2018

Owning Our Implicit Attitudes: Responsibility, Resentment, and the Whole Self

Wesley Whitaker
Claremont McKenna College

Claremont McKenna College

Owning Our Implicit Attitudes: Responsibility, Resentment, and the Whole Self

submitted to
Professor Gregory Antill

by
Wesley Whitaker

for
Senior Thesis
Fall 2017
December 4, 2017

This page intentionally left blank

Abstract

Are implicit biases something we can rightly be held responsible for, and if so, how? A variety of social and cognitive psychological studies have documented the existence of wide-ranging implicit biases for over 30 years. These implicit biases can best be described as negative mental attitudes that operate immediately and unconsciously in response to specific stimuli. The first chapter of this thesis surveys the psychological literature, as well as presents findings of real-world experiments into racial biases. I then present the dominant model of implicit attitudes as mere associations, followed by evidence that at least some implicit attitudes take on a propositional form and involve making inferences based on evidence. I then reject adopting either of these two rigid models in favor of a dispositional approach that treats implicit biases as on the same spectrum of, but adjacent to, beliefs.

I then evaluate the moral wrongdoing associated with holding explicitly prejudicial beliefs, appealing first to Kantian notions of respecting individuals as agents, then appealing to Strawson's argument that we are responsible for expressions of our will. Our status as human agents involves participating in complex and sustained interactions with others, which necessarily implies that we take part in the social practice of holding each other responsible for the quality of their will. The reactive attitudes we display in our everyday interactions indicate which features and circumstances are most important when investigating this practice. After applying this approach to implicit attitudes, I then pose the objection that their unconscious and unendorsed nature disqualifies implicit attitudes as proper expressions of our will. I develop this objection using Scanlon's account of moral responsibility, which requires the capacity to self-govern in light of principles that are generally agreed upon as good reasons for guiding interactions with one another. Finally, I critique Real Self theories that seek to arbitrarily privilege one part of ourselves in favor of the Whole Self, which privileges those features that are most integrated into our overall character.

Acknowledgements

First and foremost, I would like to thank Professor Greg Antill, not only for his endlessly patient help and guidance with this project, but also for introducing me to this topic and field within philosophy. His class on Moral Psychology was one of my favorite seminars due to the superb readings and wonderful discussion, which he was entirely responsible for. Thank you for indulging my exploration into tangents and for getting me back on track when I swerved too far in the wrong direction.

I would also like to thank Professor Andrew Schroeder for changing my life in his Introduction to Political Philosophy class that I took three years ago. It inspired me to become a philosophy major and, looking back, I cannot think of another major that would have interested me more, or allowed for as much personal growth. I must also thank the rest of the CMC Philosophy Department for being overall wonderful teachers and humans. I feel myself getting smarter anytime I talk to you and I hope I can one day find a place where I feel challenged and engaged like I have in all of my classes.

My friends and classmates have been extremely kind and helpful throughout, my time at CMC, not only academically, but also providing much needed moral support. Thank you for listening, caring, and even feeding me. You are what make this place special and I will cherish my memories with all of you.

Last but not least, I am eternally grateful to my mother for making my education possible, and for never failing to support me in my decisions. She has sacrificed so that I could have experiences she could only dream of, and I will never forget that. I cannot thank you enough for the opportunities you have helped create and I promise to make you proud.

Table of Contents

Introduction

The presence of implicit bias has been well documented and widely agreed upon among psychologists for at least 30 years. By adapting well-established scientific methods and principles from the fields of cognitive and social psychology, researchers have amassed a mountain of peer-reviewed studies that have documented instances of implicit bias regarding race, gender, sexual orientation, national origin, age, and many more characteristics. These findings have even begun to permeate the mainstream and are often cited as evidence that discrimination is alive and well even if overt forms of racism, homophobia, sexism, or ageism are not.

The first half of this paper focuses on investigating implicit attitudes as a phenomenon tracked through various psychological experiments. I will present the dominant view that implicit attitudes are the result of implicit associations made between concepts, followed by evidence suggesting that a propositional model of implicit attitude might also be useful to entertain. By surveying the psychological literature, I hope to highlight the important features of implicit attitudes without decisively taking a stance on which model best captures the full details. Instead, I offer the dispositional approach to belief as an alternative to formalistic definitions that is more useful in adjudicating claims of responsibility. While implicit attitudes are markedly different from explicit beliefs, they are best thought of as on as the same spectrum of mental states.

The second half of this paper will attempt to sketch an account of responsibility for our implicit attitudes. By first describing the moral wrongdoing committed by holding explicitly prejudicial views, I will then trace how the same sort of wrongdoing spills over

into holding implicit attitudes by using P.F. Strawson's account of responsibility as a social practice. Instead of getting hung up on theoretical preconditions, I will use Strawson to argue that we can be held responsible for things that express the quality of our will and our reactive attitudes provide useful signposts for our moral judgments. As an objection to Strawson, I will present Scanlon's reasons-based account of moral responsibility, which challenges the suggestion that implicit attitudes are true reflections of our will. Finally, I argue that we should abandon Real Self models that seek to arbitrarily privilege one part of the self, such as the part engaged in reasoning and judgment. Instead, by embracing the Whole Self, we can gain deeper, more nuanced insight into our implicit attitudes and how much our responsibility for them depends on how deeply integrated they are with our overall character.

Empirical Foundation

The definition of implicit bias that I will be using is, "an attitude that an individual harbors toward a certain subject matter, usually social groups and/or individuals within them, that operates quickly, automatically, and (typically) without his or her knowing about it" (Johnson, 2016, 1). It is worth noting that most psychological literature uses the terms implicit bias and implicit attitude interchangeably, however, because of the negative connotation in folk conversation against implicit biases, I will use implicit attitudes to describe a broad category of affective states, evaluative states, and stereotypes. The use of attitude instead of bias also serves to highlight that the phenomena in question is a mental state rather than a behavioral bias, which may actually result from underlying attitudes. I am also careful to not use the phrase implicit belief as beliefs are generally thought of as

propositional attitudes, meaning they describe the mental state of having some attitude or stance about a proposition, or the potential state of affairs in which that proposition is true. Using this loaded term begs the first question I will be attempting to answer, which is constructing an adequate account of the structure of implicit attitudes.

The study of implicit biases began as early as the 1950s, however, it was research on automatic semantic associative links in memory in the 1970s that most directly led to the discovery that prejudice or stereotyping might occur implicitly, which I will describe in more detail in the coming pages. This finding was solidified and expanded through the 1980s with the rise of indirect measures (Johnson, 2016, 2). Unlike direct measures, which rely only on self-reporting by the subject such as asking for their emotional response, indirect measures track the physical responses of a subject in response to certain stimuli, such as heart rate or the time interval it takes to perform a pairing task.

An interesting and unexpected pattern emerged when social psychologists began to compare the results of direct and indirect measures. Typically, one would expect that one's self-reported results would match the results of the indirect measures, especially in experimental settings where the subject had no to reason to deceive the researchers and claimed to be accurately reporting their feelings and mental states. Nevertheless, a pattern of divergence emerged between the two measures, indicating that subjects "harbored preferences and aversions that were in conflict with the explicitly expressed opinions provided in direct tests" (Johnson, 2016, 2). For example, respondents would have an easier time matching positive concepts to pictures of white faces than black faces, while maintaining no explicit racial bias. The longer time required to match positive concepts

with black faces than white faces was interpreted by researchers as evidence for a negative attitude towards African Americans. Further, when these findings were presented to subjects, they were unable to provide any account that might explain the divergence. This divergence and lack of conscious awareness of what led to the results of the indirect measures prompted researchers to posit that different mental constructs were at play. Unlike our conscious beliefs, which we are aware of and usually involve some kind of inference based on evidence, the results from indirect measures were thought to be explained by a process that was automatic and inaccessible.

During the 30 plus years that researchers have documented this pattern of divergence between direct and indirect measures, one of the first experiments to conclusively show the existence of implicit attitudes came from the nexus of cognitive and social psychology. Conducted by Devine in 1989, the experiment demonstrated that "social perception and memory were shown to be influenced by exposure to semantically (or stereotypically) related information" (Jost et al., 2009, 43). Described briefly, subjects participated in a menial task, during which they were exposed to either a relatively large or small proportion of words related to common stereotypes of African Americans. They were then asked to evaluate a person named "Donald" as either friendly or hostile. The findings demonstrated that subjects who were exposed to a greater proportion of words commonly associated with negative stereotypes about African Americans were more likely to judge Donald as hostile. On top of showing that subjects' social judgements could be affected by exposure to semantically laden content without their knowledge, the comparison to subjects' explicit beliefs about race – or direct measures – revealed a deep divide with the indirect measures (Jost et al., 2009, 43). Similar studies have been replicated dozens of

times, leading to strong consensus among the psychological community that the cognitive salience of a familiar stereotype can implicitly bias social judgment in stereotype-consistent ways.

Another experimental design came from cognitive psychology and used semantic pairing to investigate the nature of implicit attitudes. The main idea behind semantic pairing is that social attitudes, including prejudices and stereotypes, are "empirically captured by the degree to which they are linked through speed and efficiency to semantically related concepts" (Jost et al., 2009, 43). The idea behind this kind of experiment is to investigate the immediate behavioral reactions that result from exposure to certain stimuli. Note that the intervals used are too quick to attribute to conscious reasoning processes and therefore, the faster the behavioral response, the link between the concept and the behavior is hypothesized to be stronger. For example, researchers would prime subjects by exposing them to words that are either connotatively or denotatively correlated with stereotypes of women, such as nurse or teacher. The speed with which subjects identified female pronouns subsequently was greatly increased among the experimental group compared to the control group, which was primed with words that were not associated with gender (Jost et al., 2009, 43). These findings pair nicely with the idea that knowledge is essentially organized in our memory as semantic associations between concepts and objects that are derived from our personal experience and normative rules.

Finally, another approach to investigating implicit attitudes involved evaluative priming. Like in semantic priming experiments, the presence and strength of implicit attitudes is measured by the time it takes subjects to classify specific words into categories

after being primed by a valence. Again, the idea is that implicit attitudes work faster than explicit attitudes and should therefore take less time to be reflected in behavioral measures. However, instead of using semantic valences – like words commonly associated with stereotypes – evaluative experiments primed subjects by exposing participants to different photographs, such as either white or black faces (Jost et al., 2009, 44). Then, subjects were asked to categorize certain words as either positive or negative, as quickly as possible. Again, the time frame allowed ensured that respondents were operating on automatic, rather than controlled processes. Fazio et al. (1995) was one of the first to use this kind of model to measure implicit attitudes regarding race. The results found that when white participants classified words as positive, it took them less time when they had been primed with white faces than when they had been primed with black faces. This pattern of findings has been interpreted by many researchers as indicating an implicit positive association – or for my purposes, a positive implicit attitude – among white participants towards those who share their race (Jost et al., 2009, 44).

This research paved the way for Greenwald et al. to develop the Implicit Association Test (IAT), which has become the most widely used tool for measuring implicit bias (Jost et al., 2009, 45). Building on the basic framework of the evaluative priming experiments conducted previously, the IAT also measures implicit attitudes by measuring the latencies between being exposed to some kind of primer – either semantic or visual – and responding to a classification question. Specifically, it gauges differences in "how easy or difficult it is for people to associate individual exemplars of various social categories (whites vs. blacks, rich vs. poor, gay vs. straight, and so on) with abstract words and categories that have evaluative implications (e.g., good vs. bad, pleasant vs.

unpleasant)" (Jost et al., 2009, 45). Just like the evaluative primer experiments, subjects who take a shorter time to characterize positive words with white faces (vs. black faces), and conversely, take a longer time to characterize negative words with white faces (vs. black faces), are theorized to have an automatic preference for white faces. Someone who lacked such implicit preferences would be expected to have relatively equal response times across all categories (Jost et al., 2009, 45). The same structure has been applied to numerous other social categories beyond race, including age, disability, sexuality, and gender.

While the purpose of this paper is not to dissect the validity of such experiments, it is worth noting that Nosek and Smith (2007) have summarized the IAT exhibits construct, convergent, and divergent validity. Further, Phelps et al (2000) provided physiological evidence to support the conclusions repeatedly found in IATs assessing race. They found that IAT scores were heavily correlated with the magnitude of amygdala activation – which is the part of the brain responsible for producing fear responses – when subjects were exposed to black faces rather than white faces (Jost et al., 2009, 45). Similar results have been replicated by Mendes et al. when assessing white subjects speaking to predominately black audiences. The upshot of this that neuroscience corroborates the abductive physiological evidence gathered from experiments such as the IAT.

Given this brief overview of investigations into implicit attitudes in a necessarily sterile and controlled environment, it would be fruitful to look at cases that more closely resemble real-world scenarios. Thankfully, the literature of research into implicit attitudes is rich with cases where implicit attitude assessments, such as the IAT, were able to predict

the behavior eventually carried out by the subjects in question. Just a small subset of recent research has demonstrated the predictive validity of implicit attitude measurement tests in a variety of cases, including: hiring and resume evaluation, student behavior towards classmates and identity clubs and organizations, police officer behavior towards unarmed suspects, and treatment regimens assigned by doctors. These cases describe non-trivial scenarios where the decisions of the subjects would have serious implications for those involved. In order to actually engage in an investigation into responsibility as it applies to these mental states, we must view them in their complete, morally rich context.

Implicit Attitudes in the Real World

Published in July 2003, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination" is an attempt by Marianne Bertrand and Sendhil Mullainathan to investigate the pervasiveness of racial discrimination in labor markets. Every measure of economic success reveals significant racial inequality in the US labor market, with African Americans being twice as likely to be unemployed, according to the Council of Economic Advisers (Bertrand et al., 2004, 2). Assuming that black and white applicants have identical backgrounds in terms of work experience and skills, it would follow that employers would extend callbacks to roughly an equal proportion of black and white applicants. Such would be the case in a truly racially blind society filled with employers that followed through in their commitment to be equal opportunity workplaces and comply with the Civil Rights Act of 1964. The experiment by Bertrand and Sendhil recreated these exact conditions in order to find out if such colorblindness had been achieved by using the real world as their laboratory.

The field experiment consisted of sending resumes in response to help-wanted ads in Chicago and Boston newspapers, followed by measuring the number of callbacks each resume received for interviews. After creating a bank of artificial resumes, researchers "randomly assigned very white sounding names (such as Emily Walsh or Greg Baker) to half the resumes and very African American sounding names (such as Lakisha Washington or Jamal Jones) to the other half" (Bertrand et al., 2004, 2). The experiment consisted of responding to over 1300 employment ads in the sales, administrative support, clerical and customer service job categories, and sending nearly 5000 resumes. Their findings indicate a substantial difference in callback rates split across racial lines: "applicants with white names need to send about 10 resumes to get one callback, whereas applicants with African American names need to send around 15 resumes to get one callback" (Bertrand et al., 2004, 2-3). According to the authors, this 50 percent gap is very statistically significant, and they estimate that "a white name yields as many more callbacks as an additional eight years of experience" (Bertrand et al., 2004, 3). In addition to this crucial finding, the study also found that discrimination levels are consistent across industry and that employers located in more African American neighborhoods are slightly less likely to discriminate.

While this study provides convincing evidence that the presence of racial discrimination is a pervasive feature of searching for a job in two typical American cities, there are a few clarifications that must be made to relate the findings to our discussion of implicit attitudes. First, this study does not differentiate between explicit and implicit biases. It is entirely possible that a number of the potential employers reviewing the resumes did in fact hold explicitly racist views about the inferiority about African Americans. Nevertheless, it is nearly impossible to attribute all of the racial discrimination

captured by the study to explicit racists given the large sample size. Even if explicit discrimination could account for a portion of the findings, the very statistically significant findings indicate that there are most likely subtler forms of discrimination at play, including discrimination that would not be endorsed by the potential employers. The most plausible construction of the events of the study is that the vast majority of resumes were reviewed by people who explicitly express more or less egalitarian views and would not consider themselves as having prejudicial beliefs about any racial category. This is where my investigation into implicit attitudes becomes useful to explain this pattern of behavior.

Taken in this light, this study provides nearly a textbook example of a semantic priming experiment. The manipulated names that are heavily correlated with race serve as the semantic primer since they feature at the top of the resume and are most likely the first piece of information presented to the reader. The subjects (employers) are then directed to make a classification of either worthy or not worthy of receiving an interview. As such, this process almost perfectly mirrors the IAT, which asks how easy or difficult it is for people to associate individual exemplars of various social categories with categories that have evaluative implications. If it is easier for an employer to give a callback for an applicant with a white name (or positive valence) than it is to give a callback to a comparable applicant with an African American name (or negative valence), this is well explained by an automatic preference in favor of white applicants.

I am interested in moving beyond the behavioral data provided by this study to investigate the underlying mental processes at work. As described above, the potential employers displayed different behavior depending on whether the name at the top of the

resume was perceived as being associated to a white candidate compared to a black candidate. There is good reason to believe that this experiment tracks implicit biases because the mental attitudes being investigated operate quickly, automatically, and likely without awareness by the subject. No other information regarding race is provided, and the placement of the name (primer) at the top of the resume ensures that the information is presented quickly. The data indicate that the name had a causal effect, triggering one mental attitude for black applicants and another attitude for white candidates. This automatically formed attitude then likely colored the rest of the process of reviewing the resume. In other words, the implicit attitude interfered with the explicit reasoning process involved in evaluating candidates by making the reader more or less sensitive to certain information. Readers with an implicit bias against African Americans, for example, may see the legitimate qualifications on the resume as less impressive (or even fabricated) and judge their past employment as undemanding or requiring lower levels of skill. They might view the applicant more harshly for having gaps in work history, or look unfavorably on the school they attended. Each of these are examples of the subtle ways in which bias may be introduced into what should be an objective, reason-based process. Further, the fact that employers would offer such justifications for why they acted the way they did – as opposed to acknowledging their bias – indicates that the bias operates outside of their conscious awareness.

Building an Associational Model

As the name suggests, discussion surrounding the findings of the IAT focus primarily on the *association* between the two concepts or exemplars presented. The

commonly accepted view, by psychologists and the public alike, is that these kinds of tests measure subconscious or unconscious connections that our brains have made as the result of being conditioned to pair one with the other. When repeatedly exposed to two stimuli concurrently, our mind automatically begins to associate the two, even if they are words like thug and images of young black men. There is no higher level processing involved in the activity. We do not endorse, and might even object to the association in our conscious mind; nonetheless, habituation is a powerful tool and these responses can be deeply ingrained (Mandelbaum, 2016, 633). It is worth noting that endorsing a mental state or belief does not have to happen explicitly. To endorse a mental state simply means to acknowledge that it is an accurate description of one's mental condition. To endorse a belief is to accept that formulation as accurate to one's take on the world.

The divergence between the results of direct measures versus indirect measures led psychologists to posit the existence of implicit attitudes in the first place and this feature continues to pose an interesting challenge for researchers to explain. As previously mentioned, the first form of dissociation involves the subject's ability to consciously recognize the existence of his or her own biases. After taking various IATs, for example, subjects "expressed shocked and disbelief" when they learned that their direct and indirect results differed at all, meaning they were unaware that their behavior could reflect anything but what they consciously avowed (Johnson, 2016, 6). In other words, they could not locate the source of their biased behavior anywhere inside them, no matter how rigorously they engaged in introspection. It is worth mentioning that more recent studies have found that, among subjects who demonstrate diverging implicit and explicit attitudes, some have posited that the divergence could be attributed to learned stereotypes. Nevertheless, the

distinction between acknowledging that you could be plagued by attitudes you do not endorse is separate from finding the source of an attitude. This led to the initial hypothesis that whatever mental process or construct was responsible for the results of indirect measures must reside below the surface of what is consciously available to a given person.

The second form of divergence between implicit and explicit attitudes was recognized later after a more substantial body of evidence had been produced about a variety of target stimuli. Greenwald et al (2009) found that the degree of variance between direct and indirect measures depended on social sensitivity of the topic being evaluated. More specifically, research shows that when the topic is more socially sensitive – such as race or sexuality – the gap between the direct and indirect measures is likely larger than when the topic is less socially sensitive – such as brand preferences (Johnson, 2016, 6). Thus, the resulting behavior will be more closely aligned with what would be predicted from evaluating one's explicit commitments in cases where the attitude is less controversial.

Finally, Fazio et al (1995) provides a final way in which direct and indirect measures diverge. Using a three-pronged approach, first researchers gathered data on implicit attitudes using an indirect meaning-word matching task and then explicit attitudes using the Modern Racism Scale questionnaire. Lastly, students filled out a questionnaire composed of questions about the importance of not being perceived of as biased by society, which "was used to assess the students' motivation to control seemingly prejudicial reactions" (Johnson, 2016, 8). Their findings indicate that the correlation between one's direct and indirect measures varies depending on their motivation to control how they

might be perceived. The results between the first two assessments diverged when students were highly motivated to mask their prejudicial beliefs, but converged when students were not highly motivated (Johnson, 2016, 8). The crucial takeaway from this experiment is that the drivers of the indirect assessment results seem automatic, insofar as they cannot be controlled in order to keep up appearances of being unbiased. The results from direct assessments, on the other hand, are easy to manipulate because a subject can censor or revise their explicit beliefs depending on the context. In other words, "the less control available, the more likely we are to see results akin to those measured by indirect tests" (Johnson, 2016, 8).

Faced with this puzzle of divergence between direct and indirect measures, social psychologists posited that these findings provided evidence for the existence of dual mental constructs. In other words, they argued that there must be different mental constructs at play: one operating at the explicit level and captured by direct measures, and one operating beneath the surface which was tracked by the indirect measures (Johnson, 2016, 2). A mental construct is a combination of processes and representations, and therefore, theorists that argue for distinct mental constructs may find the distinction in the processes, representations, or both (Johnson, 2016, 15). Researchers thus hypothesized that implicit attitudes could explain the behavioral results of the indirect measures. Explicit attitudes result from reason-based and reflective processes about propositionally structured representations. Conversely, implicit attitudes were posited to be associatively generated and automatic (or the result of a reflex) about simple concepts (Johnson, 2016, 3). Theories of human information processing and cognition easily assimilated to this finding as evidence indicates that a large amount of cognition actually "occurs automatically,

effortlessly, and outside our conscious awareness" (Jost et al., 2009, 43). Models that recognize these attitudes as the result of distinct mental constructs are called dual process theories of implicit attitudes and have been the dominant view since the early study of implicit bias.

The MODE Model

One of the most developed dual-process models, the MODE model, separates spontaneous from deliberative attitude-to-behavior processes, and considers "*M*otivation and *O*pportunity to serve as the major *DE*terminants of which is likely to operate" (Fazio, 2014, 156). Essentially, the model argues that spontaneous processes are automatic and occur without any conscious reflection by the individual, as opposed to deliberative processes, which are controlled and reflective. By automatic, Fazio uses the definition used by Shiffrin and Dunmais in their characterization of automaticity: a process is automatic if encountering a stimulus activates the associated evaluation from memory without the individual's intent, and even if the individual is attempting to engage in another task (Fazio, 2014, 156). In other words, spontaneous processes require no effort on the part of the subject, and can occur even when the subject is burdened with other pursuits.

This formulation of spontaneous processes sets up a striking contrast to deliberative actions, which require both time and effort on part of the subject to engage in. Deliberative processes often take the form of weighing the costs and benefits of a certain action before picking a plan of action and thus, track the behavior recorded by direct measures, such as self-reporting (Fazio, 2014, 163). One of the ways in which this type of process differs from spontaneous processes is that it prompts the subject to engage in hypothetical

reasoning by considering their attitudes towards potential alternatives. Engaging in this type of higher level processing is taxing for the subject, requiring both sufficient opportunity and motivation. The opportunity provided is context dependent, and can include factors such as time allotted before making a decision, sufficient background or contextual data, and being able to devote sufficient attention to the task. Likewise, motivation is heavily context dependent as subjects can have a wide range of reasons to privilege deliberative processes (Fazio, 2014, 158). Making an extremely important financial decision, for example, will provide greater motivation than deciding what to order at a restaurant, and is thus more likely to engage one's reasoning faculties instead of their "gut feeling." The MODE model is one of the more widely supported dual-process approaches because it provides a solid account of which process occur depending on the conditions of motivation and opportunity.

While the MODE model presents this picture of automatic processes, there is only a cursory explanation about what and how concepts and processes are being "automatically activated." Essentially, Fazio argues that exposure to a stimulus automatically activates an attitude, often implicitly. In its most basic form, implicit evaluation occurs when "a stimulus in the environment activates the corresponding node in semantic memory, which in its turn automatically activates the node representing positive or negative valence, which in its turn influences a certain evaluative response" (De Houwer, 2014, 343). One way to spell out further details of this process is by borrowing from another well-developed dual process theory – the APE model – posited by Gawronski and Bodenhausen. According to the APE model, exposure to stimulus A immediately activates the closest mental concept to the stimulus through associative processes (Johnson, 2016, 15). Further, the activation

of concept A can trigger related concepts that are often associated – or often experienced concurrently – with the first concept. For example, the stimulus of seeing an old person would activate the concept "elderly" in the mind of the perceiver. Nearby concepts, such as "frail" or "wise", might also be activated. The degree to which they are likely to be activated is proportional to the strength of the associative connection between them, such as how many times the concepts have been co-activated (Johnson, 2016, 16).

One's mental attitude can best be conceptualized as one's reaction to the sum of the concepts that are activated by the stimulus. Thus, as in the case of the above example, the activated concepts need not elicit the same reaction. "Wise" is likely to elicit a positive reaction, whereas "frail" is likely carries a negative connotation. Depending on the strength of each of these responses, one's overall attitude might be positive, negative, or even neutral if they essentially cancel each other out. Once activated, the attitude influences how the stimulus is construed in the current situation, "either directly, as when the activated evaluation forms the immediate appraisal, or indirectly, as when it biases perceptions of the qualities of the object" (Fazio, 2014, 156). In other words, the attitude either triggers an immediate gut reaction (e.g. "yuck!") or acts a lens which the object is perceived through and affects its subsequent evaluations and interactions.

The pattern of divergence present in many of the experiments previously mentioned is well accounted for on the MODE model, and further experiments have supported findings about the role of motivation and opportunity in attitudes affecting behavior. The model leaves open the possibility that spontaneous processes can be the immediate result of exposure to a stimulus, but can then transition into deliberative processes under the right

conditions. These kinds of 'mixed processes' that involve both types of attitude-to-behavior mechanisms gives the model greater flexibility than more rigid accounts of the associational model. The impact of implicit attitudes, according to the model, will be reduced when individuals have both the "motivation and the opportunity to deliberate about the available information and, in so doing, overcome the influence of pre-existing attitude" (Fazio, 2014, 162).

Applying the framework to the job application experiment, the MODE model is most clearly explained by looking at how it accounts for a resume reviewer who harbors a negative implicit attitude, or an implicit bias, against African-Americans. Bracketing questions of opportunity and motivation, the name or stimulus at the top of the resume triggers a spontaneous process resulting in the activation of the concept "black", which is closely associated with the names chosen by researchers. For some individuals, the concept "black" may be enough to trigger an immediate negative response. For others, the activation of "black" will spread to and activate other nearby concepts that might be associatively linked, such as "thug", "lazy", "disrespectful", and so on. The negative reaction produced from the net sum of all of these concepts being activated constitutes one's mental attitude, which then colors the way in which they perceive the remainder of the resume. Being in a negative mental state, as opposed to a neutral or positive mental state, can easily result in perceiving and interpreting information differently, such as devaluing the qualifications of the applicant. This is one kind of description underlying the implicit bias posited by Bertrand and Mullainathan in their real world experiment.

Implications of the Associative Model

One of the reasons that associative models have enjoyed widespread support among social psychologists is that they allow one type of process to account for a suite of disparate phenomena. In particular, associational models of implicit bias attribute association as the driving force behind the formation and structure of mental states, as well as describing the ways certain mental states relate to others (Mandelbaum, 2016, 633). Combining these aspects provides some insight into how associations may be modified, and thus, test the hypothesis that associative structures are really what underlies implicit bias.

According to Mandelbaum, we can "infer whether a given cognitive structure is associative by seeing how certain types of information modify (or fail to modify) behaviors under the control of cognitive structures" (Mandelbaum, 2016, 634). Associative learning involves the conditioning of stimuli and responses, or even stimuli and other stimuli, through repeated patterns of reinforcement. Sometimes also called 'classical conditioning', the textbook example is Pavlov's dogs that salivated (response) at the ringing of a bell (stimulus) after repeated exposure to food (stimulus) only being offered after the bell had been rung. The only way to extinguish this kind of association-based behavior is to counter condition the two stimuli so that they are not presented together. In other words, to stop the dogs from salivating at the ringing of the bell, one must be careful to only ring the bell in the absence of food, and likewise only offer food when the bell has not been rung. The same logic can be applied to humans; sooner or later the association can be weakened or even extinguished (Mandelbaum, 2016, 634).

Extinction and counter conditioning processes not only explain successful ways to extinguish associations, they also indicate what approaches will be unsuccessful. If I crave a beer every time I finish my calculus homework because of a pattern of following such behavior, it would be silly to try to throw some kind of rational argument at me as a way to break the connection in my mind. Only after enough times of finishing my calculus homework and abstaining from drinking beer, or counter conditioning by drinking another beverage, will I eventually begin to decrease my craving. If rational argumentation, exposure to new evidence, or any other logical intervention is successful in modifying an implicit attitude, then that implicit attitude does not have associative structure (Mandelbaum, 2016, 635). Applying this framework to associatively structured implicit attitudes, Mandelbaum posits this characterization of the principle behind Associational models of implicit bias: Implicit biases (a) can be changed by changing certain environmental contingencies, and (b) can only be changed by changing certain environmental contingencies (Mandelbaum, 2016, 635). He then goes on to object to (b), presenting evidence that implicit attitudes are susceptible to logical intervention. To use terminology more widely reflected in the philosophical literature, being susceptible to logical intervention is equivalent to saying the implicit attitude is reason responsive. He goes on to argue in favor of the Structured Belief hypothesis, where implicit attitudes are propositionally structured which is necessary to account for the various ways in which they are adjusted in response to evidence and other logical interventions.

Entertaining a Propositional Model

While less common, there are advocates of the propositional model within the field of social psychology, including Jan De Houwer who published her own model in 2014. I will define a propositionally structured belief as "some attitude, stance, take, or opinion about a proposition or about the potential state of affairs in which that proposition is true." According to De Houwer, the critical difference between associational and propositionally structured attitudes is that propositional statements encode relational information (De Houwer, 2014, 344). Relational information allows for two concepts to be joined in more complex ways than a simple association. For example, consider the concepts "I" and "good". A propositional structure allows for a distinction to be drawn between the statements "I am good" and "I want to be good" – a critical difference.

To convincingly argue that implicit attitudes take on a propositional structure, at least in some cases, advocates of a propositional model must show that there are interventions besides extinction or counter conditioning that reliably modify implicit biases. A strict interpretation of the associational model would not allow for propositions at all, avoiding the question whether propositionally structured cognitive representations can be associated with each other (Mandelbaum, 2016, 638). The more charitable interpretation will relax this requirement, meaning it will allow for associations to hold between any kind of mental structure. Still, Mandelbaum details several studies that provide convincing evidence that implicit beliefs are reason responsive in three ways: implicit attitudes 1) engage with some form of balance theory processing; 2) respond to argument strength; and 3) can be the result of purely formal and symbolic learning. I will

now briefly outline each of the experiments that he takes to support each of these problems pose a problem for the associational model.

The first line of attack employs the well-recognized social and folk psychological principle that people interpret the enemy of their enemy to be their friend, which is captured by Heider's Balance Theory (Mandelbaum, 2016, 638). If I dislike Tom, and I know that Tom dislikes Jerry, then I am prone to like Jerry – at the very least I should like Jerry more than Tom. The reason I am likely to have this attitude is because I make a simple inference from the evidence that Tom dislikes Jerry. My opinion of Tom is negative; therefore, I have no reason to agree with his assessments of others. In fact, I have reason to think Tom is wrong and that Jerry deserves a positive, or at least neutral, assessment. Unlike a propositional model that readily accepts this inference, the associational model predicts the opposite outcome. When I pair a negative valence (my opinion of Tom) with another negative valence (Tom's opinion of Jerry), the resulting attitude should also be negative because I have no positive valence to which I can attribute positive reactions towards.

Gawronski et al. (2005) examined the effects of cognitive balance on implicit attitudes in 2005 by first showing participants a photo of an unfamiliar individual (CS1), which was then paired with statements that were either consistently positive or negative (Mandelbaum, 2016, 638). After conditioning participants to respond to the CS1 with the designated evaluation, researchers then introduced another unfamiliar individual (CS2) and participants were told whether the CS1 like or disliked CS2. Participants finally undertook an affective priming task to assess their implicit attitudes towards the unfamiliar individuals. At first glance, the associational model accounts for the data from the

experiment because those who were positively primed towards CS1, and were told CS1 had positive feelings towards CS2, had positive implicit evaluations towards CS2.

Nevertheless, the findings for subjects that encounter negative valences track exactly what Heider's Balance Theory would expect, which is the exact opposite of what the associational model would predict. Subjects who encountered a negatively valenced CS1 and were told that CS1 disliked CS2 actually had positive implicit evaluations of CS2 (Mandelbaum, 2016, 639). For example, if I was taught that Tom was bad and learned that Tom disliked Jerry, the experiment suggested I would automatically have a favorable implicit reaction to Jerry without ever meeting him, based solely on my knowledge of Tom and his opinion of others. The logic of 'the enemy of my enemy is my friend' holds true in this instance and its rational basis is obvious. It is then a challenge for advocates of the associational model to explain this result without reference to such logic, which is off-limits as associations have no such basis (Mandelbaum, 2016, 639).

The second way in which Mandelbaum finds support for rational processes at work in implicit evaluations is with regard to the strength of arguments and how they affected implicit attitudes. Brinol et al (2009) tested this by subjecting one group of participants to only strong arguments about hiring black professors (such as it would decrease class sizes and improve the overall quality of the faculty without raising costs), while another group were exposed to only weak arguments (such as it being trendy or would give current professors more free time). The arguments presented to each group were of equal length and mentioned "African American professors" equal amounts of times. Participants then took a race implicit association test. The data found that subjects "in the strong argument

group had more positive implicit attitudes towards African Americans than those in the weak group" (Mandelbaum, 2016, 640).

The content presented to each group was meticulously controlled to identical in all respects except for the strength of the argument. As a result, the only mechanism available to explain the discrepancy in evaluative responses is the engagement with subjects' higher-level reasoning faculties. In fact, the associational model would hypothesize that merely associating African American with professor (which is assumed to have a positive valence), should decrease the bias present in subjects (Mandelbaum, 2016, 641). Furthermore, even weak arguments contained positive content, providing another reason to expect the procedure to reduce bias. Nevertheless, only the strong arguments had any noticeable effect on reducing racial bias. These findings support the structured belief hypothesis because argument strength is exactly the kind of evidence that propositionally structured processes are equipped to deal with. Reasoning and inference play a key role in evaluating whether evidence is convincing or not.

Gregg et al. (2006) ran a series of experiments that directly tested the effects of conditioning versus reasoning on implicit attitudes (Mandelbaum, 2016, 643). Specifically, researchers were interested in probing a dual-process model of implicit bias that postulated the "existence of two complementary representational systems: a rule-based one, in which sudden transformations of serial representations (or symbols) occur, and an associative one, in which gradual transformations of connectionist representations (or weights) occur" (Mandelbaum, 2016, 643). In other words, researchers were interested in the effects of learning via assimilating the same piece of information multiple times, which they termed

"concrete learning", compared to hypothetically assuming that an object possesses particular characteristics, which they termed "abstract learning" (Mandelbaum, 2016, 643). To engage in abstract supposition, they posited, required entertaining "cognitions that were purely *formal* and *symbolic*," making them especially well-suited to explain attitudes that are "rule-based, rational, and constructed" (Mandelbaum, 2016, 643). Concrete learning, on the other hand, was well-suited to activate associational representations as it involved experiential and repeated exposure.

To test their theory, researchers created two fictional tribes: the Niffites and the Luupites. Participants were then split into a group of 'concrete learners' and a group of 'abstract learners'. The first group was then conditioned through a traditional approach – consisting of 240 rounds – by pairing strongly valenced words with each group (e.g. Niffites were paired with 'barbaric', while the Luupites were paired with 'benevolent') (Mandelbaum, 2016, 643). Abstract learners were instead asked to suppose that there were two such tribes, one that was peaceful and the other savage. Importantly, this group was subjected to no conditioning at all. All participants then took a IAT that tracked good and bad associations with the two tribes. The results found that no differences were found between the two differently conditioned groups (Mandelbaum, 2016, 644). For supporters of associational models, this poses several problems. First, the fact that an association could be generated without any conditioning and only considering a hypothetical challenges the assumptions of how such implicit attitudes are developed. Beyond this, the strength of the implicit attitude should at least be stronger for the participants who endured 240 rounds of classical conditioning, compared to the group who merely considered one sentence.

Finally, researchers replicated the experiment, this time with a twist. Participants underwent the same procedure, and then those in the abstract learning group were asked to suppose two more hypothetical groups and were told that one was equivalent to each of the previously mentioned groups (Mandelbaum, 2016, 644). The mere mention of equivalence between groups was enough to make the results of a corresponding IAT indistinguishable across category for the abstract learners. They had the same attitude strength towards the second set of hypothetical groups as the first. These results indicate that implicit attitudes can have "cognitive effects that are not predicated on chains of conditioning, but are modulated based on acknowledgement of logical equivalence" (Mandelbaum, 2016, 644). The problems for defenders of associational models became even worse when researchers then tried to counter condition the hypothetical learners through classical conditioning and the attempt failed to extinguish the original attitudes. Logical intervention, on the other hand, was successful when the researchers informed subjects that a mistake had been made earlier and the two groups had been inadvertently switched (Mandelbaum, 2016, 645).

In the words of the researchers themselves, "our first two experiments therefore empirically contradict what dual-process models can plausibly be taken as imply, namely, that automatic attitudes are relatively immune to sophisticated symbolic cognition" (Mandelbaum, 2016, 645). If the associational hypothesis were correct, intensive evaluative conditioning should "create stronger attitudes than merely giving subjects a single piece of counter attitudinal information" (Mandelbaum, 2016, 645). Nevertheless, the data shows that this is not the case, and further, logical intervention was more successful than counter conditioning. The defender of the associational model would be hard pressed to explain these phenomena on purely associational grounds; the mental processes required

are far from the kinds of automatic, non-rational processes are thought to produce the results of direct measures.

In addition to the studies referenced by Mandelbaum, Jan De Houwer provides evidence that propositional attitudes can be formed automatically through the process of task misapplication (De Houwer, 2014, 345). Further, she goes on to argue that propositional attitudes can also be stored in episodic-like memory and activated spontaneously when the correct stimulus is presented. There is no reason to think that propositional processes cannot be triggered automatically and occur outside of our conscious awareness. Such a narrow view of propositional mental states would fail to give credit for the kind of everyday tasks that involve reasoning but nonetheless feel like second nature with enough practice such as checking your mirrors while driving or dribbling in basketball. Mandelbaum goes on to argue that, beyond the evidence from experiments, the fact that purely associational models have been abandoned in other fields of psychology – such as psycholinguistics – is evidence that much of the literature on implicit bias may be working on outdated assumptions about how our mind operates (Mandelbaum, 2016, 646).

Nevertheless, the purpose of this investigation into the nature and structure of implicit attitudes is not to present a definitive, knock-down argument that implicit attitudes are either entirely associational or propositional in structure. In fact, as previously mentioned, the MODE model could possibly accommodate many of the findings that Mandelbaum points to as evidence for the presence of propositional processes by attributing sufficient motivation or opportunity on part of the subjects. Thus, the main take away of this discussion is that implicit attitudes are an extremely heterogeneous group of

phenomena and no one theory has been demonstrated to consistently explain all instances or features. Importantly, however, this more detailed account is quite different than the folk understanding of many towards implicit attitudes and biases. Rather than viewing implicit attitudes as no different than the knee-jerk reaction that occurs when a nerve is stimulated, we should appreciate that the behavior tracked by direct measures is the result of a much more complicated process that gets to the core of how our minds work.

There are four main features of implicit attitudes highlighted in the above discussion that will become extremely relevant as we turn to probing responsibility for implicit attitudes. First, both empirical research and philosophers of mind agree that one key distinction between implicit and explicit attitudes is that the prior are unconscious – we do not have access to them as we do our regular beliefs about the world. While respondents have been able to offer post-hoc explanations that they might harbor implicit biases, this prediction is quite different from being able to trace the process that resulted in the indirect measures. Second, implicit beliefs operate immediately, meaning they happen too quickly to have emanated from conscious decision-making. I am hesitant to say they happen automatically, unlike De Houwer, because research has shown they can be mediated in some circumstances. Automatic seems too restrictive and reductive to capture the more detailed picture presented above. Third, implicit attitudes are unendorsed, which follows from the above requirement that they are also unconscious. Of course, there could be implicit attitudes that closely track explicit attitudes, but because the divergence between the two is what enables their measurement and testing, the only implicit attitudes I am interested for these purposes are those one would object to as accurately representing one's mental state. Finally, implicit attitudes are – in at least some cases – reason

responsive. Several studies have demonstrated that implicit attitudes can be altered or extinguished through rational argumentation or changing evidence, rather than repeated exposure through classical conditioning. While this might hold for only a subset of implicit attitudes, my ultimate account of responsibility will not hinge on reason-responsiveness as a necessary condition and thus, the small but significant evidence presented thus far is sufficient.

The Dispositional Approach

It is easy to get bogged down in psychology experiments where researchers exchange blows, trying to find a certain kind of indirect measure that tips the scales in favor of their favored hypothesis. While this approach is fruitful for understanding the nature of and processes behind our implicit attitudes, I would now like to take a step back. The kinds of day-to-day interactions that I am interested will never be fully captured in sterile environments with carefully controlled stimuli and responses. Therefore, I will now propose an alternative method to thinking about our mental attitudes.

Consider the case of Julie the implicit racist as inspired by Eric Schwitzgebel. Julie is a white professor at a small, liberal college who staunchly expresses and defends egalitarian views. She makes conscious efforts to include authors from diverse backgrounds when preparing syllabi for her classes, attends lectures on the racial disparities in the criminal justice system, and even supports eliminating using standardized testing due to researching showing it correlates more strongly with race than actual success in school. In other words, she is as woke as can be. Nevertheless, Julie consistently displays racial prejudice in many of her spontaneous reactions (Schwitzgebel, 2010, 532). For example,

even though she explicitly denies difference in intelligence based on race, she is often surprised when her black students in class make a good point in discussion, whereas she would simply expect the same from a white student. Her bias extends into her grading of exams and participation, and even to interactions with black non-students as well. As a member of the hiring committee, she would fail to recommend a black person as most qualified or require much greater evidence than would be expected of a white candidate. Further, suppose that Julie is fully aware of this pattern of behavior and even strives to be better, deliberately attempting to overcome her bias in particular cases. Still, she cannot maintain constant self-vigilance and exercise perfect control, and her well-intentioned efforts can even be construed as patronizing condescension.

If the results of the IAT are any estimation, people with attitudes like Julie's are a large portion of the population. What is to be made of these mixed cases? Her explicit avowals of egalitarianism, and history of openly putting those beliefs into practice, indicate that she harbors no prejudicial conscious beliefs about members of other races. On the other hand, however, she consistently treats and forms opinions of others based on their race, suggesting that she harbors some form of racial prejudice. What form does this prejudice take? It does not quite amount to a belief in the same way that she *believes* that all races are equal. After all, it would be a contradiction to believe P and its negation simultaneously. Nevertheless, the phenomenon goes beyond simply having a negative affective or emotional response because she expresses some attitude towards an object – namely people who do not share her own race. Therefore, taking a snapshot of Julie's mental state will always provide an incomplete picture. If we were to ask those around her, the answer to whether or not they think she has prejudicial views will likely depend on the nature of their

interactions. Those who have seen her champion diversity programs will emphatically reject that assertion, but students of color would also be justified in asserting the presence of her prejudice as they might be well aware of her pattern of condescension or aversion towards black students.

To say that Julie is prejudiced sometimes and not others is unhelpful if we are trying to sketch a general account of moral responsibility that goes beyond discrete interactions. Further, this formulation seems detached from the way that prejudice is commonly talked about as a deeply engrained character trait that does not fluctuate by simply changing circumstances. For a better approach to understanding our attitudes, Eric Schwitzgebel argues that we should adopt a dispositional approach, which he argues is a vague label that admits of in-between cases of belief rather than a bright line boundary (Schwitzgebel, 2010, 533). To have an attitude, on his account, is to be disposed to act a certain way, or more specifically to be "apt to interact with the world in patterns that ordinary people would regard characteristic of having that attitude" (Schwitzgebel, 2013, 2). As a general account, this description of attitudes is intended to capture all propositional attitudes (believing, desiring, etc.), reactive attitudes (resenting, appreciating, etc.), and other attitudes directed towards people, things or events (loving Tim, hating jazz, etc.) (Schwitzgebel, 2013, 1).

To have a belief on the dispositional account is to be "disposed to act and react in various ways in various circumstances based on a broad dispositional base" (Schwitzgebel, 2010, 533). In other words, believing that P goes beyond having P in some metaphorical "belief box", it involves one's ability and tendency to do things of lots of different kinds. Gilbert Ryle uses the example of ice-skating and believing that the ice is dangerously thin.

In these circumstances, believing that the ice is dangerously thin is "to be unhesitant in telling oneself and others that it is think, in acquiescing to other people's assertions to that effect, and objecting to statements to the contrary" (Schwitzgebel, 2010, 534). Further, to have that belief is also to "be prone to skate warily, to shudder, to dwell in imagination on possible disasters and warn other skaters" (Schwitzgebel, 2010, 534). As Ryle puts it, believing that P is not a propensity to make only theoretical moves; it invites the believer to make certain executive and imaginative moves, as well as have certain feelings. In other words, believing that P simply involves acting as if P obtains in the right sorts of circumstances. Thus, to have a belief is to match a certain dispositional profile, all else being equal.

While this approach to thinking about beliefs may appear ad-hoc at first, its marked advantage is that it provides a more useful framework for evaluating in-between cases, or those where an agent does not *fully* match the dispositional profile of holding a certain belief. Sometimes a person will fail to possess all elements of a dispositional structure that is relevant to a belief. For example, the ice-skater could fail to warn other skaters of the thickness of the ice due to being in a bad mood, but failing to meet this expectation of the dispositional account does not reduce the fact that he believes the ice dangerously thin. It would be silly if holding a belief meant checking all of the boxes of behavior that is expected under those circumstances (Schwitzgebel, 2013, 3). We readily acknowledge that many people can share the same belief by demonstrating the same overarching pattern while failing to participate in the exact same actions of others. Therefore, the dispositional account of belief must define believing as meeting *enough*, but not all, of the relevant features of the dispositional profile.

Once we acknowledge that one can believe while only possessing part of the dispositional structure that is characteristic of that belief, we can move beyond treating beliefs as a strict binary where the agent either possesses the belief or does not. Instead, there must be some kind of continuum of belief ranging from full possession of all the relevant dispositions to possession of none of them. Thus, we can ask to what degree does someone possess a certain belief, rather than if they simply do or do not (Schwitzgebel, 2010, 534-5). For example, consider the belief that California is the best state in the United States. My neighbor and I both may hold this belief, but he may exhibit more of the behaviors that are characteristic of holding that belief, such as proudly displaying a California flag in his front yard while my yard remains barren. In adopting this view of beliefs, we reject that there are bright-line distinctions between believing P and not believing P. The dispositional approach treats belief as a necessarily vague term. Thus, when discussing cases where broad dispositions are only partly possessed, the dispositional model prompts us to tread more carefully than simply attributing or denying beliefs. If my motorcycle almost always starts, except when it has been especially cold, it does not seem right to say that it either is or is not reliable. I should instead answer with a more careful description and specify the conditions under which it is and is not likely to work (Schwitzgebel, 2010, 535). The same can apply to people. If I am always honest, except when I go on dates, it would be wrong to call me truthful *full-stop* – it would depend on the exact circumstances.

Coming back to the case of Julie, the dispositional model prompts us to reject the binary of Julie as either prejudiced or not. On the one hand, she has the appropriate dispositional structure in some respects: she is disposed to affirm both inwardly and

outwardly that all the races are intellectually equal and to admonish those who express blatantly racist opinions. On the other hand, however, she fails to possess an egalitarian dispositional structure: many of her emotional reactions and spontaneous judgements are in fact prejudiced (Schwitzgebel, 2010, 537). Julie is a cut and dry case of in-between belief where a simple attribution or denial of prejudice fails to be satisfactory. We can narrow the scope by asking if she is prejudiced in the context of a debate, where her conscious commitments will prevail. In that case, her peers would agree she expresses explicit egalitarian views. Likewise, we can confine the scenario to recruiting new professors, where it would be wrong to ignore her prejudice when a person of color is competing for a job and her implicit attitudes result in discrediting the applicant's qualifications.

Rigid definitions of belief are ill equipped to assist in the investigation of responsibility for implicit attitudes. Given that implicit attitudes are not considered completely analogous to structured beliefs, the dispositional model demonstrates that one cannot *fully* believe something if their implicit attitudes result in them failing to meet some of the relevant dispositional features. In other words, explicit commitments and implicit attitudes are both factors in the overall holding of a belief and it is their sum that creates our overall dispositional profile. Each affect different parts of our dispositional structure, therefore, when investigating questions of responsibility for believing, we can distinguish responsibility for our explicit commitments as separate from our responsibility for our implicit attitudes. Further, the dispositional model highlights that accounts of belief must admit of in-between cases where a clear binary is insufficient.

While we cannot forget the key distinctions highlighted earlier (unconscious, unendorsed, immediate, and sometimes reason-responsive), there is more in common with implicit and explicit attitudes than one might be led to believe after surveying academic literature and folk psychology. If there were a spectrum of mental states from reason-based, structured beliefs to more basic affective states like hunger or pain, implicit attitudes would fall somewhere in somewhere in between. They share a number of features with structured beliefs that also clearly distinguish implicit attitudes from other states that we certainly cannot be held responsible for. First, they are directed towards certain persons or objects. Second, implicit attitudes are more complex in how they come about: they are not purely responses to environmental or bodily triggers that occur in the background as part of our body's almost mechanical functioning. Instead, they develop over time through experience, draw on memory, and perhaps include some kind of inference based on evidence in some cases. While one of the characteristic features of structured beliefs is that they are responsive to change when presented with new evidence, we often find ourselves in situations where we are reluctant or unwilling to change our conscious commitments. Beliefs that are central to one's character, conception of self, cultural identity, and so on, are often resistant to modification. The sum of our implicit and explicit attitudes are what constitute our dispositional character, and thus our overall belief system on the dispositional model. Thus, treating implicit attitudes like conscious beliefs may prove useful in the overall investigation into accounts of moral responsibility.

Resentment and Responsibility

Before moving onto my account for responsibility for implicit attitudes, I would like to make my starting point clear. First, I am interested specifically in investigating moral responsibility, or under what circumstances we can rightly be praised or blamed. Questions of political and social responsibility are essential to any discussion involving implicit attitudes and prejudice, especially in the context of tracking the sources of and avenues to modify implicit attitudes. Systemic factors have led to the marginalization of a number of groups and identities, building on and shaping the kinds of prejudice that exist today through the images and tropes present in media we consume, the nature of our commercially driven economy, and the very structure of our government. Nevertheless, discussions of political, social, and shared responsibility often abstract away from the personal level. Without an individual and moral account, however, it is all too easy to remain apathetic by focusing at the high level, instead of looking to the kinds of everyday interactions that are no less important. By providing a positive account for responsibility of our implicit attitudes, I hope to provide a crucial first step in the path towards the overall fight against prejudice by giving individuals a stake to both improve their moral character, as well as expand into collaborative efforts towards more structural advancements.

For the purposes of this paper, I will bracket the issues of moral skepticism as well as the debate between determinism and free will as it relates to responsibility. I will assume that, as humans, we have some kind of basic moral obligations towards others. Building on the work of Scanlon, Strawson, Hieronymi, and others, I will also assume that these moral commitments extend beyond the sphere of action. Beliefs, and other kinds of conscious, reason-based mental states, as they relate to others, are parts of ourselves that we can be

held morally responsible for. For example, I will take it as a given that one can be held responsible for holding the view that some people are not actual people by nature of their skin color. By moral responsibility, I am not interested in issues of blame or liability, nor do I seek to prescribe any specific consequences—legal or otherwise—that should follow from being found responsible. Further, I am also not interested in features of responsibility that emanate from one's relationship with another, aside from their shared connection as humans. Sometimes called substantive responsibility, this would include the increased responsibility of a parent towards his child or a politician towards her constituents. In the following pages, I will consider someone morally responsible when they can rightly be praised or blamed for their action or mental state.

Narrowing in on Prejudice

To better understand the moral injustice that can result from implicit beliefs, I will demonstrate that the same sort of injustice is present regardless of if the belief or attitude is conscious or unconscious. Since beliefs or attitudes are directed towards people, viewing the interaction from the perspective of the agent whom the belief is directed toward will provide a useful perspective to identify and evaluate the moral wrongdoing. Returning to the employment study described earlier, consider the case of an explicit racist who consciously and openly acts on beliefs that black people are lazy and less intelligent than white people when sorting through applications for a job opening. For the qualified black applicants who fail to receive an interview—much less a job offer—there are tangible harms done by unfairly limiting their opportunity to employment and all of the benefits that come with it. While important, these facts alone cannot account for the kind of (justified) resentment the applicants would feel towards the racist manager. For example,

the applicant could receive a job offer the very same day as getting rejected, but would still be justified in feeling upset with the racist manager because of his views toward her.

One potential, yet extremely unsatisfying, justification for the resentment the applicant feels towards the manager is that he was irresponsible in his reasoning by letting generalizations and stereotypes guide his hiring decision. Nevertheless, even if we assume that members of the stereotyped group are more likely to have these traits than the white applicants, she is still justified in feeling resentment against the manager. She may blame him for a flaw in his reasoning or for making an inference with too little evidence to support it, but that is only a fraction of what is morally problematic in the interaction. The deeper moral injustice for which the manager is guilty of is that he failed to treat and view her like everyone else: he assessed important features of her character negatively based solely on race. There are a number of different avenues to describing the specific injustice committed by the manager, but they all appeal to the same idea that the applicant's status as a person means that one's interactions with them ought to be guided by certain, widely agreed upon principles of fairness, respect, and equality.

Despite Kant's unfortunate history of harboring racial prejudice himself, a Kantian approach might appeal to the notion of respect for persons put forth in the Humanity Formulation of the Categorical Imperative, which requires humans to treat others always as ends in themselves, not just mere means. Respect for the humanity in persons requires that we treat them with some sort of regard that recognizes and does not constrain all of the features that make us distinctively human, including our capacities to engage in self-directed rational behavior and to identify and pursue our own ends (Arpaly, 2002, 237). In

other words, our interactions with others should be governed by principles that respect their agency. Part of what is required to respect one's agency is to be judged and evaluated by the actual content of one's moral character, not by generalizations about other matters that are irrelevant, such as race. Because Kantian respect requires evaluating agents as unique individuals, rather than appealing to one's membership to certain groups or communities. While this general line of argumentation could be expanded in greater detail, there is a more straightforward path, however.

Strawson's Account of Moral Responsibility

In *Freedom and Resentment,* P.F. Strawson developed his own account of moral wrongdoing that locates the analogous principle of basic respect for human beings as the foundation for our moral obligations. He begins by describing how the long-running debate between determinists and libertarians over the concept of freedom is extremely unlikely to produce any definitive answers (Strawson, 1968, 73). Thus, approaches to responsibility that hinge on the existence of freedom to act are necessarily limited in their overall usefulness. Instead of waiting for an answer that will never arrive, he argues that we should abandoned this detached approach for one that instead looks to the "non-detached attitudes and reactions of people" engaged in commonplace interactions. Feelings and reactions such as "gratitude, resentment, forgiveness, love, and hurt feelings" will provide useful signposts for an investigation into moral responsibility because they largely "depend upon, or involve, our beliefs about [one's] attitudes and intentions" (Strawson, 1968, 74). If someone breaks my property accidentally while trying to help me, my feelings of indignation towards them will be minimal compared to the resent I would feel if they acted out of contemptuous disregard for my existence. Thus, my reactive attitude and emotion is

useful evidence as it helps track what matters to us, namely whether their attitude toward me was of "goodwill, affection, or esteem on the one hand, or contempt indifference, or malevolence on the other" (Strawson, 1968, 75). One of the benefits of this approach is that we place different value on ideals such as self-esteem, love, security, and human dignity depending on the circumstances. Our reactive attitudes are helpful insofar as they adjust to these changing circumstances and do not confine us to privileging any one of these ideals as supreme.

Drilling deeper, Strawson then focuses on resentment as a primary reactive attitude to track intuitions about questions of responsibility. Specifically, he is interested in when it is appropriate or reasonable to feel resentment, and what special conditions must be present to not feel resentment when it would otherwise be expected. Strawson has in mind situations "in which one person is offended or injured by the action of another and in which – in the absence of special considerations – the offended person might naturally or normally be expected to feel resentment" (Strawson, 1968, 77). He identifies two broad categories of considerations that could make feeling resentment inappropriate: those that encourage modifying our emotional response and those that lead us to withhold our reactive attitudes altogether because the subject cannot rightly be called an agent.

Within the group that treats the subject in question a responsible agent in typical circumstances, there are a number of pleas one might make that provide exculpating or ameliorating reasons to modify our reactive attitudes. I will use exculpating reasons to describe the kind that give reason to think that the action was in fact not a true reflection of the agent's will, but rather caused by something else entirely (Strawson, 1968, 83). For

example, if my friend cancels plans with me because he got a flat tire and cannot drive, his response of "I couldn't help it" encourages me to not interpret his cancellation as a true reflection of his will towards me. Rather, there are other circumstances, namely the nail in his tire, that are motivating his action. While he might generally be considered a fully responsible agent, these specific circumstances are beyond his control and there is nothing he did or could do to reach a different result. Thus, his exculpatory plea prompts me to withhold my reactive attitudes towards him because they are responding to other factors external of the content of his will. I may still be justified in feeling upset about not getting to see my friend, but it cannot be called resentment because it is not directed at a responsible agent.

Ameliorating reasons, on the other hand, are not as straightforward as exculpating reasons that remove the content of the will of the subject as a driving factor. The subject's action is still a reflection of his will in some way, but ameliorating reasons are the kind that provide greater context about what might be complicating his action, such as competing motivations or commitments. To clarify, consider again my friend cancelling plans, but instead of having a punctured tire, he has an assignment and "couldn't help it because he was too busy". Because his cancellation is still within his voluntary control and external factors are not directly limiting his ability to act, my reactive attitude of being upset is still in response to the content of his will.

However, we can provide context that will modify the kind of resentment I feel towards him. Consider that the assignment he is working on is something that was assigned weeks ago and he has been putting it off it until today. I would be justified in feeling

resentment towards him in this case because it reflects a certain level of indifference felt towards me. If he had felt greater good will towards me, he likely would have gotten his assignment done earlier so that he could spend time with me. On the other hand, if the assignment was unexpected, I should moderate my reaction to feel less or even no resentment. He is still an agent in both cases because his actions are freely under his control, but there are varying reasons to think that his cancellation is more or less a reflection of his true will towards me. Thus, ameliorating reasons typically the take the form of the various kinds of contextual factors we regularly consider when attributing responsibility, and the degree to which they modify our attitudes are reflections of how good of a reason they are. Still, ameliorating reasons "do not invite us to see the agent as other than a fully responsible agent" because the subject is not simply at the mercy of factors unrelated to his will (Strawson, 1968, 75)

Beyond exculpating or ameliorating considerations, Strawson identifies that sometimes we are asked to view the agent in a different light than under normal conditions because the agent is presented as "psychologically abnormal or as morally undeveloped" (Strawson, 1968, 75). 'She's just a child' or 'he is schizophrenic' are pleas that would fall under this category. Here our reactive attitudes must be withheld because it would be inappropriate to demand the same kind of attitude of goodwill given the cognitive and social limitations of the agent. Instead of engaging in the full range of attitudes involved in participation in complex human relationships, we must instead adopt a different approach, which Strawson calls taking an objective attitude toward the agent (Strawson, 1968, 76). When a child yanks a dog's tail for example, we must suspend the resentment that would be warranted if an adult had acted the same way. We instead approach the child as someone

to be coached and taught to do better. Thus, to have an objective attitude towards someone is to treat them as an object of social policy, something less than a full agent. Conversely, to adopt a participant attitude means to treat them as a fully-fledged participant in interpersonal relationships. Their lack of full agential status makes it inappropriate to feel our reactive attitudes as normal because they cannot express the same kind of ill-will that an agent can, or at the very least, facts about their person give reason to believe that their actions are somehow distorted from conveying their true attitudes.

Strawson acknowledges that we can adopt an objective attitude towards average agents, albeit for only a limited time (Strawson, 1968, 77). To adopt an objective attitude towards someone is to become blind to whatever good or ill will or indifference they possess. For example, in an attempt to not entangle myself in a dispute between my roommates, I can temporarily adopt an objective attitude towards them and their actions. I could likewise treat my partner with an objective attitude if she asked for help breaking a bad habit. It would, however, be impossible to suspend my reactive attitudes in all of my interactions within these relationships. While this might be helpful in a narrow set of cases, the objective attitude is ill-equipped to capture how we actually treat our relationships with others because we do in fact care about the contents of other peoples' wills as they are directed towards us. To have a fulfilling and productive relationship with my roommates, knowing how they feel about me and whether they show good or ill will is crucially important. My romantic relationship with my partner places an even greater value on knowing the will and intentions of both parties involved because the very relationship is predicated on feelings of good will towards one another.

To suspend one's reactive attitudes in the context of interpersonal relationships would fundamentally alter the relationship so significantly that it would fail to resemble the normal human interaction that we now recognize. The relationship would devolve to one that is purely detached and eventually dissolve altogether. It would require failing to see those around us as full agents capable of possessing a wide range of attitudes and pursuing their own projects. We would not be able to feel the kinds of reactive attitudes that are characteristic of every day human experience, such as feeling gratitude toward a helpful friend or resentment towards someone who disrespects us. Even though we can adopt the objective attitude temporarily, that makes it no less ridiculous to think that treating others as cognitively or morally abnormal should be the default condition. Instead, we should respect how emotionally rich our interactions with one another actually are by recognizing the value of our reactive attitudes in informing our relationships, and reserve the objective attitude only for cases when it is appropriate.

The kinds of reactive attitudes discussed thus far, such as resentment and gratitude, are essentially reactions to the quality of others' wills towards us as evidenced by their actions. When someone shows indifference or lack of concern, my feeling resentful follows from their disregard in my interest as a full agent. Thus, our feelings are built on, and reflect, "an expectation, and demand for, the manifestations of a certain degree of goodwill on the part of other human beings" towards us (Strawson, 1968, 78). There are analogous feelings associated when we perceive an injury done to someone else. When my friend is rude towards a waiter, my feeling disapproving of their action is a reaction to the quality of their will directed towards the waiter. By stepping away from our personal claims of how we think we ought to be treated by others, we remove our stake from the game, thus

removing self-interest as the source of our reactive attitudes. Instead, they take on a generalized form about the kind of goodwill we expect to be manifested in all human beings as they interact with anyone who can experience moral indignation – which we now think of as all humans, minus those who qualify as abnormal as described above.

Having considered our reactive attitudes as evidence of the demands we make of others in their interactions with us and with others, there is another feature which, once highlighted, will complete the account of reactive attitudes serving as a basis for moral appraisal. Beyond the logical connection between our demands on others, there are "self-reactive attitudes associated with demands on oneself for others" (Strawson, 1968, 80). By recognizing others as agents and thus rightly having certain demands of good will, our status and commitment to being an agent necessarily implies that we can rightly ask the same of ourselves, highlighting the human connection we share. Our demands of a certain level of good will from others directed towards us and towards other people imply that others place similar demands of good will on us. Pulling these features of reactive attitudes together, we arrive at a standard of morality that approximates the Golden Rule of which we are familiar: we should treat others the same way that we would ask them to treat us. These demands of certain forms of inter-personal regard come from our shared membership to the human community of agents in constant interaction with one another.

As witnessed in the non-generalized attitudes, there are some cases where we see someone in a different light than typical agents. Perhaps their picture of the world is a delusion or completely unintelligible to any kind of conscious reasoning. In any case, we see their actions as wholly lacking in the moral sense because they are not able to engage

like typical agents with others and sustain deep, interpersonal relationships. This would include those who are cognitively disabled in certain ways, as well as young children. Our inclination to feel resentment towards someone like this is inhibited because all of our reactive attitudes are inhibited. It is inappropriate to place the same kinds of demands on them as we do on typical agents. The same holds for generalized reactive attitudes that take on a moral character: we take the objective stance and see them as an object of training and coaching. Taken in this light, the abnormal person is not "seen as a morally responsible agent", as one who participates in moral relationships or is part of the moral community (Strawson, 1968, 86). Consider the case of a young child that is still learning about what behavior is expected of him and of others. Our relationships with him is one of training and coaching when he does something problematic. We do not try to blame him, but rather explain why his action was wrong and what he can do better in the future. He does not comprehend that others would be upset with him for his actions because they show his disregard for their well-being Thus, we must view people like him as inappropriate objects of moral responsibility, at least while they are young. Nevertheless, as described earlier, these cases are limited and it would be absurd to adopt this as the default position. Most people then, barring these extreme cases, are appropriate objects of moral responsibility insofar as they participate in and sustain characteristically human interactions.

Strawson set out to provide an account of responsibility that did not rely on the truth or falsity of the determinist thesis. His main critique against existing approaches to responsibility is that they occur in the arena of intellectual debate, instead of looking to how we experience morality as a fact of everyday life emanating from the feelings we have as a result of our interpersonal relationships. They rest on the assumption that to hold

someone responsible, they must meet certain objective conditions of being responsible, and those conditions themselves are justified (Strawson, 1968, 72). Strawson rejects that such objective conditions exist, arguing that holding people responsible is just something that we naturally do and thus, it needs no external justification. Because holding people responsible is part of our social practice, adopting the objective attitude as a default position is unsatisfying because it sanitizes our interactions of what we generally consider important. If I am wronged, the purpose of making a moral assessment of the one who wronged me is not simply to treat him as someone whose actions should be corrected. Holding someone responsible only as a means to control their behavior is an affront to our commitment to treating individuals as agents. Thus, my moral assessment of wrong doing goes deeper by providing justification for withholding good will towards the offender (Strawson, 1968, 77). He did not demonstrate good will towards me, but reaction of resentment contrasts with his lack of feeling resentment towards me when I rightly blame him for wrongdoing.

Our reactive attitudes serve as helpful tools to adjudicate whether responses are appropriate, not because they are grounded in an objective theoretical schema, but because they directly track what is important in the practice itself. They prompt us to investigate the nature of our interaction with others and test what reasons count as legitimate mitigating factors. If the subject in question is able to participate in human interactions, that is sufficient to view them as a morally responsible agent. They regularly participate in the practice of being held responsible and holding others responsible because that simply is part of what human interaction requires: being sensitive to whether someone treats you with the appropriate regard, is indifferent, or even expresses ill-will. On Strawson's view,

we are responsible for things that are reflections our will, and morality is a useful part of our social practice that provides guidance about what kind of will we should express towards others.

Applying Strawson

Returning to the case of an explicitly racist manager making decisions on which applicants to extend interviews to, Strawson's account provides a convincing explanation of the resentment the applicant would rightly feel against the manager. His explicitly prejudicial views in this situation are crystal clear expressions of his ill-will towards people of certain ethnicities. By holding beliefs about black people that are lazy or unintelligent, he demonstrates the content of his will as directed towards them, and is one of disrespect, loathing, and one might rightly say even hateful. As someone who participates in complex social relationships and is not cognitively underdeveloped, he is rightly treated as a full agent. Further, there are no exculpating reasons to think the black applicant is reacting to factor external to the manager's will. Likewise, there are no real ameliorating factors that would lead her to temper her resentment. Therefore, we can hold the racist manager fully responsible for his explicit prejudice on Strawson's account, but note that this judgment does not rest on specific theoretical judgments. Rather, he is responsible by meeting the minimum requirements of agency and blameworthy by the nature of his ill-will directed towards a certain racial group. Returning to the investigation into implicit attitudes, I will now describe an everyday situation to test Strawson's framework against. If the requirements of agency and harboring ill-will are present, then the argument for holding an implicitly prejudiced person responsible will follow a similar path as holding an explicitly prejudiced person responsible.

Consider the case of Michael, a college undergraduate who happens to be black. Upon boarding the subway one day, he realizes that a lady sitting across from him, Julie the Professor, quickly checks to make sure her purse is zipped shut and clutches it more tightly against her side while glancing at Michael. This sort of action is characteristic of our professor outlined earlier. While Julie would explicitly condemn prejudice, she is nonetheless susceptible to acting in ways that are in line with having negative attitudes towards people of color. This case is precisely one of those times: she formed a negative implicit attitude upon seeing Michael enter the train.[1] We are provided with good evidence that this attitude exists as her behavior reflected the main relevant features of implicit attitudes: immediate, seemingly unconscious, and unendorsed. Therefore, we have good reason to believe that these circumstances are faithful to the understandings outlined earlier of how implicit attitudes operate.

Having filled in the scenario from Julie's perspective, Strawson also prompts us to inquire into Michael's take on the interaction. For Michael, this kind of interaction is all too familiar. As a young black man of average build, who also liked to wear hoodies, he was used to white people acting like Julie. The sideways glances on the subway and being followed by security when shopping were just other examples of the frequent kind of glimpses of behavior that send the message that his presence makes them uncomfortable, on edge even. Upon processing that it had happened again, his immediate emotional

---

[1] I am being necessarily vague about the content of the attitude because the exact structure is irrelevant for the time being. While certainly compatible with a dispositional model to belief (or belief-adjacent phenomena), this account leaves space for other psychological theories to fill in the precise details.

reaction is of resentment toward Julie. By making her nervous, he feels like he does not belong on the subway, and blames Julie for thinking negatively of him just because of his skin color. These are his immediate, unclouded reactions, which are prime sources of evidence into moral responsibility on Strawson's account. It should also be noted that there could be cumulative effect from such interactions taking place somewhat frequently. The negative feelings of alienation and unhappiness with himself would only amplify the emotional blow this interaction produces.

The first step in assessing Julie's moral responsibility for her implicit attitude is to first clarify that she is someone who were are not expected to withhold our reactive attitudes towards in general. As described earlier, Julie is a social, community-engaged professor at a liberal arts college. She has all of the necessary capacities to reason, as well as sustain meaningful social relationships. No matter how you might scrutinize her, she is by no means considered abnormal in any relevant respect. Thus, she seems to easily satisfy the first requirement that is the type of person that regularly engages in the types of interactions regarding trust, compassion, disagreement, and guilt. Other people regularly make demands of her, and she makes similar claims in return, all while regularly exercising her participatory reactive attitudes. These include basic expectations of fair treatment, respect and general good or indifferent will from those around her. We can go further still by reaffirming that it would be inappropriate to adopt the objective approach in this situation. Julie clearly does not merit that treatment due to psychological abnormality or moral underdevelopment, and is likewise not merely an object of social policy. She is a fully developed, reasonable, and responsible agent, satisfying all of the relevant criteria to be considered a morally responsible agent on Strawson's view.

Reflection of Will

Just from the basic facts of the scenario as described, Julie's status as an agent indicates that she has a will to potentially be reflected in her actions, but it is not immediately clear it is an actual reflection of her will. Her actions make Michael feel like he did not belong, and conveyed her own discomfort, uneasiness, or even fear, resulting from his presence. Even with all of her explicitly egalitarian beliefs, Michael is justified in his feelings of resentment towards her as his emotions are a reaction to basic norms of respect and fairness being violated.[2] Nevertheless, there may be deeper factors about the nature of implicit attitudes that may give reason to doubt that Julie actually harbors ill-will towards Michael in this instance. While his interpretation of the events justifies his reactive attitudes in response, is there another reason why his reactive attitudes should be suspended in this case?

The first kind of plea that Julie would make that Michael should suspend his reactive attitudes is likely that there is some kind of exculpatory consideration – that her reaction and any ill-will he felt as a result was due to a factor beyond her control. This is a strong start because it comports well with the psychological literature on the nature of implicit attitudes, highlighting that they are unconscious, unendorsed, and not the result of Julie's reasoning faculties. They are not "within our control" in the same way that explicit beliefs are in that we can acknowledge them and even work to change them if we so choose.

---

[2] For a moving discussion on a similar case, I suggest reading the introduction from Patricia Williams' *The Death of the Profane* for a detailed account of what it feels like to be a person of color and automatically perceived as dangerous.

Further, as an exculpatory rather than ameliorating reason, this type of plea would likely let Julie off the hook so to speak in terms of responsibility in this case. Nevertheless, this kind of appeal to lack of control or involuntariness is not exactly the kind of consideration that Strawson has in mind.

In rejecting the approach to responsibility based on theoretical preconditions, he emphasizes that the kinds of exculpatory reasons he is interested in are not just analogs for such judgments that are external to the practice of holding people responsible. This is because, even if voluntariness is generally important to our judgments of responsibility, there are in fact cases where the interaction and nature of a relationship justify attributing responsibility. Consider someone like the Nazi leader Richard Spencer: his beliefs about people of color are not voluntary in the strict sense – they are not "up to him" in that he can simply pick to have them or not. These beliefs are central to his identity and thus deeply held and resistant to change, even if repeatedly presented with evidence to the contrary. Still, we are likely to feel greater resentment towards Spencer -- someone who embraced Nazism as an adult -- than our explicitly racist manager. If the manager developed his beliefs as a result of his sheltered upbringing in a homogenous place and prejudiced parents, we might even find his beliefs less voluntary and under his control than Spencer's, but we are still nonetheless apt to treat him with less resentment than Spencer.

Returning to the example described earlier of my friend who cancelled our plans due to his tire getting punctured, this serves as a great example of an exculpatory reason to withhold my reactive attitudes of resentment. The reason, beyond that it was out of his voluntary control, is that his reason for action did not emanate from him at all: the

unpredictable world sabotaged our plans. I cannot feel resentment for him because his will is still good, but it is blocked by some rogue factor. Put simply, the action did not come from his will at all, but was a consequence of something external to him entirely. We should try to ask if the same can be said of Julie: did her implicit attitudes resulting in clutching her purse emanate from her? At first glance, it seems like the answer to this question is an obvious 'yes'. There is no rogue influence out in the external world that caused Julie to act in that way. It might be argued that stereotypes and presentations in the media of young black men as dangerous might qualify as this kind of external consideration, but that cannot be part of a direct causal chain to Julie's action. They might have a role to play in explaining the origin of her implicit attitudes, but those attitudes are just what we are trying to assess responsibility for. Further, while implicit beliefs may be unconscious, unendorsed, and not typically the result of reasoning, none of these factors locate implicit attitudes as something unattached to persons – they are located *somewhere* in the mind of the subject. Julie's implicit attitudes are exactly that: hers, not anyone else's.

At this point, Julie might switch strategies and reformulate her plea that the implicit attitudes that prompted her to act are not *really* hers because they do not reflect her *real* self. She might argue that they do not accurately represent the real her because she would not endorse them. Instead, they belong properly to her unconscious self, the one acting on instinct instead of reason. After all, we should not forget her strong egalitarian beliefs and the number of ways in which she has acted upon them. While implicit attitudes might constitute *some* part of her, it could not properly serve as the part that we are interested in when it comes to questions of moral responsibility. Thus, her exculpating plea is that her implicit attitudes could not be reflections of her will because the side of her they reflect is

not her true self. This kind of response presents a much more interesting challenge to Strawson's account and can be highlighted through T.M. Scanlon's approach to moral responsibility.

Scanlon's Reasons-based Account

Scanlon's account to moral responsibility offers a worthy challenge to Strawson while sharing in the overarching project of treating moral appraisal as essentially concerned with "the quality of an agent's will" (Scanlon, 1986, 167). Another similarity between the two accounts is the rejection of morality as simply a system of restraints that we accept in order to gain protection against the harmful conduct of others. In other words, moral sanction is seen as something more than a sanction meant to make others comply with external rules and it carries extra force in that it relies on a kind of internal regulation (Scanlon, 1998, 268). Scanlon and Strawson also agree that not everyone can be seen as a proper object of moral appraisal. It is inappropriate to pass moral judgment on a young child, someone who is sleep-walking, or under hypnotic suggestion for example. Nevertheless, the separate justifications for why these subjects like these are not treated as morally responsible agents cuts to the heart of their disagreement about the role of morality.

While Strawson would agree that we should withhold moral appraisal of small children or those under hypnosis, he would argue that this is because they depict agents who are incapable of participating in sustained human relationships – they are unable to be sensitive to and react to the qualities of the wills of others. Scanlon's justification, on the other hand, is a decisive departure from this line of argument. He would argue that these cases are united in that they describe agents who lack a non-moral capacity, namely the

capacity for critically reflective, rational self-governance (Scanlon, 1986, 174). By critically reflective, he means the ability to "reflect and pass judgments upon one's actions and the thought processes leading up to them". By rational, he means being "sensitive to reasons and the ability to weight them" (Scanlon, 1986, 174). This capacity is the kind of theoretical precondition that is external to the practice of holding people responsible that Strawson rejects. As described earlier, he argues that passing moral judgments on others is just a fact of life, something we naturally do by nature of our interactions with one another, and these judgments are reflected in our reactive attitudes.

Scanlon's contractualist account of moral responsibility does acknowledge the reactive attitudes Strawson draws on, but does not privilege their importance in assessing responsibility. Reactive attitudes "can explain why moral judgments would normally be accompanied by certain attitudes, but these attitudes are not the basis of its account of moral judgment" (Scanlon, 1986, 167). The important distinction, highlighted here, is that having certain attitudes – including resentment or indignation – are not sufficient for moral appraisal. Rather, one's *judgment* that another's behavior is morally faulty serves as the basis for making moral judgments. To judge that another person's behavior is morally faulty is, "to believe that there is a divergence of this kind between the way that person regulated his or her behavior and the kind of self-regulation that mutually acceptable standards would require" (Scanlon, 1986, 167). This combines the precondition of having the capacity for critically reflective, rational self-governance with reasoning about what principles would widely be agreed upon by others to guide interactions. Thus, morality is a system of "co-deliberation" based on the basic desire to regulate one's behavior "according to standards that others could not reasonable reject insofar as they, too, were

looking for a common set of practical principles" (Scanlon, 1986, 166). Moral reasoning, then, is the process of working out which principles each of us could be expected to employ as a basis for deliberation and to accept as a basis for criticism.

To clarify these abstract points, consider again the case of the explicitly prejudiced manager who refuses to even extend an interview to applicants with names commonly associated with African Americans. We can also assume he also possesses the necessary capacities to be sensitive to reasons, interrogate himself, and generally modify his behavior when given good reason. He thus qualifies as an appropriate agent to hold morally responsible. His action is blameworthy because he "fails to take account of or knowingly acted contrary to a reason that should, according to any principles that no one could reasonably reject, have counted against his action" (Scanlon, 1998, 271) In other words, his action violated the principles of respect and fairness, which are principles that we generally agree on as reasons that should guide our behavior with one another. What makes the action wrong and not merely harmful is that he flouted the kinds of requirements that flow from another's status as an agent. He failed to provide an adequate justification for violating the principles that regulate our default conduct with others. Instead, he viewed his prejudicial beliefs as a better source of reasons for action than the principles of respect and fairness. Further, the moral community has decided through the process of co-deliberation that prejudicial beliefs are not good reasons, while respect and fairness are strong reasons that should be considered based on our status as rational agents.

One of the main points highlighted by the above is example is the emphasis that Scanlon puts on the reasons that one acts. When someone act for the right reasons – those

that come from co-deliberation about what we owe each other as rational agents – their action is morally praiseworthy; when someone acts for the wrong reasons – those that ignore the reasons given by others – their action is morally blameworthy. When Scanlon uses the term reason, he is referring to normative reasons, which he defines as "the type that give a justification, or cite the reasons counting in favor of something" (Scanlon, 1998, 19). If I ask your reasons for believing that the sky is blue, for example, a normative reason is a justification as to why you think you should believe it. This is to contrast with descriptive or operative reasons that merely describe how it came about.[3] In other words, normative reasons provide the 'why' and operative reasons provide the 'how'.

There are some mental states that are able to track reasons, which Scanlon calls "judgment sensitive attitudes" (Scanlon, 1998, 20). These are things like "beliefs, intentions, hopes, fears, and attitudes such as admiration, respect, contempt, and indignation", but do not extend to basic states such as hunger or tiredness. Judgment sensitive attitudes can be defined as "attitudes that an ideally rational person would come to have whenever that person judged there to be sufficient reasons for them and that would …"extinguish" when that person judged them not to be supported by reasons of the appropriate kind" (Scanlon, 1998, 20) We can rightly ask if someone is justified in feeling any of the attitudes in the first group because they should be able to provide an account of why it is good to have that attitude. The only response that could be given for the latter category, however, would be a descriptive account of how it came to be that the person felt

---

[3] Since Scanlon uses reason to mean normative reasons, I will do the same for the remainder of this paper.

such a way. According to Scanlon, "the idea of judgment-sensitivity helps to isolate the sense in which attitudes can be things we are responsible for" because we can attribute them to people who can properly be asked to defend them (Scanlon, 1998, 21). Thus, we are responsible only for what we can point to reasons that count in favor of having that attitude. Our "moral self" must then be distinguishable from the rest of ourselves that contains all kinds of other attitudes that we can only provide descriptive accounts for.

Julie now has sufficient philosophical firepower to defend her second plea that her implicit attitudes are not a reflection of her real self. Appealing to Scanlon, she would likely say that she cannot be held responsible for Michael's feelings of resentment because she cannot provide the right kind of reasons – the kind that justify why she has the attitude. Instead, she is likely only able to point to a descriptive account of how she came to harbor such an implicit negative attitude towards black men like Michael. She could, for example, point to the stereotypes perpetuated in the media of black men as dangerous. This reason is categorically not a normative reason, however. She does not see anything good in this reason that justifies her having it. In fact, she can cite many reasons why it is a bad thing. Therefore, her implicit attitudes are not a true reflection of her moral self because they do not come from the part of her that is engaged in the process of evaluating and deciding upon which normative reasons are good ones. This approach comports with Frankfurt's hierarchical approach to the self, which privileges "higher order capacities", such as reasoning and rational self-governance, above more basic capacities. Such an idea can even be seen as originating from Plato's distinction between one's Reason and Desire (Arpaly, 199, 171).

The Inadequacy of Real Self Theories

Susan Wolf coined the term "real self theories" in 1993 to describe frameworks like Frankfurt's that seek to privilege one part of the self over another. The main idea is that some actions flow from the Real Self, and thus belong more profoundly to the agent than actions flowing from the 'fake' self (Arpaly, 1999, 165). Various thinkers disagree about what exactly constitutes the Real Self, but someone like Scanlon would use some kind of account that justified the special status of the Real Self as being open to moral assessment based on their ability to respond to normative reasons. The Real Self approach has some obvious intuitive advantages. For example, if I am hungry, and irritable because of it, it is natural to argue that my irritableness did not come from any reason to express ill will. It actually came from my 'fake self', the side of me that is controlled by basic processes like hunger. Nevertheless, attempting to break ourselves into fragments seems somewhat arbitrary and could get us off the hook too easily. Further, Real Self theories ask us to impute moral praise and blame in ways that shock our intuitions.

To argue against Real Self models, Nomy Arpaly claims that they fail to generate the proper moral appraisal in cases of inverse *akrasia*, using Huckleberry Finn as an example. In the classic Twain novel, Finn is presented with the opportunity to turn Jim in and is confronted with two conflicting desires: one of them he identifies as his weakness, which makes him want to help Jim, and the other he identifies with his conscience, which makes him want to turn Jim in to the authorities (Arpaly, 1999, 161). Beyond these two desires that appear to be on equal footing, he also has a deeper attitude in that he believes his desire to turn Jim in is commendable, and thus wants this to become his will. He says that he wants to "follow through on his conscience." Nevertheless, Finn is ultimately

unable to turn his desire to follow his conscience into action and ends up helping Jim rather than turning him in. Note that in this case, Finn has been raised and educated in such a way to see segregation and racism as legitimate reasons for action, while disregarding respect and fairness as illegitimate reasons to guide actions towards black people. Thus, his reason-responsive Real Self endorses turning Jim in, while his fake self encourages him to help Jim.

What should we make of Finn in this case? Using a reasons-responsive Real Self model would lead us to judge Finn as morally blameworthy because immoral reasons of prejudice are what motivate his conscious commitments. Even though he fails to act on them, he would be no less blameworthy because he still expresses ill will towards Jim by endorsing those prejudicial reasons. His inaction cannot redeem him into someone that deserves any kind of moral praise and we are prompted to treat him just as blameworthy as if he had turned Jim in. This result, however, seems to fly in the face of our intuitions. Finn should not be treated the same as an explicit racist who feels no desire to help Jim (Arpaly, 1999, 163). If we accept the Real Self view that privileges one's judgments, this is the conclusion that we must reach. Thus, we need an approach to moral responsibility that gives Finn *some credit* for his non-judgment sensitive attitudes that encourage him to help Jim. Nevertheless, it would be perhaps even more ridiculous to adopt a view of the Real Self that privileged one's non-rational appetite. Doing so would give Finn too much credit and paint him as some kind of praiseworthy egalitarian. As with all complicated cases, we must somehow embrace the messy middle and sketch an account that gives proper credit to the agent based on all relevant factors, rather than a specific and arbitrary fragment.

Towards The Whole Self

According to Arpaly, one's character is not dependent on any structure of the self, but still depends on the extent to which one's self is embodied in the action. In other words, the Whole Self theory holds that, other things being equal, "an agent is more praiseworthy for a good action, or more blameworthy for a bad action, the more the morally relevant psychological factors underlying it are integrated within her overall personality" (Arpaly, 1999, 172). Thus, a continuum is formed of motives, where those that fit better with the agent's character are privileged above others for purposes of assigning blame and praise. Consider for a moment that beliefs, attitudes, emotions, desires, moods, schemas, stereotypes, etc., are all in the same class of psychological entities that produce action. For clarity, I will refer to these simply as beliefs and desires. A belief or desire is deeply integrated to the extent that it satisfies two conditions: 1) it is *deep*; and 2) it does not *oppose* other deeply held attitudes. (Arpaly, 1999, 173).

Each of these conditions require further explanation. A belief or desire is deep "insofar as it is a powerful force in determining the actor's behavior, deeply held, deeply rooted" (Arpaly, 1999, 173). Deep beliefs are also resistant to revision, as is indicated by someone being reluctant to revise deeply held beliefs that are central to their identity or perspective of the world. Deep desires are those that are readily satisfied from many options, rather than being pressured into it by limited options or insufficient opportunity to decide. Beliefs and desires oppose one another when they cannot be held at the same time, or be true simultaneously. For example, my belief that my favorite team is the best is more deeply held when it does not conflict with desires to support other teams they compete against. Arpaly also argues that only the morally relevant beliefs and desires should be

evaluated by how deeply integrated they are. There can be completely irrelevant reasons that are deeply integrated, but might not have any causal role in prompting action. For example, “the integration of a sexist man's belief that "women belong in the home" affects the blame he merits for refusing to hire women, but the integration of his belief that the best way to avoid hiring women is to throw out their applications is not itself morally relevant” (Arpaly, 1999, 173).

Using Arpaly’s general framework for the Whole Self provides a way to convincingly respond to Julie’s plea that she should not be held morally responsible in the case of clutching her purse: her implicit attitudes that motivated the action were not a reflection of her true self, and thus not a reflection of her true will. On the one hand, Julie has conscious beliefs that everyone is equal and should be treated with the same level of respect, dignity and fairness. They are deeply held as evidenced by her commitment to putting them into action through her support for affirmative action policies, membership to the diversity and inclusion advisory board, and her inclination to reproach those who disagree. Further, they are unlikely to be subject to easy revision as they are sincerely held. On the other hand, however, she has negative implicit attitudes towards people of color that regularly manifest themselves in her behavior towards students of color in class, interactions with black workers at the store, and even clutching her purse upon Michael’s entering the train. Beyond their widespread integration into all kinds of daily interactions, they are also deep insofar as they are extremely difficult to revise and alter, despite Julie’s best attempts. Nonetheless, Julie makes a conscious effort to try to limit this unendorsed type of behavior by reminding herself to be extra cautious when grading her black students’ assignments, for example.

Her conscious commitments and implicit attitudes are directly opposed to one another, so it is difficult at first pass to use opposition as a means for assessing how well integrated these attitudes are to her overall character. Still, we could point to Julie's other conscious commitments, such as supporting equality among genders and sexual orientations. Assuming that her commitment to egalitarianism is deep, she will certainly voice strong convictions to the principles of fairness and respect for difference. Further, she also has a deep desire to only act in ways that respect those principles, as well as the desire to eventually rid herself of such attitudes. Thus, the Whole Self model, as applied to Julie, arrives at a conclusion that is not a dramatic departure from a Real Self model: in some sense, Julie's actions are less than fully her own. The reason is not simply that her actions are caused by an implicit attitude that conflicts with her conscious commitments, but given her overall character – including beliefs and other attitudes – the action is somewhat poorly integrated.

We can contrast Julie with someone who lacks the above-and-beyond commitment to egalitarianism that she exhibits. Consider James, another professor in Julie's department. Like Julie, his actions regularly manifest negative implicit attitudes towards people of color. He often views the contributions of black students as less valuable during discussion, grades their papers more harshly, and is more likely to view black men on the subway as aggressive. Unlike Julie, however, he does not embrace egalitarianism as whole heartedly. He thinks that concerns about representation of people of color on the faculty are overblown, is unconvinced after being presented with multiple arguments about the presence of systemic racism, and opposes multifamily units in his neighborhood because they attract people he thinks of as lower class. To be fair, he still explicitly condemns

racism and thinks of himself as an egalitarian who cares about racial equality. Nonetheless, he does not put his convictions into action in the same way that Julie does, nor is his commitment to equality as deep and pervasive for all groups. He might harbor prejudice against homosexuals because of his religious beliefs, for example.

If one were to accept the Real Self model, we would have to treat Julie and James as essentially identical cases – after all, both act contrary to their reasons when reacting to Michael on the train. The strength of the Whole Self model is that it enables us to identify a meaningful distinction between them. James' commitment to egalitarianism is much less deeply rooted in his overall character than Julie's, as evidenced by having fewer conflicting beliefs and desires, as well as fewer actions to support the depth of his commitment to egalitarianism. Likewise, his negative implicit attitudes are more deeply integrated, as evidenced by day to day interactions and lack of concern to correct the behavior. Thus, when James recoils on the subway at the sight of Michael, his action is a truer reflection of his will than Julie's clutching her purse is a reflection of her will.

Conclusion

Returning at last to Strawson's account of moral responsibility, I argue that the above approach of using the Whole Self rather than the Real Self is the most faithful way to approach the question of what counts as a reflection of our will. The degree to which an attitude is integrated into our overall character is proportional to the degree to which it reflects our true will. Strawson's entire account is carefully built around whether or not certain theoretical preconditions pertain, such as the truth of the determinist thesis. By treating responsibility as a social practice that needs no external justification, his project is

decidedly different than Scanlon's account of contractarianism. Instead of trying to provide a comprehensive account of morality as a system of rules that have been arrived at through co-deliberation between rationally reflective, self-governing agents, Strawson is content to accept the practice as it plays out in real life. His approach prompts us to not forget that morality is best understood by participating in the complicated and messy interpersonal relationships that make us characteristically human. Thus, his approach is well suited to evaluate claims of moral responsibility for implicit attitudes. The case offered above of Michael is something that plays out many times every day, and the reactive attitudes felt by people in his position should be used as critical evidence to identifying and investigating such a subtle and pervasive moral wrong.

In the end, the investigation into moral responsibility into implicit attitudes can hinge on fundamental issues, such as the role of morality itself. Strawson's account is attractive precisely because it was built around the determinist thesis, knowing that it presented an issue that could never be resolved. Ultimately, he argued that we are too thoroughly committed to our relationships as we currently understand them, and that even if it were true that none of our actions were freely determined, we would be unable to act as if that were so (Strawson, 1968, 77). Providing an account of moral responsibility for implicit attitudes, at first, seems like a similarly daunting task. Nonetheless, this paper attempts to lay out an argument to persuade those who are sympathetic to Strawson's view that we expect a minimum level of good will from those we interact with.

Works Cited

Arpaly, Nomy. "Moral Worth." *The Journal of Philosophy* 99.5 (2002): 223-45. Web. November 11, 2017.

Arpaly, Nomy, and Timothy Schroeder. "Praise, Blame and the Whole Self." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 93.2 (1999): 161-88. Web. November 5, 2017.

Bertrand, Marianne, and Sendhil Mullainathan. *Are Emily and Greg More Employable than Lakisha and Jamal? a Field Experiment on Labor Market Discrimination*. Working Paper 9873 Vol. Cambridge, Massachusetts: National Bureau of Economic Research, 2004. Web. September 9, 2017.

De Houwer, Jan. "A Propositional Model of Implicit Evaluation." *Social and Personality Psychology Compass* 8.7 (2014): 342-53. Web. October 12, 2017.

Fazio, Russel H., and Michael A. Olson. "The MODE Model: Attitude-Behavior Processes as a Function of Motivation and Opportunity." *Dual-Process Theories of the Social Mind.* Eds. J. W. Sherman, B. Gawronski, and Y. Trope. New York, New York: Guildford Press, 2014. 155. Web. September 22, 2017.

Gawronski, Bertram, and Galen V. Bodenhausen. "The Associative–Propositional Evaluation Model: Theory, Evidence, and Open Questions." *Advances in Experimental Social Psychology* 44 (2011): 59-127. Web. September 22, 2017.

Hieronymi, Pamela. "Responsibility for Believing." *Synthese* 161.3 (2008): 357-73. Web. August 12, 2017.

Johnson, Gabrielle M. *Implicit Bias: A Literature Review*. Working Paper ed., 2016. Web. August 12, 2017.

Jost, John T., et al. "The Existence of Implicit Bias is Beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of Ten Studies that no Manager Should Ignore." *Research in Organizational Behavior* 29 (2009): 39-69. Web. August 14, 2017.

Mandelbaum, Eric. "Attitude, Inference, Association: On the Propositional Structure of Implicit Bias." *Noûs* 50.3 (2016): 629-58. Web. August 12, 2017.

Scanlon, T. M. "The Significance of Choice." *The Tanner Lectures on Human Values.*Brasenose College, Oxford University, 1986. 151-216. Web. November 23, 2017.

---. *What we Owe to each Other*. Cambridge, Massachusetts: Belknap Press of Harvard University Press, 1998. Print.

Schwitzgebel, Eric. "Acting Contrary to our Professed Beliefs Or the Gulf between Occurrent Judgment and Dispositional Belief." *Pacific Philosophical Quarterly* 91 (2010): 531-53. Web. October 17, 2017.

---. "A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box." *New Essays on Belief.* Ed. N. Nottelmann.Palgrave, 2013. 1. Web. October 17, 2017.

Strawson, P. F. "Freedom and Resentment." *Studies in the Philosophy of Thought and Action.* Great Britain: Oxford University Press, 1968. 71-96. Print.