2016

Dismantling the Novelty and Mystery in Implicit Bias: A New Perspective

Ying Huang
Bard College

Dismantling the Novelty and Mystery in Implicit Bias: A New Perspective

Senior Project Submitted to
The Division of Social Studies
of Bard College

by

Ying Huang

Annandale-on-Hudson, New York

Dec 2016

This project is dedicated to the Bard Philosophy Department, and to my parents.

Acknowledgments

I want to first express my special gratitude to Kritika Yegnashankaran, my lifetime role model, friend and teacher. Without her, this project would not be possible. I can't express how much I have learned and been influenced by her as a philosophy student and as a person. There was not a time where I was not amazed by her insight and astuteness. Without her, philosophy would never be such an adventure for me. Her generosity and patience as a professor, her sense of responsibility as an advisor, and her wholehearted support as a friend are what have brought me to this stage.Without her, I would not be the person I am now.

I also want to express another special gratitude to Jay Elliot. Thank you so much for your Lives of Animal class, a class which really pushed to me to think about who I am and how to confront difficult ethical questions with courage. It is because of your class that I started to think about what it means to live as a student of philosophy.

I owe much to my other two board members, Justin Hulbert and David Shein. Justin, who is always there for me when I need help relating to cognitive psychology, and who has greatly broadened my understanding of it. David, who still helps me with some of my writings despite all his busy administrative lives.

Of course, to the philosophy department; without your help and guidance I would not have grown to the person I am today. Daniel Berthold, Robert Martin, Garry Hagberg, Thomas Bartscherer, and Marco Dees, thanks so much for your courses and guidance over the years. Learning philosophy from you is inspiring and challenging. I won't say how much I have grasped Plato, Hegel and Wittgenstein, but I am certainly still trying. Thank you, Marco, for showing me how much fun metaphysics can be! You have all taught me what it means to be a better person.

Though my parents, Cheng Qiyun and Huang Xiaohong, won't be able to read this project, I still want to express my gratitude. I owe you so much as an Asian female kid who was born in China. Thank you for always trying to provide me with the best education. Thank you for sending me abroad since I was 15. Thank you for always standing by me regardless of my study of History in Singapore, or Philosophy in the states. Thank you for recognizing the value of liberal arts education and believing in me.

Kyra Middeleer, Leonardo Santoso, Justin Shin and Alexander Breindel; Thanks so much for the fun times we have spent together. Special thanks to Justin, who is always there to help whenever I need to run an argument, and for his amazingly sharp insights. Justin, doing philosophy would not be as much fun with your crazy thought experiments!

Table of Contents

Chapter One: Introduction

The metaphysics of implicit bias has attracted much attention from psychologists and philosophers, due to its epistemic and moral significance. For instance, philosophers concerned with individuals' moral responsibilities wonder to what extent the etiology and mechanism of implicit bias are out of conscious control. Most of the discussions presume implicit bias to operate according to a homogenous mechanism.

In this chapter, I will give a general overview of the philosophical debate over implicit bias, an introduction to the most prominent test of implicit bias in psychology: the Implicit Association Test (IAT), as well as an overview of the psychological debate over the distinction between implicit and explicit attitude.

1. Past Philosophical Debate

1. 1 The Associative Model

The standard associative model takes implicit bias to result from one's sociocultural learning history. This model suggests that two or more representations are often paired or combined together in one's learning history; when one of them gets activated, others also get automatically activated. The term "representation" here is a general term, and it can refer to concepts or propositional attitudes such as beliefs. For example, according to the associative model, it's more likely for someone back in the 1920s, where feminist theories are still in their cradles, to associate "female" with "family."

Tamar Gendler: *Alief*

Tamar Gendler in her 2008 paper "Alief and Belief" advances the associative model, and introduces a provocative new term "*a*lief" for implicit associations. She argues that aliefs are distinct from all other mental states, and defines alief as"...a mental state with associatively linked content that's *representational*, *affective* and *behavioral*, and that's activated...by features of the subject's internal or ambient environment " (Gendler 2008, 642).

Alief is called so because it's "*a*ssociative, *a*utomatic, and *a*rational. As a class, aliefs are states that we share with nonhuman *a*nimals; they are developmentally and conceptually antecedent to other cognitive attitudes that the creature may go on to develop. And they are typically also *a*ffect-laden and *a*ction generating" (Gendler 2008, 641). This mental state is sui generis because it's not explained adequately by the current belief-desire folk psychology. What is so special about alief? Gendler argues that once an alief is activated, it consists of three crucial components: 1) the representation of some object, concepts or circumstances; 2) the experience of some emotional state; and 3) the readying of some action (Gendler 2008, 643). Once one of these mental states gets activated, others would also consciously or unconsciously get activated. By 'consciously or unconsciously,' Gendler argues that the activation of an alief state can either be initiated by a perception, regardless of whether it's conscious or unconscious, or by a non-perceptual thought.

Take Gendler's own example. She describes how people are scared and hesitate to walk across the 70-feet high glass Skywalk over the Grand Canyon. On the one hand, people must consciously believe that the skywalk is safe and properly engineered, otherwise they wouldn't

even step onto it. On the other hand, even if they believe that the skywalk is safe, they still hesitate to walk across it. According to Gendler's alief model, in the case of the skywalker, the input of subjects' visual systems is the 70-feet-high edge of a cliff. This particular visual input activates emotions such as fear and anxiety, and meanwhile activates actions such as hesitation and withdrawal. Thus, the alief model explains the strange phenomenon that the tourists are afraid of walking across the glass skywalk even if they believe it's safe to do so.

1.2 The Doxastic Model

Unlike the Associative Model, philosophers who advocate the Doxastic Model argue that implicit bias is essentially a belief or belief-like state. As introduced above, belief is a particular kind of representation. For people like Eric Mandelbaum, belief is a special type of representation because it involves inferential transitions.

Eric Mandelbaum: Implicit Bias as a form of Belief

Eric Mandelbaum takes a strong position that implicit bias is a form of belief that necessarily entails propositional contents, and operates in conjunction with unconscious inference (Mandelbaum 2013 & 2016).

In arguing against the associative model, Mandelbaum first makes a distinction between the associative learning process and the associative cognitive structure. By having the distinction, he argues that "theorists may think that implicit biases are acquired through some form of associative learning, but to infer from that to the idea that an associative structure has been acquired is unwarranted" (Mandelbaum 2016, 634). The notion of cognitive structure or mental process plays the most crucial role in his understanding of the metaphysics of implicit bias.

According to Mandelbaum, for those who hold the position that *associative* cognitive structure underpins implicit bias, they necessarily have to be committed to holding what he calls the AIB Principle, which can be presented as following.

> AIB Principle: Implicit biases (a) can be changed by changing certain environmental contingencies and (b) can only be changed by changing certain environmental contingencies, i.e., by extinction or counterconditioning. (Mandelbaum 2016, 635)

That is, Mandelbaum argues that whoever thinks that implicit biases operate on an associative cognitive structure has to commit herself to the position that implicit biases can never be modulated in a reason-responsive way, but only by counterconditioning. Mandelbaum then cites different examples to show that people with implicit bias attitude do respond inferentially, which indicates that some kind of propositional structure is necessary for implicit bias. If implicit bias necessarily requires a propositional structure that is truth apt, Mandelbaum then concludes that implicit bias is essentially a form of propositionally structured beliefs that undergo unconscious inferential process.

To see why Mandelbaum insists that unconscious inferential process, and thus propositional structure, is involved in those states that someone like Gendler would categorize as "associative cognitive structure," I will give some of his examples. One example cited by Gendler in her paper and then reinterpreted by Mandelbaum in his paper is Paul Rozin's poison experiment (Rozin et al. 1986). In Rozin's experiment, participants are first shown two empty bottles, then witness the experimenter filling both with sugar. Participants are then asked to paste the labels 'Sucrose' and 'Sodium Cyanide' to each of the bottles based on their personal preferences. In what follows, the two bottles are emptied and filled with water. As a result, participants are more hesitant to drink from the bottle that was formerly labelled as 'Sodium Cyanide' by themselves.

The conclusion reached by Gendler is that even though the participants believe that both bottles contain sugar, they *alieve* "cyanide, dangerous, avoid." The alief that gets activated here is belief-discordant.

Mandelbaum argues that the content of this particular alief, i.e. "CYANIDE, DANGEROUS, AVOID" (Gendler 2008, 648), is merely *structurally* associative. Mandelbaum points out that the affective and behavioral contents of this putative alief are not just general reactions towards the concept 'cyanide'; rather, the affective and behavioral contents are specific reactions toward the specific bottle that was formerly labelled as 'cyanide.' That is, if the contents are merely structurally associative, the participants have no reason to generally show more hesitancy against the 'cyanide' bottle. Given that the participants' behavior seem to bind to the 'cyanide' bottle, Mandelbaum postulates that alief contains "a content more akin to THAT [demonstrative standing in for the bottle] DANGEROUS CYANIDE AVOID" (Mandelbaum 2013, 203). He further argues that a syntactic structure for the content "THAT DANGEROUS CYANIDE AVOID" has to be present in order to guide the participants to avoid *that* particular bottle. Without much further explanation, Mandelbaum suggests that the only possibility remains is that the content of the putative alief is a propositional content, e.g. 'That is dangerous cyanide, avoid it,' which binds with the target bottle. This propositional content is then processed in conjunction with the participants' unconscious inference.

1.3 the Intermediate Positions

Between those two ends of spectrum, some other philosophers take intermediate positions along the spectrum. For instance, Eric Schwitzgebel argues that implicit bias lies in a state of

in-between belief (Schwitzgebel 2012); Neil Levy takes the stance that even though implicit bias is not as certain as belief, it's still some form of patchy endorsements that reveals itself occasionally in certain circumstances (Levy 2015 & 2016).

Eric Schwitzgebel: Implicit Bias as an In-Between Belief

In his paper "Self-Ignorance," Schwitzgebel gives an example of a sexist whose name is Ralph. Being a college professor, Ralph, on the one hand, is ready to argue "coherently, authentically, and vehemently for equality of intelligence" (Schwitzgebel, 194); on the other hand, he constantly thinks that his male students look brighter and is constantly more surprised by the bright comments made by his female students. Ralph himself might or might not have noticed the incoherence among his behaviors. Schwitzgebel argues that Ralph is not the best authority in this case to know whether he has sexist dispositions, because others might have observed his sexist behaviors to which he might be completely oblivious.

For Schwitzgebel, even if Ralph authentically believes in the equality of intelligence, it is not enough to prove that he is not a sexist. Belief is not only manifested in one's linguistic expressions, but also one's behaviors such as one's spontaneous emotional reactions and implicit assumptions. According to Schwitzgebel, Ralph is in an in-between state given the incoherence among his various dispositions, so it is neither right to say that Ralph believes in the equality of intelligence, nor right to say that he does not believe so.

Neil Levy: Implicit Bias as a Patchy Endorsement

Though Levy agrees with Mandelbaum that implicit bias consists of propositional structure, he disagrees with Mandelbaum to the extent that he thinks the propositional structures that constitute implicit bias are not stable or continuous enough to be called beliefs. Instead, he calls them 'patchy endorsements.' It is 'endorsement' in the sense that implicit bias does have propositional structure that renders truth conditions. It it 'patchy' because it might fail to respond to the semantic contents of other mental states systematically.

Take the Rozin experiment series as an example again. As introduced above, Mandelbaum argue, as an objection to Gendler, that the best interpretation for the participants' reactions is that there is propositional-content-driven mental processes occurring in conjunction with unconscious inference. Four years after the initial Rozin experiment, another similar experiment is conducted. The experimental description is cited below.

> Subjects faced two empty brown 500 ml bottles. In the presence of the subject, the experimenter opened a container of "Domino" cane sugar, and poured some into each bottle, so that about ¼ of each bottle was filled. The experimenter informed subjects that she was pouring sugar into each bottle. The experimenter then presented the subject with two typed labels. One had not sodium cyanide, not poison written on it, with a red skull and cross bones preceded by the word not. The other label had sucrose, table sugar typed on it. The subject was invited to put one label on each bottle, in any way he or she chose. The experimenter then set out two different colored plastic cups, one in front of each bottle, and poured unsweetened red (tropical punch) 'Kool-Aid' from a glass pitcher into both, until they were about half full. Now, using separate, new plastic spoons for each bottle, the experimenter put a half spoonful of powder from one sugar bottle into the glass standing in front of that bottle, and repeated this with the other glass for the other sugar bottle. (Rozin et al 1990, 583)

The former glass is labeled as "sucrose, table sugar" and the latter one is labeled as "not sodium cyanide, not poison." The results show that the participants are still reluctant to drink from the glass that is labeled as "not sodium cyanide, not poison" even though they witness that only

sugar powder is put into the glasses. If we apply the previous interpretation from Mandelbaum to this experiment, it seems that the participants should be responsive to the propositional content such as “not sodium cyanide, not poison, do not need to avoid.” That is, following Mandelbaum’s argument, unconscious inference should have been made with the content such as “not sodium cyanide,” “not poison” and “do not need to avoid,” and thus guides the participants’ behaviors in the way that they would not avoid the target bottle, which is not the case of this experiment result. Hence, this experiment is used by Levy to argue that the very mechanism suggested by Mandelbaum fails to respond to labels that are stated in a form of negation. The failure to process or respond to negation then puts a challenge for Mandelbaum’s strong position that implicit bias is essentially a belief state. Though it might not always be the case that unconscious inference is employed, Levy argues that some level of content-driven interaction is still present.

2. Psychological Indirect Measures

Similar to the states of play in Philosophy, Psychologists also disagree on the models of states involved in Implicit Bias. Some psychologists believe that we can have both explicit and implicit attitudes. By attitude, psychologists generally refer to a disposition that often involves propositional contents. For instance, Michael Hogg and Graham Vaughan in their book *Social Psychology* define attitude as "a relatively enduring organization of beliefs, feelings, and behavioral tendencies towards socially significant objects, groups, events or symbols" (Hogg and Vaughan, 150). Similarly, in the very beginning of their book *The Psychology of Attitudes*, Alice

Eagly and Shelly Chaiken define attitude as "..a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor" (Eagly and Chaiken, 1). That is, psychologists take attitudes to be dispositional, and sometimes involve representations.

Explicit attitudes refer to those attitudes, usually evaluations of individuals' stance on other individuals, objects or events, which can be avowed directly by individuals. This type of attitude can be understood as personal level attitudes in philosophy. Implicit attitudes, on the other hand, often refer to attitudes in which individuals are not aware of, which can be understood as sub-personal level attitudes in philosophy. There are several indirect methods, such as IAT (Implicit Association Test), AMP (Affect Misattribution Procedure) and GNAT (Go/No-go Association Task), developed by psychologists to measure the strength of individuals' implicit attitude of which IAT is generally considered as the most reliable and well-known one.

2.1 Implicit Association Test (IAT)

What IAT tests directly is an individual's reaction time, and what IAT aims to test indirectly is the strength of automatic association between different concepts, such as the Black or females, and evaluations, such as violent or weak. An IAT test consists of several stages, and below is a typical procedure of an IAT test.

The test I am citing in this chapter is on the association between gender and career/family. In the first stage, participants are asked to sort the word presented in the middle of the screen to either the category on the left or the category on the right.

Figure 1.

For instance, as shown in Figure 1, participants are expected to answer whether the word ‘home’ belongs to the category of family or career. Similarly, the second stage of the test is where participants are asked to sort the name presented in the middle of the screen to either the gender category on the left or the one on the right.

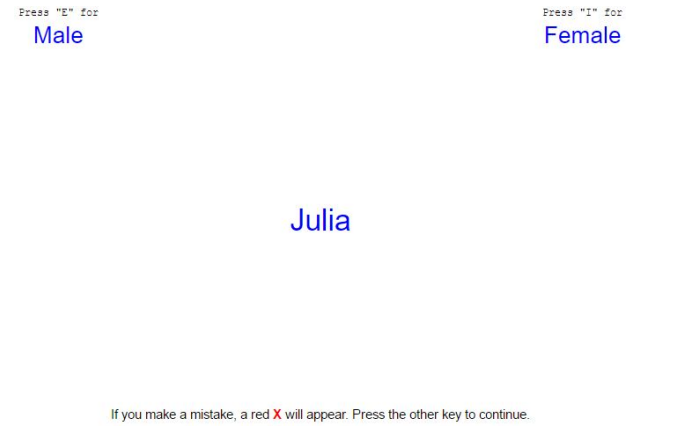

Figure 2.

In the third stage of the test, different categories are combined together. As shown in Figure 3, participants are asked to sort the word presented in the middle, e.g. Professional, to either the combined category ‘Male or Family’ on the left, or the combined category ‘Female or Career’ on the right.

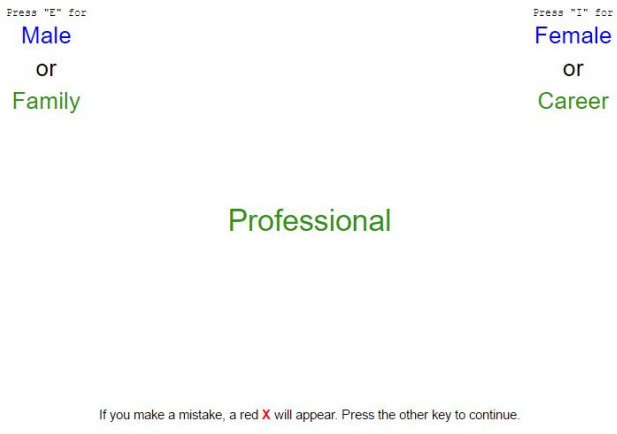

Figure 3.

The placements of categories are then switched in the fourth stage. As shown in Figure 4, the category ‘career’ is now placed on the left hand side.

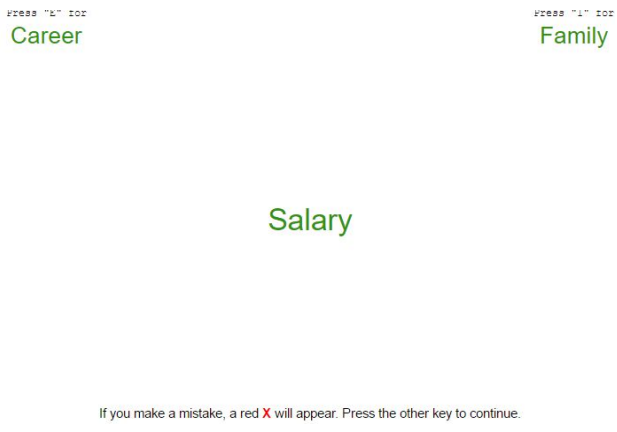

Figure 4.

In the final stage, following the placement of categories in the fourth stage, categories are combined together. Similar to the third stage, participants are asked to sort the word in the middle to either the combined categories on the left or the one of the right.

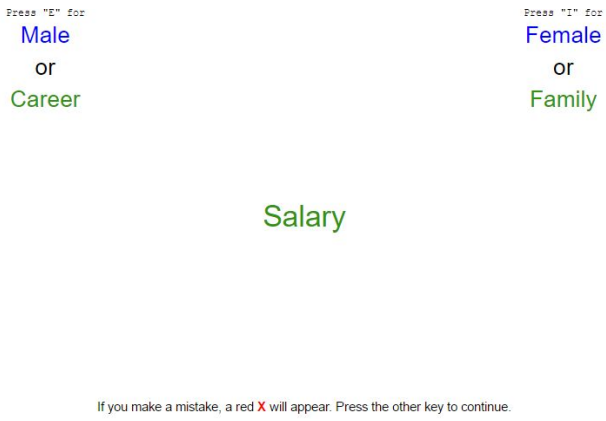

Figure 5.

Each stage consists of several similar word pairs like these. The average difference in terms of reaction time indicates whether the participant has a preference for one combination over the other. For instance, if it takes in average a shorter time for the participant to choose the male/career combination than the female/career combination for words 'professional,' then we say that the participant has a preference for and thus a stronger association between the word 'male' and 'career.'

2.2 Psychological Debates over the Model of Attitude

However, psychologists who hold different views on the models of attitudes might disagree on how we should interpret the results of IAT. Specifically, whether psychologists endorse or

deny the difference between implicit and explicit attitudes depends on how they interpret the interplay of explicit and implicit attitudes.

Psychologists such as Russell H. Fazio and Tamara Towles-Schwen propose the MODE model (Motivation and Opportunity as Determinants), which takes attitude to be the product of a single-system process. According to MODE, whether motivations and opportunities are involved in a process will determine whether the attitude getting expressed is spontaneous or deliberative (Fazio & Towles-Schwen, 1999). Spontaneous attitudes refer to those automatic and immediate evaluations of an object, a concept or an event from the perspective of an individual. Deliberative attitudes refer to those less immediate attitudes of individuals after they have thoughtfully taken into accounts the benefits and disadvantages when they evaluate a particular concept or an issue. That is, without much involvement of motivation and opportunity, the attitude of an individual that gets activated would only be a result of an automatic and spontaneous activation, instead of a consciously deliberative one. This automatic activated attitude will then shape one's evaluation of an object or guide one's behavior.

In other words, to what extent an individual's spontaneous or deliberative attitude governs her behavior depends on the degree of motivation and opportunity. For instance, if an individual lacks specific intensive motivation to finish a task, her behavior is more likely to be governed by her automatic and spontaneous attitude; and vice versa. Under the MODE model, the results of IAT can be taken as individuals' spontaneous attitudes with the lack of motivation and opportunity. However, it seems that, following this line of reasoning, the MODE model fails to account for cases where individuals do have a strong motivation to not to be considered as having implicit bias when taking IAT tests. For instance, there could be cases where an

individual tries as hard as she can to choose the word pairs as quickly as possible in order to not to be considered as having racist or sexist implicit bias. According to the MODE model, the results of the IAT test can be seen as expressing the individual's deliberative attitude, given the presence of motivation and opportunity. However, what IAT tests is exactly the reaction time difference. We might end up facing a result where the individual's attitude revealed by IAT contradicts what she avows before the IAT test, and both attitudes are considered as explicit attitudes by the MODE model.

Psychologists like Bertram Gawronski and Galen V. Bodenhausen, on the other hand, endorse the position that the distinction between our implicit and explicit attitudes is a result of a dual-system process. For instance, Gawronski and Bodenhausen propose the APE model, i.e. Associative – Propositional Evaluation model (Gawronski and Bodenhausen 2006 & 2011). According to the APE model, our implicit and explicit evaluations are the product of two distinct processes that are qualitatively different. Specifically, what gives rise to our implicit attitude is an associative process, and what gives rise to our explicit attitude is a propositional process. They argue that "associative processes are further specified as the activation of mental associations on the basis of feature similarity and spatiotemporal contiguity; propositional processes are defined as the validation of activated information on the basis of logical consistency" (Gawronski & Bodenhausen, 73).

The associative process and the propositional process, according to the APE model, do not operate in isolation. It is possible that the logical consequence of one's propositional evaluation is inconsistent with the corresponding propositional evaluation of the associative attitude. The inconsistency between one's implicit attitude and explicit attitude shown by the IAT can thus be

seen as a result where one's automatic associative attitude is not accepted in one's propositional evaluative stance. For instance, it might take less amount of time for a participant to choose the combination of 'male' and 'career' than the combination of 'female' and 'career,' which indicates a stronger association between 'male' and 'career.' The result might contradict one's explicit propositional evaluation that both females and males are suitable for career. Because of this dual-system process, Gawronski and Bodenhausen remain silent on which one of the two evaluations is an individual's genuine evaluation; instead, they argue that one of the evaluations has to be abandoned once the inconsistency is pointed out to prevent cognitive dissonance.

2.3 What Exactly is IAT About?

In the previous discussion of this subsection, we have seen that on the one hand, the IAT test claims that it is to test the strength of *automatic* association between concepts or evaluations; on the other hand, psychologists who hold different views about the models of attitudes disagree on how we should interpret the result of IAT tests. Moreover, the disagreement among psychologists over the models of attitudes also differs from the disagreement among philosophers. To have a better understanding of the IAT test despite all these interdisciplinary disagreements, I want to now press on the notion of "automatic."

The term "automatic" is more commonly used in the literature from psychology than from philosophy, and is sometimes used interchangeably with "implicit" and "unconscious." What exactly is denoted by the notion of "automaticity" in psychology, however, is far from homogenous. Psychologist John Bargh, for example, argues in his paper "The four horsemen of automaticity: Awareness, efficiency, intentions and control" (Bargh 1994) that the use of the

term “automaticity” does not follow a strict definition. That is, when psychologists use the term “automaticity,” they use it to describe different mental phenomenon that might not necessarily fall under the exact same category. Bargh argues that we should not understand the psychological use of the term “automaticity” under an all-or-none assumption. Instead, as indicated in the title, Bargh argues that there are four distinct aspects of the notion of automaticity: awareness, efficiency, intention and control. “Awareness” refers to the state in which a subject is not aware of whatever mental process that is going; “intention” refers to the fact that the subject herself does not activate or initiate the mental process; “controllability” means that the alteration or pause of a mental process might be beyond one’s personal level control once it is initiated; and “efficiency” means that the mental process requires minimal attentional effort. The four aspects need not all to be involved in order for a mental phenomenon to be categorized as ‘automatic.’ That is, there can be a mental process categorized in psychology as “automatic” if it is under some degree of personal level control yet initiated without one’s intention. There can also be a mental process categorized as “automatic” that is both beyond one’s intention and control, and yet under one’s awareness.

Similarly, in Philosophy, even though the term “unconscious” is used more often than the other two, ‘unconscious’ is still used as an umbrella term to refer to those mental states or processes that are either beyond one’s conscious control or initiated without one’s personal level intention. Moreover, ever since Ned Block’s paper “Some Concepts of Consciousness” (Block, 2002) there is generally a consensus in philosophy of mind to separate phenomenal consciousness from access consciousness. Phenomenal consciousness refers to the qualitative aspect of an experience, i.e. what it is like to be in a particular state. Access consciousness refers

to the state in which information is cognitively registered and broadcast for further use in reasoning or action. Often the term "unconscious" is used in philosophy when the phenomenal consciousness aspect is missing.

Though it's uncertain how much the notion of 'automaticity' and 'unconscious' from both disciplines overlap, Bargh's study does leave the room for different possible interpretations of the IAT test. For instance, even though it's generally agreed that the activation of stereotypes can be unintentional and efficient, to what extent it is beyond one's conscious control is a question. In what follows, the notion of "implicit" in "Implicit Association Test" also requires further examination. That is, the experimental design reflects the unintentional and efficient aspects of implicit bias, but necessarily the awareness and controllability aspects.

Chapter Two: Scrutinizing the Etiologies

In the previous chapter, I offered a brief overview of the debate over the metaphysics of implicit bias among philosophers and psychologists. As Mandelbaum points out, most philosophers working on implicit bias assume "that there is a monolithic phenomenon to be investigated" (Mandelbaum 2016, 631). In this chapter, I will first show that the different cases presented by different philosophers as having a homogenous etiology are indeed heterogeneous. I will then argue for at least a tripartite distinction in terms of the etiologies of implicit bias. By etiology, I mean the causal mechanism of implicit bias. For instance, for Gendler, the causal mechanism of implicit bias is thus a sui generis state with an associative structure; for Mandelbaum, the causal mechanism of implicit bias is propositional contents in conjunction with unconscious inference.

1. The Mix of Examples Cited

When philosophers and psychologists talk about implicit bias, the general strategy they take is to find incoherence among dispositional behaviors. This type of incoherence is often made obvious by the implications of one's behavior that one is not aware of. Though all the philosophers above try to analyze cases of behavioral incoherence, are they talking about the same type of incoherence? Specifically, do the examples have the same etiology?

The cases cited by Gendler are examples where 1) the formation of beliefs seems to require less effort, and 2) the inconsistent behaviors can either be deliberative or non-deliberative, i.e. it is not necessary that the behaviors in Gendler's cases are a result of personal level decision. In

Gendler's examples, even though the tourists believe that the glass skywalk is safe, they still hesitate to step onto it; and even though that the participants in Rozin's experiments witness both glasses or bottles being filled with sugar, they still hesitate to drink from the glass that is labeled as 'poison' or even 'not poison.' The formation of the subjects' beliefs might be effortless. For instance, the tourists will not even step onto the skywalk if they do not already believe in the safety of the skywalk; the experience of witnessing the sugar powder from the same sake being poured into the two bottles or glasses immediately renders and justifies the belief that it's only sugar in both bottles or glasses. Even though the subjects have beliefs that can be formed or justified immediately, they nevertheless behave in an opposite way, either deliberately or non-deliberately. In the skywalk case, the tourists' hesitation or withdrawal might be completely beyond their control and intention, which is not a result of deliberation. However, in the Rozin experiment, it seems that the participants have to deliberate a decision to walk to one of the glasses.

Schwitzgebel's Ralph the sexist case, however, has a more obvious asymmetry between what the subject avows explicitly and how the subject behaves. Unlike the effortless formation of beliefs in Gendler's cases, the formation of Ralph's belief about the equality of intelligence seems to require more personal level deliberation. For instance, Ralph's belief about intelligence equality might be a result of him reading a feminist theory or reasoning with someone else. Similar to Gendler's skywalker example, Ralph might either be consciously aware or not aware of his inconsistent behaviors, though Schwitzgebel himself takes Ralph to be unaware of his states.

Above we have seen that the examples of behavioral incoherence cited by philosophers such as Gendler and Schwitzgebel differ from each other based on different aspects: for instance, whether the formation of belief is effortless, whether the initiation of inconsistent behaviors is a result of deliberation, and whether the subject is aware of her behavioral incoherence.

Although there are different types of behavioral incoherence, I want to propose that implicit bias characteristically involves a special type of behavioral incoherence: conflict between an individual's behaviors and her normative meta-beliefs. I take the assertions we express against racism or sexism to be a higher-order normative belief, which is usually expressed in the form of "It's wrong to think/be P." For instance, "it's wrong to be sexist" or "It's wrong to think that males are always superior." Normative meta-beliefs can either be possessed by a group of individuals or be unique to an individual.

The term 'bias' is often attributed to a prejudice and stereotype in favor of or against a certain group of people, which is normatively loaded, i.e. taken as 'ought-to-be' or 'ought-not-to-be.' It should be noted that the normatively-loaded implicit bias should still be distinguished from a general type of racial generalization. For instance, in general people with light colored skin are more easily to get sunburned because darker colored skin generally produces more melanin, a dark brown to black pigment, to protect skin. If a white skin person is reminded by her dark skin friend to get more suncream before going to the beach, we would not call that racial bias or implicit bias because no normatively loaded phenomena is involved here.

Gendler's skywalk case, though a case of behavioral incoherence, does not involve any normatively loaded phenomena, i.e. there is no normative standard as to whether we should be afraid of height or not. Schwitzgebel's case, Ralph the Sexist, is clearly a case that concerns

normatively loaded phenomena, i.e. being sexist is usually taken to be something that ought not to be. That is, even if Gendler's example is a case of behavioral incoherence, we would not really attribute the term 'bias' to it. However, for Ralph the Sexist case, clearly people would say that Ralph has implicit bias against females.

Thus, I take the 'bias' or stereotype against a socially stigmatized group as something that 'ought not to be'; and I take 'implicit bias' to be the umbrella term for whatever state in which the subject is not consciously aware of her stereotype and prejudice.

2. Scrutinizing Implicit Bias Cases

2.1 What's Happening Explicitly

A higher-order normative belief has the following features: 1) the formation of such beliefs is often driven by non-epistemic motives, such as social norms or emotional reactions; 2) people's endorsement of normative beliefs vary in degrees; 3) this normative belief governs our behaviors, emotions and other first order beliefs, i.e. behavior or disposition *needs* to be habituated in accordance with such beliefs. The term '*needs*' here emphasizes a first person effort, i.e. we often subjectively feel the pressure to conform our dispositions or behaviors to our normative beliefs. Empirical questions like "to what extent can habituated behaviors or dispositions override our ingrained propensities successfully?" cannot be answered solely from an a priori perspective.

It should be noted that it is possible to abandon a normative belief and keep the dispositions or behaviors when there is inconsistency. Consider the following case. Peter is homosexual and has been brought up in an environment where homosexuality is considered as immoral. Initially,

Peter tried to act 'straight.' After reading LGBTQ theories and participating in LGBTQ movements, Peter finally abandoned his prior meta-belief and accepted his own sexuality. In this case, the prior higher-order normative belief can also be abandoned by the subject.

The notion of 'habituation' suggests a time-consuming process from belief-formation to behavioral coherence. That is, it's not that once we form a normative belief, every dispositions of ours will immediately match the belief sometimes. Sometimes it takes time for us to recognize the inconsistency between our behaviors and normative meta-beliefs, as well as to reshape our dispositions or behaviors accordingly.

2.2 Distinct Etiologies

We have seen how examples of implicit bias are treated as homogenous, but now I will argue for their heterogeneous nature. Consider the following three cases.

Case 1: Your friend Alice explicitly and authentically argues against sexism. One day during your conversation after reading a rape case, Alice suggests that females should learn to protect themselves by staying home at night or dressing up "properly." You then explain to Alice that 1) 'proper dressing code' is a socially constructed standard; 2) thinking that it's ok for males to hangout at late night but not females is itself sexist and problematic; 3) the only reason for a rape case to happen is that there is a rapist, nothing else. After arguing and reasoning with you, Alice replies that "Oh, I have never thought about that."

Case 2: Bob believes that he has zero tolerance for racism, and he is active in movements like "Black Lives Matter." In his daily activities, he almost never shows any instance of bias against black people. However, after trying out an IAT test, Bob is startled to be told that he has

racial bias to the extent that he does automatically associate the term "African American" with "Bad" or "Violence".

Case 3: Carol used to live in a diverse community, and she is friends with a lot of black people. She never experienced any unusual or negative emotion or affect when she interacted with black people, and she genuinely believes that it's wrong to make generalizations about people from other races. Carol then moves into another town and she is aware that 80% of the crimes in town are committed by black males. Since then, *whenever* she encounters a black male at night on the street, she starts to feel anxiety and fear.

I propose that these cases of implicit bias do not have the same etiologies, and will now point to the differences. Note that this list of etiology is not exhaustive, and while the members are distinct from each other, they are not necessarily mutually exclusive. It might very well be that hybrid etiologies sometimes underlie particular instances of implicit bias.

2.2.1 Ignorance

As mentioned above, the extent to which one's behaviors are consistent with one another is a result of long-term reflection and adjustment. Cases of behavioral incoherence often occur during the habituation process due to ignorance. By ignorance, I mean not being aware of the consequences or implications of a concept or belief.

One intrinsic problem with the P-is-Wrong type of belief, especially if P is socially or culturally constructed, is that the concept or proposition P is in itself highly abstract and complicated. Consider the following two beliefs: 1) it's wrong to open the door; 2) it's wrong to be sexist. Belief 1 is much more straightforward to follow, i.e. once we have formed the belief

"it's wrong to open the door," all we have to follow is not to open the door, and nothing else. However, it's not the case for belief 2. Belief 2 can have numerous logical consequences that the subjects are not aware of, and some logical consequences do not immediately follow from the normative belief. The subject cannot be expected to be fully aware of all the logical consequences of a normative belief like "Sexism is wrong," especially when the contradiction between her prior dispositions and the normative belief require a relatively complex proof. That is, unlike Type 2 belief, the normative belief "it's wrong to be sexist" does not provide any direct and immediate guidance as to how we should behave.

Take *Case 1* where Alice's occasional expressions contradict her explicit belief as an example. The replies like "I have never thought about that" suggest some degree of ignorance from Alice; and the fact that responses are from different angles also suggests the complex consequences entailed by the term 'sexism' itself. For instance, the second response "thinking that it's ok for males to hangout at late night but not females is itself sexist" is not immediately entailed by the normative belief 'sexism is wrong.' To see why Alice's suggestion contradicts her own belief, she would be expected to at least go through the following steps of reasoning.

1. Sexism is wrong.
2. Females should stay indoor at night in order to be safe.
3. Premise 2 suggests a higher level of desirability for males to stay outside at night and be safe, but not females.
4. Premise 2 suggests a double standard in terms of expectation between the two genders, which is clearly sexist.
5. Premise 2 contradicts premise 1, and thus premise 2 should be abandoned.

It's possible for someone to authentically endorse a higher-order belief without herself fully going through the complex reasoning or realizing why her prior dispositions contradict her beliefs, until someone else points out.

For topics like ‘sexism’ or ‘racism’ in which the conclusion of academic debate is still up in the air, e.g. it’s still under debate whether affirmative action is racist or a remedy of racism, it’s conceivable for people to be ignorant of the implications and consequences of their behaviors or expressions during the process of habituation. The constant reflection of people’s behavior or dispositions can be initiated either by themselves or by others. Thus, I take ignorance to be one of the etiologies of implicit bias. The case of ignorance happens when the subject is not aware of the logical implications of what she avows intentionally, and therefore does not behave in accordance with her avowed higher-order normative belief.

It should noted that in *Case 1* though Alice intentionally avows that “females should learn to protect themselves by staying home at night or dressing up ‘properly’,” we wouldn’t call it an explicit bias from Alice. A distinction between intentionally avowing a sentence and being consciously aware of the implications of the sentence should be made. What is being intentionally avowed by Alice is the sentence per se to express her concern about females’ safety at night, not the implications of her expression. So long as one can intentionally avow a sentence P without consciously aware of the racist or sexist implications of P, cases of ignorance should be counted as implicit bias instances.

2.2.2 Automatic Association

I suggest that the most plausible etiology for case 2 is automatic association, or automatic categorization. Similar to Case 2, participants of the IAT test are often shocked to see that they do automatically associate a certain group of socially stigmatized people with some negative stereotypes, even though they are not aware of their implicit bias or the manifestation of their

implicit bias in daily lives. The manifestation of automatic association in daily cases might be that people unintentionally associate females with family, or associate males with higher intellectual capacity in areas like science and mathematics.

As Gendler and most psychologists work in the area of implicit bias have suggested, automatic association is a result of repeated patterns of response that might be unintentional, uncontrollable and beyond one's conscious awareness.

As introduced in the previous chapter, Eric Mandelbaum denies that associative cognitive structure underpins implicit biases; instead, he argues that implicit biases necessarily entail a propositional structure because they are reason-responsive. To support that argument, Mandelbaum cites the following experimental study: Two groups of people participate in a race IAT intervention. One group of participants are asked to read arguments that strongly encourage the hiring of African American professors. The strong argument makes a connection between the rise of the number and quality of professors without corresponding tuition increase if African American professors are hired. Another group of participants are asked to read a weaker version of the arguments, arguing that the hiring of African American professors could become trendy and allows more free time for the currently hired professors (Brinol et al.,2009). The experiment then shows that the participants in the strong argument group showed more positive implicit attitudes toward African Americans than the ones in the weak argument group. Mandelbaum infers from these experimental results that the so-called automatic association proposed by Gendler and other psychologist is actually reason-responsive, i.e. reading the strong argument about the positive benefits for hiring African American professors helps strengthen the

association between African Americans and positive evaluations, which inevitably undermines the very nature of associative structure.

It should be noted here that there are at least two ways to interpret the experiment results: 1) as Mandelbaum has argued, the metaphysics of implicit bias is essentially a belief state that is processed in conjunction with unconscious reasoning; 2) after reading the strong arguments, the participants put in more personal level effort, which requires some form of reasoning, to suppress the potential automatic association they have while taking the IAT test. The second possibility suggests that the participants themselves do not go through unconscious reasoning when taking the IAT test; rather, they themselves become more cautious when taking the test to not show strong association between African American and negative evaluations.

One crucial aspect of IAT to keep in mind is that the algorithm of IAT aims at the *average difference* in terms of reaction time between different sets of combination. That is, if it takes more time in average for one to react to the combination of 'women/career' than the combination of 'women/family,' we say that one has a slight/moderate/strong association between women and family. However, the result can be easily manipulated if the participant puts in more personal effort to remind herself not to associate 'women' with 'family' or 'black' with 'bad.' If the participant constantly reminds herself what shouldn't be associated together while looking at each combination, it might take more time in average for her to make her choice, but drastically changes the *average difference* in terms of reaction time. Though participants are reminded to press the keyword as fast as possible, there is no guarantee that the participants would not go through some process of personal level reasoning while making their decisions. Thus, what might have happened to the Brinol experiment is that, the participants take more time making

their decisions after reading the strong arguments due to a personal level intervention, which inevitably results in an average smaller difference between the reaction times to different combinations. In other words, what the second possibility suggests is that the evaluations the participants make are no longer effortless but more deliberative, i.e. each evaluation they make reflects their personal level meta-belief. If that is the case, then it rejects Mandelbaum's suggestion that automatic association is necessarily reason-responsive and thus a belief state.

What is being argued by Mandelbaum here is that the etiology of automatic association is not involved in the above experiment. The rejection of his argument hence leaves the possibility of the alteration of automatic association open. What kind of methods can be used for alteration and whether certain forms of intervention have a permanent or temporary effect on the associative structure are still unknown, precisely because the exact mechanism of automatic association is still unknown.

2.2.3 Unconscious Reasoning

I argue that the etiology of *Case 3* is unconscious reasoning, specifically unconscious inductive generalization, rather than automatic association.

The fear and anxiety Carol has is a manifestation of her attitude towards the black people she has encountered in the new town, i.e. she might unconsciously suspect the black man who walks towards her to be potentially dangerous or violent. Carol almost never experienced any fear or anxiety about black people before she moved into the new town. As someone who genuinely believes in anti-racism, Carol also constantly avoids making false generalizations about people from other racial groups. Given that 1) Carol constantly reminds herself not to make false

generalizations, and 2) she is aware of the criminal rates in her new town, the only reason left is that her fear and anxiety arose due to an unconscious generalization. That is, the fear and anxiety that she experiences might be a result of the following unconscious inductive reasoning.

1. 80% of crimes in town are committed by black males.
2. A black male is walking towards me.
3. It's highly possible that this black male is going to commit a crime.

The unconscious reasoning process I have listed above is a result of the subject being aware of the general statistics or background of the black people in town, and can be processed unintentionally, effortlessly with its initiation beyond the subject's conscious awareness. This type of reasoning is different from the deliberative reasoning that philosophers rely on, precisely because it blurs the line between deliberative reasoning and intuition. There are at least two reasons as to why such reasoning process can be effortless and almost automatic: 1) the proof requires very few steps of premises; 2) the repeated performance of reasoning would also make the process more and more effortless.

2.3 What's Wrong?

The three etiologies I have listed above, especially the last two etiologies, might seem perfectly acceptable in other circumstances. For instance, automatic association and effortless inductive reasoning sometimes do help us navigate through the world around us. Above we have tried to get clear of the nature of our explicit beliefs. In this subsection, I will show that the problem of implicit bias does not necessarily lie on the level of mechanism. Instead, once we realize that our anti-sexism or -racism beliefs are themselves normatively-loaded, we do have a duty to intervene or at least be more careful about our sub-personal mechanism when it comes to

racial, gender or sexual orientation issues. Whether we choose to intervene or not and whether we choose to abandon our prior dispositions to conform to our normative beliefs or not reflect our evaluative stance, which might in turn render moral responsibility.

For instance, as Gendler points out in her paper "On the Epistemic Costs of Implicit Bias, (Gendler 2011)" the ability to make rapid, automatic and effortless categorization in the form of automatic association is fundamental to how we make sense of the world and how we navigate through the world. The term 'automatic' here refers to a cognitive process that is unintentional, beyond our conscious control and requires minimal attentional resource. That is, the general ability to automatically categorizing our surroundings is actually essential to our survival and cognitive development. For instance, most of the time we see an object in the shape of a cup, we automatically associate the property of being usable for drinking liquid with it.

However, the very essential capacity of us should be called into attention when it comes to racial and gender issues. The employment of automatic association in normatively-loaded matters generates both epistemic and moral concern. From the epistemic perspective, our normal automatic categorization is justified because there is indeed a connection between a certain type of object and a certain type of property. In short, our automatic categorization or association is justified because we are certain in some circumstances that all Fs are Gs. However, it is not the case in racial, gender and sexual orientation issues. It is too disreputable to consider race as a subspecies, and thus groundless to categorize people from 'a race' with some particular, especially negative, traits, which might in turn give rise to discrimination and hate crimes. As compared to the non-normatively-loaded automatic categorization which helps us navigate through everyday lives, such 'ought-not-to-be' automatic association is more likely to result in

the infringement of certain groups of people's human rights, and thus should be called into reevaluation.

Or another example from the etiology of unconscious reasoning is the following case. 80% of the time when Douglas visits Diagon Alley, it is raining in Diagon Alley. As a result, Douglas takes a raincoat with him whenever he visits Diagon Alley. The decision of Douglas is not necessarily a result of deliberative intention, and such decision can be a product of unconscious generalization from Douglas: 1) 80% of the time during my visit in Diagon Alley it is raining; 2) it's highly likely that it is going to rain in Diagon Alley; and thus 3) a raincoat is needed. The mechanism for this case is structurally similar to the mechanism for *Case 3*, but the judgment or decision made by Douglas is more acceptable than the one made by Carol, precisely because *Case 3* is normatively loaded.

With that said, I am arguing that at least three different mechanisms can give rise to the cognitive phenomena of implicit bias and the very mechanisms themselves might be perfectly acceptable in circumstances where no normatively-loaded phenomena are involved. That is, an otherwise typical and useful cognitive mechanism has devastating outcomes when it comes to racial, gender and sexual orientation issues, or anything else that is normatively loaded. Thus, the higher-order normative beliefs we have against racism or sexism render us a special moral duty to be careful and reflective during the employment of these sub-personal mechanisms.

3. Categorizing Examples Cited by Philosophers in terms of Etiologies

After having listed the possible etiologies that give rise to implicit bias, I am going to argue in this section that some philosophers are talking past each other because the examples they have

cited in their arguments can actually be categorized under the different etiologies that I have listed above.

For instance, as I have argued in section 2.2.2, it is too quick to take for granted that the test of IAT provides accurate measurement of our automatic association. On the one hand, we have Gendler and other psychologists who assume the homogeneity of implicit bias as automatic association; on the other hand, we have Mandelbaum who insists that implicit bias is essentially a form of belief. Though they disagree on the metaphysics of implicit bias, they all take the IAT test to be a fair indication of one's degree of implicit bias. However, as I have argued above, it is possible to manipulate the result of IAT with some personal level conscious effort to slow down the average reaction time while reducing the average difference between different reaction times. In this case, what Mandelbaum takes to be reason-responsive propositional content is actually automatic association intervened by personal level deliberative reasoning.

The Rozin examples cited by Mandelbaum and Levy, which I have discussed in Chapter One, in arguing against Gendler's Alief model are also problematic. Both Mandelbaum and Levy argue that what the Rozin examples show is that implicit bias necessarily involves unconscious inferential processes and thus propositional content. However, it should be noted that the experimental design of the Rozin experiments is radically different from the IAT test. The IAT test is designed based on the background assumption that automatic association is acquired as a product of one's long-time learning history. That is, one need not necessarily to go through inferential steps in order acquire the association or categorization, i.e. one can simply takes the association for granted as part of one's learning history. The experimental design of the Rozin experiments, however, does not involve any form of learning history that the participants can

simply take for granted for, which makes it extremely troubling to be used as an example to either argue for or argue against the mechanism of implicit bias.

Moreover, the interpretation from Mandelbaum also renders his own position self-contradictory. One the one hand, it seems immediately justified that the participants in the Rozin experiments believe that both glasses contain only sugar, simply because they have witnessed the whole procedure of pouring sugar. On the other hand, if Mandelbaum takes the participants' hesitation to drink from the glass labeled as 'poison' to be a result of unconscious belief in conjunction with unconscious inference, it seems necessary to follow that the participants must have implicitly believed that the glass labeled as 'poison' does contain poison. What follows is that the participants believe in both P and not P.

Switzchegebel's Ralph the sexist case, however, seems to be a vague example as compared to examples cited by other philosophers. The case of Ralph can be seen as a result of both his ignorance and automatic association. It's not clear to us whether Ralph has realized the implications of his emotional reactions to his female students. If he is not even aware of his reactions and no one else has ever pointed it out to him, we say that ignorance plays a role in his implicit bias. Regardless of whether Ralph is aware of his reactions or the implications of his reactions, his emotional reactions seem to be automatic, unintentional and out of his conscious control, which is a case that can be categorized under Gendler's Alief model. That is, the auditory input of his female students' bright comments automatically activate his reaction such as surprise, and his egalitarian avowals are his higher-order normative belief.

The table below would give a clear picture as to how different philosophers talk past each other when citing different examples for their positions.

		Nature of Relata	
		Non-Propositional Representation	Propositional Representation (e.g. Belief)
Nature of Relations	Association	Gendler	*Empirical*
			Levy: Patchy
	Reason-Responsive	╳	Mandelbaum

As we can see from the table, Gendler takes implicit bias to be non-propositional representations associated with each other; Mandelbaum takes implicit bias to be propositional representations in conjunction with inference; and Levy takes implicit bias to be propositional representations that are not always reason-responsive. It seems that non-propositional representations by definition are not reason-responsive; and possibilities of propositional representations associated with each other can be explored in further empirical studies.

Chapter Three: Moral Responsibility

In this chapter, I aim to understand we can be morally responsible for actions caused by our implicit bias. Does the etiology of the bias, and the degree to which we can be aware of it or control it, matter?

Contemporary philosophers who endorse the 'Quality of Will' view argue that conscious awareness of the reasons of our actions is not a necessary condition for one's moral responsibility. I will first introduce the term 'Quality of Will,' followed by an argument from George Sher, a proponent of the QoW view, which argues that one's dispositions and traits constitute one's moral responsibility regardless of the presence of conscious awareness of the reasons of actions. I will then present a counter argument from Neil Levy, who makes a distinction between the expression of one's will and the reflection of one's will, and argues that a mere reflection of will is not sufficient for moral responsibility. After showing shortcomings of both arguments that need to be addressed, I will finally propose a new account of moral responsibility for actions due to implicit bias.

1.The 'Quality of Will' View

1.1 P.F.Strawson: the Quality of Will Manifested in Actions

Among the discussions of moral responsibility, the term 'quality of will' is first introduced by P.F.Strawson in his paper "Freedom and Resentment" (Strawson 1962). Strawson argues that the reason as to why we have reactive attitudes, such as resentment, towards others after they perform some actions is that we hold people responsible or blameable if their actions fail to meet a certain standard of good will. I take what Strawson and his followers mean by 'will' as a

certain set of dispositions. These dispositions often shape and guide how we act. As Strawson himself put it, "in general, we demand some degree of goodwill or regard on the part of those who stand in these relationships to us, though the forms we require it to take vary widely in different connections. The range and intensity of our reactive attitudes towards goodwill, its absence or its opposite vary no less widely" (Strawson, 7). For instance, if someone cuts in line without apology, we might feel a sense of resentment because in general we expect people to apologize or at least explain the reasons for their behaviors.

The problem with reactive attitudes is that, as Strawson has pointed out, if we only have reactive attitudes towards a particular behavior of someone, our attitude about the relation between that person and her behaviors would be restricted by the action per se when it is performed. In what follows, Strawson then argues that besides the reactive attitudes, we also have objective attitudes when we interact with people. Objective attitudes and reactive attitudes are not mutually exclusive, though they are indeed opposed to each other. By adopting objective attitudes, we shift our attention from the *actions* to the *agents* themselves. Our adoption of the objective attitude, according to Strawson, is "a consequence of our viewing the agent as incapacitated in some or all respects for ordinary interpersonal relationships" (Strawson, 13). For instance, we might initially have a reactive attitude towards someone who cuts in line; our reactive attitude might then get altered by our objective attitude, e.g. we realize that she who cuts in line suffers from some mental illness and thus has difficulty understanding basic social etiquette and responsibility.

Some interpret Strawson's argument here to be an argument that emphasizes the priority of 'being held responsible' over 'being responsible.' That is, some interpret Strawson's argument in

the way that his argument entails the proposition that 'being responsible is defined in relation to the practice of holding responsible." It is not clear from the paper "Freedom and Resentment" itself that Strawson himself holds the strong position that we cannot talk about the concept of moral responsibility unless we bring in the practice of holding responsible. Moreover, the practice of hold responsible discussed in "Freedom and Resentment" is derived from different agents' psychological contribution, e.g. whether we have reactive attitudes towards someone's behaviors depends on how we ourselves judge the situation. Moreover, the concept of the 'practice of holding responsible' is in itself problematic, which will lead us to a debate between moral realism and antirealism. With that said, the rest of my discussions would simply take the side that one can be morally responsible for one's behaviors independent from a third person's judgment and psychological contributions, i.e. one's actual moral responsibility can be independent from the practice of holding responsible.

Though the innovative notion of 'Quality of Will' is introduced under the context of holding someone to be morally responsible, we need not necessarily always understand the notion of 'Quality of Will' in relation to a third person perspective. To be more specific, Strawson himself concludes that "the reactive attitudes...are essentially reactions to the quality of others' wills towards us, as manifested in their behaviour: to their good or ill will or indifference or lack of concern" (Strawson, 15). The term 'Quality of Will' is still useful in the discussion of moral responsibility in the sense that it helps us discuss one's moral responsibility in relation to one's dispositions, while at the same time bypassing the tricky awareness condition.

1.2 George Sher: Responsibility without Awareness

In *Nicomachean Ethics* Book III, Aristotle argues that "it is the voluntary ones that are praised and blamed" (Aristotle, 35), and he defines involuntary actions as "what [come] about by force or because of ignorance" (Aristotle, 35). The notion of 'force' or 'compulsion' is often discussed in relation with free will. Traditional debates over moral responsibility have been focusing on the notion of "compulsion," e.g. whether we have free will, or whether free will can be compatible with determinism. The notion of 'ignorance,' though generally neglected in traditional debates, has now attracted attention from contemporary philosophers in their discussions on the epistemic conditions required for moral responsibility, e.g. whether conscious awareness is required for one to be morally responsible or blameable. As mentioned in the beginning of this chapter, the difficulties some philosophers have encountered in arguing about one's moral responsibility for one's actions due to implicit bias is that implicit bias as a cause might be out of one's conscious control and conscious awareness. For instance, some philosophers thus take the position that morally responsible actions are only consciously performed, or controlled.

Together with other philosophers like T.M.Scanlon and Angela Smith, George Sher is seen as a proponent of the 'Quality of Will' account of moral responsibility. In this chapter, I will only take Sher's account as a representative. As a proponent of the 'Quality of Will' view, Sher has tried many arguments to tackle the problem of conscious control and awareness, and argued that individuals might still be responsible for their actions despite the absence of awareness or control. For example, in his book *Who Knew? Responsibility Without Awareness* (Sher 2009), George Sher rejects the naive view that an agent is not responsible for whatever she is not aware

of, which he calls as 'the Searchlight View.' Instead, he proposes a disjunctive view of the epistemic condition that is required. Presented as follows.

> FEC (Full Epistemic Condition): When someone performs an act in a way that satisfies the voluntariness condition, and when he also satisfies any other conditions for responsibility that are independent of the epistemic condition, he is responsible for his act's morally or prudentially relevant feature if, but only if, he either
>
> (1) is consciously aware that the act has that feature (i.e., is wrong or foolish or right or prudent) when he performs it; or else
>
> (2) is unaware that the act is wrong or foolish despite having evidence for its wrongness or foolishness his failure to recognize which
>
> (a) falls below some applicable standard, and
>
> (b) is caused by the interaction of some combination of his constitutive attitudes, dispositions, and traits; or else
>
> (3) is unaware that the act is right or prudent despite having made enough cognitive contact with the evidence for its rightness or prudence to enable him to perform the act on that basis. (Sher, 143).

The sentence "[an agent's act] is caused by the interaction of some combination of his *constitutive attitudes*, *dispositions*, and *traits*" suggests that even if an agent is not aware of the wrongness of her action, there is still some way to connect her agency with her wrongdoing. Sher argues in Chapter Six that one of the factors that explains why some agents fail to realize the wrongness of their actions is exactly something to do with their own psychological states. For Sher, to come to know what an agent is like is precisely to know what kind of psychological states or dispositions she has. For instance, in Chapter Two Sher gives an example about how a babysitter feeds a baby with alcohol in order to calm down the baby, which inevitably leads to alcohol poisoning. Sher argues that in this case, the babysitter is for responsible for alcohol poisoning, not due to her failure as a lack of information. Instead, Sher argues the babysitter is

responsible because of her 'flawed patterns of thought' (Sher, 90), i.e. it is common sense that delicate babies should not be fed with toxic alcohol.

If the previous example sounds like an example that argues for moral responsibility based what an agent should have known, which is demanding and problematic for some philosophers, we can look at another example Sher gives in Chapter Two. Joliet is afraid of burglars and she is alone in the house while her husband and son are gone for holiday. She grabs a gun when she hears movement in her kitchen, and ends up shooting her own son, whom she thought is an intruder. Sher argues that Joliet is responsible for her son's death because "that the tendency to panic that prevents Joliet from recognizing that she should not pull the trigger is itself part of what makes her the person she is" (Sher, 92). Here, Sher seems to take the propensity to panic as part of Joliet's dispositional properties that constitute who Joliet is. Mistakenly shooting her own son is a manifestation of Joliet's very own dispositional character trait and thus Joliet is still responsible for her action, even though what she did is beyond her intention and probably awareness.

2. Neil Levy's Objection

2.1 The Expression/Reflection Distinction

To reiterate, the Quality of Will position in general holds that agents are morally responsible for actions that manifest their wills and who they are as agents, regardless of whether they are aware of their motivating reasons and attitudes. Neil Levy objects to this position, and argues that if an agent is not aware of her motivating attitude, then the attitude only reflects, instead of expresses, her will.

Levy's objection to the QoW view is supported by a distinction he makes between actions that *express* a will and actions that merely *reflect* a will. For Levy, an action expresses a will when the action is caused by the will or attitude in the right way. An action merely reflects a will when it conveys, but is not caused in the right way, by some will or attitude that the agent actually has. Levy further elaborates that the difference between the expression and reflection of an action lies in whether the link between actions and attitudes is *accidental*. That is, if the link is only an accidental one, then the actions merely reflect the attitudes that the agent actually has. (Levy 2011, 7) According to Levy, we should understand the term 'accidental' by asking ourselves whether the attitude manifested is 'an *isolate* in [one's] cognitive economy' (Levy 2011, 10). For instance, according to Levy's own analogy, we don't count a justified belief as knowledge if this belief is merely accidentally justified, i.e. this particular belief might be an isolate within a web of beliefs or even at odds with the rest justified beliefs. For instance, imagine a case in which a child believes that all cokes are poisonous and happens to believe that the coke her dad is drinking is not poisonous simply because he is drinking it, we would not necessarily conclude that the child has knowledge about coke. Similarly, if one attitude that is manifested in our actions is disconnected from all our other attitudes, Levy argues that this attitude only shares an accidental link with our wills and thus does not reflect who we are.

How should we understand the notion of 'isolate' in relation to the frequency of the manifestation of particular attitudes? Later in the text, Levy argues that it's one thing that one incident that reveals a particular attitude happens while it's another thing that *patterns* of actions or behaviors that reveal the same attitude happen. Here Levy seems to suggest that if an agent performs an action that 'reflects' some abnormal or isolate attitude of the agent only one or two

times, then we have to conclude that this attitude is only an isolate in the agent's cognitive structure, and thus her evaluative stance. Being a rare incident is part of what Levy understands as an isolate in one's cognitive structure though it does not exhaust the list of what count as 'an isolate.'

To see what Levy means, I will use some of his own examples to elaborate. For instance, Levy argues that if an agent is drunk, i.e. she is thus temporarily cognitively impaired, whatever she says or does cannot clearly express her attitudes or dispositions. In this case, even though there is *an* attitude that is expressed, it is not legitimate for us to claim that the attitude that is expressed is the agent's attitudes. Levy argues that what is entailed by the QoW view is that the "actions express the agent's practical identity, or at very least a significant part of it: if an agent is essentially good, for example, but occasionally has a passing malicious thought, an action that is caused by a malicious impulse and which is very out of character for the agent may fail to express the quality of her will" (Levy 2011, 8). The phrases used by Levy here seem to suggest that we are able to draw a clear line to conclude that a particular agent is *essentially good* and thus bracket whatever actions that are malicious and occasional under the category of "an isolate in one's cognitive structure."

The previous example used by Levy might not be a good one since it involves the use of alcohol as disinhibitor, so I will give another example from Smith that is discussed by Levy. Smith gives an example in which a woman forgets her friend's birthday, and argues that the woman is morally responsible for the lapse because it expresses her quality of will. The reason given by proponents of the QoW view is that our lapses "can reasonably be taken to reflect *a lack of appreciation* for the significance of the events in question" (Arpaly, 29). Levy rejects

Smith's argument because he thinks that Smith is too quick to conclude that the omission is exactly due to a lack of appreciation. Levy argues that Smith's argument is built upon the conceptual normative claim that "our attitudes have a rational connection with what occurs to us unbidden, what we notice and what we neglect" (Levy 2011, 11), which is too empirically demanding. Given we are not ideal rational agents, it is possible that our conscious and unconscious states are inconsistent with each other sometimes; and according to Levy, some of the unconscious mechanisms we have are not the ones that we actually endorse. Note that Levy seems to interpret the QoW proponents' position as such that the causal link between one's attitudes and actions entails that one endorses whatever attitudes that cause the actions. In what follows, Levy argues that it is possible that the lapse is caused by some unconscious attitudes that we as agents do not personally endorse, i.e. the lapse might not necessarily express who we are but merely reflect *some* attitude.

2.2 Against Levy's Objection: the Complexity and Inconsistency of Dispositions

As elaborated in the previous section, the main objection from Levy against the QoW account is that he thinks in a lot of cases where there is a lack of conscious awareness, actions often do not express who the agents are even though they do reflect some types of attitudes or dispositions. However, the line drawn by Levy between accidental cases and non-accidental cases is problematic because Levy's argument relies on the assumption that we can be certain about what kind of people the agents *essentially* are based their dispositions, and then categorize those attitudes or actions that are rare and at odds with their dispositions and agencies as an isolate. As mentioned above, Levy thinks that if an agent is *essentially good*, then the occasional

malicious thoughts are merely isolate. Whatever categorized as an isolate thus only reflects but not expresses one's will. However, an immediate question can be raised about the notion of one being *essentially good*: by '*essentially good,*' is Levy suggesting a strong position regarding human nature that people don't change over time in their traits or dispositions?

In this section, I will thus present an argument against the sharp distinction that Levy is trying to draw between the mere reflection and expression of one's will, by focusing on the complexity of our dispositions. I will argue that given the complexity and inconsistency of our dispositions, it is difficult to tell right at the moment whether an action reflects or expresses one's will. Whether we should reach the conclusion such that our dispositions are not as consistent as we used to think or that there is no clear cut between the reflection and expression of dispositions or will, Levy necessarily faces serious objections with his attempt of drawing the sharp distinction.

To first see what I mean by the complexity of dispositions, I will give few examples where a particular disposition can be manifested differently in emotions, thoughts and actions. For instance, take Jane Austen's characters as an example, Gilbert Ryle has argued in his *The Concept of Mind* that,

> When Jane Austen wished to show the specific kind of pride which characterised the heroine of 'Pride and Prejudice', she had to represent her actions, words, thoughts and feelings in a thousand different situations. There is no one standard type of action or reaction such that Jane Austen could say 'My heroine's kind of pride was just the tendency to do this, whenever a situation of that sort arose' (Ryle, 31).

Someone might have a disposition of pride and manifest it in different ways. For instance, the disposition of pride can be manifested by one refusing to perform in front of the public due to the

fear of embarrassment, or by one looking down on others, or by one feeling anxious when one loses to others.

Second, dispositions are not only complex, but often inconsistent with one another. For example, we can imagine a case where a student is thought to be responsible by her professors and fellow classmates, e.g. she always finishes her readings before class, never submits her assignments late, and almost never breaks any of her promises; however, it's totally conceivable that the same student never turns off the light in her room and never separates recyclable trash from landfill, etc. Or consider another example where a student is considered to be peaceful and easygoing by her fellow classmates. For instance, whenever there is a confrontation, she is always the one who takes one step back and acts as if she is totally fine. She seems to be a peaceful, considerate and pleasant presence in class. However, it is also conceivable that she always takes one step back simply because she wants to leave a good impression; whenever after a confrontation, she would always redirect her hostility to her pet. Moreover, she will remember everyone who has had confrontation with her, even though she tries to hide her antagonism. In this two cases, agents' manifestations of their dispositions are often at odds with each other. Moreover, if someone is only peaceful and considerate in the public to avoid confrontation, there is no fact of the matter whether this person does not necessarily have the disposition of being considerate or that the inconsistency in her behaviors is only an isolate of her cognitive structure.

Third, we should also keep in mind that it's one thing that we have certain dispositions, it's another thing that those dispositions get manifested. Dispositions can only be manifested if there is a stimulus condition being present or it reaches a certain threshold. Drawing to Sher's example of the mom who accidentally killed her son, it is possible that even though she always has a

disposition of being afraid of burglars, she never manifested any intense fear of burglars simply because the stimulus condition is missing, e.g. her husband and son had never left her alone at house while going out for a week-long trip. In this case, according to Levy's argument, there is indeed no pattern of this type of manifestation of her will. Nonetheless, it would be unjustified to conclude that simply because this is the first time, or probably the only time, of this type of manifestation, this manifestation is thus accidental, i.e. this action merely reflects her will instead of expressing her will.

Forth, sometimes one particular instance of manifestation that is at odds with an agent's practical agency does reveal something special and even genuine about the agent. For instance, imagine a case where Bob is always considered as a coward because he is afraid of darkness, water and fire. Bob's classmates always make of fun of him for not daring to turn on the bunsen burners in science class. However, one day when his neighbour's house gets on fire and Bob sees that a dog is crying in the burning house, he runs into the house to rescue the door despite his fear of fire. His classmates are surprised and amazed by his actions, and tell him that "You are not as cowardly as we used to think." If, following Levy's reasoning, we should consider Bob's constant fear of bunsen burners as part of his dispositions and practical identifies, then we ought to conclude that the dog rescuing case is merely an isolate in Bob's cognitive structures, which does not seem to be the right verdict here. We could also consider the dog rescuing case as a case where a certain threshold is present, e.g. rescuing a dog from a burning house is much more normatively loaded than using a bunsen burner in a science class. It is possible that Bob had never encountered a burning house with some animals inside, i.e. it is possible that the threshold

is never met, but it does not follow that Bob does not have the disposition to be brave simply because the braveness is never manifested due to the missing threshold.

Above I have given different examples where 1) people's dispositions are often inconsistent with each other (discussed in the first objection), and 2) the lack of patterns of manifestation should not simply be considered as an isolate in one's cognitive structure (discussed in the second, third and fourth objections). As mentioned in the beginning of this chapter, Levy's sharp distinction between the expression and reflection of one's will is based on the assumption that we can draw a clear line between one being *essentially good* and one's occasional malicious thought, which is problematized by the above examples. Given 1) the complexity and inconsistency of one's dispositions and 2) the required threshold or stimulus condition for the manifestation of dispositions, we can conclude that it is impossible to draw a clear distinction between what is reflected and what is expressed. As a result, Levy's objection to the QoW position should be rejected.

3. A New Account of Moral Responsibility: Idiosyncratic Evaluative Stance and Meta-Reflection

Though Levy's objection to the QoW account might not be a successful one, he is right in pointing out that some of the claims made by philosophers from the QoW camp are too empirically demanding. For instance, as argued by Levy in his objection to Smith's case where a woman forgets her friend's birthday, it is an empirical question to what extent certain types of information or even dispositional belief are easily accessible or retrievable. It's also an empirical question whether different people have relatively similar capacities in retrieving certain

information in order to make some decisions. That is, as implied by Levy, a reasonable account of moral responsibility should not be empirically demanding, and should not conflate empirical questions with conceptual inquiries.

Now I want to propose a new account of moral responsibility that helps distinguish different moral responsibility in relation to the three etiologies of implicit bias discussed in the previous chapter, i.e. the degree of moral responsibility differs from each other comparatively. The new account is structurally similar to the Quality of Will view in the sense that I take the actions caused by implicit bias to reflect *something* of the agent, i.e. an agent's idiosyncratic evaluative stance and personal effort for high-order reflection on her actions. I propose that the new account that emphasizes the Idiosyncratic evaluative stance and meta-Reflection (IRA for short) should be stated as follows.

> IRA: One is morally responsible for one's actions caused by implicit bias *if and only if* (1) the internalization of implicit bias is a result of one's idiosyncratic evaluative stance; and (2) once others point it out to the agent, she does not take initiative to meta-reflection or personal level intervention.

By 'idiosyncratic evaluative stance,' I mean the particular evaluative stance an agent takes on a particular issue; this particular evaluative stance is shaped by unique personal history and potentially unique. By "meta-reflection," I mean the ability to reconsider and rationally reflect upon the reasons and causes of one's actions. The notion of 'rational reflection' is important in the sense that if the option of meta-reflection is *completely* left out for an agent, it is questionable whether the agent enters the realm of moral responsibility, i.e. whether the agent is autonomous in performing her actions.

In what follows, I will assess the three etiologies distinguished previously in light of IRA and determine whether individuals are morally responsible for actions arising from bias with the

etiology. What I will be arguing with IRA is that this account allows room for personal effort in terms of reflection and intervention, and the role of personal effort is crucial in understanding why and how different etiologies leads to different degrees of moral responsibility comparatively.

3.1 Personal level Ignorance

I argue that what an agent avows explicitly reflects her evaluative stance to some degree, regardless of whether she is aware of the implications of what she has avowed. To understand why one's avowals still reflect her evaluative stance even though she is not aware of the implications, we should be careful about the distinction between one's intention in expressing her stance and one's awareness about the implications of assertions.

For instance, take Alice from *Case 1* in Chapter Two as an example, it is possible that the very expression that 'females should not be expected safe by staying outside at night' does reflect Alice's first order evaluative stance on females' desirability to stay safe outside. That is, given that Alice cares about the safety of females, she might think that staying inside at night is a good solution to prevent rape cases. Though the reasons as to why Alice said so might be that she is not aware of the implications or that she is not aware of other options in dealing with the situation, it does not necessarily undermine the fact that Alice does take a stance on the desirability of female staying safe at night when she avows the sentence. Similarly, consider another example. Olivia makes a suggestion to her friend Julia that Julia should reconsider her decision in attending philosophy graduate programs because the current climate in philosophy is not female-friendly, even though Olivia argues in other circumstances the climate in terms of

gender equality in academia should be improved. The reason as to why Olivia still makes the suggestion might be that she thinks that she cares about her friend Julia and does not want Julia to go through some rough years in academia. Olivia might not be aware of the inconsistency in her expressions and beliefs, but it does not necessarily undermine the fact that she does take a stance in whether Julia should attend graduate programs. With that said, I argue that the unawareness of the implications and inconsistency does not rule out the possibility that one's avowal can express one's evaluative stance on some particular issues.

However, it would be too demanding to conclude from here that since Alice takes an undesirable stance on the desirability of females' safety, she is then morally responsible for her avowals. This is because, as argued in previous chapter, it would be too epistemically demanding to require an agent to be aware of every possible implications of her first-order beliefs or stances. For instance, if Alice lives in a society where the majority of people share her stance, she might not be aware of the inconsistency until someone else points it out to her. That is, it's possible that even someone has the potential to reflect upon her own bias, she doesn't even have the ability to do so in her own cultural or social environment.

However, once someone else has pointed out to her the incoherence among her behavioral dispositions, then Alice does have a moral duty to adjust her first order belief precisely because now she does have a direct and immediate access to her evaluative stance. That is, for a rational agent, once someone reasons with her why the stance she takes on a particular issue is incompatible with her higher order evaluative stance, the agent should be reason-responsive by abandoning one of the stances. Though in the previous chapter I have argued that it takes time for one's dispositions to be adjusted to conform to one's higher order beliefs, it is not the case for

first order beliefs. Once the reasons are available for a rational agent, she does have the immediate ability to deliberate a stance, e.g. in this case, abandoning one of the premises.

If Alice fails to do so, we would say that her failure does raise a genuine doubt as to what extent she really has or holds her higher-order anti-sexism belief, which will then put her in a position in which she is morally responsible for what she avows.

3.2 Automatic Association

In the same paper where Levy objects to the QoW account, Levy argues that having implicit bias is not the same thing as the agent being committed to the bias's content. Even though it is possible for our implicit biases to be altered, we should still keep in mind that "(1) our enculturated attitudes, acquired early and encoded in patterns of responses that are deeply ingrained, are resistant to change, and (2) having explicit beliefs with a contrary content is not sufficient to make a great deal of difference to our attitudes" (Levy 2011, 12). Here Levy seems to suggest that implicit bias attitudes are often enculturated attitudes that are largely shared within society and culture, and often end up being reason-resistant. However, that fact that an attitude can be enculturated doesn't rule out the possibility that different individuals can endorse a particular attitude to different degree.

To argue why it still leaves room for automatic association to reflect one's evaluative stance, I will present an agreement of Jules Holroyd. In her paper 'Responsibility for Implicit Bias, (Holroyd 2012)' Jules Holroyd rejects one of Jennifer Saul's arguments for not holding

individuals responsible for their automatic association between socially stigmatized groups and negatively connotated words/concepts. Saul's arguments presented as follows.

> 1. Individuals cannot be held responsible for cognitive states or processes whose causal etiology lies wholly in factors out of their control.
>
> 2. Living in a sexist and racist culture is out of an individual's control.
>
> 3. Having implicit biases (against e.g. women, black people) results solely from living in a sexist and racist culture.
>
> 4. Therefore, individuals cannot be held responsible for their implicit biases. (Holroyd, 279)

Holroyd argues that there are two ways to interpret premise 3:

> 3a. Having implicit bias (i.e. having certain cognitive associations) results solely from living in a sexist and racist culture.
>
> 3b. Being influenced by implicit biases (i.e. manifesting them in behaviour and judgement) results solely from living in a sexist and racist culture.

Premise 3b, according to Holroyd, should be rejected. Several experiments have shown that individuals vary significantly in the extent to which implicit bias show up, such as their response time in IAT (Amodio et al 2003). Acknowledging that different individuals might have been exposed to different experiences and social norms, Holroyd nonetheless argues that "the extent to which we manifest biases may rather be a function of other cognitive states we have, and over which we plausibly have control" (Holroyd, 280).

Other studies done by Devine and her colleagues (Devine 1989; Amodio et al 2003) suggest a significant difference between implicit bias (in the form of automatic association) manifested among people who see behaving in a non-prejudiced way as important *in itself*, and those who see it as important because of social pressure, as well as those who see non-prejudiced behaviour as of little importance. That is, individuals who hold non-prejudiced behaviors to be important *in itself* tend to display less bias in different implicit bias tests. It's uncertain whether the

negative associations per se are weaker for individuals who see non-prejudiced behaviors important in itself, or that these individuals have developed more effective mechanism to regulate the manifestation. It's also uncertain whether the mechanism to regulate the manifestation is from a conscious personal level effort or a sub-personal system.

Regardless, if there is a strong correlation between the personal level explicit attitudes and the manifestation of implicit bias, it leaves room open to argue that "the manifestation of implicit bias [in the form of automatic association] would appear to be a function of the agents' attitudes, values and beliefs" (Holroyd, 281). That is, the internalization of automatic association between a certain group of people and stereotypes might be a result of how an agent takes a particular stance when interacting with an attitude that is pervasive in a society.

However, as mentioned above, it still remains uncertain whether professional help is needed for intervention in the case of automatic association and whether the intervention is temporary or permanent. For instance, Patricia Devine and her colleagues (Devine et al 2012) have argued that a 12-week multifaceted intervention designed by psychologists not only decrease the target group's participants' bias propensity, but also raises their awareness and concerns about discrimination. The result suggests that professional help might provide effective intervention in reducing bias. However, it is not the case that every individual has the access to professional intervention. Brandon Stewart and Keith Payne (Stewart and Payne 2008) argue that implementation intentions, i.e. "specific plans linking a behavioral opportunity to a specific response" (Stewart and Payne 1332), can help reduce automatic association. At first glance, it seems to be a positive response to whether individuals themselves can intervene and alter their existing implicit biases. However, given their experiment designs of providing participants with

counter-stereotypical response strategy when the participants are tested by IAT, it is an open question as to what extent the so-called malleability of automatic associated bias is permanent.

Though it is uncertain as to what extent an implicit bias in terms of automatic association can be completely and successfully altered, it does not mean that an individual does not have the moral duty of being more conscious and thoughtful about one's automatic association. That is, even though it is unknown whether an agent can successfully get rid of her implicit bias caused by automatic association, she is morally responsible for her bias so long as she does not try to be more cautious about her propensities once she is aware of them.

3.3 Unconscious Reasoning

It is not clear to what extent one's employment of unconscious reasoning or inference per se reflects one's evaluative stance. For instance, Ellen, who is a logic major and has habituated relatively complex patterns of reasoning, might unconscious employ different forms of unconscious reasoning, e.g. unconscious induction, more frequently than Fiona, who is less educated and is not familiar with logical reasoning. As argued in Chapter Two, the very mechanism of unconscious induction per se is acceptable in most of our daily cases so long as they do not involve any normatively loaded phenomena. That is, it is conceivable that due to the patterns of reasoning that Ellen has developed, she is more likely to make unconscious false generalization when it comes to racial or gender issues. However, to what extent she is morally responsible for her false generalizations should not be assessed merely based on her employment of unconscious generalization.

That is, as I have argued in the previous two subsections, a meta-reflection aspect should be taken into account when assessing Ellen's moral responsibility. Once Ellen has realized that the usual patterns of reasoning cannot be justified in racial or gender cases, even if the pattern itself is well supported by well-researched statistics, she has a duty to intervene. Unlike the etiology of automatic association, the mechanism of unconscious reasoning should be more reason-responsive. For instance, in cognitive behavior theory, the thought record is one of many useful and fundamental tools for subjects to deal with their social anxiety. The basic strategy of thought record is to help a subject identify her negative habituated thoughts, identify the link between the subject's thoughts and her emotions, and then examine whether there are sufficient evidences to support such type of habituated thought or reasoning. Once it has been identified that there is no sufficient evidence in support of her automatic and often negative thought, the subject is encouraged to intervene with the question "Do I really need to believe what comes to my mind now?" when she encounters the same type of negative thoughts in future. The strategy of thought record is often found effective in cognitive behavior theory, which suggests that unconscious, especially habituated, reasoning should be reason-responsive.

Chapter Four: Possible Applications

In the previous chapters, I have distinguished different etiologies that lead to different degrees of moral responsibility. In this chapter, I will analyze a particular product of implicit bias, i.e. how implicit bias leads to object misidentifications. The term "object misidentifications" is an umbrella term for the misjudgment or false perception of objects. Cases of police officers firing on unarmed civilian often happen because police officers misidentified the objects in unarmed people's hands as weapons. I argue that two etiologies could both result in object misidentification via different mechanisms: automatic association could causally alter our visual perceptual experience, and habituated reasoning could cause us to make false judgement that is independent from our visual perceptual experience. I will then discuss the possible moral responsibilities by applying the conclusions from previous chapters.

1. Racial Bias and Riots

Cases of police misidentifying the objects in people's hands often happen in the states, and often lead to fatal outcomes. These fatal outcomes, rightly, trigger nation-wide protests or even riots against racism.

For instance, on August 9, 2014, a 18-year-old black man Michael Brown was fatally shot by the police officer Darren Wilson. Witness reports for the shooting of Michael Brown, the tragedy that triggered the Ferguson riots, differ as to what exactly Michael Brown was doing with his

hands, as well as whether he was approaching the police officer, who fired a total of 12 bullets at Brown, when he was shot.

Or consider the shooting of Keith Lamont Scott on September 20, 2016, another tragedy that triggered the Charlotte riot. According to the police, Scott was with a gun and he failed to listen to the police's order to drop his gun (Shoichet). However, according to Scott's daughter Lyric Scott, Keith Scott was only reading a book in his car, which the police denied. The police asserted that they did find a gun, but not any book, in Scott's car. According to attorney Justin Bamberg, it's uncertain whether Scott was holding a gun or not; moreover, it's even impossible to tell what Scott was holding in his hands after viewing police videos.

Another more detailed example is that of Amadou Diallo, an African immigrant, killed on February 4, 1999 by four New York City police officers after 41 shots. The four involved police officers were assigned to the Street Crimes Unit, i.e. one of their tasks was to take illegal guns off the street. Upon further investigation, no weapons were found in the victim Amadou Diallo. Instead, there were only a pager and a wallet lying beside Mr. Diallo's body. According to Diallo's neighbours, he had no criminal record during his two-year stay in New York City, and he was often described as a shy and hard working 22-year-old young man. "It was unclear yesterday why the police officers had opened fire on the man at 12:44 A.M. in the vestibule of his building at 1157 Wheeler Avenue in the Soundview section" (Cooper), the police inspectors told the New York Times.

It seems that all the cases above involve some kind of object misidentification, i.e. the police officer mistook whatever in the victim's' hands as weapons. The source of their mistake could be the cultural and social stereotypes that associate the black people with characteristics such as

"aggressive" and "violent," which are internalized by police officers either voluntarily or involuntarily. The first possible explanation for the police officers' frequent object misidentification is that the implicit bias they have against black people causes them to perceive the victims with some kind of weapons, which is a case of cognitive penetration, a term I will discuss later. The second possible explanation is that the police officers are habituated to reason that black males are more likely to hold weapons in their hands and hence make false judgments, regardless of what they have seen actually. In the rest of this chapter, I will analyze and explain the two possible mechanisms, as well as the difference in terms of the moral responsibility police officers have for their actions.

2. Automatic Association and Cognitive Penetration

2.1 Visual Perceptual Experience and Cognitive Penetration

Before discussing the possibility of cognitive penetration by implicit bias, I will first introduce what I mean by visual perceptual experience and cognitive penetration in this section.

2.1.1 What is Visual Perceptual Experience?

Philosophers generally make a distinction between perceptual experience and perceptual judgment. I use visual perceptual experiences here to refer to the *phenomenally conscious* states we are in when we visually perceive something. That is, I take the term visual perceptual experience as the *final product* of whatever internal perceptual processes that occur. In the rest of this chapter, I will also use "experiences" as an abbreviation when referring to "visual perceptual

experiences." Given that visual perceptual experiences are our phenomenally conscious states, each experience has its unique phenomenal character. To say that an experience has its phenomenal character is to say that there is something it is like to be in a particular experience uniquely. For instance, the phenomenal character of the experience of perceiving a white male is what it is like to perceive that white male. The phenomenal character of the experience of looking at a black female is what it is like to look at that black female.

Given that it is phenomenally different for us to perceive a white male than to look at a black female, it follows that the representational properties of the two different experiences are also different. Representational properties, by definition, are those properties conveyed to us via our experiences. It is possible that the representational properties in one's experiences are entirely determined by the subject's properties in situations like hallucination; it is also possible that the representational properties in one's experiences are partially determined by the subject's properties in tandem with the mind-independent objects. The most important thing to keep in mind is that 1) representational properties are not the properties of mind-independent objects, instead, they are those properties that are presented to us by our experiences, 2) representational properties can be partially determined by both properties of mind-independent objects and the properties of the subjects themselves.

The notion of representational properties is crucial to understand the content in our visual perceptual experience. For instance, in her book *The Contents of Visual Experience*, Susanna Siegel defends the view that our experiences have contents, i.e. the "Content View," and argues that for an experience to have content is for it to have an accuracy condition. An accuracy condition is a condition under which the content of experiences can be verified. Her argument to

defend the Content View is that the contents of experiences derive from the representational properties of those experiences. That is, what the content of one's experience reflects is what it is like for one to be in that particular experience. In defending the Content View, she provides a framework called Argument from Appearing to test the content of our visual perceptual experiences.[1]

To make it clearer as to what Siegel and I mean by representational properties, I will give a few examples. Suppose I am reading a paper and suddenly I hallucinate a pink elephant. At this point of time, what it is like for me to be in this experience is what it is like for me to see a pink elephant. The representational properties in my experience are thus the color, shape and natural kind properties of the pink elephant even though there is no real external pink elephant. Or suppose I am playing with my black cat and suddenly I only see a white cat in front of me. What it is like for me to be in such an experience is what it is like for me to see a white cat. Thus the representational properties in my experience are the color, shape and natural kind properties of the white cat even though what is really in front me is a black cat. Or suppose that a pine tree looks different to me before and after I have carefully studied the nature of a pine tree. If the same tree looks different to me, it suggests that the properties that are conveyed to me via my

---

[1] Susanna's Argument from Appearing is summarized as below.

- P1: All visual perceptual experiences present clusters of properties as being instantiated.
- P2: If an experience E presents a cluster of properties F as being instantiated, then: Necessarily: things are the way E presents them only if property-cluster F is instantiated.
- P3: If necessarily: things are the way E presents them only if property-cluster F is instantiated, then: E has a set of accuracy conditions C, conveyed to the subject of E, such that: C is satisfied in a world only if there is something that has F in that world.
- P4: If E has a set of accuracy conditions C, conveyed to the subject of E, such that E is accurate only if C, then: things are the way E presents them only if property-cluster F is instantiated, then: E has a set of accuracy conditions C*, conveyed to the subject of E, such that E is accurate iff C*.
- Conclusion: All states of seeing objects having properties have contents. (Siegel, 45)

visual perceptual experiences are also different, i.e. higher level natural kind property is present after I gain the recognitional disposition of a pine tree in this case.

With that said, the ontology of experiences within Siegel's framework, which is also the ontology I am endorsing in this section, suggests that representational properties are whatever properties that are conveyed via one's experience to oneself, i.e. the properties of the *product* of internal perceptual processes. It seems necessarily to follow that a change in one's phenomenal state leads to a change in the representational properties. That is, phenomenology of one's experience necessarily determines whatever properties that are represented in one's visual perceptual experience within Siegel's framework.

It is also worth noting the historical distinction between perceptual experience and perceptual judgment before we proceed. Take the Hermann Grid Illusion as an example (figure 6).

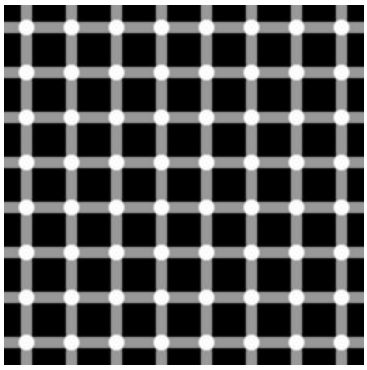

Figure 6

After being told that there are indeed no grey blobs at the intersections of the white grids, we might stop making a visual perceptual judgment about the grey blobs. However, the very judgment itself does not necessarily alter our visual perceptual experiences of the grey blobs.

2.1.2 What is Cognitive Penetration?

In her 2012 paper "Cognitive Penetrability and Perceptual Justification," Siegel defines cognitive penetration as follows:

> If visual experience is cognitively penetrable, then it is nomologically possible for two subjects (or for one subject in different counterfactual circumstances, or at different times) to have visual experiences with different contents while seeing *and attending to* the same distal stimuli under the same external conditions, as a result of differences in other cognitive (including affective) states. (Siegel 2012, 205-6)

According to Siegel, "cognitive penetrability is a thesis about the etiology of experience contents" (Siegel 2012, 207). That is, cognitive penetration examines what causes the content of our experiences.

Is it possible that the examples of police officers shootings reviewed above are cases of cognitive penetration? If we assume so, then the police officers literally *perceived* weapons in the victim's hands, instead of simply judging or guessing that they saw weapons. Automatic association, in this case, might play a causal role in the contents of their experiences. The counterfactual situation would be that, if the representation of a black man approaching with something in his hand does not automatically activate the visual representation of a weapon, then the lack of this particular automatic association would not cause the police officers to perceive a weapon. If every other variable in the opening examples stays the same, and automatic association is what alters the content of experiences, we say that it is a case of cognitive penetration. What does not count as a case of cognitive penetration? For instance, if whatever was in Mr. Diallo's hand did not look like a weapon to the police officers, but they nonetheless

guessed so without conforming to their experiences, we say that it is not a case of cognitive penetration but rather a case of false judgment.

The possibility of cognitive penetration is not limited to extreme cases like the opening examples, but also appears in our everyday life. For instance, imagine a case in which Alice and Bob quarrelled and stopped talking to each other for a few days. One day when they happen to see each other, Bob looks upset to Alice, and Alice thinks that Bob probably is still mad at her. According to Bob, he already forgets about the quarrel and is not upset at all the time they meet. However, the representational emotional property that is presented in Alice's visual perceptual experience is upset. It seems that in this case, Alice's experience is altered, i.e. cognitively penetrated, by her belief that Bob is mad at her. As discussed above, this case also confirms the position that if we want to talk about a subject's *phenomenally conscious* state, it necessarily follows that phenomenology determines the representational properties, in this case, the emotional expressions that get represented in experiences.

2.2 Possible Cases of Cognitive Penetration

2.2.1 Experimental Review

It is not clear whether the opening examples could be cases of cognitive penetration. However, in the rest of this section, I will try to prove that at least it is possible for some experiences to get cognitively penetrated by implicit bias, by closely examining a psychology experiment. In their 2003 paper "Facing Prejudice: Implicit Prejudice and the Perception of Facial Threat," Kurt Hugenberg & Galen V. Bodenhausen argue that "people's *earliest* perceptions of the faces of others are not immune to stereotypic biases" (Hugenberg and

Bodenhausen, 643). When discussing visual processing, cognitive psychologists distinguish different stages, e.g. early, intermediate and late vision. Early perceptions often refer to the detection of basic features such as motion, color and border. Early stage is often not related to phenomenal consciousness.

In conducting their experiments, they ask 24 European American participants to detect the offset and onset of facial hostility when they observe four movies with computer-generated faces. They are asked to observe one movie each time, each movie contains a facial morph that either is either black or white. The faces of the black and white targets are constructed by the software Poser 4™. The facial structures and expressions, such as the positions and shapes of mouths and eyes, are constructed to be identical. The ethnicity of the face is indicated by hair style, skin tone and hair color. The targets' faces morph from unambiguous hostility to unambiguous happiness.

Figure 7

By unambiguous hostility, it means that the constructed faces at the beginning are rated by the software as "substantially more hostile than the end-point happy expressions, regardless of raters' level of implicit or explicit prejudice" (Hugenberg and Bodenhausen, 641). In between the two stages of unambiguous hostility and unambiguous happiness, the targets' facial

expressions are designed to be ambiguous, somewhere between hostility and happiness, as seen in the figure. In their first experiment, the participants are asked to press a button when they no longer perceive the initial hostility in the targets' faces. Time is recorded for each movie respectively. All the participants are also asked to perform IAT afterwards to test whether they have high prejudice or not. The findings are that "participants higher in implicit prejudice [indicate] that hostility offset occurred later for Black faces than did lower-prejudice participants" (Hugenberg and Bodenhausen, 642).

The results of the above experiment can be explained by two possibilities: 1) the participants perceived longer hostility on the black faces due to their implicit bias; or 2) it took the participants longer time to judge what they had seen when they saw the black faces. To further test which possibility holds, Kurt Hugenberg & Galen V. Bodenhausen ask the participants to detect the onset of hostility in facial expressions in their second experiment. All the materials and designs in the second one are identical to the first one, except that the targets' faces morph from unambiguous neutral expressions to unambiguous hostile expressions. The findings are that "individuals high in prejudice [perceive] the onset of hostility much earlier for Black faces than did low-prejudice participants" (Hugenberg and Bodenhausen, 642).

Once the second possibility of the participants taking longer time to judge the onset or offset of emotions is ruled out, it seems that the only possibility remains is that the same emotion displays are *perceived* differently due to different degrees of implicit bias the participants have, i.e. implicit bias, particularly automatic association in this case, alters people's visual perceptual experience. Following Gendler's Alief model, the associated concepts such as "black people, hostile" and associated visual representational content of emotional categorizations are

unconsciously activated by the participants' visual input of the black people's faces. The question that remains then becomes: how exactly does automatic association cognitively penetrate one's visual perceptual experiences of emotions.

2.2.2 Emotions as Representational Properties in Experiences

To argue for the possibility of cognitive penetration, we thus need to argue that emotion is indeed *perceptually perceived*. The Hugenberg experiment reviewed above takes the position that the *earliest perceptions* of others' faces can be altered by prejudice, even though the general consensus among psychologists suggests that early vision processing only involves the detection of basic features such as color and motion. Similarly, philosophers also disagree on what properties can get represented in experiences. Some argue that only lower-level properties can be represented, while others argue that higher-level properties can also be represented. Lower-level properties refer to those properties such as size, shape, color, location and etc. In contrast, higher-level properties might include properties such as natural kind properties, causal properties, perceptional relation properties and etc. The distinction between higher-level properties and lower-level properties lies in the sense that a higher-level property such as natural-kind property can be seen as a combination of different lower-level properties. For instance, when perceiving a banana, the lower-level properties get perceived include it being yellow, crescent, big or small and being at certain location. That is, to say that only lower-level properties being perceived means that whatever gets represented in experience is exhausted by the representation of sensory qualities. In contrast, to say that higher-level properties such as natural kind property also gets represented in the experience of a banana suggests that the

property of being a banana, instead of just the separate properties of being yellow and being crescent, is also represented.

Or take the Hugenberg experiment as an example. If only lower level properties are represented, then we say that the participants only perceive the color and the shape of the computer-generated faces, together with the positions and shapes of the organs on the faces. If that's the case, we say that the participants do not visually perceive emotions. Rather they make judgments based on their perceptions, and so, cognitive penetration does not occur. However, if higher level properties such as emotion are also represented in our experiences, then participants not only perceive the location of the target's face, but also visually perceive different emotions ranging from hostility to happiness. In this case, cognitive penetration occurs because visual perceptual experiences themselves are altered by implicit bias, independent from the participants' judgments.

In arguing for a possible mechanism of the perception of emotions, I will first introduce a constructivist theory in psychology, the Conceptual Act Theory (CAT). What CAT hypothesizes is that emotion is perceived or experienced in a way in which the subject categorizes the physical changes on other people's faces or within herself via emotion concept knowledge. This act of categorization is called "situated conceptualization" (Barrett, 4). They refer to "situated conceptualization" as the process in which concept knowledge makes ambiguous sensations meaningful because the conceptualizations are highly context-dependent, i.e. the cultural and social background of a subject influences the way a subject categorizes her experiences.

The term "constructivism" indicates that people in the CAT camp are in a position against some traditional platonic understandings of emotion or perception. They argue that whatever

physical or mental states we are in do not have intrinsic functions as "emotions" or "perceptions." Instead, those physical states can be seen as the consequence of a *constantly modified constructive* process during which the stored concept knowledge makes sensory inputs meaningful. That is, emotions are not biologically hardwired; instead, we learn to categorize emotions within a particular social or cultural background. The sensory inputs themselves can be *ambiguous*. By concept knowledge, CAT refers to "the rich cache of instances that populate what someone 'knows' about different categories" (Lindquist et al, 2). The categories can be seen as fuzzy boundaries that are also highly culture-dependent. For instance, "black cat" can be considered as a threatening symbol of evil omens in Europe and Anglo-America, and thus be categorized under the umbrella of "fear"; "black cat" can also be considered as the symbol of protection and fortune in some regions in Japan, and thus be categorized under the umbrella of "sense of security," e.g. people in Japan buy black maneki-neko (the lucky cat) if they feel threatened. People from these two different cultures can have different emotional experiences if the concept knowledge that makes their sensations meaningful vary in culture.

Moreover, given the constructivist position of CAT, it is plausible to derive from their own argument that one's categorization of emotion can stay stable for a certain period of time if she stays in the same social or cultural environment within that period of time.

Lindquist's research might also be compatible with the possibility that the categorization of emotions constitutes our perceptual experiences. In their 2015 paper "The Role of Language in Emotion: Predictions from Psychological Constructionism," Lindquist and her colleagues have explored several cognitive science studies and argue that "language, once connected to certain perceptual representations that become stored as concept knowledge, alters ongoing adult

perception by selecting certain sensations for conscious awareness while suppressing other sensations from conscious awareness" (Lindquist et al, 10). That is, situated conceptualizations can operate at a subpersonal level that does not require the subject's conscious effort to constantly categorize how she feels or what she sees. Phrases such as "selecting certain sensations for conscious awareness" suggests the possibility of automatic selection, which leaves the possibility that CAT supports the position that emotions as higher-level properties can be perceived in one's experiences.

As argued above, CAT leaves the possibility of emotion being visually *perceived* open, I will now borrow Siegel's argument for rich content to further argue why emotions can be perceived as higher-level properties in one's visual perceptual experience. .

Based on Siegel's framework of the Content View, her Rich Content View argues that higher-level properties such as natural kind properties, causal properties and perceptional relation properties can be represented in our experiences. The significance of her Rich Content View is that it opens possibilities and provides a framework for the position that our visual perceptual experience can be cognitively penetrated by other mental states.

The method proposed by Siegel to test whether certain higher-level properties are represented in our experiences is called the method of Phenomenal Contrast.[2] The method itself is based on

---

[2] Her strategy can be summarized as follows. Let E1 be the target experience with the target higher level property and E2 be the contrasting experience.

- P1: E1 and E2 differ in phenomenology.
- P2: thus there is a phenomenological difference between E1 and E2.
- P3: E1 and E2 differ in content.
- P4: E1 and E2 do not differ in content with respect to their lower level properties.
- P5: E1 and E2 differ in content with respect to some property other than those which are a part of lower level properties.
- C: Properties other than those which can be a part of the lower level contents of experiences can be a part of the contents of experience.

assuming the infallibility of introspection, i.e. whatever information we gain through introspection is always infallible. To put it simply, the method is to imagine two experiences, i.e. a target experience and a contrasting experience, that differ phenomenally, form a target hypothesis that purports to explain the representation of a certain type of higher level property in the target experience, and see whether the target hypothesis provides the best explanation. The best explanation should thus be that the two experiences differ phenomenally because the target experience has the target content while the contrasting experience does not. The explanation that is "best" rules out other possible alternatives. These possible alternatives could include cases such as 1) the experiences differ phenomenally because there is a difference in non-rich contents, e.g. lower-level properties; 2) the experiences differ phenomenally because there is a difference in nonrepresentational properties, e.g. raw feels that are not representational; and 3) the experiences differ phenomenally because there is a difference in non-visual experiences, such as judgment.

To reiterate, we have assumed that the situated conceptualization of emotions during the period of experiment is stable, and we have discussed that CAT opens the possibility of unconscious regulation of sensory information that makes emotions represented in experiences. I will now proceed to argue that emotion can indeed be visual-perceptually experienced as a higher-level property by following Siegel's method of phenomenal contrast.

As we have already established, the perception of emotions undergoes a constructivist process in which concept knowledge plays a significant role in helping people categorize different sensations. Imagine a case in which Alice migrates to a foreign country where the culture is radically different from her own country's. She has a neighbour, Bob, who always

looks at her with a contorted face. As a new member, Alice visually perceives the shape and positions of Bob's eyes and mouth, without being able to recognize what they mean. Once after the constructivist process of situated conceptualization, Alice can then recognize that actually Bob's contorted face expresses his curiosity. That is, in this case, Alice gains recognitional dispositions of Bob's facial expressions. Does this example involve any phenomenological change that is caused by the representation of emotion? To apply Siegel's method, let E1 be the target experience in which Alice visually perceives the emotion as curiosity, i.e. Alice sees the facial expression and categorizes the expression as curiosity visually. Let E2 be the contrasting experience in which Alice sees the same expression without visually categorizing the expression as curiosity.

It seems intuitively obvious that E1 differs from E2 phenomenally. The first type of alternative I am going to rule out is the objection that the phenomenological difference between E1 and E2 is not a difference in visual perceptual content, but only a difference in either cognitive phenomenology or background phenomenology exclusively. That is, the familiarity that Alice gains via a recognitional disposition is not reflected in Alice's sensory experiences at all. The possibility of a change in background phenomenology should be immediately ruled out because the phenomenological change we are discussing here is a *local* change in perceiving Bob's face, i.e. it is implausible that every time Alice sees Bob she changes her mood or is drunk.

In terms of cognitive phenomenology, the change in phenomenology due to occurrent thoughts should first be ruled out. It is not necessary that Alice always explicitly entertains the propositional attitude such as "*that face* looks familiar" when undergoing the phenomenon. The

demonstration "that face" here specifically refers to a demonstrative thought when Alice sees Bob's curious face. For instance, Alice can simply infer the phenomenological change by recalling how Bob's face looks different to her before and after she gains the recognitional disposition.

The change in phenomenology due to commitment-conferring attitudes should also be ruled out. Suppose Bob's family has bought a newly-developed artificial intelligence that is designed as Bob's duplicate in every aspect, and let's call it Bod. One day when Alice sees Bod, she is informed that Bod is not Bob. Given that the propositional attitude "*that face* looks familiar" refers to a specific demonstrative thought when Alice sees Bob, we would suppose that Alice ceases to dwell on the attitude that she is looking at a face she has been familiar with for a long time. Nevertheless, Bod's contorted face should still look exactly the same to Alice after she gains the recognitional attitude, i.e. Alice would still experience something different phenomenally when looking at Bod. In this case, we can say that Alice's familiarity with Bob's hostile face is not exclusively due to a commitment-conferring attitude, because the visual stimuli in Alice's experience is not sensible to the mere difference between Bod and Bod.

What I have argued above suggests that at least the possibility of the experiences differing phenomenally because of *judgment* should thus be ruled out. Another alternative I am going to rule out is the objection that the phenomenological difference is not due to the recognitional disposition of *emotion* as a natural-kind property, but a curiosity-shape-Gestalt switch. The term curiosity-shape-Gestalt here refers to a general shape complex that captures the shared shape property of curious facial expressions in this community. This complex shape property may include the shape of eyes, mouth and proportional position relationship among the organs. That

is, under this possibility, when we say that Alice gains recognitional disposition after the process of situated conceptualization, we say that Alice gains recognitional disposition of this shape-Gestalt instead of the emotional property. Now let's imagine a counterfactual situation. In another possible world, the exact same contorted face of Bob expresses an emotion of repulsion, instead of curiosity. It is plausible to assume that the phenomenological change Alice experiences as recognizing Bob's face as curiosity is different from her experience of recognizing Bob's face as repulsion, due to the different connotations of repulsion and curiosity. In this case, the general shape Gestalt alternative is too abstract to help pinpoint the specific phenomenological change we are discussing here, and thus should also be eliminated.

So far, after eliminating other crucial alternatives, we have established that a phenomenological change in Alice's experiences before and after she gains a recognitional disposition is due to the representation of emotion as a *natural kind* property in her experience. That is, once after one gains recognitional disposition of a certain emotional property, that particular emotion can be represented in one's visual perceptual experience.

2.2.3 Penetration by Automatic Association

We have seen from the experiment that the facial designs of the black faces and the white faces only differ in terms of their skin color, hair color and hair style. The rest of lower-level properties such as the shape and position of their eyes, eyebrows and mouths remain the same. That is, the faces of the black target and white target share a general shape-Gestalt within the same time frame in the four movies. However, faces that share the same shape-Gestalt are perceived differently by the participants in the experiments. As argued above, alternatives such

as raw feels, occurrent propositional attitudes and background mood could not provide a satisfying explanation for the phenomenological change. The only apparent explanation left is that the facial displays are perceived differently by the participants because different emotional properties are represented in the participant's' experiences. Given that the black and white faces share the same shape-Gestalt, the only plausible explanation for the representations of different emotional properties is that *people learn to categorize the emotional displays of ethnical groups differently due to implicit bias*.

To reiterate, the phenomenology of one's experience can be partially constituted by the properties of mind-independent objects, and also partially constituted by the properties of the subjects themselves. Following the constructivist mechanism introduced, the participants' "conceptual knowledge," though might be endorsed by them unconsciously, is what makes their perception of facial displays meaningful. The United States, where the experiments are conducted, do have a long history of racial discrimination. Descriptions like "Black people are aggressive," "Black people are easily irritated," are "Black people are more hostile than white people," are indeed often overheard. The concept knowledge in the case of the experiment consists in associated concepts such as "black people" and "hostile" or "aggressive." This kind of concept knowledge need not be accurate, nor does it need to be based on the subjects' personal experiences. The concept knowledge can simply be a legacy of a society that is historically and currently high in racial discrimination. The participants themselves might not consciously commit to the associations, i.e. it is possible that, growing in a particular culture or society, the participants are unconsciously or implicitly endorsing the position supported by the associations. Even though the positions are not consciously endorsed by the subjects, they can

still become stored information that later regulate the subjects' perceptions. Accordingly, the participants might learn to categorize the same emotional displays in white and black people differently, due to implicit bias caused by automatic associations.

2.3 Cognitive Penetration in Object Recognition

Above I have argued for the possibility that our visual perceptual experiences of others' emotions can get cognitively penetrated by automatic association. One might question whether the same argument can be used for object recognition, because the emotional expressions along the shape continuum share much more perceptual similarity than objects like books, guns and wallets. The special aspect of the police officers' cases is that, as compared to the Hugenberg experiment, these officers aren't making object identifications in a vacuum. For instance, besides the perpetrator's skin color, the situational context might also be highly relevant, e.g. whether it is late at night in a higher-crime area, whether the police officer's job is to react swiftly to illegal gun possession, and whether the victim's facial expression or body language does send out a signal of hostility. Moreover, the location as well as the distance between the police officers and the victims might also be relevant.

To what extent our capacity for categorical perception can be undermined by distance and pressure is an empirical question. However, if there is room for long-distance under-pressure object miscategorization in our visual perceptual experiences, then it is possible that the police officers' experiences of perceiving guns are cognitively penetrated by their internalization of the automatic associations such as "black people, guns, dangerous."

3. Habituated Reasoning and False Judgment

As argued in the above sections, there are two possibilities to explain the experimental results: 1) implicit bias due to automatic association alters the subject's experience, which is a case of cognitive penetration; 2) implicit bias due to habituated reasoning only alters the subject's judgment of her experience, but not the experience itself. I have argued for the former possibility in the previous sections, i.e. the role of implicit bias is not to alter the perceptual content itself but only the subject's judgment. But what about the second alternative explanation?

The alternative explanation suggests that implicit bias operates independent from the product of our perceptual processes. One possible etiology for this alternative explanation is that based on the policemen's past experiences and judgments, they have developed habituated reasoning along the lines that 1) when a black men behaves suspicious with something in his hand, 2) it's highly possible that the black man is holding a gun in his hand, and thus 3) some actions should be taken.

That is, there exists a discrepancy a subject's judgment and her experience: the subject does not really perceive a gun in the victim's hands but nonetheless judges so. A discrepancy between one's experience and one's judgment is possible in the case of the opening examples, especially the Diallo case because the jobs of the four policemen is to react swiftly to any potential danger. The police officers might indeed have experienced numerous cases in which their suspicions of illegal gun possession are confirmed. The successful cases of gun confiscation then reinforce their habituated reasoning that if a black people is behaving suspiciously at late night on the street, it is highly possible that she or he possesses a gun illegally. That is, when the police

officers saw the victim, due to the special circumstances, it is possible for their perceptual experiences to be overridden by their habituated reasoning; and thus, they ended up guessing and judging that they saw a gun without confirming their experiences.

4. The Difference In Terms of Moral Responsibility

Above I have discussed two possibilities where implicit bias could give rise to police officers' object misidentification: 1) their visual perceptual experiences are being causally altered, or cognitively penetrated, by automatic association; 2) they make false judgement based on habituated reasoning that is independent from their visual perceptual experiences.

4.1 First Pass: the Difference between Experience and Experience-Independent Judgment

Prima facie, there is already a difference in terms of the justification of judgments based on experience between the two possibilities. The cognitive penetration case is a case where perceptual judgment is made and thus actions are performed based on one's visual perceptual experience; the false judgment by habituated reasoning case, however, is a case where one makes a judgment without conforming to one's visual perceptual experience. One might argue that for positions like police officers, they should have been more cautious in not making false judgments since a single false judgment might easily cause the death of an innocent man. In this sense, simply deferring to their habituated reasoning and judgment without conforming to their actual visual perceptual experiences is irresponsible of them.

To what extent the difference in terms of judgments' justification makes a difference in one's moral responsibility is still questionable. The position which holds that one's visual perceptual

experience provides a prima facie justification for one's judgment and belief is called Perceptual Dogmatism. That is, this position holds that if our visual perceptual experiences are such and so, then we are justified to believe that the world is exactly such and so. If Perceptual Dogmatism is true, then the prima facie difference between the two possibilities holds. That is, the case of false judgment cannot be justified by the visual perceptual experiences, while the judgment based on the cognitively penetrated experiences can be immediately justified. However, the Perceptual Dogmatism position, argued by philosophers such as Susanna Siegel and Nicolas Silin, is challenged by the possibility of cognitive penetration.

The problem seen by philosophers like Siegel and Silin is that, Perceptual Dogmatism position places no requirement on the etiology of experiences that are used to justify one's judgment. That is, it also immediately follows from Perceptual Dogmatism that, if our visual perceptual experiences are causally influenced or altered by our beliefs and desires, these seemingly unjustified experiences still provide prima facie justification for the beliefs that follow. For instance, as Susanna Siegel (Siegel 2012) argues, the perceptual justification some people have based on their visual perceptual experience as seeing black people as aggressive and hostile falls under a notorious and pernicious epistemic circularity. That is, on the one hand, people will readily see black people as hostile if they believe that black men are more likely to be aggressive and hostile than white men; on the other hand, they also rely upon their visual perceptual experiences as an evidence to confirm their beliefs. As a result, their reasoning and justification end up falling into a vicious circle. In cases like this where one's visual perceptual experiences are polluted and altered by one's cognitive states, philosophers argue that the

justification provided is thus downgraded. That is, the difference in terms of justification alone does not prove sufficient for assessing the moral responsibility.

4.2 Second Pass: Sensitivity to Reason and Meta-Reflection

I propose that the moral responsibility account discussed in Chapter Three provides a better distinction in terms of moral responsibility than the mere justification of judgment. I have argued in Chapter Three that two factors constitute one's moral responsibility for actions caused by implicit bias: 1) whether the etiology reflects one's idiosyncratic evaluative stance and 2) once implicit bias is called into attention, whether the agent is actively engaged in higher-order reflection.

The urgency of intervening in implicit bias should be called into attention to police departments and police officers after so many incidents of object misidentification have happened. Working in a position where a slight false judgment could cause the death of another innocent person, police officers have the duty to be constantly engaged in meta-reflections on their implicit bias propensities. The crucial difference in terms of moral responsibility thus lies in the meta-reflection part. As I have predicted in Chapter Three, the etiology of automatic association case is the least reason-responsive and thus the least sensitive to one's meta-reflection. Habituated reasoning by its very nature should be sensitive to meta-reflection and such forms of meta-reflection should be practiced by the police officers. My conclusion, then, is that police officers who make false judgments due to habituated reasoning process, perhaps formally instituted by departmental training and performance review, are more

responsible for their actions than those whose experiences have been cognitively penetrated by automatic association.

5. The Problem of Racial Profiling

Above I have applied the distinction among etiologies and the boundary of moral responsibility to a growing type of tragedy: the police officers misidentified the objects in the non-white people's hands and thus caused the death of unarmed non-white people. All the examples I have cited are examples where the unarmed non-white people are suspected for street level crimes by police officers. With that said, I want to draw a special attention to one of the most problematic discriminatory practices in the United States of America: Racial Profiling. Racial Profiling is the practice by many departments in the states, including the police departments, to target individuals for suspicions of crime based on their races, religions, or even nationalities. According to the "Guidance Regarding the Use of Race By Federal Law Enforcement Agencies" issued by the U.S. Department of Justice in 2003

> "Racial profiling" at its core concerns the invidious use of race or ethnicity as a criterion in conducting stops, searches and other law enforcement investigative procedures. It is premised on the erroneous assumption that any particular individual of one race or ethnicity is more likely to engage in misconduct than any particular individual of another race or ethnicity.
> Racial profiling in law enforcement is not merely wrong, but also ineffective. Race-based assumptions in law enforcement perpetuate negative racial stereotypes that are harmful to our rich and diverse democracy, and materially impair our efforts to maintain a fair and just society. (Guidance regarding, 1)

In their newly published "Guidance For Federal Law Enforcement Agencies Regarding The Use of Race, Ethnicity, Gender, National Origin, Religion, Sexual Orientation or Gender Identity," the U.S. Department of Justice further clarifies that "Law Enforcement Officers May Never Rely

on Generalized Stereotypes, But May Rely Only on Specific Characteristic-Based Information" (Guidance for Federal, 5).

On the one hand, we have the Supreme Court which holds law enforcement racial profiling as a serious violation of the Constitution, i.e. all individuals are accorded equal protection of laws, and the Department of Justice which constantly calls into attention the malpractice of law enforcement racial profiling. On the other hand, we constantly see tragedies happen as a result of the malpractice of law enforcement racial profiling. Given that the law enforcement racial profiling practice has been adopted by police departments and officers for a long time, it is possible that the negative products of such practice require serious and constant effort from the police departments to overcome what they have already internalized.

There are at least two central problems with the practice of racial profiling. First, the very agenda and practice of racial profiling makes it difficult to trace the moral responsibility and blameworthiness of the malpractice of the law enforcement police officers. The police officers are called law enforcement police officers because they are supposed to perform their duties in a policing capacities for public purposes. It is possible that for those police officers who did not have high implicit bias, they internalize habituated reasoning precisely because of the departmental practice of racial profiling. Second, the very structure of racial profiling reinforces the causal mechanism of implicit bias, whereas the police officers are of those who should pay special attention to their implicit bias.

As a result, concrete actions from the police departments should be taken to help their law enforcement officers to either alter or override their implicit bias that might be reinforced by their long-term malpractice of racial profiling. Even if it is still uncertain what kind of method

can help alter one's implicit bias permanently or temporarily, police departments should collaborate with different cognitive scientists to keep themselves updated.

Chapter Five: Conclusion

In the previous chapters, I have argued that implicit bias does not have a homogenous etiology. Instead, at least three distinct etiologies can give rise to what we call implicit bias. Moreover, I have argued that the problem with implicit bias does not lie exactly on the level of mechanism, but rather with the fact that the phenomenon is normatively loaded. We all have the sense that a stereotype against a socially stigmatized group of people is something that ought not be. The belief we have against sexism or racism is thus a higher-order normative belief, which by our own lights should provide some degree of guidance as to how we should behave. Thus, we have a special duty to reflect upon our implicit bias once it is made obvious to us, even though the success of intervention depends on the etiology of the bias and can only be assessed empirically. As a result, one is morally responsible for her actions caused by implicit bias if and only if 1) the etiology of that particular implicit bias instance reflects her evaluative stance and 2) she as an agent fails to put in personal level effort to reflect on her implicit bias once it is made clear to her.

Fortunately, despite all the tragedies reviewed in Chapter Four, there have been more and more governmental interventions in counter-implicit bias trainings. For instance, on June 27, 2016, the Department of Justice announced its department-wide counter-implicit bias training, i.e. “it will train all of its law enforcement agents and prosecutors to recognize and address implicit bias as part of its regular training curricula” (Department of Justice). According to the Department of Justice, the new training will over 28,000 department employees. Through the new training, they will learn how to recognize and address their own implicit bias.

Works Cited

Amodio, David M., et al. "Individual Differences in the Activation and Control of Affective Race Bias as Assessed by Startle Eyeblink Response and Self-Report." *Journal of Personality and Social Psychology*, vol. 84, pp. 738-53.

Aristotle. *Nicomachean Ethics*. 2nd ed., Hackett Publishing Company, 1999.

Arpaly, Nomy. *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford UP, 2004.

Bargh, John A. "The Four Horsemen of Automaticity: Awareness, Intention, Efficiency, and Control in Social Cognition." *Basic Processes*, edited by Robert S. Wyer and Thomas K. Srull, Lawrence Erlbaum Associates, 1994, pp. 1-40.

Barrett, Lisa Feldman. "The Conceptual Act Theory: A Précis." *Emotion Review*, vol. 6, no. 4, 2014, pp. 292-97.

Block, Ned. "Some Concepts of Consciousness." *Philosophy of Mind: Classical and Contemporary Readings*, edited by David Chalmers, 2002, pp. 206-19.

Briñol, Pablo, et al. "Changing Attitudes on Implicit versus Explicit Measures: What Is the Difference?" *Attitudes: Insights from the New Implicit Measures*, edited by Richard E. Petty et al., Psychology Press, 2009, pp. 285-326.

Cooper, Michael. "Untitled post." *The New York Times*, 5 Feb. 1999, www.nytimes.com/1999/02/05/nyregion/officers-in-bronx-fire-41-shots-and-an-unarmed-man-is-killed.html.

"Department of Justice Announces New Department-Wide Implicit Bias Training for Personnel." *The Department of Justice*, 27 June 2016, www.justice.gov/opa/pr/department-justice-announces-new-department-wide-implicit-bias-training-personnel.

Devine, Patricia G. "Stereotypes and Prejudice: Their Automatic and Controlled Components." *Journal of Personality and Social Psychology*, vol. 56, 1989, pp. 5-18.

Devine, Patricia G., et al. "Long-Term Reduction in Implicit Race Bias: A Prejudice Habit-Breaking Intervention." *Journal of Experimental Social Psychology*, vol. 48, no. 6, 2012, pp. 1267-78.

Eagly, Alice, and Shelly Chaiken. *The Psychology of Attitudes*. Cengage Learning, 1993.

Fazio, Russell H., and Tamara Towles-Schwen. "The MODE Model of Attitude-Behavior Processes." *Dual-Process Theories in Social Psychology*, edited by Shelly Chaiken and Yaacov Trope, Guilford Press, 1999, pp. 97-116.

Gawronski, Bertram, and Galen V. Bodenhausen. "Associative and Propositional Processes in Evaluation: An Integrative Review of Implicit and Explicit Attitude Change." *Psychological Bulletin*, vol. 132, no. 5, 2006, pp. 692-731.

---. "The Associative-Propositional Avaluation Model: Theory, Evidence, and Open Questions." *Advances in Experimental Social Psychology*, vol. 44, 2011, pp. 59-127.

Gendler, Tamar Szabó. "Alief and Belief." *Journal of Philosophy*, vol. 105, no. 10, 2008, pp. 634-63.

---. "On the Epistemic Costs of Implicit Bias." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, vol. 156, no. 1, 2011, pp. 33-63.

Guidance for Federal Law Enforcement Agencies regarding the Use of Race, Ethnicity, Gender, National Origin, Religion, Sexual Orientation or Gender Identity. *U.S. Department of Justice*, www.justice.gov/sites/default/files/ag/pages/attachments/2014/12/08/use-of-race-policy.pdf. Accessed 8 Dec. 2014.

Guidance regarding the Use of Race by Federal Law Enforcement Agencies. *U.S. Department of Justice*, 15 Dec. 2010, www.justice.gov/sites/default/files/crt/legacy/2010/12/15/guidance_on_race.pdf.

Hogg, Michael, and Graham Vaughan. *Social Psychology*. Pearson Education, 2013.

Holroyd, Jules. "Responsibility for Implicit Bias." *Journal of Social Psychology*, vol. 43, no. 3, 2012, pp. 274-306.

Hugenberg, Kurt, and Galen V. Bodenhausen. "Facing Prejudice: Implicit Prejudice and the Perception of Facial Threat." *Psychological Science*, vol. 14, no. 6, 2003, pp. 640-43.

Levy, Neil. "Expressing Who We Are: Moral Responsibility and Awareness of Our Reasons for Action." *Academia*, 2011.

---. "Implicit Bias and Moral Responsibility: Probing the Data." *Philosophy and Phenomenological Research*, vol. 93, no. 3, 2016, pp. 1-24. *Wiley Online Library*, onlinelibrary.wiley.com.

---. "Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements." *Noûs*, vol. 49, no. 4, 2015, pp. 800-23.

Lindquist, Kristen A., et al. "The Role of Language in Emotion: Predictions from Psychological Constructionism." *Frontiers in Psychology*, vol. 6, 2015.

Mandelbaum, Eric. "Against Alief." *Philosophical Studies*, vol. 165, no. 1, 2013, pp. 197-211.

---. "Attitude, Inference, Association: On the Propositional Structure of Implicit Bias." *Noûs*, vol. 50, no. 3, 2016, pp. 629-58.

Rozin, Paul, et al. "The Sympathetic Magical Law of Similarity, Nominal Realism and Neglect of Negatives in Response to Negative Labels." *Psychological Science*, vol. 1, no. 6, 1990, pp. 383-84.

---. "Operation of the Laws of Sympathetic Magic in Disgust and Other Domains." *Journal of Personality and Social Psychology*, vol. 50, no. 4, 1986, pp. 703-12.

Ryle, Gilbert. *The Concept of Mind*. Routledge, 2009.

Schwitzgebel, Eric. "Self-Ignorance." *Consciousness and the Self: New Essays*, edited by JeeLoo Liu and John Perry, Cambridge UP, 2012, pp. 184-97.

Sher, George. *Who Knew? Responsibility without Awareness*. Oxford UP, 2009.

Shoichet, Catherine E. "Keith Lamont Scott: What We Know about Man Shot by Charlotte Police." *CNN*, 22 Sept. 2016, www.cnn.com/2016/09/22/us/keith-lamont-scott.

Siegel, Susanna. "Cognitive Penetrability and Perceptual Justification." *Noûs*, vol. 46, no. 2, 2012.

---. *The Contents of Visual Experience*. Oxford UP, 2010.

Stewart, Brandon D., and Keith Payne. "Bringing Automatic Stereotyping under Control: Implementation Intentions as Efficient Means of Thought Control." *Personality and Social Psychology Bulletin*, vol. 34, 2008, pp. 1332-45.

Strawson, Peter F. "Freedom and Resentment." *Proceedings of the British Academy*, vol. 48, 1962, pp. 1-25.