *16*(8), 772–781. https://doi.org/10.1093/scan/nsaa072

Park, B., & Young, L. (2020). An association between biased impression updating and relationship facilitation: A behavioral and fMRI investigation. *Journal of Experimental Social Psychology*, *87*, 103916. https://doi.org/10.1016/j.jesp.2019.103916

Park, S. A., Miller, D. S., & Boorman, E. D. (2021). Inferences on a multidimensional social hierarchy use a grid-like code. *Nature Neuroscience*, *24*(9), 1292–1301. https://doi.org/10.1038/s41593-021-00916-3

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2017). Spontaneous neural encoding of social network position. *Nature Human Behaviour*, *1*(5), 0072. https://doi.org/10.1038/s41562-017-0072

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2018). Similar neural responses predict friendship. *Nature Communications*, *9*(1), 332. https://doi.org/10.1038/s41467-017-02722-7

Pennebaker, J. W. (1997). Writing About Emotional Experiences as a Therapeutic Process. *Psychological Science*, *8*(3), 162–166. https://doi.org/10.1111/j.1467-9280.1997.tb00403.x

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. 26.

Pennebaker, J. W., & Chung, C. K. (2013). Counting little words in big data. *Social Cognition and Communication*, 25–42.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, *54*(Volume 54, 2003), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Piazza, J., Sousa, P., Rottman, J., & Syropoulos, S. (2019). Which Appraisals Are Foundational to Moral Judgment? Harm, Injustice, and Beyond. *Social Psychological and Personality Science*, *10*(7), 903–913. https://doi.org/10.1177/1948550618801326

Popal, H., Wang, Y., & Olson, I. R. (2019). A Guide to Representational Similarity Analysis for Social Neuroscience. *Social Cognitive and Affective Neuroscience*, *14*(11), 1243–1253. https://doi.org/10.1093/scan/nsz099

Poppenk, J., Köhler, S., & Moscovitch, M. (2010). Revisiting the novelty effect: When familiarity, not novelty, enhances memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1321–1330. https://doi.org/10.1037/a0019900

Proulx, C. M., Helms, H. M., & Buehler, C. (2007). Marital Quality and Personal Well-Being: A Meta-Analysis. *Journal of Marriage and Family*, *69*(3), 576–593. https://doi.org/10.1111/j.1741-3737.2007.00393.x

Pryor, J. B., LaVite, C. M., & Stoller, L. M. (1993). A Social Psychological Analysis of Sexual Harassment: The Person/Situation Interaction. *Journal of Vocational Behavior*, *42*(1), 68–83. https://doi.org/10.1006/jvbe.1993.1005

Rajapakse, T. (2020). *Simple Transformers* [Python]. https://simpletransformers.ai

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), e2024292118. https://doi.org/10.1073/pnas.2024292118

Reeder, G. D., & Coovert, M. D. (1986). Revising an Impression of Morality. *Social Cognition*, *4*(1), 1–17. https://doi.org/10.1521/soco.1986.4.1.1

Reeder, G. D., Messick, D. M., & Van Avermaet, E. (1977). Dimensional asymmetry in attributional inference. *Journal of Experimental Social Psychology*, *13*(1), 46–57. https://doi.org/10.1016/0022-1031(77)90012-9

Reeder, G. D., & Spores, J. M. (1983). The attribution of morality. *Journal of Personality and Social Psychology*, *44*(4), 736–745. https://doi.org/10.1037/0022-3514.44.4.736

Reis, H. T., Maniaci, M. R., Caprariello, P. A., Eastwick, P. W., & Finkel, E. J. (2011). Familiarity does indeed promote attraction in live interaction. *Journal of Personality and Social Psychology*, *101*(3), 557–570. https://doi.org/10.1037/a0022885

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rosenberg, M. (1965). Rosenberg self-esteem scale. *Journal of Religion and Health*.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rothbart, M., Fulero, S., Jensen, C., Howard, J., & Birrell, P. (1978). From individual to group impressions: Availability heuristics in stereotype formation. *Journal of Experimental Social Psychology*, *14*(3), 237–255. https://doi.org/10.1016/0022-1031(78)90013-6

Roy, M., Shohamy, D., & Wager, T. D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in Cognitive Sciences*, *16*(3), 147–156. https://doi.org/10.1016/j.tics.2012.01.005

Russell, D., Peplau, L. A., & Ferguson, M. L. (1978). Developing a measure of loneliness. *Journal of Personality Assessment*, *42*(3), 290–294.

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*(6), 995–1008. https://doi.org/10.1037/0022-3514.91.6.995

Saegert, S., Swap, W., & Zajonc, R. B. (1973). Exposure, context, and interpersonal attraction. *Journal of Personality and Social Psychology*, *25*(2), 234–242. https://doi.org/10.1037/h0033965

Sahi, R. S., & Eisenberger, N. I. (2021). Why Don't You Like Me? The Role of the Mentalizing Network in Social Rejection. In M. Gilead & K. N. Ochsner (Eds.), *The Neural Basis of*

Mentalizing (pp. 613–628). Springer International Publishing. https://doi.org/10.1007/978-3-030-51890-5_32

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv:1910.01108 [Cs]*. http://arxiv.org/abs/1910.01108

Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., & Wright, B. (2020). Sarcasm detection using machine learning algorithms in Twitter: A systematic review. *International Journal of Market Research*, *62*(5), 578–598. https://doi.org/10.1177/1470785320921779

Saxe, R., & Kanwisher, N. (2004). People Thinking about Thinking People: The Role of the Temporo-Parietal Junction in "Theory of Mind." In *Social Neuroscience*. Psychology Press.

Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C., & Norman, K. A. (2018). Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nature Communications*, *9*(1), 3920. https://doi.org/10.1038/s41467-018-06213-1

Schiller, B., Baumgartner, T., & Knoch, D. (2014). Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. *Evolution and Human Behavior*, *35*(3), 169–175. https://doi.org/10.1016/j.evolhumbehav.2013.12.006

Schiller, D., Freeman, J. B., Mitchell, J. P., Uleman, J. S., & Phelps, E. A. (2009). A neural mechanism of first impressions. *Nature Neuroscience*, *12*(4), 508–514. https://doi.org/10.1038/nn.2278

Schöne, J. P., Parkinson, B., & Goldenberg, A. (2021). Negativity Spreads More than Positivity on Twitter After Both Positive and Negative Political Situations. *Affective Science*, *2*(4), 379–390. https://doi.org/10.1007/s42761-021-00057-7

Schuck, N. W., & Niv, Y. (2019). Sequential replay of nonspatial task states in the human hippocampus. *Science*, *364*(6447), eaaw5181. https://doi.org/10.1126/science.aaw5181

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, *61*(2), 195–202. https://doi.org/10.1037/0022-3514.61.2.195

Sebanz, N., & Knoblich, G. (2009). Prediction in Joint Action: What, When, and Where. *Topics in Cognitive Science*, *1*(2), 353–367. https://doi.org/10.1111/j.1756-8765.2009.01024.x

Sedikides, C., & Gregg, A. P. (2007). Portraits of the Self. In M. Hogg & J. Cooper, *The SAGE Handbook of Social Psychology: Concise Student Edition* (pp. 93–122). SAGE Publications Ltd. https://doi.org/10.4135/9781848608221.n5

Sedikides, C., & Gregg, A. P. (2008). Self-Enhancement: Food for Thought. *Perspectives on Psychological Science*, *3*(2), 102–116. https://doi.org/10.1111/j.1745-6916.2008.00068.x

Sels, L., Cabrieto, J., Butler, E., Reis, H., Ceulemans, E., & Kuppens, P. (2020). The occurrence and correlates of emotional interdependence in romantic relationships. *Journal of Personality and Social Psychology*, *119*(1), 136–158. https://doi.org/10.1037/pspi0000212

Settanni, M., & Marengo, D. (2015). Sharing feelings online: Studying emotional well-being via automated text analysis of Facebook posts. *Frontiers in Psychology*, *6*, 1045.

Sharot, T., Rollwage, M., Sunstein, C. R., & Fleming, S. M. (2023). Why and When Beliefs Change. *Perspectives on Psychological Science*, *18*(1), 142–151. https://doi.org/10.1177/17456916221082967

Shen, X., & Ferguson, M. J. (2021). How resistant are implicit impressions of facial trustworthiness? When new evidence leads to durable updating. *Journal of Experimental Social Psychology*, *97*, 104219. https://doi.org/10.1016/j.jesp.2021.104219

Shen, X., Tokoglu, F., Papademetris, X., & Constable, R. T. (2013). Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *NeuroImage*, *82*, 403–415. https://doi.org/10.1016/j.neuroimage.2013.05.081

Sherman, J. W., & Bessenoff, G. R. (1999). Stereotypes as Source-Monitoring Cues: On the Interaction Between Episodic and Semantic Memory. *Psychological Science*, *10*(2), 106–110. https://doi.org/10.1111/1467-9280.00116

Shulman, G. L., Astafiev, S. V., McAvoy, M. P., d'Avossa, G., & Corbetta, M. (2007). Right TPJ deactivation during visual search: Functional significance and support for a filter hypothesis. *Cerebral Cortex (New York, N.Y.: 1991)*, *17*(11), 2625–2633. https://doi.org/10.1093/cercor/bhl170

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. https://doi.org/10.1038/s41562-018-0425-1

Silver, B. M., & Ochsner, K. N. (2024). Changes in Online Moral Discourse About Public Figures During #MeToo. *Affective Science*, *5*(4), 346–357. https://doi.org/10.1007/s42761-024-00250-4

Simchon, A., Guntuku, S. C., Simhon, R., Ungar, L. H., Hassin, R. R., & Gilead, M. (2020). Political depression? A big-data, multimethod investigation of Americans' emotional response to the Trump presidency. *Journal of Experimental Psychology: General*, *149*(11), 2154–2168. https://doi.org/10.1037/xge0000767

Sisco, M. R., Bosetti, V., & Weber, E. U. (2017). When do extreme weather events generate attention to climate change? *Climatic Change*, *143*(1–2), 227–241. https://doi.org/10.1007/s10584-017-1984-2

Somerville, L. H., Heatherton, T. F., & Kelley, W. M. (2006). Anterior cingulate cortex responds differentially to expectancy violation and social rejection. *Nature Neuroscience*, *9*(8), 1007–1008. https://doi.org/10.1038/nn1728

Somerville, L. H., Kelley, W. M., & Heatherton, T. F. (2010). Self-esteem Modulates Medial Prefrontal Cortical Responses to Evaluative Social Feedback. *Cerebral Cortex*, *20*(12), 3005–3013. https://doi.org/10.1093/cercor/bhq049

Soral, W., & Kofta, M. (2020). Differential Effects of Competence and Morality on Self-Esteem at the Individual and the Collective Level. *Social Psychology*, *51*(3), 183–198. https://doi.org/10.1027/1864-9335/a000410

Springer, A., de C. Hamilton, A. F., & Cross, E. S. (2012). Simulating and predicting others' actions. *Psychological Research*, *76*(4), 383–387. https://doi.org/10.1007/s00426-012-0443-y

Stangor, C., & Ruble, D. N. (1989). Strength of expectancies and memory for social information: What we remember depends on how much we know. *Journal of Experimental Social Psychology*, *25*(1), 18–35. https://doi.org/10.1016/0022-1031(89)90037-1

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences*, *108*(19), 7710–7715. https://doi.org/10.1073/pnas.1014345108

Staresina, B. P., Alink, A., Kriegeskorte, N., & Henson, R. N. (2013). Awake reactivation predicts memory in humans. *Proceedings of the National Academy of Sciences*, *110*(52), 21159–21164. https://doi.org/10.1073/pnas.1311989110

Staresina, B. P., Henson, R. N. A., Kriegeskorte, N., & Alink, A. (2012). Episodic Reinstatement in the Medial Temporal Lobe. *Journal of Neuroscience*, *32*(50), 18150–18156. https://doi.org/10.1523/JNEUROSCI.4156-12.2012

Stepanikova, I. (2012). Racial-Ethnic Biases, Time Pressure, and Medical Decisions. *Journal of Health and Social Behavior*, *53*(3), 329–343. https://doi.org/10.1177/0022146512445807

Stewart, A. E. B., Vrzakova, H., Sun, C., Yonehiro, J., Stone, C. A., Duran, N. D., Shute, V., & D'Mello, S. K. (2019). I Say, You Say, We Say: Using Spoken Language to Model Socio-Cognitive Processes during Computer-Supported Collaborative Problem Solving. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–19. https://doi.org/10.1145/3359296

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences*, *115*(37), 9210–9215. https://doi.org/10.1073/pnas.1807222115

Strauss, J. P., Barrick, M. R., & Connerley, M. L. (2001). An investigation of personality similarity effects (relational and perceived) on peer and supervisor ratings and the role of familiarity and liking. *Journal of Occupational and Organizational Psychology*, *74*(5), 637–657. https://doi.org/10.1348/096317901167569

Sykora, M., Elayan, S., & Jackson, T. W. (2020). A qualitative analysis of sarcasm, irony and related #hashtags on Twitter. *Big Data & Society*, *7*(2), 205395172097273. https://doi.org/10.1177/2053951720972735

Tambe, A. (2018). Reckoning with the Silences of #MeToo. *Feminist Studies*, *44*(1), 197. https://doi.org/10.15767/feministstudies.44.1.0197

Tambini, A., & Davachi, L. (2019). Awake Reactivation of Prior Experiences Consolidates Memories and Biases Cognition. *Trends in Cognitive Sciences*, *23*(10), 876–890. https://doi.org/10.1016/j.tics.2019.07.008

Tamir, D. I., & Mitchell, J. P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proceedings of the National Academy of Sciences*, *107*(24), 10827–10832. https://doi.org/10.1073/pnas.1003242107

Tamir, D. I., & Thornton, M. A. (2018). Modeling the Predictive Social Mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences*, *113*(1), 194–199. https://doi.org/10.1073/pnas.1511905112

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Thornton, M. A., & Mitchell, J. P. (2017). Consistent Neural Activity Patterns Represent Personally Familiar People. *Journal of Cognitive Neuroscience*, *29*(9), 1583–1594. https://doi.org/10.1162/jocn_a_01151

Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019). The Social Brain Automatically Predicts Others' Future Mental States. *Journal of Neuroscience*, *39*(1), 140–148. https://doi.org/10.1523/JNEUROSCI.1431-18.2018

Thorp, J. N., Gasser, C., Blessing, E., & Davachi, L. (2022). Data-Driven Clustering of Functional Signals Reveals Gradients in Processing Both within the Anterior Hippocampus and across Its Long Axis. *Journal of Neuroscience*, *42*(39), 7431–7441. https://doi.org/10.1523/JNEUROSCI.0269-22.2022

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology*, *66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two Key Steps in the Evolution of Human Cooperation: The Interdependence Hypothesis. *Current Anthropology*, *53*(6), 673–692. https://doi.org/10.1086/668207

Tompary, A., & Davachi, L. (2017). Consolidation Promotes the Emergence of Representational Overlap in the Hippocampus and Medial Prefrontal Cortex. *Neuron*, *96*(1), 228-241.e5. https://doi.org/10.1016/j.neuron.2017.09.005

Trope, Y. (1998). Dispositional bias in person perception: A hypothesis-testing perception. In J. M. Darley & J. Cooper (Eds.), *Attribution and social interaction: The legacy of Edward E. Jones.* (pp. 67–126). American Psychological Association. https://doi.org/10.1037/10286-002

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1), 178–185. https://doi.org/10.1609/icwsm.v4i1.14009

van Baar, J. M., Halpern, D. J., & FeldmanHall, O. (2021). Intolerance of uncertainty modulates brain-to-brain synchrony during politically polarized perception. *Proceedings of the National Academy of Sciences*, *118*(20), e2022491118. https://doi.org/10.1073/pnas.2022491118

van Kesteren, M. T. R., Beul, S. F., Takashima, A., Henson, R. N., Ruiter, D. J., & Fernández, G. (2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: From congruent to incongruent. *Neuropsychologia*, *51*(12), 2352–2359. https://doi.org/10.1016/j.neuropsychologia.2013.05.027

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage*, *48*(3), 564–584. https://doi.org/10.1016/j.neuroimage.2009.06.009

van Schie, C. C., Chiu, C.-D., Rombouts, S. A. R. B., Heiser, W. J., & Elzinga, B. M. (2018). When compliments do not hit but critiques do: An fMRI study into self-esteem and self-knowledge in processing social feedback. *Social Cognitive and Affective Neuroscience*, *13*(4), 404–417. https://doi.org/10.1093/scan/nsy014

Vetter, P., Butterworth, B., & Bahrami, B. (2011). A candidate for the attentional bottleneck: Set-size specific modulation of the right TPJ during attentive enumeration. *Journal of Cognitive Neuroscience*, *23*(3), 728–736. https://doi.org/10.1162/jocn.2010.21472

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Visconti di Oleggio Castello, M., Halchenko, Y. O., Guntupalli, J. S., Gors, J. D., & Gobbini, M. I. (2017). The neural representation of personally familiar and unfamiliar faces in the distributed system for face perception. *Scientific Reports*, *7*(1), 12237. https://doi.org/10.1038/s41598-017-12559-1

Vuletich, H. A., & Payne, B. K. (2019). Stability and Change in Implicit Bias. *Psychological Science*, *30*(6), 854–862. https://doi.org/10.1177/0956797619844270

Wagner, D. D., Chavez, R. S., & Broom, T. W. (2019). Decoding the neural representation of self and person knowledge with multivariate pattern analysis and data-driven approaches. *WIREs Cognitive Science*, *10*(1). https://doi.org/10.1002/wcs.1482

Wahlheim, C. N., & Zacks, J. M. (2024). Memory updating and the structure of event representations. *Trends in Cognitive Sciences*, *0*(0). https://doi.org/10.1016/j.tics.2024.11.008

Wallace, P. (2015, November 9). *The Psychology of the Internet*. Higher Education from Cambridge University Press; Cambridge University Press. https://doi.org/10.1017/CBO9781139940962

Wegner, D. M., & Giuliano, T. (1980). Arousal-Induced Attention to Self. *Journal of Personality and Social Psychology*, *38*(5), 719–726.

Wessels, N. M., Zimmermann, J., Biesanz, J. C., & Leising, D. (2020). Differential associations of knowing and liking with accuracy and positivity bias in person perception. *Journal of Personality and Social Psychology*, *118*(1), 149–171. https://doi.org/10.1037/pspp0000218

Willis, J., & Todorov, A. (2006). First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychological Science*, *17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Wojciszke, B. (2005). Morality and competence in person- and self-perception. *European Review of Social Psychology*, *16*(1), 155–188. https://doi.org/10.1080/10463280500229619

Wojciszke, B., Brycz, H., & Borkenau, P. (1993). Effects of information content and evaluative extremity on positivity and negativity biases. *Journal of Personality and Social Psychology*, *64*(3), 327–335. https://doi.org/10.1037/0022-3514.64.3.327

Wyer, N. A. (2010). You Never Get a Second Chance to Make a First (Implicit) Impression: The Role of Elaboration in the Formation and Revision of Implicit Impressions. *Social Cognition*, *28*(1), 1–19. https://doi.org/10.1521/soco.2010.28.1.1

Xiong, Y., Cho, M., & Boatwright, B. (2019). Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of Twitter during the #MeToo movement. *Public Relations Review*, *45*(1), 10–23. https://doi.org/10.1016/j.pubrev.2018.10.014

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635

Yu, W., Zadbood, A., Chanales, A. J. H., & Davachi, L. (2024). Repetition dynamically and rapidly increases cortical, but not hippocampal, offline reactivation. *Proceedings of the National Academy of Sciences*, *121*(40), e2405929121. https://doi.org/10.1073/pnas.2405929121

Yzerbyt, V. (2018). The Dimensional Compensation Model: Reality and strategic constraints on warmth and competence in intergroup perceptions. In *Agency and Communion in Social Psychology*. Routledge.

Yzerbyt, V. Y., Rogier, A., & Fiske, S. T. (1998). Group Entitativity and Social Attribution: On Translating Situational Constraints into Stereotypes. *Personality and Social Psychology Bulletin*, *24*(10), 1089–1103. https://doi.org/10.1177/01461672982410006

Zacharias, C. (2018). *Twint* [Python]. https://github.com/twintproject/twint

Zadbood, A., Nastase, S., Chen, J., Norman, K. A., & Hasson, U. (2022). Neural representations of naturalistic events are updated as our understanding of the past changes. *eLife*, *11*, e79045. https://doi.org/10.7554/eLife.79045

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*(2, Pt.2), 1–27. https://doi.org/10.1037/h0025848

Zajonc, R. B. (2001). Mere Exposure: A Gateway to the Subliminal. *CURRENT DIRECTIONS IN PSYCHOLOGICAL SCIENCE*, *10*(6).

Zaki, J. (2013). Cue Integration: A Common Framework for Social Cognition and Physical Perception. *Perspectives on Psychological Science*, *8*(3), 296–312. https://doi.org/10.1177/1745691613475454

Zaki, J. (2014). Empathy: A motivated account. *Psychological Bulletin*, *140*(6), 1608–1647. https://doi.org/10.1037/a0037679

Appendix A: Chapter 1 Supplemental Materials

Public Figure	**Date of Allegation**	**6 months prior to allegations**	**Initial 3 weeks after allegations**	**1 year after allegations**
Al Franken	11/16/17	10841	137350	3400
Alex Jones	2/28/18	18736	26983	31191
Andy Dick	10/31/17	490	3361	429
Aziz Ansari	1/13/18	1470	28572	443
Ben Affleck	10/10/17	2703	19761	3836
Bob Weinstein	10/17/17	3	2072	43
Brett Ratner	11/1/17	161	14306	65
Bruce Weber	1/13/18	361	3091	387
Bryan Singer	12/4/17	347	8058	611
Charlie Rose	11/20/17	1593	44274	764
Chris Hardwick	6/14/18	316	17336	275
Cody Wilson	9/19/18	135	4841	90
Dustin Hoffman	11/1/17	738	6519	504
Ed Westwick	11/7/17	1655	8866	149
Eric Greitens	1/10/18	343	3441	99
Eric Schneiderman	5/7/18	489	15684	105
Garrison Keillor	11/29/17	470	12531	243
George HW Bush	10/25/17	1639	13837	1319
George Takei	11/10/17	3752	15295	2198
Glenn Thrush	11/20/17	767	4405	126
Harvey Weinstein	10/5/17	423	265282	7850
James Franco	1/11/18	6152	21013	1825
James Levine	12/3/17	97	3228	60
James Toback	10/22/17	5	7532	12
Jeremy Piven	10/31/17	246	3897	132
John Conyers	11/20/17	153	30419	149
John Lasseter	11/21/17	182	6381	134
Junot Diaz	5/4/18	276	4026	91
Kevin Spacey	10/29/17	1041	131284	4133
Les Moonves	7/27/18	226	12175	244
Louis CK	11/9/17	1980	73558	1259
Mario Batali	12/11/17	512	9472	158
Mark Halperin	10/25/17	289	10067	65
Marshall Faulk	12/11/17	375	2877	529
Matt Lauer	11/29/17	483	111214	1142
Morgan Freeman	5/24/18	5015	47761	393
Morgan Spurlock	12/14/17	185	3353	57
Oliver Stone	10/12/17	2014	4438	594
Roy Price	10/12/17	236	5041	606
Ryan Lizza	12/11/17	48	2119	38
Ryan Seacrest	2/26/18	1074	13786	1428
Scott Baio	1/27/18	2139	10729	953

Stan Lee	1/9/18	7272	12002	10008
Steve Wynn	1/27/18	244	24836	686
Steven Seagal	11/9/17	901	4162	758
Sylvester Stallone	11/16/17	1679	4638	2717
Tavis Smiley	12/13/17	145	7474	67
TJ Miller	12/19/17	2335	4161	751
Tom Brokaw	4/26/18	642	9181	218
Trent Franks	12/7/17	154	9083	42
Total		**83532**	**1245772**	**83376**

Table A.1.1: Public figures and Tweets included in the dataset.

Appendix B: Chapter 2 Supplemental Materials

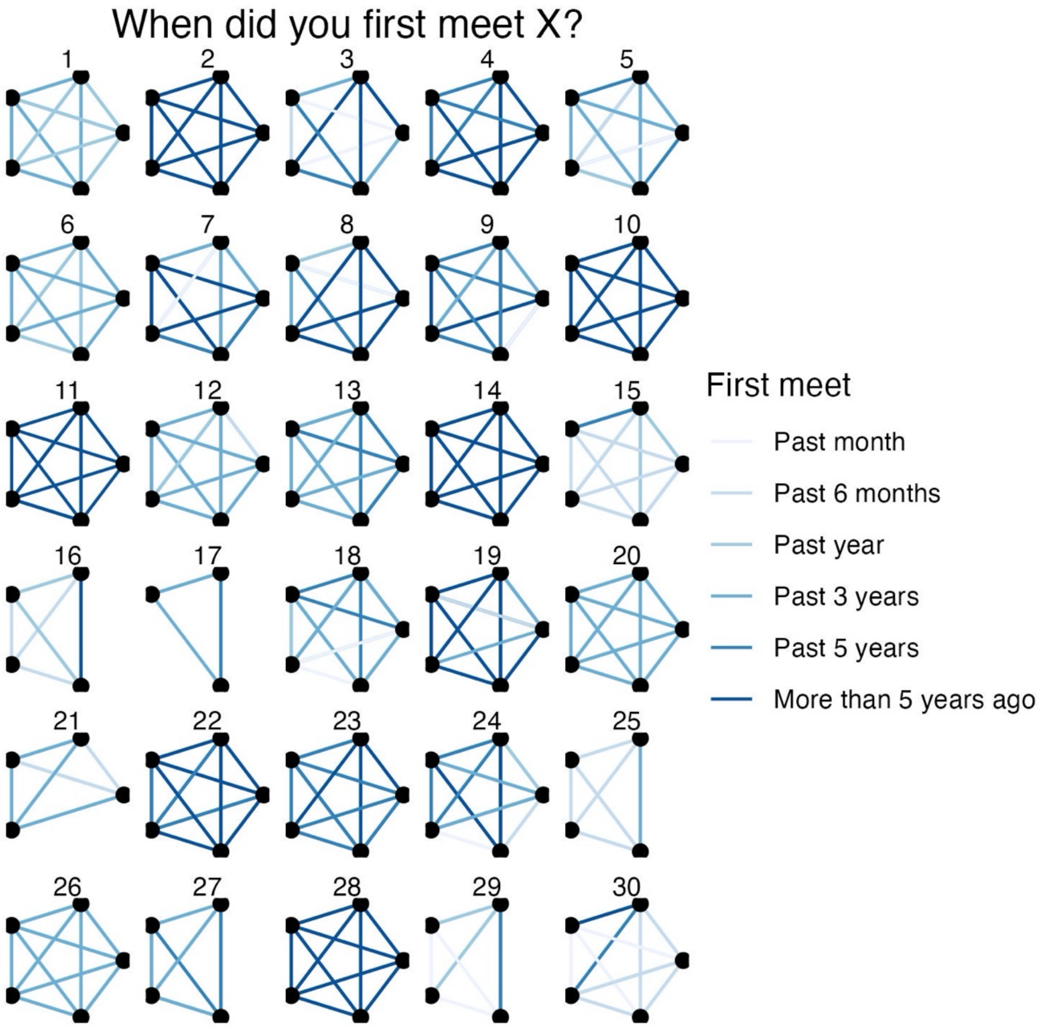


Figure A.2.1. Representations of friend networks based on when they had first met. Each network represents a participant group, and each black dot is a participant. Darker blue lines signify participants who have known each other longer, while lighter blue lines signify participants who have met more recently.

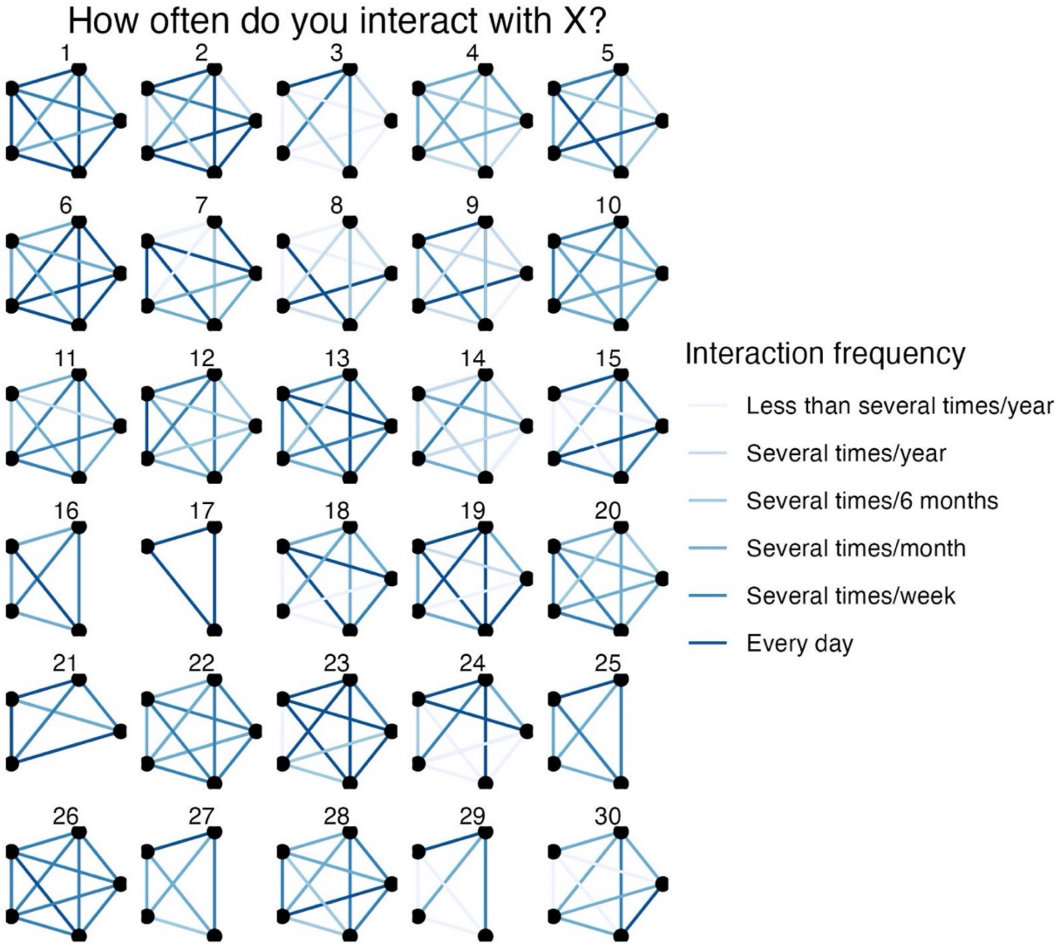


Figure A.2.2. Representations of friend networks based on how often they interact. Each network represents a participant group, and each black dot is a participant. Darker blue lines signify participants who interact more frequently, while lighter blue lines signify participants who do not interact as often.

Person perception dimension	Question	Predictor variables	Outcome variables
Competence	Q1	Timepoint	Trait perception
Sociability	Q1	Timepoint	Trait perception
Competence	Q1	Timepoint	Trait perception absolute difference
Sociability	Q1	Timepoint	Trait perception absolute difference
Competence	Q2	Similarity, liking, familiarity	Trait perception
Sociability	Q2	Similarity, liking, familiarity	Trait perception
Competence	Q3	Puzzle solving performance	Trait perception
Sociability	Q3	Team collaboration performance	Trait perception
Competence	Q3	Similarity, liking, familiarity	Puzzle solving PAB
Sociability	Q3	Similarity, liking, familiarity	Team collaboration PAB
Competence	Q3	Puzzle solving PAB	Trait perception
Sociability	Q3	Team collaboration PAB	Trait perception

Table A.2.1. A breakdown of all hypothesis-driven models that we ran.
All models were Bayesian multi-level models. Participant and group were grouping variables and the predictor variables were also random effects. Q1: Does an unfamiliar and challenging group activity lead to altered perceptions of friends' traits? Q2: How are perceptions of a friend's traits influenced by aspects of our relationship to them (i.e., relational factors)? Q3: What is the relationship between perceptions of actions and perceptions of traits? Timepoint as a predictor variable includes three timepoints: Between one week and one day before the escape room (pre-game), immediately after the escape room (post-game), and one week after the escape room (one week later).

Appendix C: Chapter 3 Supplemental Materials

fMRIPrep preprocessing

Anatomical data preprocessing: The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al. 2010), distributed with ANTs 2.3.3 (Avants et al. 2008), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a *Nipype* implementation of the antsBrainExtraction.sh workflow, using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, Zhang, Brady, and Smith 2001). Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1, Dale, Fischl, and Sereno 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (Klein et al. 2017). Volume-based spatial normalization to one standard space was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template.

Functional data preprocessing: First, a reference volume and its skull-stripped version were generated. Susceptibility distortion correction was omitted. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve and Fischl 2009). Co-registration was configured with six degrees of freedom. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al. 2002). BOLD runs were slice-time corrected to 0.46s (0.5 of slice acquisition range 0s-0.92s) using 3dTshift from AFNI (Cox and Hyde 1997).

The BOLD time-series were resampled onto fsaverage6, and then resampled onto their original, native space by applying the transforms to correct for head-motion. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions, Power et al. (2014)) and Jenkinson (relative root mean square displacement between affines, Jenkinson et al. (2002)). FD and DVARS were calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al. 2014). The three global signals were extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor, Behzadi et al. 2007).

Principal components were estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components were then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM and combined CSF+WM) were generated in anatomical space. This mask is obtained by dilating a GM mask extracted from the FreeSurfer's *aseg* segmentation, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99. For each CompCor decomposition, the *k* components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration.

The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al. 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

Supplemental references

Abraham, Alexandre, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gael Varoquaux. 2014. "Machine Learning for Neuroimaging with Scikit-Learn." *Frontiers in Neuroinformatics* 8. https://doi.org/10.3389/fninf.2014.00014.

Avants, B.B., C.L. Epstein, M. Grossman, and J.C. Gee. 2008. "Symmetric Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain." *Medical Image Analysis* 12 (1): 26–41. https://doi.org/10.1016/j.media.2007.06.004.

Behzadi, Yashar, Khaled Restom, Joy Liau, and Thomas T. Liu. 2007. "A Component Based Noise Correction Method (CompCor) for BOLD and Perfusion Based fMRI." *NeuroImage* 37 (1): 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042.

Cox, Robert W., and James S. Hyde. 1997. "Software Tools for Analysis and Visualization of fMRI Data." *NMR in Biomedicine* 10 (4-5): 171–78. https://doi.org/10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L.

Dale, Anders M., Bruce Fischl, and Martin I. Sereno. 1999. "Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction." *NeuroImage* 9 (2): 179–94. https://doi.org/10.1006/nimg.1998.0395.

Esteban, Oscar, Ross Blair, Christopher J. Markiewicz, Shoshana L. Berleant, Craig Moodie, Feilong Ma, Ayse Ilkay Isik, et al. 2018. "FMRIPrep." *Software*. Zenodo. https://doi.org/10.5281/zenodo.852659.

Esteban, Oscar, Christopher Markiewicz, Ross W Blair, Craig Moodie, Ayse Ilkay Isik, Asier Erramuzpe Aliaga, James Kent, et al. 2018. "fMRIPrep: A Robust Preprocessing Pipeline for Functional MRI." *Nature Methods*. https://doi.org/10.1038/s41592-018-0235-4.

Fonov, VS, AC Evans, RC McKinstry, CR Almli, and DL Collins. 2009. "Unbiased Nonlinear Average Age-Appropriate Brain Templates from Birth to Adulthood." *NeuroImage* 47, Supplement 1: S102. https://doi.org/10.1016/S1053-8119(09)70884-5.

Gorgolewski, K., C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, and S. Ghosh. 2011. "Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python." *Frontiers in Neuroinformatics* 5: 13. https://doi.org/10.3389/fninf.2011.00013.

Gorgolewski, Krzysztof J., Oscar Esteban, Christopher J. Markiewicz, Erik Ziegler, David Gage Ellis, Michael Philipp Notter, Dorota Jarecka, et al. 2018. "Nipype." *Software*. Zenodo. https://doi.org/10.5281/zenodo.596855.

Greve, Douglas N, and Bruce Fischl. 2009. "Accurate and Robust Brain Image Alignment Using Boundary-Based Registration." *NeuroImage* 48 (1): 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060.

Jenkinson, Mark, Peter Bannister, Michael Brady, and Stephen Smith. 2002. "Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images." *NeuroImage* 17 (2): 825–41. https://doi.org/10.1006/nimg.2002.1132.

Klein, Arno, Satrajit S. Ghosh, Forrest S. Bao, Joachim Giard, Yrjö Häme, Eliezer Stavsky, Noah Lee, et al. 2017. "Mindboggling Morphometry of Human Brains." *PLOS Computational Biology* 13 (2): e1005350. https://doi.org/10.1371/journal.pcbi.1005350.

Lanczos, C. 1964. "Evaluation of Noisy Data." *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis* 1 (1): 76–85. https://doi.org/10.1137/0701007.

Power, Jonathan D., Anish Mitra, Timothy O. Laumann, Abraham Z. Snyder, Bradley L. Schlaggar, and Steven E. Petersen. 2014. "Methods to Detect, Characterize, and Remove Motion Artifact in Resting State fMRI." *NeuroImage* 84 (Supplement C): 320–41. https://doi.org/10.1016/j.neuroimage.2013.08.048.

Satterthwaite, Theodore D., Mark A. Elliott, Raphael T. Gerraty, Kosha Ruparel, James Loughead, Monica E. Calkins, Simon B. Eickhoff, et al. 2013. "An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data." *NeuroImage* 64 (1): 240–56. https://doi.org/10.1016/j.neuroimage.2012.08.052.

Tustison, N. J., B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. 2010. "N4ITK: Improved N3 Bias Correction." *IEEE Transactions on Medical Imaging* 29 (6): 1310–20. https://doi.org/10.1109/TMI.2010.2046908.

Zhang, Y., M. Brady, and S. Smith. 2001. "Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm." *IEEE Transactions on Medical Imaging* 20 (1): 45–57. https://doi.org/10.1109/42.906424.