Gabbrielle M Johnson

June 10th, 2019

First chapter: Introduction to Implicit Bias, Routledge

The Psychology of Bias: From Data to Theory

Chapter Overview: What's going on in the head of someone with an implicit bias? Attempts to answer this question have centered on two problems: first, how to explain why implicit biases *diverge* from explicit attitudes and second, how to explain why implicit biases change in response to experience and evidence in ways that are sometimes *rational*, sometimes *irrational*. Chapter 1 introduces data, methods, and theories to help us think about these questions. First, the chapter briefly outlines the features of good, explanatory psychological theories built on empirical data, and the pitfalls they must avoid. Next, it presents an overview of the empirical data relevant to two main questions: implicit-explicit divergence and rationality. Finally, it surveys the theories intended to provide psychological explanations for those empirical data, providing examples of each. The chapter ends with some summarizing reflections, and in particular it confronts the possibility that bias is in fact a mixed-bag of many different sorts of psychological phenomena, making one unified psychological explanation misplaced.

1. Introduction

What's going on in the head of someone with an implicit bias? Often psychologists answer this question by saying such a person has an *unconscious mental association.* On this view, when we

say someone has an implicit bias against the elderly, for example, we're saying they quickly, automatically, and unconsciously associate someone's being **elderly** with, say, that person's being **frail**, **forgetful**, or **bad with computers**. This view of implicit bias comes very naturally to us, as we're used to our minds making associations quickly, automatically, and without our conscious awareness. For example, when I say **salt**, you automatically think **pepper;** or when I say **hip**, you think **hop**; when I say **Tweedledee**, you think … . You don't need to deliberate about what comes next; you just know. These characteristics of the *associationist* picture may help us explain one of the most vexing aspects of implicit bias: *divergence*. Divergence occurs when our unconscious mental states differ, or diverge, from our consciously-held mental states. Consider, for example, an individual who, when asked, says women are just as capable of succeeding in leadership roles as men are but continues to act in ways that seem at odds with that sentiment. They might, for example, rate a male applicant for a leadership role more favorably than a female applicant with equally impressive credentials. In this case, we say that this individual's explicit (or consciously accessible) beliefs about gender *diverge* from their implicit attitudes. At the conscious level, the person believes men and women are equally capable; however, at the implicit level, the person has a bias against women.

The associationist view provides a natural explanation for divergence: this person's conscious beliefs diverge from their unconscious beliefs because there are two distinct and independent mental constructs involved at each level of consciousness. At the conscious level, this person has deliberately considered evidence and is convinced that men and women are equal. However, at the unconscious level, this person's automatic, reflex-like processes lead them to associate **male** with **leader**. Because distinct and independent constructs operate at each

level, we get different results depending on which level the individual relies on at any given time. I call this approach, which distinguishes between different kinds of states or processes for explicit and implicit attitudes, a *dual-construct model.* Dual-construct models excel at explaining *divergence*, or the differences between explicit and implicit attitudes.

Dual-construct models—like the associationist picture above—have gained favor among psychological accounts of implicit biases. However, more recently, interesting studies exploring the malleability of implicit biases—that is, our ability to change implicit attitudes—suggest that the operation of mental processes at the two levels might not be so different after all. In particular, our implicit attitudes sometimes change when confronted with reasons to do so. I'll call this newly emerging data *rationality of bias*, or *rationality* for short.

The term 'rationality' is intended as a term of art here, which I'll use to pick out particular features of mental states that I'll explain in more detail later. This notion of rationality is intentionally more robust than you might initially think when hearing the term. For example, you might think that my having an association between salt and pepper is rational—it makes sense that I would think of one right after the other since they often appear together. However, this sort of superficial rationality won't be enough to capture the unique features I want this technical notion to capture. I'll return to this point and explain exactly what unique features I have in mind in section 3.3.

The possibility that implicit biases might ever be rational, albeit rarely, is surprising for the dual-construct model, which predicts that rational and deliberative processes are unique to the explicit level and, thus, entirely absent at the implicit level. Even stronger, the fact that rational processes might be in operation at both the explicit and implicit levels suggests that the

dual-construct model, which attributes to each level distinct and independent kinds of states and processes, might be mistaken. Instead, one might think the evidence for rationality suggests that implicit attitudes are just like, or at least similar to, ordinary explicit *beliefs*, save for one kind of belief is unconscious while the other is conscious. I'll call these sorts of approaches *belief-based models,* because they equate the kinds of constructs leading to explicit and implicit attitudes. Belief-based models excel at explaining *rationality*, or similarities between explicit and implicit attitudes.

In this chapter, I discuss these two fact patterns—*divergence* and *rationality*—in detail. I begin by reviewing standard assumptions about psychological theories more generally, such as what they aim to do and how we evaluate them. Following this preliminary discussion, I review the empirical data indicating patterns of divergence and rationality, and I examine how the two main approaches—dual-construct models and belief-based models—are each sufficient to deal with one of the fact patterns, but struggle to explain both. I'll then look at views that attempt to carve out a middle ground between dual-construct and belief-based models. These views argue that implicit biases constitute a unique kind of mental construct, which is not easily explained by either standard dual-construct or belief-based models.

2. What is a Psychological Explanation?

Roughly, *psychology* is a scientific discipline that aims to explain an intelligent creature's behavior in terms of that creature's mental states and processes. In other words, psychologists look to a creature's state of mind in order to understand why they acted the way that they did.

One fundamental assumption among most psychologists today is that humans have *mental states* that represent the world as being a certain way and that those representations of the world affect how they think and act in it. For example, you might explain your roommate's going to Chipotle for lunch using her *belief* that Chipotle makes the best guacamole. Of course, this belief might turn out to be false or your roommate might have gone to Chipotle for a different reason. But the kind of explanation you gave is what psychologists understand and expect.

This psychological methodology of building theories that explain by making reference to distinctively mental states—beliefs, desires, fears, intentions, etc.—is an example of what philosophers of science beginning with Thomas Kuhn (1962) call *a paradigm*. Within any paradigm, scientists take certain fundamental assumptions for granted as shared among members of a scientific community—in this case, the assumption that humans have mental states in the form of representations.

Alternative to this methodological paradigm was a different approach made popular by B.F. Skinner called *behaviorism.* It claimed that psychology should only study objective, observable physical stimuli and behavioral responses, and not concern itself with subjective, private mental states. In its most radical form, behaviorism claimed that *all* behaviors of intelligent creatures can ultimately be explained in this way, without ever needing to mention internal, distinctively mental states. Although no longer popular, behaviorism made several important contributions to the methodology of psychology.

One contribution is a general suspicion toward the ease of relying on mental state explanations (See Dennett 1978: 56 citing Skinner 1971: 195). The fear is that we can't explain an unknown fact—why your roommate went to Chipotle for lunch—by using an equally

mysterious object—her internal *belief* about Chipotle. Because her belief is a mental state, it is observable only by her and no one else. So we weren't really explaining anything at all, merely replacing one mysterious fact with another.

This worry is sometimes called the 'homunculus fallacy'. The word 'homunculus' (plural 'homunculi') is Latin for 'little person'. A theory that commits this fallacy attempts to explain some intelligent behavior by way of positing some equally intelligent cause of that behavior. The idea is that this was tantamount to positing a little person inside the head of the first intelligent creature whose own behavior goes unexplained.

This same basic idea is depicted humorously in the 2015 Pixar film *Inside Out*. In this film, the perspective switches between that of a young girl, Riley, and the five personifications of the basic emotions that live in her head: Joy, Sadness, Fear, Disgust, and Anger. These five little people (or homunculi) inside Riley's head operate a control center that influences all of Riley's actions. According to the film, the explanation for why Riley acted the way she did—for example, why, when her parents feed her broccoli, she frowns, gags, and swats the vegetable away—is that there is a little person in her head prompting those reactions. In this case, Disgust finds broccoli disgusting.

If you're like me, you might ask: if we looked into the head of these little people, would there be more, even smaller people inside their heads? Of course, the film never shows us what's inside any of their heads. You might then wonder if the film has really provided any explanation of Riley's actions, or if instead it has merely pushed the explanation of her behavior back a level. We can apply a point made by Skinner (1971: 19) and say the whole purpose of introducing the little people seems to be to help us understand why Riley acts how she does. But without

providing an explanation of why the little people in Riley's head act the way they do, we've failed to explain anything.

Over time, behaviorism itself was criticized for purporting to provide explanations without actually doing so, and there was a return to theories that unabashedly allowed for reference to mental states (Fodor 1981: 6). On such views, the way to avoid the homunculus fallacy is to slowly replace complex mental phenomena with combinations of simpler, more intelligible mental phenomena (Fodor 1968: 629). The hope is that eventually we arrive at an analysis constituted entirely by simple, elementary states (for example, thoughts and concepts) and the processes that combine them (for example, logical rules). We'll call any collection of states and processes that enters into such an analysis a *mental construct*. Crucially, the explanation of how these states operate can be given without any reference to intelligent behavior.

And thus we return to the modern-day paradigm that explains behavior by reference to mental states. This paradigm has come to dominate theories of cognitive science and psychology, and is tacitly present in the theories of bias to follow. However, we should not forget the lessons of behaviorism. You should continue to ask yourself as we move through the theories: has this explanation rendered important parts of the psychological picture more understandable, or has it merely posited a convenient, but equally mysterious mental process? In other words, has it provided a genuine explanation or has it *merely* pushed the entire explanation back a level to equally intelligent homunculus-like states?

3. Empirical Data of Social Bias

At the onset of our investigation, we're faced with several questions. What are the data surrounding social bias? In what ways do methods of testing for social bias differ from one another? What patterns emerge from these data?

3.1 Direct and Indirect Measures

Before the early 1970s, tests for social bias took a direct route: if a psychologist wanted to know if someone had a bias against a particular social group, she would ask her subjects directly. Such tests are called *direct measures*. Let's focus on the case of racial attitudes in the United States. One of the earliest examples of a direct measure was a test created by Katz & Braly (1933) that asked 100 Princeton students to read through a list of 84 adjectives and write down those that they think best characterized a particular race or ethnicity. Characteristic of the time, the results indicated pervasive negative racial biases. The majority of participants in the study paired African Americans with traits like *superstitious* and *lazy*, while pairing Germans with traits like *scientifically-minded* and *industrious*.

Over time, the social landscape of the United States changed dramatically. The Civil Rights Movement of the 1950s and 60s strived to establish racial equality across the country, and ushered in a new public standard that discriminatory opinions about African Americans were socially unacceptable. During this time, direct measures began to show a decline in negative racial bias. However, although overt expressions of racist ideology were curbed, the pervasive and destructive effects of racism were still painfully evident. It seemed that people still harbored racist opinions, opinions that influenced their beliefs about and actions toward people of color;

it's just that either those individuals stopped wanting to admit those opinions to others or, more curiously, those opinions were not obvious even to them.

This prompted the emergence of *indirect measures* (sometimes called "implicit" measures) that do not rely on asking subjects to report their attitudes. Today, the most famous and widely-used indirect measure is the Implicit Association Task (IAT) first developed by Greenwald et al. (1998). (The following discussion will attempt to describe the test in detail. However, the easiest way to understand how the test operates is to take it for yourself, which you can do online at https://implicit.harvard.edu/implicit/ in the span of approximately 10 minutes.) The IAT asks participants to quickly sort stimuli appearing in the middle of a computer screen to either the left or the right depending on the categories on each side. There are always two categories on either side, forming compound categories for each side, but the compound categories change during the test.

For example, in race IATs, compound categories combine race categories, e.g., black and white, with valences, e.g., good and bad. The stimuli to be sorted into these compound categories are representations of members of one of those four categories: photos of black faces, photos of white faces, positively-valenced words like 'happy' or 'love,' and negatively-valenced words like 'sad' or 'hate.' Subjects are asked to complete several rounds or "blocks" of these sorting tasks, with the compound categories changing from round to round. The *congruent blocks* of sorting tasks combine race categories with their stereotypical valence categories: with "black and bad" on one side and "white and good" on the other. In the *incongruent blocks*, the stereotypical attribute categories are swapped: "black and good" on one side and "white and bad" on the other. The tests also switch the sides of the congruent and incongruent categories on different rounds,

attempting to eliminate any dominant-hand advantage in certain blocks, as well as to randomize the order in which the incongruent and congruent blocks are presented, attempting to eliminate any conditioning effects from getting used to the test.

The results of these tasks reveal that most Americans, including some African Americans, are quicker and make fewer mistakes (e.g., sorting to the wrong side) when sorting stimuli in the congruent blocks (See, for example, Banaji & Greenwald 2013: 47 and Axt et al. 2014: 1806). I'll call results of this sort *demonstrating a positive preference* toward white faces or *demonstrating a negative preference* toward black faces. (These labels don't make any assumptions about individuals' psychologies. They are just about the behavioral responses: whether a given participant paired African American faces faster and more accurately with positive words, negative words, or—as is true for some participants—neither.)

Now the key question motivating psychological theories of bias is this: what's going on inside the head of someone who demonstrates a positive preference toward White faces on an IAT? The favored response among the creators of the test is that it measures specific *mental constructs* where the states involved are simple concepts and the relevant process is *association*. In the case of the Race IAT, the test measures whether particular race concepts are *associated* with valence concepts. Just like in the salt and pepper example mentioned in the introduction, what it means for the racial concept **black** to be associated with the valence concept **bad** is just for mental activations of the concept **black** to reliably cause mental activations of the concept **bad**. That is, when you think of one (the concept activates), you think of the other (the other concept activates). Crucially, this theory assumes that two concepts being associated makes it easier to sort examples of them together, and harder to sort examples of each separately. So, if

black is associated with **bad**, then it will be easier to sort examples of black faces (which activate the concept **black**) and examples of negative words (which activate the concept **bad**) to the same side than to opposite sides, and the same will be true of white faces and positive words if their concepts are associated.

There are other examples of indirect measures, for example semantic priming (Banaji and Hardin 1996) and evaluative priming (Fazio et al. 1995), the Go/No-Go Associations Task (Nosek and Banaji 2001), the Sorting Paired Features Task (Bar-Anan et al. 2009), the Weapons Identification Task (Correll et al. 2002), the Extrinsic Affective Simon Task (De Houwer 2003), and the Affect Misattribution Procedure (Payne et al. 2005). But roughly, all of these tests rely on similar theoretical assumptions: that certain behavioral patterns (like quick sorting) occur as a result of the subjects' mental constructs being composed a certain way (like being made up of associated concepts).

With the basics of direct and indirect measures out of the way, we're now in a position to explore various patterns that emerge in the data they provide.

3.2 Divergence

Our first fact pattern is *divergence*: results of indirect measures are often at odds with, i.e., diverge from, results of direct measures for the same subject. In what follows, I'll explain two ways in which results of these measures diverge. The discussion that follows focuses only on *empirical data,* namely, observations of behavior; we will discuss *psychological theories* that attempt to explain that data in the following section. As we advance through the observational evidence, however, it will be a fruitful exercise for readers to hypothesize about plausible

explanations, and then weigh their hypotheses against the psychological models presented in subsequent sections.

Reportability

The first aspect of divergence is the most striking. It involves an individual's ability to report on the results of indirect and direct tests. Early explorations of implicit bias suggested that subjects who demonstrated a positive white preference on indirect measures could not report harboring preferences or aversions that would explain such behavior. In fact, when people were confronted with the fact that their indirect measure results indicated bias, many avowed egalitarians expressed shock and disbelief. In their book *Blindspot*, Mahzarin Banaji and Tony Greenwald report several instances of the "disturbing" feeling one gets when confronted with the IAT evidence indicating an implicit bias (Banaji and Greenwald 2013: 56-58). These cases include a gay activist who finds out he harbors negative associations toward the gay community and a writer whose mother is Jamaican finding out he harbors pro-White biases, stating the revelation was "creepy", "dispiriting", and "devastating." Even the authors report the first-personal shock of finding out they have biases. Banaji's experience is described as "one of [her] most significant self-revelations" (Banaji and Greenwald 2013: 57).

However, the claim that individuals are *always* unable to predict their results of indirect tests has received criticism. Some empirical studies indicate that when participants were pressed to offer a prediction about how they will perform on an indirect test, their predictions were mostly accurate (Hahn et al. 2014). In addition, other studies have shown that when interviewed after an IAT, most subjects could accurately report how they fared on the test and, moreover,

many attributed their biased IAT performance to racial or stereotypical associations (Monteith et al. 2001: 407). In some studies, merely telling subjects to attend to their "gut feelings" was enough to increase their predictions of biased results (Hahn and Gawronski 2019).

Motivational and Social Sensitivity

The second way in which results on indirect and direct tests appear to diverge involves motivational and social sensitivity.

Regarding motivation, studies demonstrate that the more a subject describes themselves as motivated, the more their pro-white results from direct tests, e.g., self-reports, goes down while results from indirect tests, e.g., IATs, remain unchanged. To demonstrate this, researchers conducted a trio of tests (Fazio et al. 1995). In addition to a timed indirect measure somewhat similar to the IAT called the Evaluative Priming Task (EPT) and a direct measure, researchers provided a third set of questions gauging how motivated the subjects were to avoid being prejudiced or appearing prejudiced to others.

The researchers found that the correlation between subjects' performances on the indirect measure and their scores on the direct measure varied depending on how high they indicated their motivation to appear non-prejudiced was (Fazio et al. 1995: 1024). In situations where they claimed to be not highly motivated, their results from indirect and direct measures matched up (either both exhibited a preference or neither did). In situations where the subjects claimed they were highly motivated, their results on the other two tasks were often mismatched, with the results on the direct measure often indicating no preference, and the results of the indirect measure indicating a positive white preference.

Similar results were found with respect to the social sensitivity of the subject matter. When researchers such as Greenwald et al. (2009) and Kurdi et al. (2018) produced meta-analyses—taking a large group of studies about the correlation between direct and indirect measures and analyzing their overall average effects—they found that direct measures correlate with results from indirect measures differently depending on how socially sensitive the topic is. For example, for topics that are not socially sensitive, like consumer preference, subjects' direct and indirect preferences align, whereas for topics that are socially sensitive, like gender and sexual orientation preferences, their direct and indirect preferences were much less correlated.

Perhaps the most interesting finding from these studies, and a crucial aspect of divergence, is about *predictive validity*. Researchers use the term *predictive validity* to pick out the degree to which they're able to predict some phenomenon on the basis of some other phenomenon. In the case of implicit bias, many researchers are interested in predictive validity between direct and indirect measures on the one hand and real-world discriminatory behavior on the other. Meta-analyses differ about the predictive validity of implicit and explicit measures. Greenwald and colleagues found that, on topics where indirect and direct measure results diverge, self-reports of racial bias showed lower correlations with biased behavior compared to correlations between IAT scores and biased behavior.

All meta-analyses of the predictive power of direct and indirect measures score both measures in the "small" to "small-to-medium" range. Even critics of the IAT grant that it has some predictive power, but the question of amount is a matter of ongoing debate and research (Brownstein, this volume; Brownstein et al. 2019). Importantly, many studies appear to support the predictive validity of IAT and other indirect measures. Real-world examples of this predictive

validity include results from a Swedish study presented by Rooth (2007), which found that discriminatory hiring practices for applicants with Arab/Muslim sounding names were well predicted by IAT measures. Additionally, studies presented by Jacoby-Senghor and colleagues (2016) and Fazio and colleagues (1995) indicated that subjects' results from indirect measures correlate well with their perceived friendliness toward African Americans. Moreover, similar results have been seen with respect to racial biases in the treatment of patients by emergency room and resident physicians (Green et al. 2007) and racial biases in the accuracy of simulated shooting tasks (Payne 2001). (However, again, some critics argue that these results are exaggerated. See Brownstein, this volume, and Brownstein et al. 2019.)

Before moving on to the theories that attempt to resolve the puzzle of divergence, I first want to examine the other major fact pattern in the empirical data: the puzzle of the *rationality of bias*.

3.3 Rationality of Bias

As with the puzzle of divergence, the puzzle of *rationality* is multifaceted. This puzzle arises from data suggesting that results from indirect tests can actually be manipulated based on what I'll call *rational intervention*s. A rational intervention is an attempt to intervene on a person's implicit attitudes that relies on the informational content of the intervention, i.e., the reasons they present, rather than mere repeated exposure to the intervention (known in psychology as "classic conditioning"). The two indications of rationality for bias that we'll focus on are responsiveness to the rational interventions of simple instructions and strength of evidence.

Sensitivity to Simple Instructions

Initially it seemed like intensive training was necessary to change how people performed on indirect measures, indicating that change was a result of conditioning rather than rational intervention (See, for example, Kawakami et al. 2000). However, recent studies have shown that some indirect measure results can be changed by one-off, simple instructions, indicating that changes might be the result of more rational interventions after all.

The first relevant study involves an experiment presented by Gregg et al. (2006: 9; additionally discussed in Mandelbaum 2015: 15-17.) In this experiment, participants were given hypothetical narratives about two fictional tribes. The narrative about the first tribe was positive while the narrative about the second was negative. Participants were then given an IAT, the results of which indicated that participants *demonstrated a positive preference* toward the first tribe and *demonstrated a negative preference* toward the second. Experimenters then did something strange: they told the participants that due to an unforeseen error (like computer malfunction), participants had been given incorrect information about the two tribes. Participants were then instructed to swap the previous narrative assignments. Participants were then asked to retake the IAT. Surprisingly, the results exhibited a shift: subjects still *demonstrated a positive preference* toward the first tribe, but to a far lesser degree than before the intervention (Mandelbaum 2015: n.21; See also De Houwer 2006a; 2014).

Sensitivity to Strength of Argument

A second common data pattern cited by theorists interested in the puzzle of rationality is the relationship between results on indirect measures and the strength of evidence being presented to subjects.

The most relevant study for this point is presented by Brinol et al. (2009; additionally discussed in Mandelbaum 2015: 12-13). Here, researchers present two experiments aimed at testing this relationship, one involving vegetable preferences and one involving race. In the experiments, participants were split into groups and presented with one of two arguments—a strong argument or a weak argument—in favor of some conclusion regarding the benefit of the target stimulus. For example, in the experiment involving attitudes toward vegetables, one group of participants was given a persuasive argument (regarding the beneficial health effects of diets that include vegetables) while the other group was given an unpersuasive argument (regarding the popularity of vegetables at weddings). The participants were also given IATs before and after being exposed to the arguments. Interestingly, only those participants that were exposed to the strong arguments demonstrated any change in IAT performance (demonstrations of positive preference toward vegetables were increased). The experiment with arguments involving race showed similar results.

4. Theories

Now that we've seen the data for social bias, new questions emerge. How can we explain these results? What is the best way to explain why indirect measures diverge from explicit measures? If implicit biases are just associations, then why does performance on indirect measures sometimes shift in apparently rational ways?

4.1 Dual-Construct Theories

According to dual-construct theories, we can explain the differences between direct and indirect measure responses by positing separate mental constructs that independently give rise to results on each kind of test. Let's call the mental constructs that give rise to results of indirect measures like the IAT results *implicit constructs*, and the mental constructs that give rise to results of direct measures like self-reports *explicit constructs*.

Consider again the first quality of divergence, reportability. One hypothesis for why subjects can report on their results of direct tests but not indirect tests is that they have conscious access to their explicit constructs, but not their implicit constructs. Psychologists and philosophers disagree about whether implicit constructs are really unconscious or just not always noticed, which would explain the data that subjects are not *always* incapable of predicting and reporting IAT results (De Houwer 2006b: 14-16; Fazio and Olson 2003: 302-303; and Holroyd and Sweetman 2016: 80-81). Even if they are somewhat conscious, they might be less easy to access consciously and to report on than explicit constructs, which still accounts for the divergence in reportability.

Adopting a dual-construct theory similarly helps explain the other quality of divergence: motivational and social sensitivity. First, it seems reasonable to assume these are two versions of the same phenomenon—socially sensitive contexts are a kind of motivational context because subjects are motivated to be and to appear egalitarian (i.e., to value equal treatment for members of different social groups). We can then explain the relevant differences by postulating dual constructs: deliberate explicit constructs and automatic implicit constructs. In situations where a

subject is highly motivated to express egalitarian attitudes, their motivation can influence the operation of deliberative constructs, but they have no control over automatic constructs. Notice also how this explanation and the previous one about unconsciousness relate to one another: the control someone has over some mental construct might be affected by the degree to which they're consciously aware of it.

Let's walk through an example of a dual-construct model. The Associative-Propositional Evaluation (APE) Model presented in Gawronski and Bodenhausen (2006; 2014a; 2014b) is one of the most developed dual-construct models. (For another example of the dual-construct approach, see Fazio and Olson's MODE model, presented in Fazio 1990 and Fazio and Olson 2003.) The theory suggests that implicit constructs and explicit constructs involve two distinct *processes*: *associative processes* and *propositional processes*, respectively.

Imagine Sounak sees his elderly neighbor Carol for the first time. When his mother asks him what he thinks of his new neighbor, he responds that he likes her and is happy to have an elderly individual in the neighborhood. However, some of his behaviors indicate he's less warm to the idea; for example, he tends to cross the street whenever he sees Carol outside. According to the APE model, although his explicitly held beliefs indicate his warm feeling toward Carol, his mental associations tell a different story.

According to APE, when Sounak sees Carol, this activates the concept **elderly** in his mind. This *associative activation* then spreads by way of associative processes to other mental concepts, e.g., **wise**, **frail**, and **forgetful** might all activate. Some of these concepts might have a positive valence (like being wise), but some of them might have a negative valence (like being frail). Some of Sounak's responses, like his crossing the street, are a direct result of these

valences. Since the overall valence of the activating concepts is negative (wise is positive, while frail and forgetful are negative), Sounak has what Gawronski and Bodenhausen (2014b: 449) call a negative "spontaneous evaluative response." This response is what causes him to cross the street; it's also what is measured by indirect measures like the IAT.

But what about his report to his mother that he's glad Carol moved to the neighborhood? APE is able to account for these responses, too. According to the model, the overall valence of the activating concepts goes through another process called *propositionalization*, which produces in Sounak a sentence-like thought of the form "I don't like Carol." This thought is treated more deliberatively than the spontaneous evaluative response is. Crucially, this thought is evaluated against all of Sounak's background beliefs, which include things like his belief that he likes elderly individuals and that Carol is an elderly individual. Of course, if he likes elderly individuals and Carol is an elderly individual, then it stands to reason that he likes Carol, making this new thought "I don't like Carol" inconsistent with his other beliefs. So, according to APE, he rejects this new sentence-like thought while leaving intact his background beliefs that indicate he's glad an elderly individual moved to the neighborhood. This is what leads to divergence: his mental concepts and associative processes that lead to the spontaneous response (i.e., the implicit construct) indicate a negative response toward Carol, while his sentence-like thoughts and propositional processes that lead to his rejection of the thought that he dislikes Carol (i.e., the explicit construct) indicate a positive response toward Carol.

APE is tailormade to reflect the ways that motivation and social sensitivity lead to divergence. It claims that the more socially sensitive the domain, the more likely people are to

engage in deliberative processes, whereas socially insensitive domains rely on the spontaneous, non-deliberative processes.

4.2 Belief-Based Theories

Belief-based theories take as their primary data the apparent *rationality* of results on both indirect and direct tests and claim these similarities occur because the underlying constructs for each are of the same belief-like type, namely, both are *sentence-like structures* that involve *rational processing*.

One of the most developed and popular versions of this view in social psychology is named 'the propositional model,' where *proposition* is a representation with a sentence-like structure (De Houwer 2014). This model has two core assumptions supporting the conclusion that implicit constructs involve rational processes and sentence-like representations: first, changes by rational interventions are the result of rational processes and second, only sentence-like representations can be changed by rational processes. Since indirect measure results can be changed by rational interventions (as demonstrated by the rationality data), it follows that first, the constructs measured by them—implicit constructs—must involve rational processes (by assumption one), and second, that they must be composed of sentence-like representations (by assumption two). According to De Houwer (2014: 346), a sentence-like structure is necessary for representing *relational information*, and relational information is necessary for *rational* interventions. Since simple associations between concepts are not able to capture the relational information that sentence-like structures are able to capture, the processes performed on them cannot be *rational*.

This argument is complicated, and so running through an example will help. Consider a belief with the content **Rahul loves Priya**. This belief is propositional; its structure is sentence-like. It also captures how Rahul and Priya are *related*. An association, remember, exists between concepts. So, the complex **Rahul loves Priya** would be a combination of three singular concepts **Rahul**, **loves**, and **Priya**, that are all associatively linked. But then the associationist model can't distinguish between complexes that are built out of the same constituents but contain different relational information, for example **Rahul loves Priya** versus **Priya loves Rahul**. The relational information conveyed by these two complexes are very different. We form different *rational* conclusions depending on which we believe, and one might be true while the other is false (much to Rahul's chagrin). So it's important that we can distinguish between them. To do that, this theory argues, we need to combine the concepts in a particular order with a sentence-like structure. Mere associative bundles just won't do.

Because the associationist model can't account for this difference, representations involved in rational processes can't be modeled by associations. Since, as we've seen from the data, some implicit constructs *are* affected by rational interventions, then according to the assumptions of this theory, they must be sentence-like mental states rather than mere associations between concepts (see also Mandelbaum 2015).

4.3 Problems

Both dual-construct and belief-based theories excel at answering their fact patterns of choice; however, each falter in resolving the other's preferred fact pattern. Dual-construct models do well in explaining divergence data, but run into difficulties in explaining rationality data due to

assuming that implicit constructs involve only associations. Likewise, belief-based models explain rationality with implicit constructs being sentence-like representations that respond to rational processes. However, if implicit biases *are* so belief-like, then why don't they look like beliefs in many respects pertaining to divergence—that is, why would they be relatively unconscious or automatic in ways that our explicit beliefs are not?

This is not to say that either theory cannot explain the other's data. For example, proponents of the APE model claim it's possible that some propositional information can affect spontaneous evaluative responses by a sort of "spillover" from the rational processes (Fazio and Olson 2003: 302). Likewise, proponents of the belief-based model can account for divergence by stipulating that the mind is made up of many conflicting, fragmented, and redundant sentence-like thoughts, some of which are unconscious and automatic, some of which are not, and some of which cause positive reactions toward individuals while others cause negative reactions toward those same individuals (Mandelbaum 2015: 20-23.).

The problem with these sorts of concessions, however, is that they appear ad-hoc. 'Ad-hoc' applies to parts of theories "made up on the fly" when problematic evidence comes in, rather than predicted in advance. Consider a famous example studied by psychologist Leon Festinger and colleagues involving a doomsday cult that had predicted the end of the world by way of a great flood on December 21st, 1954. The prophecy stated that "the chosen" among the believers would be saved from the destruction of the flood at precisely midnight the evening before by a spaceman piloting a flying saucer, who would then escort them to safety. Unsurprisingly to us now, midnight came and passed with no arrival of a spacecraft. The flood also failed to come to fruition. When faced with the apparent disconfirmation of the prophecy,

rather than reasoning that the cult teachings were false, believers rationalized the failures as consistent with the cult teachings after all. They reasoned *ad-hoc* that it was due to the true faithfulness of the cult members themselves that the flood was avoided (Festinger et al. 1956). Perhaps the real trouble with concessions like these is that they appear to lack explanatory efficacy. Returning to theories of implicit bias, although it's true that spillover effects, mixed processes, and fragmentation *can* result in the relevant fact patterns, none of these theories offer an account of *why* these effects occur when they do, they merely stipulate them by decree. Without filling in this story, these changes to the theories in order to account for both fact patterns are explanatorily inert, as in the homunculus fallacy discussed in section two. More and more research is investigating these questions and it's important that research draw on theories to make predictions rather than trying to explain the data with 20/20 hindsight.

5. Meeting in the Middle

So far, we've looked at views of implicit bias that fit into two basic camps: dual-construct theories and belief-based theories—as well as the problems with each. In what remains, we will briefly survey views that attempt to carve out space in the middle. These views will often attribute to implicit bias constructs some characteristics that are shared by the familiar constructs discussed above (e.g., associations and beliefs), but also characteristics unique to implicit bias.

5.1 Unique States and Processes

Some views attempt to carve out middle ground by treating implicit biases as being pretty similar to beliefs, but differing from them in important ways. I'll begin by surveying prominent views of this type, then present a prominent criticism.

One view that adopts the uniqueness approach is Schwitzgebel's *In-between belief* view (2002, 2013). Schwitzgebel's view is importantly different from the other views we've discussed in that it characterizes beliefs not in terms of their representational contents, but rather as tendencies to behave in various ways when faced with various physical stimuli (in this way, it's similar to behaviorism, discussed in the second section). So, rather than viewing my belief that I should take the garbage out when it's full as a mental state with the propositionally-structured content **I should take the garbage out when it's full**, we instead view it as my tendency to take the garbage out when confronted with the overflowing can. But such a view can run into problems when faced with cases where an individual seems disposed to a mixed bag of behaviors. The data on divergence above is an example: often, individuals are disposed to act in ways that make it seem like they have one belief about members of a particular social group, but they're also disposed to act in ways that make it seem like they harbor the opposite belief. (Recall also the case of Sounak and his elderly neighbor.) For these reasons, Schwitzgebel introduces the notion of *in-between cases of believing* for cases where it seems an individual doesn't fully believe, but doesn't not believe either (Schwitzgebel 2002, 260). Implicit biases, according to Schwitzgebel, are cases of in-between believing. (See also Levy 2015 and Machery 2016, 2017.)

Another, very different uniqueness approach is Tamar Gendler's *alief* view (Gendler 2008, 2011). Gendler argues that we cannot capture implicit constructs with any of the familiar

categories of psychological explanation, such as associations or beliefs. Instead, she argues, implicit constructs, as well as other "more deviant" arenas of human life, such as phobias and superstitions, should be explained by a new psychological kind called *aliefs*, which are a three-part mix of thoughts, feelings, and behavioral impulses. When a person's implicit bias construct is activated, then, that person not only has particular representational components activated (e.g., concepts like **elderly** and **frail**), but they're also prone to experience certain feelings (like being sad or scared) as well as exhibit certain behaviors (like avoidance). These *aliefs* are often at odds with a person's *beliefs*. (See her example of walking across the Grand Canyon on the transparent Skywalk: although you might *believe* that it's safe, you might simultaneously *alieve* that it's not.) Regarding implicit biases as *aliefs*, we might believe one thing regarding individuals of a particular social group, while also harboring a-rational aliefs that cause us to automatically exhibit behaviors diverging in various ways from the behaviors we would expect based on those beliefs. (See also Madva and Brownstein 2018 and Brownstein 2018.)

The criticism often directed at uniqueness approaches is that they, like the additions to the associative and belief-based models above, appear *ad-hoc*. They deal with the problems above—namely, that the collection of properties harbored by implicit bias constructs makes it so that they don't neatly fit any models of standard, familiar mental constructs, like beliefs or associations—by merely postulating a new kind of state that has all and only the relevant properties, and thereby, they fix the problem by fiat. Worries of the homunculus fallacy loom large here. That is, it's not clear that postulating these unique states really explains the operation of implicit bias constructs rather than just pushing the explanation back a level.

6. Concluding Remarks

At this point, we've surveyed many views on the topic, all attempting to account for various aspects of implicit bias operation. In fact, more and more research is coming out on bias that continues to complicate the overall picture. As methods for studying bias become more sophisticated, so too does our understanding of how bias operates in the minds of individuals. Given the variety, readers might be skeptical that there is even a unified phenomenon to be studied under the heading of implicit bias research. Holroyd and Sweetman (2016) raise this possibility. If this is right, it would explain why some data surrounding implicit bias operation just can't be explained using one, monolithic psychological explanation. Instead, we would need a variety of different theories.

The purpose of psychological theorizing around implicit bias, then, would be to search for different explanations, describing in what instances they're apt, investigating what, if anything, unifies them and, importantly, doing all this while ensuring that such explanations are *genuinely explanatory.* Such a view paints a picture of the psychology of bias in which there's still a lot of work to be done; but leaves open that many of the views we've discussed here might eventually find a home together, constituting different and important aspects of the overall picture.

References:

Axt, J. R., Ebersole, C. R., and Nosek B. A. (2014). The rules of implicit evaluation by race, religion, and age. *Psychological Science*, 25(9): 1804-1815.

Banaji, M. R. & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. New York: Delacorte Press.

Banaji, M. R. and Hardin, C. D. (1996). Automatic Stereotyping. *Psychological Science*, 7(3): 136–141.

Bar-Anan, Y., Nosek, B. A., and Vianello, M. (2009). The Sorting Paired Features Task: A Measure of Association Strengths. *Experimental Psychology*, 56(5):329–343.

Brinol, P., Petty, R., and McCaslin, M. (2009). Changing Attitudes on implicit versus Explicit Measures: What is the Difference? In Petty, R., Fazio, R. H., and Brinol, P., editors, Attitudes: Insights from the New Implicit Measures. *Psychology Press*, New York.

Brownstein, M. (2018). *The Implicit Mind: Cognitive architecture, the self, and ethics.* New York: Oxford University Press.

Brownstein, M., Madva, A., & Gawronski, B. (2019). Understanding implicit bias: Putting the criticism into perspective. Manuscript submitted for publication.

Correll, J., Park, B., Judd, C. M., and Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, 83(6):1314–1329.

De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental psychology*, 50(2):77.

De Houwer, J. (2006a). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, 37(2):176–187.

De Houwer, J. (2006b). What are implicit measures and why are we using them? In Wiers, R. W. and Stacy, A. W., editors, *The handbook of implicit cognition and addiction*, pages 11–28. SAGE Publications.

De Houwer, J. (2014). A Propositional Model of Implicit Evaluation: Implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342–353. https://doi.org/10.1111/spc3.12111

Fazio, R. H. (1990). Multiple Processes by which Attitudes Guide Behavior: The Mode Model as an Integrative Framework. In *Advances in Experimental Social Psychology* (Vol. 23, pp. 75–109). Elsevier. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0065260108603184

Fazio, R. H., Jackson, J. R., Dunton, B. C., and Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: a bona fide pipeline? *Journal of personality and social psychology*, 69(6):1013.

Fazio, R. H., & Olson, M. A. (2003). Implicit Measures in Social Cognition Research: Their Meaning and Use. *Annual Review of Psychology*, *54*(1), 297–327. https://doi.org/10.1146/annurev.psych.54.101601.145225

Festinger, L., Riecker, H. W., & Schachter, S. (1956). *When Prophecy Fails.* Minneapolis: University of Minnesota Press.

Fodor, J. A. (1968). The Appeal to Tacit Knowledge in Psychological Explanation. *The Journal of Philosophy*, 65(20):627.

Fodor, J. A. (1981). Introduction: Something of the State of the Art. In *Representations*, pages 1–31. Massachusetts: The MIT Press.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., and Bodenhausen G. V. (2014a). The Associative-Propositional Evaluation Model: Operating Principles and Operating Conditions of Evaluation. In *Dual-Process*

Theories of the Social Mind, edited by Jeffrey W. Sherman, Bertram Gawronski, and Yaacov Trope, 188–203. New York: The Guilford Press.

Gawronski, B., & Bodenhausen, G. V. (2014b). Implicit and Explicit Evaluation: A Brief Review of the Associative-Propositional Evaluation Model: APE Model. *Social and Personality Psychology Compass*, *8*(8), 448–462. https://doi.org/10.1111/spc3.12124

Gendler, T. S. (2008). Alief and belief. *The Journal of Philosophy*, *105*(10), 634–663.

Gendler, T. S. (2011). On the epistemic costs of implicit bias. *Philosophical Studies*, *156*(1), 33–63. https://doi.org/10.1007/s11098-011-9801-7

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., and Banaji, M. R. (2007). Implicit Bias among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients. *Journal of General Internal Medicine*, 22(9): 1231– 1238.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., and Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97(1):17–41.

Gregg, A. P., Seibt, B., and 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.

Hahn, A. & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgement. *Journal of Personality and Social Psychology*, 116(5): 769-794.

Hahn, A., Judd, C. M., Hirsh, H. K., and Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, 143(3):1369–1392.

Holroyd, J. and Sweetman, J. (2016). The Heterogeneity of Implicit Bias in Brownstein M. & Saul J., (eds.) *Implicit Bias and Philosophy Volume I: Metaphysics and Epistemology*, New York: Oxford University Press.

Jacoby-Senghor, D. S., Sinclair, S., & Shelton, J. N. (2016). A lesson in bias: The relationship between implicit racial bias and performance in pedagogical contexts. *Journal of Experimental Social Psychology*, 63, 50–55. https://doi.org/10.1016/j.jesp.2015.10.010

Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology,* 28(3), 280-290.

Kawakami, K., J. Moll, S. Hermsen, J. F. Dovidio, and A. Russin. 2000. Just Say No (to Stereotyping): Effects of Training in the Negation of Stereotypic Associations on Stereotype Activation. *Journal of Personality and Social Psychology* 78(5): 871-888.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Kurdi, B., A. Seitchik, J. Axt, T. Carroll, A. Karapetyan, N. Kaushik, D. Tomezsko, A. Greenwald, and M. Banaji (2018). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist.* https://doi.org/10.1037/amp00000364

Levy, N. (2015). Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. *Noûs*, *49*(4), 800–823. https://doi.org/10.1111/nous.12074

Machery, E. (2016). De-Freuding implicit attitudes in Brownstein M. & Saul J., (eds.) *Implicit Bias and Philosophy Volume I: Metaphysics and Epistemology*, New York: Oxford University Press.
Machery, E. (2017). Do Indirect Measures of Biases Measure Traits or Situations? *Psychological Inquiry*, 28(4):288–291.
Madva, A., & Brownstein, M. (2018). Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind. *Noûs* 52: 611-644. https://doi.org/10.1111/nous.12182
Mandelbaum, E. (2015). Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Noûs*. https://doi.org/10.1111/nous.12089
Monteith, M. J., Voils, C. I., and Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition*, 19(4): 395– 417.
Nosek, B. A. and Banaji, M. R. (2001). The Go/No-Go Association Task. *Social Cognition*, 19(6):625–666.
Payne, B. K., Cheng, C. M., Govorun, O., and Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89(3):277–293.
Payne, B. K. (2001). Prejudice and perception: the role of automatic and controlled processes in misperceiving a weapon. *Journal of personality and social psychology*, 81(2):181.
Rooth, D. O. (2007). *Implicit discrimination in hiring: Real world evidence*. Discussion Paper 2764. Bonn, Germany: Institute for the Study of Labor.
Skinner, B. F. (1971), *Beyond Freedom and Dignity.* New York: Knopf.

Recommendations for additional reading:

For a more technical overview of many of the psychological theories discussed in this chapter:

- Brownstein, M. 2015. Implicit bias. In: Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/spr2015/entries/implicit-bias/.

For more on the notion of a paradigm within scientific theorizing:

- Kuhn, Thomas 1962. *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

For More on the relationship between behaviorism, the homunculi fallacy, and methodologies in psychological explanation:

- Watson, John B. 1913. “Psychology as the behaviorist views it,” *Psychological review* 20.2.
- Fodor, Jerry A. 1968. “The Appeal to Tacit Knowledge in Psychological Explanation.” *The Journal of Philosophy* 65 (20): 627.

For an introduction to mental heuristics, system one vs system two, and more on the unconscious, automatic mind more generally.

- Kahneman, Daniel 2011. *Thinking, Fast and Slow*. Macmillan.

A set of study questions or your article [*indicates recommended for book]:

*What is a psychological explanation? Imagine that my roommate comes home from school, stomps across the living room to her bedroom, and slams the door. How might we explain her actions using mental representation paradigm? What objections would behaviorists make to this explanation?

*What is the homunculi fallacy and why is it bad for psychological explanation?

What's the primary difference between indirect and direct measures for implicit bias? Why is it important to have indirect measures of bias? Do you think having indirect measures is less important in the study of preferences in other domains, like what kind of soda someone likes to drink or what genre of movie they like best?

*In most of the psychological theories discussed, there was a focus on mental representations, i.e., mental states that represent the world as being a certain way. However, apart from Gendler's Alief model, there was very little talk about other mental states, like affective or emotional states, that might affect how biases operate. How might these fit into the model we've discussed so far? Do you think they're an important aspect of how we think about and act toward others?

*We've been talking about the processes subserving implicit bias as a self-contained mental construct. One of the great criticisms of behaviorism, however, is that it cannot account for *mental interaction*, that is effects that are the joint result of many mental causes. So, try to think about implicit bias in the context of a complete psychology that has perceptions, inferences, actions, desires, problem-solving abilities, is embodied, etc. (See chapters 2 and 5, this volume.) How might our understanding of implicit bias and its construct change when we think about it in this domain?

In section three we were introduced to the difference in direct and indirect measures. In section four we made the decision to call whatever gives rise to results on indirect measures an implicit construct, and whatever gives rise to the results of direct measures an explicit construct. How do these constructs relate to what we intend to pick out with the terms "implicit bias" and "explicit bias"? What other ways might we define "implicit bias" and "explicit bias"? Do these definitions presuppose aspects of a psychological theory that we might want to avoid?

It is sometimes argued that a psychological theory that uses psychological states and processes with which we're already familiar is better than a psychological theory that posits entirely new states and processes—call this *the principle of parsimony*. What are reasons for adopting the principle of parsimony? Of the various theories of implicit bias that we've looked at, which would the principle of parsimony cause us to prefer? How do we balance maximizing the principle of parsimony with the complexity of data?

Psychologists Amos Tversky and Daniel Kahneman are famous for work demonstrating that the human mind is prone to adopt a wide-variety of shortcuts and heuristics (see chapter 4 for more on the view of biases as "shortcuts"). They argue that the mind is composed of two distinct "systems": system one involves the many mental shortcuts and is responsible for our fast, automatic behaviors while system two involves more rational processes and is responsible for our slow, deliberate behaviors. How does this picture fit with the theories of implicit biases that we've been discussing? What implications might this have for the existence of other implicit constructs outside of the social domain? How might this change the way we theorize about the psychology of bias?

A list of potential glossary terms:

Mental Representation
Mental Construct
Implicit Construct
Explicit Construct
Divergence
Rationality of Bias (a.k.a. Rationality)
Dual-Construct Model
Belief-Based Models
Association
Proposition
In-between Belief
Alief

Web Resources (podcasts videos, movies):
ONLINE RESOURCES

- Harvard's Project Implicit, which hosts online versions of the Implicit Association Test (IAT). Found at: https://implicit.harvard.edu/implicit/ (IAT Test)
- Buster Benson's "Cognitive Bias Cheat Sheet". Found at: https://betterhumans.coach.me/cognitive-bias-cheat-sheet-55a472476b18

PODCASTS

- David McRaney. *You Are Not So Smart.* Available at: https://youarenotsosmart.com/podcast/- A Podcast about unconscious mental reasoning and biases in thought.

- Alix Spiegel. “The Culture Inside” *Invisibilia*. National Public Radio, Season 3, Episode 3, Available at: https://www.npr.org/programs/invisibilia/532950995/the-culture-inside

MOVIES
Inside Out
Eternal Sunshine of the Spotless Mind (a movie depicting how our mental representations in the form of memories can sometimes represent the world inaccurately or impartially, how they relate to emotion, and how that affects our actions)