THE LANDSCAPE OF BIAS:

MANIFESTATIONS ACROSS PERCEPTION, ALGORITHMS, AND TESTIMONIAL EXCHANGES

by

DE YANG

(Under the Direction of RENE JAGNOW)

ABSTRACT

Bias is deeply embedded in human perception, decision-making, and social interactions, shaping the ways we interpret the world, develop artificial intelligence systems, and assess the credibility of others. This dissertation examines three interrelated forms of bias—perceptual bias, algorithmic bias, and bias in testimonial exchanges—to provide a comprehensive understanding of how bias manifests at both individual and systemic levels. It first argues that bias is not solely a product of top-down cognitive influences but can emerge from the visual system's unconscious assumptions, shaping perception before conscious thought occurs. By drawing on research in vision science, this dissertation highlights how perceptual bias can operate independently of cognitive states, challenging conventional distinctions between perception and belief. It then explores the parallels between implicit bias in human cognition and algorithmic bias in artificial intelligence, critically evaluating current models that fail to account for the fluctuating nature of human judgment. While both forms of bias arise from pattern-based learning, this study argues that human implicit bias is more variable than algorithmic bias, resisting simplistic, rule-based interventions. By integrating insights from psychology, machine learning, and philosophy, it proposes a new framework for understanding how biases are encoded, reinforced, and mitigated

in both human and computational decision-making. Finally, it examines testimonial injustice as a systemic issue, wherein individuals from marginalized groups are unfairly deemed less credible based on identity prejudice. While some philosophers suggest that cultivating individual virtues—such as open-mindedness and credibility assessment skills—can counteract this form of epistemic injustice, this dissertation argues that such efforts are insufficient. Because testimonial injustice is embedded in broader social structures, meaningful solutions must focus on institutional reforms, including changes in legal, educational, and professional settings that shape credibility judgments. Through a synthesis of philosophical analysis and empirical research, this dissertation contributes to contemporary discussions on the nature of bias, its ethical and epistemic consequences, and the most effective strategies for mitigation. It shows the necessity of interdisciplinary approaches that address both the cognitive mechanisms underlying bias and the structural factors that perpetuate social injustice.

INDEX WORDS: Social bias, perceptual bias, implicit bias, algorithmic bias, testimonial injustice, attentional direction, cognitive penetration, visual assumption

THE LANDSCAPE OF BIAS:

MANIFESTATIONS ACROSS PERCEPTION, ALGORITHMS, AND TESTIMONIAL EXCHANGES

by

DE YANG

BA, China University of Political Science and Law, China, 2012

MA, Renmin University of China, China, 2015

MA, Georgia State University, 2018

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2025

THE LANDSCAPE OF BIAS:

MANIFESTATIONS ACROSS PERCEPTION, ALGORITHMS, AND TESTIMONIAL EXCHANGES

by

DE YANG

Major Professor:	René Jagnow
Committee:	Aaron Meskin
	Sarah Wright

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2025

ACKNOWLEDGEMENTS

I am deeply grateful to my advisors for their insightful guidance and support throughout my time in the Philosophy PhD program at the University of Georgia, as well as for their invaluable assistance with my dissertation. I also extend my appreciation to my fellow graduate students, conference participants, anonymous referees, and everyone who offered feedback on my work as it developed—our discussions have greatly enriched this project. Lastly, my heartfelt thanks go to my friends and family for their unwavering support, which not only carried me through the PhD journey but also brought light and joy to my life.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iv

LIST OF FIGURES vii

CHAPTER

1 Introduction and Literature Review 1

2 Visual Assumption and Perceptual Social Bias 6

Abstract 7

Introduction 8

Perceptual Social Bias and Top-down Influence 9

Visual Assumptions Without Top-down Influence 17

Visual Assumptions in Perceptual Social Bias 25

Concluding Remarks 34

3 Implicit Bias, Algorithmic Bias, and the Variability of Human Judgments 37

Abstract 38

Introduction 39

Two Notions of Bias 41

Johnson and the kNN Model of Implicit Bias 44

The Variability of Implicit Bias 48

The Variability of Human Judgments: A Bias Attenuator or Amplifier? 57

Concluding Remarks: the Cost of Reduced Bias 63

4 The Limits of Individual Virtue: Testimonial Injustice as Anomalous Aggregate Pattern 67

Abstract 68

Introduction 69

Testimonial Justice as a Solution to Testimonial Injustice 70

Testimonial Injustice as a Systemic Pattern 74

Testimonial Justice as an Institutional Virtue 89

5 Conclusions 95

REFERENCES 98

LIST OF FIGURES

Page

Figure 2.1: Concave footprint illusion from Morgenstern et al. (2011) 17

Figure 2.2: Müller-Lyer Illusion 17

Figure 2.3: Carpentered World Explanation for the Müller-Lyer Illusion 18

Figure 3.1: A Simple Example of kNN where k=5 45

CHAPTER 1

Introduction and Literature Review

Recently, social bias has been a central concern in philosophy, psychology, and the social sciences, with its implications reaching into nearly every aspect of human interaction and decision-making. From the subtle influences of implicit bias on our judgments to the systemic inequalities resulting from algorithmic decision-making, the study of social bias has revealed the ways in which our perceptions, attitudes, and actions are shaped by prejudices. This dissertation seeks to contribute to this rich and evolving field by exploring three distinct yet interconnected dimensions of social bias: perceptual bias, implicit bias, and bias in testimonial exchanges. Each of these dimensions highlights a unique mechanism through which bias operates. Together, they offer a comprehensive understanding of how social bias manifests and persists in both individuals and social systems.

The study of social bias is not merely an academic exercise; it has profound implications for social justice, equity, and the functioning of democratic societies. For instance, implicit bias—unconscious attitudes or stereotypes that affect our understanding, actions, and decisions—has been shown to influence outcomes in critical areas such as criminal justice, healthcare, education, and employment (Greenwald & Banaji, 1995; Payne et al., 2005). Similarly, algorithmic bias, which arises when machine learning systems reproduce or exacerbate existing social inequalities, has become a pressing concern as artificial intelligence increasingly shapes decision-making in areas like hiring, lending, and policing (O'Neil, 2016; Chouldechova, 2017). Testimonial injustice, a form of epistemic injustice where a speaker's credibility is unjustly deflated due to identity

prejudice, further highlights the ethical and epistemic harms of bias, particularly for marginalized groups (Fricker, 2007). By examining these three dimensions of bias, this dissertation aims to deepen our understanding of how bias operates across different domains and to identify effective strategies for counteracting its harmful effects.

Over the past few decades, research on social bias has expanded significantly, with scholars from various disciplines investigating its causes, consequences, and potential remedies. In philosophy, the study of bias has been informed by work in epistemology, ethics, and social philosophy. Scholars like Miranda Fricker (2007) and Tamar Gendler (2008) have explored the epistemic and ethical dimensions of bias. In psychology, research on implicit bias has drawn on cognitive and social psychology to reveal the unconscious mental processes that shape our attitudes and behaviors (Greenwald & Banaji, 1995; Payne et al., 2005). In computer science and data science, the study of algorithmic bias has focused on understanding how biases in data and algorithms can lead to unfair or discriminatory outcomes (O'Neil, 2016; Chouldechova, 2017). This dissertation is situated within this broader intellectual landscape, drawing on insights from these disciplines to address pressing questions about the nature of bias and its impact on human cognition and behavior.

The dissertation is structured in the manuscript style, consisting of three standalone papers united by the overarching theme of social bias. Each paper examines a specific way in which bias is manifested in our social life, exploring its underlying mechanisms, its implications for social justice, and the challenges it poses for traditional interventions. The first paper investigates the role of visual perception in the formation of social bias, arguing that bias can arise not only from cognitive states but also from the visual system itself. The second paper explores the relationship between implicit bias in human cognition and algorithmic bias in artificial intelligence systems,

highlighting the shared mechanisms underlying these two forms of bias. The third paper examines testimonial injustice as a systemic pattern, advocating for institutional solutions to address this form of epistemic injustice. Together, these papers provide a multifaceted exploration of social bias.

The first paper, presented in Chapter 2, explores how bias can arise at the level of visual perception itself. Most discussions of social bias focus on cognitive processes, assuming that bias comes from beliefs or attitudes. However, this chapter shifts attention to perception, arguing that bias can emerge even before conscious thought plays a role. Research in vision science suggests that our visual system makes unconscious assumptions that can distort how we see the world. For example, classic studies like the Müller-Lyer illusion and the carpentered-world hypothesis show that people perceive the same physical stimuli differently depending on prior visual experiences. Extending this idea to social perception, studies show that people misperceive objects and individuals based on race and other social categories—such as the well-documented finding that people are more likely to mistake a harmless object for a weapon when it is associated with a Black individual. If bias is operating at the level of perception itself, then interventions aimed at changing people's explicit beliefs will have limited success. Instead, institutional solutions—such as standardized training for law enforcement and medical professionals or adjustments to visual decision-making protocols—could help mitigate the impact of perceptual bias in high-stakes situations.

The second paper, presented in Chapter 3, examines the connections between implicit bias in human cognition and algorithmic bias in artificial intelligence. These two types of bias are often studied separately, but Johnson (2020a; 2020b)'s k-Nearest Neighbor (kNN) model of implicit bias shows that they share underlying mechanisms. Both function by picking up on patterns from

past experiences—whether through lived experiences in the case of human cognition or through training data in the case of AI. However, this chapter of my dissertation critiques the kNN model for failing to account for the fluctuating nature of bias and must appeal to extra-algorithmic mechanisms to do so. The chapter implies that rather than treating bias as a purely technical issue that can be solved with better data, we need to develop institutional safeguards that recognize bias as a shifting and context-dependent phenomenon.

The third paper, presented in Chapter 4, focuses on testimonial injustice—a form of bias that affects how people's credibility is assessed. Some theorists, like Miranda Fricker, argue that testimonial injustice can be countered through individual virtues, such as becoming more reflective and open-minded in how we assess credibility. While cultivating these virtues is important, this chapter argues that they are not enough on their own. People do not always have conscious control over whom they trust or find credible, and biases often operate at an automatic level. This means that efforts to combat testimonial injustice need to go beyond individual moral development and focus on structural reforms. Institutions can play a critical role in reducing testimonial injustice by implementing policies that ensure diverse representation in decision-making processes and by establishing legal mechanisms for individuals to challenge biased credibility assessments. By shifting the focus from personal responsibility to systemic change, this approach provides a more effective way of addressing testimonial injustice and promoting epistemic fairness.

Finally, Chapter 5 offers a conclusion that ties together the themes of the dissertation and discusses the implications of this research for future studies and for practical efforts to address social bias in society.

CHAPTER 2

VISUAL ASSUMPTIONS AND PERCEPTUAL SOCIAL BIAS[1]

[1] Yang, D. Accepted by *Philosophical Psychology*. Reprinted here with permission of the publisher.

ABSTRACT:

Siegel recently distinguishes between seven possible ways in which our perceptual access to social information can be biased by flawed practice of either individuals or social structures, two of which, namely attention and cognitive penetration, imply that it is the content of perception, as opposed to that of judgments, that is biased. Both attention and cognitive penetration, however, rely on cognitive states imposing top-down influences on perceptual states. As such, perceptual bias resulting from them is to a large extent merely a derivation of cognitive bias. In this paper, I propose another way in which our perception can be biased, namely, as the result of faulty assumptions made by the visual system. Furthermore, I argue that in contrast to cognitive penetration and attentional direction, perceptual bias arising in this way is fundamentally perceptual and does not depend on inputs from one's cognitive system. This sort of perceptual bias, if it exists, would have important implications for how we conceptualize social bias and pose special challenges to traditional interventions designed to counteract bias.

I. Introduction

Our visual perception of the world is sometimes biased. This happens when there is a systematic discordance between our perception and the reality it is supposed to represent. Perceptual bias is prevalent in our daily life but is largely harmless. Yet, the vast amount of empirical research in the last two decades reveals there are instances of perceptual bias that bear social significance. The findings of this research include, but are not limited to, the following: a tool is more likely to be misrecognized as a gun when it is in the hand of a black man than when it is in the hand of a white man (Payne, Shimizu & Jacoby 2005); people perceive black men as bigger and more physically threatening than young white men (Wilson, Hugenberg & Rule, 2017); facial expressions of pain are less noticeable on black faces than on white faces (Mende-Siedlecki, Qu-Lee, Backer & Van Bavel, 2019); faces are more likely to be misidentified as male when they are angry and as female when they are happy (Korb et al., 2022). I call these findings and the related real-life phenomena "perceptual social bias." Although empirical evidence has unambiguously shown the prevalence of perceptual social bias in our daily life, it is still a matter of controversy as to how the biased perception happens. More specifically, empirical evidence is inadequate to decide, in the alleged cases of perceptual bias, whether it is one's perceptual experience or some post-perceptual process that is biased.

In a recent paper, Siegel (2020) identified seven *possible* ways in which our visual processing of social information can be biased, two of which imply that it is perceptual experience *per se*, as opposed to some post-perceptual processes related to perception, that is biased, namely, cognitive penetration and attentional direction. It is important to note, however, that both cognitive penetration and attentional direction rely on top-down feedback from the cognitive system, although in different ways. As such, even if there are cases of perceptual social bias resulting from cognitive penetration and attentional direction, they would still be dependent upon biased cognitive

states and hence are merely derivations of cognitive bias. In this paper, however, I propose that there is yet another way in which perceptual social bias may arise, in addition to cognitive penetration and attentional direction. According to this proposal, perceptual social bias can be the result of faulty assumptions made by the visual system. I will argue that in contrast to cognitive penetration and attentional direction, the kind of bias resulting from this mechanism is *fundamentally* perceptual in the sense that it does not depend on inputs from one's cognitive system, which would have important implications for how we counteract social bias.

My plan for this paper is as follows. In the second section, I introduce Siegel's suggestion that perceptual experience can be biased as the result of cognitive penetration and attentional direction and argue that they both require perception being modulated by cognitive states in a top-down manner. In section three, referring to the broadly construed carpentered-world explanation of the Müller-Lyer illusion, I explain how assumptions that govern the operations of the visual system can lead to systematically inaccurate visual representations of the world. In section four, drawing on relevant empirical research, I argue that the account introduced in section three may also explain perceptual social bias. I will end my paper by briefly discussing the implications of this mechanism and argue that it highlights an alternative way of conceptualizing social bias, in addition to the now popular conceptualization that focuses primarily on the distinction between explicit and implicit bias.

II. Perceptual Social Bias and Top-down Influence

Traditionally, social bias was defined in terms of mistaken conscious attitudes and beliefs we have about a social group. For example, people often knowingly associate black individuals with being more hostile, men with being more aggressive, women with being more appeasing, etc.

This is commonly referred to as explicit social bias. Since the 1990s, however, an increasing number of psychological studies have revealed that even when people do not admit consciously endorsing stereotypical associations, their implicit endorsement of these biases can be shown by their behaviors and decision-making. A self-professed egalitarian, for example, might unreflectively clutch her bag when her black neighbor passes by, but not when her white neighbor does the same. Similarly, an HR manager who vows to treat all job candidates equally may nonetheless show a preference for candidates with traditionally white names over traditionally ethnic-sounding names. This kind of bias is referred to as implicit social bias.

Research on both explicit and implicit social bias is abundant in psychology and philosophy. Most of the research, however, focuses on the impact social biases have on high-level functions such as thinking, decision-making, and action. It was only until the recent decade or so that researchers began to investigate the possibility that low-level functions like perception can also be biased. I briefly introduced some of the studies devoted to examining this possibility in section one. Those studies, however, are ambiguous between perceptual experience *per se* and some post-perceptual processes being biased. To appreciate the difference, consider the experiment of Payne and his colleagues (2005) on the weapon identification task, which is also the most extensively studied research paradigm in this field. The participants were briefly shown images of either a gun or a hand tool (wrench, plier, etc.) in the experiment. Then they were asked to report in less than a second whether that image contained a weapon or a tool. Before they started with the actual task, participants were primed with a face that was either white or black. It turned out that the priming had a significant effect on their reports: the participants were more likely to misidentify a hand tool as a gun when they were primed with a black face than a white face. Does the experimental result suggest, when misidentification happens, that the participants illusorily see

guns where hand tools were present? Not necessarily. Because the experimental result is also compatible with the possibility that the participants simply misjudged or misinterpreted tools as weapons. This would still involve visual perception, but only insofar as it grounds their judgments or interpretations of the visual stimuli. If this is what happened in the experiment, the participants did not really "see" guns where tools were present, except in a loose or metaphorical sense. This is a theoretically less interesting possibility but compatible with the experimental results. At this point, we have two competing accounts regarding what was going on with the participants in Payne et al.'s experiment. Call the first interpretation the *perceptual account* and the second interpretation the *cognitive account*. The experimental results are compatible with both accounts.

To adjudicate between the two accounts, Stokes and Payne (2010) conducted a follow-up experiment. The experiment adopted the same procedure as the original one, except that it allowed the participants to correct their reports after each trial with no time pressure. It turned out that the participants almost always gave the correct answer during the correction phase, even if their initial reports were false. This speaks strongly in favor of some version of the cognitive account. That is to say, the participants accurately represented the target object visually but somehow misrepresented it in their high-level judgment. After all, if they misperceived rather than simply misjudged the target object, it would be unclear on what ground they were able to correct their initial false reports.

Note, however, that Stokes and Payne's follow-up experiment provides only limited support for the cognitive account. Even if we put aside the common critique that the experimental settings lack ecological validity, there remains another problem. It is unclear to what extent the results derived from their particular experimental setting apply to other experimental settings. It is even unclear whether Stokes and Payne's results can be replicated in similar but slightly different

experimental settings. Correll and his colleagues (2015) employed an empirical setup that was similar to the one used by Stoke and Payne. But instead of using clear images, they used blurry ones. The results showed that the participants could not correct their reports even given unlimited time to respond, suggesting that they might have misperceived tools to be guns. Because otherwise, it is hard to explain why they could not correct their reports when there was no time pressure. [2]

Given the wide variety of research findings, it is unlikely that there exists a single mechanism underlying the divergent phenomena of perceptual social bias. That is to say, the cognitive account may be the best explanation for some research findings, whereas the perceptual account may be the more plausible interpretation for yet others. But in this paper, I will focus on exploring the possibility of the perceptual account, which is less discussed in the literature. Given that some paradigmatic cases of social biases are themselves higher-level cognitive states, there is no mystery that judgments or other post-perceptual processes can be biased. The perceptual account, however, opens up a theoretically more interesting possibility and has far-reaching implications for how we should understand the role of perception in our mental life.

Before I proceed, a caveat about the perceptual account needs to be made. In the discussion on visual perception, it is always risky to posit properties other than shape, size, color, etc. in one's perceptual representation. While most philosophers and psychologists agree that low-level properties like shape, size, color, etc. can be represented in perceptual experience, there is little agreement about whether high-level properties like being a gun, being a person, etc. can be thus represented. Since the perceptual account implies that one sees (illusorily) a gun in the weapon

---

Notes:

[2] A possible explanation for why the participants cannot correct their reports is that the images, due to their blurriness, are ambiguous between weapons and tools. As a result, the statistical regularity associating black faces with weapons biased their visual systems to interpret the images as weapons. This is the proposal I will develop further in sections three and four.

identification task, one might be tempted to object to the account on the grounds that it implies a controversial claim about what our perceptual experience can represent. I think, however, that need not be the case. Although the perceptual account implies that the participant 'sees' a gun, it does not imply that she represents a gun as such. It is compatible with the perceptual account that she merely represents (inaccurately) the low-level properties (shape, size, color, etc.) of the object such that it looks like a weapon when in fact it is a tool. In this case, the participant's perceptual representation is still restricted to low-level properties, and the above objection to the perceptual account is therefore unmotivated. In what follows in this paper, I will continue to speak of things like "perceivers represent guns in their visual experience" for the sake of simplicity. But in so doing, I do not claim that their visual experience represents guns as such.

It is ultimately an empirical question as to whether the perceptual or cognitive account is true in a particular scenario or experimental setting. But philosophers can nonetheless contribute to the relevant research by speculating on *possible* ways in which perceptual social bias arises and reflecting on their implications for our understanding of the relevant phenomena. Based on the broad distinction between the perceptual account and the cognitive account, Siegel (2020) distinguishes further between different varieties of the two accounts. Siegel suggests seven possible ways in which perception can be biased, two of which imply it is the content of perceptual experience, as opposed to the content of judgment, that is biased, and hence presumably fall under the category of the perceptual account. One is through cognitive penetration, and the other is through attentional direction. If cognitive penetration is what happens to the subjects in the weapon identification task, then "the pliers look to the subjects exactly like a gun, due to the influence on the perceptual experience of a cognitive state activated by the black prime" (Siegel 2020, p. 103). Whereas if attentional direction is what happens, then "the pliers look somewhat like a gun because

the state activated by the black prime directs the subject's attention to features of the pliers that are congruent with being a gun (metallic), and away from features incongruent with being a gun (shape)" (ibid).

According to this characterization, both cognitive penetration and attentional direction involve the activation of certain cognitive states, where cognitive states include beliefs, judgments, desires, etc. So how exactly are they different from each other? To answer this question, it will be helpful to consider the connection between cognition and perception. Almost everyone agrees that perception can influence cognition. My seeing it raining outside, for example, can dispose me to believe that it is raining. But is it equally obvious that the influence can happen in the opposite direction, namely, perception being influenced by cognition? That depends on the kind of influence we are talking about. Consider Macpherson (2017) 's example. You believe that you have an exam tomorrow. The belief causes you to be stressed, which in turn causes a migraine. Then the migraine causes you to experience flashing lights in your visual field. This is apparently an example of cognition influencing perception, and examples like this are ubiquitous in everyday life. But they are philosophically rather uninteresting because there is nothing puzzling about how this kind of influence takes place.

Consider another example: Suppose you are looking at a painting and your attention is captured by a particular feature. As you turn your attention to this feature, you see more of its details. Presumably, in this case, your desire to see the painting in detail influences the content of your experience. This is an example of attentional direction. What we commonly refer to as attention covers a wide variety of distinct phenomena. For this reason, some philosophers and psychologists have even gone so far as to reject the usefulness of the very concept of attention (Anderson 2023; Hommel et al. 2019). But for the current discussion, it suffices to characterize

attention as a process of selection, whereby a perceiver directs limited cognitive resources (sometimes unconsciously) towards an object(s) or feature(s) in the external environment.[3] This is exactly what happens in the example above and cases like that. The existence of cases in which this notion of attention direction alters the content of perception is relatively uncontroversial. But these cases do not count as cognitive penetration in the strict sense.[4] Following Pylyshyn (1999), theorists engaged in the debate typically hold that a genuine case of cognitive penetration must satisfy two requirements, namely, *semantic coherence* and *directness*. Semantic coherence requires that the contents of the penetrating cognitive states stand in rational or content-preserving relation to those of the penetrated perceptual states. Whereas directness requires that the influence cognitive states exert on perceptual states is direct and is not mediated by some other states. The first example above clearly does not meet the requirement of semantic coherence: your belief of having an exam tomorrow is not semantically relevant to your seeing flashing lights. The second

[3] Thanks to an anonymous reviewer for bringing to my attention the wealth of empirical and philosophical research on attention. Two things need to be noted here. First, whereas it is common to characterize attention as a process of selection, Fazakas and Nanay (2021) have advance an alternative view suggesting that attention is actually a process of amplification rather than selection. But this need not undermine my characterization of attention because their characterization of attention is almost exclusively at the neural level, while mine is mostly at the subject level. Furthermore, even if their characterization is also reflected at the subject level, this alternative function of attention can also cause the alteration of perceptual contents, which is all I need at this point.
Second, as a concept that comprises a variety of distinct phenomena, attentional phenomena can be divided in various, cross-cutting ways in terms of their subject-level functions and underlying neural mechanism. One important distinction relevant to the current discussion is between internal and external attention. Internal attention refers to how we focus inwardly to process and generate mental interpretations of this information, whereas external attention refers to the way we attend to relevant sensory information in our environment (Chun et al. 2011). Another important distinction is covert attention and overt attention, which are typically considered subtypes of external attention. Simply put, covert attention is defined as paying attention without moving the eyes and overt attention is defined as selectively processing one location over others by moving the eyes to point at that location (Itti & Koch 2000). My characterization of attention in the main text focuses exclusively on external attention because the current discussion is on outer perception. But it may cover both overt and covert attention since both subtypes of attention can result in the alteration of perceptual content.

[4] Note that there are also philosophers suggesting that certain forms of attentional direction satisfy the requirement of directness and hence count as cases of cognitive penetration (see Macpherson, 2012; Stokes, 2018). Whether this suggestion is reasonable or not, however, is not the concern of this paper, because my aim in this paper is argue for a possible mechanism for perceptual social bias that is distinct from both attentional direction and cognitive penetration. As such, the plausibility of my proposal does not depend on a specific relationship between attention and cognitive penetration.

example does not satisfy the directness requirement. Though arguably, your desire to see the painting in more detail is semantically coherent with your actually seeing it in more detail, the influence is only indirect. That is, your desire enabled you to focus on the details of the painting through the accommodation of the lens in your eyes.

To understand what cognitive penetration really amounts to, contrast this with an example that is conceivable but may not actually exist: your desire to see the painting more clearly exerts a direct influence on your visual perception, bringing the details of the painting into view even if your gaze direction remains the same. It is cases of this sort that proponents of cognitive penetration purport to identify.

Unlike attentional direction, it is a matter of controversy whether instances of cognitive penetration really exist. But it is beyond the scope of this paper to deal with this issue, so I do not take sides on it. For the current purpose, it suffices for us to note from the discussion above that cognitive penetration and attentional direction are both top-down processes, processes in which cognitive or higher-level mental states influence perception.[5] In the case of cognitive penetration, cognitive states influence perceptual states directly. Whereas in the case of attentional direction,

[5] Even though they are not as extensively studied as top-down attention, there also exist attentional phenomena that are purely bottom-up, without being mediated by some higher cognitive state. The existence of bottom-up attention is common in everyday life and also well-established in cognitive psychology. It occurs when stimuli are salient because of their inherent properties relative to the background (Katsuki & Constantidinis 2014). For example, a perceiver's attention is immediately drawn to a green dot when it is surrounded by all red dots. Examples like this are abundant and there may also be cases of perceptual social bias that can be accounted for by this kind of attention mechanism. But this does not undermine my argument for the existence of an additional mechanism of perceptual social bias. First, the kind of attention mechanism Siegel has in mind when she talks about attention being a possible explanation for perceptual bias is probably top-down. That is because bottom-up attention, being induced directly by external stimuli and not through the mediation of some internal states of the perceivers, does not sit well with Siegel's characterization of the attention direction which states that "the state activated by the black prime directs the subject's attention to features of the pliers that are congruent with being a gun (metallic), and away from features incongruent with being a gun (shape)" (2020, 103). Second, in the case of bottom-up attention, a question arises as to what makes the stimuli salient to a perceiver. One is that the salience is innate, just like in the above example: the perceiver's visual systems are built in such a way that green and red create a strong visual contrast when put in adjacent to each other. The other possibility is that the salience of the stimuli relative to the background is learned from one's perceptual history. If either possibility is true, then perceptual social biases caused by this bottom-up attention (if they exist at all) would be compatible with the additional mechanism I propose in section 3 and 4.

the influence of cognitive states on perceptual states is only indirect, that is, mediated by other mental or physical states. But despite the difference, top-down influence is implicated in both cognitive penetration and attentional direction.

More empirical evidence is needed to determine whether cognitive penetration or attentional direction accounts for particular research findings regarding perceptual social bias. But if they do, the sort of perceptual social bias they account for would largely be derived from cognitive bias because both cognitive penetration and attentional direction require that the content of perceptual experience be influenced by biased high-level cognitive states in a top-down manner. In the next two sections, however, I will argue that there is yet another possible mechanism underlying perceptual social bias, which does not rely on top-down influence.

III. Visual Assumptions without Top-down Influence

A recurring theme in vision science has been the postulation of perceptual tendencies or assumptions to explain the relationship between visual experience and the physical reality it represents. On the one hand, visual stimuli are ambiguous in the sense that, in principle, the 2D images projected on our retina correspond to an indefinite number of possible 3D arrangements of surface colors, shapes, illumination, etc. The same retinal image may correspond to a pig, only the rear half of a pig, a wax imitation peccary, a tapir, etc. But on the other hand, the resulting perceptual state is not ambiguous (if we exclude ambiguous figures); that is, we are typically quite confident about what we visually represent and are rarely confused. How is this possible? According to a popular view in vision science, this is the result of our visual systems making prior assumptions about the world we perceive. The most well-known example is the assumption of light-from-above. When estimating 3D shapes from shading, the human visual system resolves the

ambiguity by assuming light usually shines from overhead. Figure 1 depicts a concave footprint illuminated from the bottom of the page. Yet, in the absence of visual cues indicating the actual lighting source, the visual system defaults to the assumption of light-from-above and thus interprets the figure as convex. Admittedly, this assumption functions merely as a heuristic for the functioning of the visual system and can be overridden by lighting cues that are even barely perceptible (Morgenstern, Murray & Harris, 2011). But still, light-from-above is the default assumption built into the human visual system. People with normal vision usually cannot help but view the image in this way unless stronger countervailing lighting cues are available.

Figure 2.1 Concave footprint illusion from Morgenstern et al. (2011)

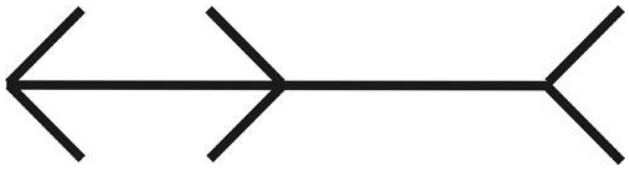

Figure 2.2 Müller-Lyer illusion

Perceptual assumptions or heuristics like light-from-above are particularly useful in explaining optical illusions (Mamassian & Landy, 1998). But how does the human visual system acquire these assumptions? Though recently challenged, it is commonly believed that the assumption of light-from-above primarily reflects innately specified mechanisms, or "natural constraint" (Scholl, 2005). But there are also visual assumptions that are mainly learned through experience. An excellent example to illustrate this is the Müller-Lyer illusion. The Müller-Lyer illusion is an optical illusion consisting of two horizontal line segments that end with arrowheads

pointing either inwards or outwards (figure 2.2), named after its discoverer, German sociologist Franz Carl Müller-Lyer in 1889. The two segments are of equal length. However, due to the presence of the arrowheads, people tend to perceive the segment with the arrowheads pointing inward to be longer than the other one.

But despite its prevalent effect on human visual perception, it has long been known that everyone is not equally susceptible to its illusory effect. Collecting data from a sample consisting of Europeans, Africans, and the Philippines, Segall and his colleagues (1963) found that the Europeans were more susceptible to the Müller-Lyer illusion than the other groups of participants.

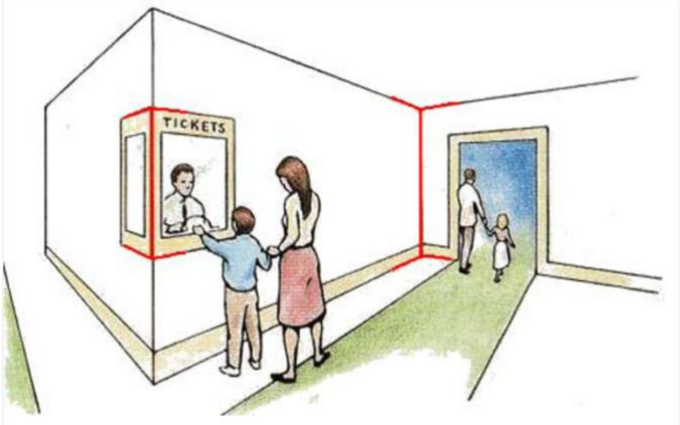

Figure 2.3 Carpentered world explanation for the Müller-Lyer illusion

To explain people's different susceptibility to the illusion, Segall put forth the influential "carpentered world" hypothesis. Living in highly carpentered environments full of artifacts constructed from straight lines and right angles, Europeans have grown accustomed to seeing corners everywhere, as a result of which their visual systems develop a habit of interpreting the arrowheads in the Müller-Lyer shapes as far and near corners: arrows pointing outwards indicate nearer corners, and inward-pointing arrows indicate corners farther away. These corners are often reliable indicators of size differences in natural environments. But they become misleading when displayed on two-dimensional planes, resulting in the Müller-Lyer illusion. Figure 2.3 depicts a

carpentered environment in which arrowheads function as corners. The highlighted segment on the left looks considerably shorter than the one on the right. This is a perfectly accurate representation of the actual size of the two segments in natural three-dimensional environments. But once placed on a two-dimensional plane, the two segments would actually be of equal length even though they look different. This illusory experience happens because the perceivers' repeated exposure to the kind of scenarios depicted in figure 2.3 has embedded certain assumptions or heuristics in their visual systems such that their brains override the retinal information which says both segments are equal in length.

Note that Segall and his colleagues' original study suffered from serious methodological flaws: since the sample consisted of subjects from three different ethnic groups, it left open the possibility that their ethnicity, rather than the ecological environment they inhabited, explained the difference in their Müller-Lyer susceptibility.[6] But multiple studies in subsequent years, ruling out possible confounds, replicated the experimental results of Segall's experiments (Deręgowski, 2013; Nijhawan, 1995; Petersen & Wheeler, 1983). These studies thus provided solid evidence for the carpentered world hypothesis.

To the extent that the Müller-Lyer illusion can at least partially be accounted for by the carpentered world hypothesis, two implications can be drawn for our current discussion. First, the carpentered world hypothesis implies that the visual assumptions making us susceptible to the illusion are, to some extent, learned from and shaped by prior visual experience rather than innately determined. According to the carpentered world hypothesis, visual percepts generated by the Müller-Lyer shape are determined empirically by the image-source relationships one has been

[6] This is not a mere logical possibility, as Jahoda (1971) 's and Berry (1971) 's studies suggested that retinal pigmentation was actually correlated with people's susceptibility to the illusion, and retinal pigmentation was correlated with ethnicity.

exposed to over accumulated experience. The reason why people experience the Müller-Lyer shape differently is that they have different perceptual histories. Those growing up in rectilinear environments were repeatedly exposed to Müller-Lyer shapes at relatively early stages in life. In their living environments, arrows pointing outwards indicate nearer corners, and inwards-pointing arrows indicate corners farther away. This, in conjunction with the more fundamental perspectival assumption that objects closer to the perceiver project larger images on her retina, has the effect that the length of the segment contained between two inward-pointing arrows is increased and that the length of the segment contained between two outward-pointing arrows is decreased.

More recent attempts to account for the Müller-Lyer illusion in the spirit of the carpentered world hypothesis invoked the notion of visual statistical learning, which describes the extraction of statistical regularities from visual environments across time or space (Howe & Purves, 2005). According to this view, one of the key functions of our visual systems is to calculate the joint or conditional probabilities of shapes co-occurring during the viewing of complex visual configurations based on the visual experience of similar stimuli in the past.[7] [8] In light of this account, what the visual system does when visually encountering a Müller-Lyer shape is to figure out, given the segments attached with inward and outward arrows, whether the two segments are more likely to be equal or different in length. Perceivers who grow up in a rectilinear carpentered environment are far more likely to see stimuli in which segments attached with arrows pointing outwards are shorter than those pointing inwards than stimuli in which the two segments are equal

[7] This sort of view come in different forms. Gregory (1997), strictly following the carpentered world hypothesis, suggested that the arrows in the Müller-Lyer shape were perspective drawings of corners. Howe and Purves (2005) take what they called the *wholly empirical* approach and extended the Müller-Lyer shape to all objects in natural scenes, rather than just corners. Nour and Nour (2015) applied Bayesian statistics to analyze the visual system's capacity to process visual information.

[8] Characterized in terms of probability distributions, some of the things people commonly refer to as visual assumptions (i.e., the assumption of light-from-above) are not really visual assumptions in the strict sense. Rather, they should be considered as visual hypotheses, which are only components of visual assumptions.

in length. This information is encoded in their visual systems, as a result of which they construct a visual representation that reflects this likelihood. Perceivers brought up in an environment that is not rectilinear, on the other hand, have seldom been exposed to this kind of visual cues. Their visual systems thus fail to learn to make inferences about the lengths of segments conditional on the presence of those arrows. Consequently, in their perception of the Müller-Lyer shape, the arrows pointing inwards or outwards, presumably resembling far and near corners in the carpentered world, have little interference with the images of the two segments projected on their retinas. This explains why they are less or even not at all susceptible to the illusion.

The second implication of the carpentered world account of the Müller-Lyer illusion is the existence of a mechanism responsible for systematically inaccurate perceptual processing that is different from both attentional direction and cognitive penetration. Even if one is convinced and believes that the two segments in the Müller-Lyer illusion are equal in length, her visual experience would still represent them as different, meaning that her visual representation is independent of her cognitive states. But as I have explained in section 2, both attentional direction and cognitive penetration presuppose top-down influences from cognitive states. Hence, the mechanism responsible for the Müller-Lyer illusion, namely, visual assumptions learned from the perceptual world, must be different from both of them.

But how could visual assumptions possibly be independent of higher-level cognitive states? The very notion of a visual assumption implies that vision involves inferences. As such, one might be tempted to think that visual assumptions must themselves be higher-level cognitive states, and hence for them to play a role in visual processing, top-down influence has to take place. However, this reasoning ignores the possibility that visual assumptions can be genuinely perceptual — that they are developed and encoded *within* the visual systems. First, vision is more

than just the passive reception of information. Rather, the visual system actively participates in the processing of visual inputs and is "smart" enough to carry out a lot of inferential processes (Kanizsa 1985; Pylyshyn 1999). This is thus compatible with the view of our minds being modular (Fodor, 1983). As Munton (2019) points out, "Much statistical learning is posited to take place *within* the visual system. Equally, non-visual information may influence the inputs and outputs to an early visual module without contravening the purported informational encapsulation of the visual system" (p. 138). Informational encapsulation constrains the information visual modules can access: they can access only visual inputs, but not higher-level cognitive states. However, informational encapsulation says nothing about what perceptual modules can do with the visual inputs they receive (Firestone & Scholl 2016). As such, the visual system can make inferences about visual inputs based on visual assumptions that are innate or learned from experience *even if* the mind is modular in the stringent Fodorian sense.[9]

Second, the reason why visual assumptions can be developed and encoded within the visual system is that visual assumptions, unlike assumptions we make use of in syllogistic inference, need not have propositional content. In fact, many visual assumptions are simply not structured in such a way that can be properly characterized in terms of propositions. Instead, they are better characterized in terms of the probability distribution of possible representations accounting for

[9] Thanks to an anonymous reviewer for pointing this out. To clarify my point, I do not hereby assert the truth of the Fodorian modularity theory. Actually, as an anonymous reviewer has pointed out, the Fodorian view has been challenged and invalidated by massive empirical evidence and theoretical considerations. I agree that the theory has largely been rendered implausible, despite some recent attempts to defend it (Firestone & Scholl 2016). However, I do not take a stance on this debate in my paper. My claim is that even if the stringent Fodorian modularity theory were true, my view would still be valid because the visual system's ability to make inferences need not depend on inputs from the cognitive system. Conversely, if the mind does not adhere to a modular framework (either Fodor's framework or Carruthers' less stringent modular framework), this would not undermine my view either. Because even if the visual system can be influenced by inputs from the cognitive system, it does not rule out the possibility that the visual system can also make inferences on its own.

visual inputs.[10] When presented with a stimulus, the image projected on the retina (i.e., visual input) gives rise to an indefinite number of possible representations (call each possible representation a visual hypothesis) to account for the stimulus. What the visual system does is calculate, based on the visual input, the likelihood of each visual hypothesis accurately representing reality and choose the one that maximizes the likelihood. In the case of the Müller-Lyer illusion, a perceiver's visual encounter with a Müller-Lyer shape automatically generates an indefinite number of visual hypotheses regarding the relative lengths of the two segments. The perceiver's visual system then calculates the likelihood of these hypotheses. But for the sake of simplicity, suppose that the visual input gives rise to only two visual hypotheses, one representing the two lines of the shape as equal in length and the other representing the inward-pointing segment as longer than the outward-pointing segment. If no arrows were attached to the segments, the perceiver would see them as equal in length. This is because the images projected on the perceiver's retinas are equal in length, and hence the probability of the former hypothesis being accurate is greater than its alternative. However, the presence of the arrows changes the probability distribution. As a result of visual statistical learning from past experience, the arrows attached to the segments add more weight to the hypothesis that the segments are different in length. Consequently, the perceiver's visual system overrides the size of the retinal images, decides that the two segments are more likely to be different in length, and then represents the relative lengths of the two segments in this way in the visual experience. Indeed, this involves a good amount of inferential processes. But insofar as these inferences, based on the probability distribution of visual hypotheses, are statistical rather than semantic, there is nothing preventing them from happening

[10] This characterization of visual assumptions is directly implicated in the Bayesian approach to visual perception (Nour and Nour, 2015; Scholl, 2005). But it is also compatible with views that are not explicitly Bayesian (see Gregory, 1997; Howe & Purves, 2005).

within the visual system. Therefore, they need not be top-down in the sense that implies higher-level cognitive states exerting influence on lower-level perceptual states.

IV. Visual Assumptions in Perceptual Social Bias

The last section shows how visual assumptions learned from prior visual experience can modulate current visual perception by referring to the example of the Müller-Lyer illusion. In this section, I will argue that this mechanism may also be applied to explaining some types of perceptual social bias. If my argument is successful, another item would be added to Siegel's list of the many ways in which one's processing of perceptual information can be biased by the flawed practice of both individuals and social structures.

The mechanism, as characterized in the last section, suggests that the more often perceivers are exposed to a certain kind of stimulus in the past, the more likely, when a specific visual input is given, they are to "see" that stimulus when the input is different but sufficiently similar to it. In the case of the Müller-Lyer illusion, their frequent exposure to corner cues in carpentered environments leads to the formation of an assumption that disposes their visual systems to overestimate the lengths of segments with inward arrows and to underestimate the length of segments with outward arrows. In consequence, when corner cues (i.e., the arrows) are available, the perceiver "sees" the two segments in a way that does not represent reality.

Now we are ready to apply the analysis to perceptual social bias. In this section, I will take the study of weapon identification by Correll and his colleagues (2015) as my paradigmatic example, though mutatis mutandis, the analysis can be applied to other phenomena of perceptual social bias. But before I proceed, a caveat needs to be made concerning the ontological status of the concept of race. While the traditional naturalist conception of race, which treated race as

reflecting biological foundations that separate humanity into discrete groups, has long been repudiated, there is still a debate concerning the ontological status of race among philosophers. Mallon (2006) famously distinguished between three main metaphysical positions about race in the relevant philosophical discussions. Racial population naturalism argues that although humanity cannot be separated into static and discrete groups as conceived of by traditional naturalism, races are still real in the sense that they are biologically grounded populations. In contrast, racial constructionism denies race has any biological foundations and instead holds that it is constructed through racialized social practice. Nonetheless, constructionism maintains the reality of race (though as a social, rather than biological reality), which puts it in opposition to the third metaphysical position about race, racial skepticism, which holds that race is not real and human races do not exist at all. Note that the debate on the ontological status of race is still ongoing and little consensus has been reached among philosophers.[11] So in this paper, I will refrain from taking a stance on the debate and am not committed to any specific metaphysical position about race. Despite this, I will use terms such as "black" and "white" individuals for illustrative purposes only.[12] But I remain neutral as to whether these terms designate biologically grounded social groups, racialized social groups, or something else.

Recall that in their experiments, the participants were presented with images in which a man (either black or white) held an object in his hand, and their task was to identify whether the

[11] Mallon suggests that the disagreements between the three positions are just illusory because they "share a broad base of agreement regarding the metaphysical facts surrounding racial or racialized phenomena that suggests their views are complementary parts of a complex view incorporating biological, social, and psychological facts" (2006, p.527). But see Hochman (2017) for an objection against Mallon's suggestion.

[12] I am grateful to an anonymous reviewer for pushing me to clarify these points. Also, it might seem that my use of these terms cannot really be neutral with respect to the three metaphysical positions about race because it cannot be compatible with racial skepticism, which denies the existence of race altogether. But note that often, racial skepticism is accompanied by an alternative social categorization to account for the *apparent* reality of race (e.g. Blum 2010). As such, it can still be meaningful to talk about black and white individuals even according to racial skepticism, though these terms are just proxies for some alternative social categories.

object was a gun or a tool by pressing buttons on keyboards. The image was blurry such that although the race of the man was easily recognizable, the object was not. It turned out that the participants were more likely to mistake tools for guns when they were in the hands of black men than when they were in the hands of white men. According to the explanation proposed above, this is what happened in the experiment:

> When presented with an image of a man (black or white) holding a tool, the visual input gives rise to an indefinite number of visual hypotheses (that object being a wrench, a hammer, a gun, etc.). The probabilities of each of them accurately representing reality are then calculated by the perceiver's visual system. Due to the perceiver's perceptual history of being more frequently exposed to black men with guns than white men with guns, more weight is given to the gun representation when the man in the image is black. As a result, the perceiver's visual system decides, compared to when the face is white, that it is more likely that the object in the hand of the person is a gun in the presence of a black face. Therefore, other things being equal, the perceiver is more likely to misrepresent the object as a gun when the man is black than when he is white.

This story is well in line with the accounts of perceptual social bias suggested by Munton (2019) and Neemeh (2020). I refer to this account as the simple probabilistic account of the learning of visual assumptions.[13] However, an apparent problem arises for this simple account: it is improbable that the participants' encounters with stereotypical associations (e.g., black people with guns) in real life are frequent enough to embed assumptions in their visual systems, especially considering that most samples in the experiments consist of young college students (e.g., Correll, Wittenbrinks, Crawford & Sadler, 2015, study 2; Payne, Shimizu & Jacoby, 2005). They probably have little chance to interact with black

[13] Note that Munton (2019) does not intend to propose a *realistic* explanation. Rather, her aim is to show that in a highly racialized society, even if our visual systems function ideally such that they perform perfect statistical learning, they might still result in biased perception and ground prejudiced beliefs.

people holding guns and may not even have seen guns at all in real life. Consequently, as Roberts says, "most of the students would have been unlikely to have encoded a statistical regularity between the two based on their perception of real-world objects" (2021, 4551). The above account does not seem adequate even in experiments where the sample consists of a broader population (e.g., Correll, Wittenbrinks, Crawford & Sadler, 2015, study 1). Admittedly, data reveal that black people account for a higher percentage of the offenders in violent crimes than other racial groups combined. For example, in 2019, 51.2% of all homicide offenders were black (JJDP n.d.). However, I doubt that the participants in those experiments have witnessed many violent crimes committed by black people with guns in real life. But if my doubt is valid, how is it possible for their visual systems to develop an assumption that associates black people with guns?

One possibility is suggested by Roberts (2021), who claims that images might have played an essential role in the process of visual statistical learning. Roberts' primary focus is on how images contribute to the objectification of women in perception, but her suggestion can easily be extended to other forms of perceptual social bias, including people's perceptual tendency to associate black people with guns. Most ordinary people are unlikely to have witnessed a lot of violent crimes committed by black people with guns in real life. Nevertheless, we are frequently exposed to media coverage of violent crimes. In the coverage, black people are disproportionately depicted as either perpetrators or victims of violent crimes, both of which tend to link black people with guns. For example, in a recent report, researchers found that mugshots were used in coverage of 45% of cases involving black people accused of crimes, whereas the number is only 8% when it comes to cases involving white defendants (Equal Justice Initiative, 2021). This biased depiction of black people is also reflected in Hollywood movies. By analyzing a database of 160,000 acting

credits from 26,000 major US movie releases, Zachary Crockett, a former Vox staff, found that "gang member" and "thug" roles were predominantly played by black actors — 62% of all actors who were credited as "gang members" and 66% who were credited as "thugs" are black (Crockett, 2016). These are just a few examples among many that show how the association between black people and guns is established through imagistic representations. Given that these images are prevalent in daily life and likely account for most ordinary people's perceptual encounters with black people carrying guns, it seems plausible that they are responsible for the development of visual assumptions that associate black people with guns in perception, as Roberts suggests.

Roberts' suggestion overcomes the inadequacy of the simple probabilistic account of visual statistical learning by relocating the source of visual statistical learning from real-life experience to image perception. But in addition to Roberts' suggestion, I argue that another revision, focusing on the process instead of the source of visual statistical learning, can also be made to the simple probabilistic account to accommodate the fact that people do not frequently encounter visual stimuli that associate black people with guns in real life. I now turn to this revision.

Recall that according to the simple probabilistic account of visual statistical learning, one's visual system calculates the probability of a visual hypothesis accurately representing reality on the basis of one's perceptual history, namely, one's visually encountering scenes in accordance with that hypothesis in the past. It assumes that one's visual system is statistically rational in the sense that the likelihood of a visual hypothesis is proportional to the frequency of one's exposure to image-source relations conforming to that hypothesis. As such, each visual encounter with people holding guns, whether black or white, would be equally weighted and encoded in a perceiver's visual system. This assumption, however, has been called into question by some recent psychological and neurophysiological research.

The weights assigned to visual encounters with stimuli can be influenced by various factors. Jones and his colleagues (2006) reported that primacy and recency effects impact perceptual processing: in some tasks, visual processing is biased toward earlier stimuli, while in other tasks, it is biased toward recent stimuli. Geisler and Kersten (2002) suggested that prior knowledge regarding the reliability of the information sources plays a role in adjusting the relative weights assigned to them. Hence, my perception of something in a normal situation is likely to be assigned more weight than something I see in a desert because the latter might just result from a mirage. Another factor I want to highlight that may influence the distribution of weights to visual encounters is the affective valence of the stimuli, due to its relevance to the current discussion. Drawing on two lines of empirical research, I argue that stimuli like a black man holding a gun are often processed more efficiently by our visual systems than, so to speak, a white man with a gun. As a result, our visual encounter with the former may be weighted more than the latter, and hence they do not contribute equally to the probability calculation of the visual system.

The first line of research concerns people's affective ratings of black people, especially black males. In a racialized society like the U.S., it is hardly surprising that negative affects often accompany ordinary people's perception of black men. Empirical findings from the last two decades also confirmed this. In Amodio and Hamilton (2012) 's study, participants were told by experimenters that they would have a discussion on social issues with either a white or a black partner. Information concerning the race of the discussion partner was tacitly conveyed to them by disclosing the partner's name. Although the discussion never actually happened, merely informing participants of the possible discussion was enough to manipulate their anxiety: participants who believed they would discuss with a black partner showed significantly greater anxiety than those with a white partner. In one of Shapiro et al. (2009)'s experiments, participants viewed an online

slide show in which pairs of male faces appeared briefly on a screen in succession and were then asked to rate each in terms of how threatening the person came across. Due to sensory adaptation, when an angry white male face was paired with a white male face wearing a neutral expression, the neutral face was perceived as less threatening. This effect, however, did not take place when the two faces were black. In Trawalter et al. (2009)'s study, participants were briefly presented (30 ms) with the faces of black men. In the meanwhile, their pattern of attention was recorded using a dot-probe detection paradigm. The researchers found that their patterns of selective attention were very much like those when exposed to pictures of evolved threats such as spiders and snakes. The three studies are just a few of the vast literature on affective responses generated by the perception of black men. The research is varied. But it unanimously suggests that negative emotions often accompany people's perception of black men and black faces alone are enough to trigger feelings of fear in the perceivers.

The second line of research concerns the well-known phenomenon that our visual systems process emotionally significant stimuli, especially threatening or fear-related stimuli, more efficiently than neutral stimuli. Whereas the former category includes stimuli such as spiders, pictures of mutilations, angry faces, or words like death and murder, the latter includes flowers, clownfish, neutral faces, or words like table, lamp, etc. Empirical research vindicating this phenomenon is enormous. Consider Soares and Esteves (2013) 's study. Participants were presented with displays for brief durations under conditions of high perceptual load (each of the displays contains 4-8 different objects). Their task was to detect specific objects from the displays. The results showed that the participants were faster at detecting fearful than neutral objects. Furthermore, the results showed that their detection was also more accurate when asked to identify the former compared to the latter. Also, consider studies on binocular rivalry. Binocular rivalry is

a visual phenomenon in which two different images are presented simultaneously to each eye. Rather than perceiving a stable, single amalgam of the two stimuli, the perception of someone with normal binocular vision would alternate between them as they compete for perceptual dominance. This competition, however, becomes one-sided dominance when the two stimuli do not have the same affective valence. Alpers et al. (2005), for example, found that when presented with a neutral and a threatening image, the latter would usually predominate over the former in this rivalry, indicating that our visual systems give priority to the processing of fear-related stimuli.

The above-mentioned are just a few of the studies on the efficient processing of emotionally valenced, especially negatively valenced visual information. There is still an ongoing debate concerning the underlying mechanism of this phenomenon. Some hold that affectively-valenced visual stimuli are processed more efficiently because they receive additional neural representation, while others suggest that it is the result of more attentional resources being allocated to emotionally charged visual stimuli.[14] But details of the debate need not bother us here. For the current purpose, it suffices for us to say that both accounts point in the same direction: affect-laden information, especially negatively valenced information, is treated by our visual systems differently than neutral stimuli. Consequently, it is encoded in our visual systems such that it is weighted more than neutral information. This corollary makes evolutionary sense because

[14] Advocates of the first account suggest the existence of a visual pathway that is devoted specifically to the processing of affect-laden visual stimuli. Several lines of research indicate that the amygdala is involved in the encoding of affective stimuli but not in that of neutral stimuli (Amaral, Price, Pitkanen & Carmichael, 1992). Based on these lines of research, some then identified a subcortical neural pathway conveying visual information (LeDoux, 1986). Unlike the better-known ventral and dorsal pathways, this pathway is specialized in processing emotionally valenced, especially negatively valenced visual information. As a result of this additional visual pathway, the processing of affective stimuli, especially fear-related stimuli, is enhanced.
However, the existence of such an additional visual pathway is contentious: some studies suggest that this pathway is not functional in primates (Pessoa & Ungerleider, 2004). So an alternative account proposes that threatening stimuli are processed more efficiently not because they are enhanced but rather because they are prioritized in visual processing. Alpers et al.(2005)'s study described in the main text is an example supporting this account, which indicates that when attentional resources are limited and multiple stimuli compete for perceptual dominance, our visual systems would automatically allocate attentional resources to fear-related stimuli.

the ability to rapidly and accurately detect threatening stimuli confers enormous adaptive advantage. But this corollary, taken together with the first line of research on negative affects accompanying the perception of black men, leads to what Neemeh (2021) calls "bootstrap hell." It opens up the possibility that, in calculating the probability of possible representations, if perceivers hold strong sentiments against black people, their past visual encounters with black men holding guns or images of this sort would be weighted more than those with white men even if they do not in fact encounter the former more frequently than the latter.[15] As a result of this unduly weighting, even a few visual encounters with black people carrying guns (either in real life or in images) would be sufficient for their visual system to develop a robust assumption associating black people with guns.

As an anonymous reviewer has correctly pointed out, most of my considerations above are based on empirical studies with US-American participants. While this is true, I did it for legitimate reasons. This is because the paradigmatic example I have been using in this section to explain how perception can be shaped by biased visual assumptions, namely the weapon identification task, reflects a stereotypical association between black people and guns that is the most prominent in the US. It is unclear whether and to what extent this association is also held by people from different socio-cultural backgrounds. Despite this, my proposed account of perceptual bias is not limited to a specific group of people. Even if people with different socio-cultural backgrounds do not have a strong tendency to associate black people with guns to the extent that their visual

[15] One might expect that the perception of both a black man and a white man carrying a gun would trigger a feeling of threat, but only to a greater extent in the former case. But studies suggest that sometimes seeing a white man with a gun, rather than triggers a negative emotion, triggers a feeling of security (Hayes, Fortunato & Hibbing, 2021).

processing is influenced by it, it is not unreasonable to expect that they would have learned other forms of associations between a social group and some object or attribute from their environments, which leads to distorted visual processing. My proposed account may then be used to explain how this distortion occurs — the associations are learned through repeated exposure to stimuli conforming to them (maybe in conjunction with some other mechanisms adjusting the weights of the stimuli) and then developed into biased visual assumptions. Exploring cross-cultural differences, as well as similarities, in perceptual biases is both interesting and crucial for understanding their origins. But this topic has only recently begun to draw attention from researchers (Fiske 2017). Much more empirical work needs to be done before we can gain useful insights into how stereotypical associations differ across socio-cultural contexts and how they influence perceptual processing from a philosophical perspective.

V. Concluding Remarks

In the last two sections, I have advanced a *possible* explanation of how perception can be biased as the result of flawed visual assumptions. It adds another item to Siegel's list of the many ways in which our access to social information can be biased. However, my project is largely speculative, and it ultimately depends on empirical research to decide whether the suggested mechanism is responsible for particular instances of perceptual social bias. Nonetheless, my proposed mechanism, if exists, would point to a kind of social bias that is fundamentally perceptual, which has important implications for how we conceptualize social bias.

Unlike perceptual bias resulting from cognitive penetration and attentional direction, the kind of social bias resulting from faulty visual assumptions is fundamentally perceptual in the sense that it can directly modulate perceptual experience and can persist even in the absence of

corresponding cognitive states. I call this kind of social bias *exclusively perceptual bias*. Note, however, that exclusively perceptual bias does not necessarily constitute a different kind of social bias, in addition to the familiar distinction between explicit and implicit bias. Exclusively perceptual bias is surely distinct from explicit bias to the extent that the latter is conceptualized as conscious attitudes or beliefs we hold toward a certain social group. As such, they are necessarily cognitive states and are hence must be different from exclusively perceptual bias. But on the other hand, the nature of implicit bias is a more controversial issue, probably due to the fact that it can only be probed through indirect measures. Traditionally, implicit bias has been conceptualized as simple associations between concepts or images. But in the last decade, this simple associationist view has been challenged, and a plethora of alternative ways to conceptualize implicit bias have been proposed: Gendler (2008) proposes to treat implicit bias as what she calls alief; Levy (2015) claims that implicit bias is patchy endorsements; Mandelbaum (2016) suggests that implicit bias is best understood as unconscious belief; Nanay (2021) conceptualizes implicit bias in terms of mental imagery. These proposals are incompatible with each other but are all backed by some empirical evidence. This fact leads Holroyd, Scaife, and Stafford (2017) to suggest that it is impossible to have a unified account of implicit bias, and it should rather be considered as a term that covers a heterogeneous set of mental or behavioral phenomena. If Holroyd and her colleagues' suggestion is correct, then it is very likely that a subset of the phenomena researchers commonly classify as implicit bias actually coincides with what I call exclusively perceptual bias.

So the significance of exclusively perceptual bias is not that it would constitute a distinct kind of social bias, as Neemeh (2020) seems to suggest. Rather, it is important because it points to the need to conceptualize social bias in terms of an alternative distinction, namely, that between perceptual and cognitive bias. This distinction has been long been neglected given that the

traditional conceptualization of social bias relies almost exclusively on that between explicit and implicit bias. This neglect may have led researchers to ignore features that are unique to some forms of bias and results in their failure to come up with effective interventions to counteract them. If one has a false belief, the most effective way to change her belief would be to present her with evidence showing that her belief is false. But this method probably does not work to correct one's inaccurate perception — no amount of evidence would change how one perceives an object. Similarly, training or interventions working effectively to counteract cognitive bias may not work equally well against bias that is fundamentally perceptual. But to come up with effective interventions to tackle perceptual bias, the necessary first step would be to recognize its perceptual nature.

CHAPTER 3

IMPLICIT BIAS, ALGORITHMIC BIAS,

AND THE VARIABILITY OF HUMAN JUDGMENTS[16]

[16] Yang, D. To be submitted to *Philosophy & Technology*.

ABSTRACT:

Research on implicit bias in human cognition and algorithmic bias in AI systems has gained significant traction due to its implications for social justice and ethical decision-making. While implicit bias refers to unconscious attitudes influencing human judgment, algorithmic bias arises when machine learning models produce discriminatory outcomes. Despite their similarities, these two areas of research have largely developed in isolation. This paper critically evaluates Gabbrielle Johnson's proposal that implicit and algorithmic biases share underlying mechanisms, particularly through the k-nearest neighbors (kNN) model. While Johnson's model provides important insights into understanding the underlying mechanisms of implicit bias, it fails to account for the variability of human judgment. This paper argues that extra-algorithmic mechanisms influence bias in ways that can either amplify or attenuate discrimination. The analysis highlights the trade-offs between bias reduction, predictive accuracy, and fairness, ultimately highlighting the complexity of addressing bias in both human and algorithmic decision-making.

I. Introduction

Over the last decade, two prominent areas of research have gained significant traction in philosophy and relevant empirical science: implicit bias in human cognition and algorithmic bias in artificial intelligence (AI) systems. Both fields have garnered considerable attention due to their profound implications for social justice, ethical practice, and broader societal concerns. Implicit bias research focuses on the unconscious attitudes and stereotypes that influence human judgments and behaviors, often in ways that contradict one's consciously held beliefs. This area of research explores the subtle, often unintended ways in which our mental processes can lead to discriminatory outcomes, even when we explicitly denounce these prejudiced views (Greenwald & Banaji, 1995; Greenwald, McGhee & Schwartz, 1998). Algorithmic bias, on the other hand, investigates how computational models and machine learning algorithms can produce decisions or predictions that systematically disadvantage certain social groups. This area of research examines the ways in which seemingly objective technology can perpetuate or even exacerbate existing social inequalities (Chouldechova, 2017; O'Neil, 2016).

A notable and widely cited example of algorithmic bias is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) system, a risk management tool used by the U.S. courts to assess the likelihood of a defendant re-offending after release. Studies by ProPublica have demonstrated that COMPAS exhibits significant racial bias. The system is more likely to classify Black defendants as having a high risk of recidivism, while White defendants are more likely to be classified as having a low risk, even when controlling for factors such as prior criminal history and offense severity (2016). This pattern of decision-making mirrors biases observed in human parole officers, who are more likely to grant parole to White defendants than to Black defendants, reflecting implicit biases that associate Black individuals with higher

criminality (American Bar Association, 2022). This parallel suggests that implicit bias and algorithmic bias may be more closely connected than previously recognized, potentially driven by similar underlying mechanisms.

Despite the connections, research on implicit bias and algorithmic bias has largely evolved in isolation.[17] This separation is understandable, given the apparent difference between human cognition and AI algorithms. But this division has also limited the opportunity for cross-disciplinary insights that could enrich our understanding of bias in both human and algorithmic decision-making. Recent works by Gabbrielle Johnson, however, help bridge the gap between these two lines of inquiry (2020a; 2020b). In her works, Johnson draws upon the *k-nearest neighbors* (kNN) model from machine learning to develop a new account of implicit bias, suggesting shared mechanisms underlying both forms of bias.

In this paper, I hope to critically evaluate Johnson's efforts to link the emergence of bias in human decision-making to that in AI algorithms. The paper is structured as follows: In Section 2, I distinguish between two notions of bias—epistemic and discriminatory—and identify the latter as the concept most relevant to discussions of both implicit and algorithmic bias. In section 3, I introduce Johnson's kNN model of implicit bias. In Section 4, I argue that while Johnson's model offers valuable insights into the shared mechanisms underlying both human and AI decision-making, it fails to fully account for the dynamic and fluctuating nature of implicit bias. To address

[17] The study of implicit bias emerged from social psychology and cognitive science, focusing on unconscious attitudes and stereotypes that influence human decision-making. Pioneering work by Greenwald & Banaji (1995) introduced the concept of implicit social cognition, while subsequent research, such as that by Payne et al. (2017), explored how implicit biases operate at both individual and systemic levels. This body of research relies heavily on experimental methods to measure unconscious biases and their effects on behavior. In contrast, research on algorithmic bias originated in computer science and data science, examining how machine learning systems can perpetuate or exacerbate social inequalities. Foundational critiques, such as O'Neil's (2016) analysis of biased algorithms in predictive policing and Eubanks' (2018) exploration of automated decision-making in social services, highlight the societal impacts of algorithmic bias. This field employs technical analyses of datasets, algorithms, and their outputs, often focusing on fairness metrics and the ethical implications of automated systems.

this limitation, I propose that additional mechanisms beyond the kNN algorithm must be considered. In Section 5, I discuss the effects of these extra-algorithmic mechanisms—specifically, whether they attenuate or amplify biases acquired through purely algorithmic processes— and argue that their impact can vary, potentially leading to either outcome. Finally, in Section 6, I conclude this paper by discussing the implications of cases where extra-algorithmic mechanisms reduce bias, arguing that even though these mechanisms can potentially reduce biases in decision-making, this achievement involves significant trade-offs, particularly between *Individual* and *Group Fairness*.

II. Two notions of bias

At its core, bias involves a *systematic*, as opposed to *random*, departure from a norm or standard of correctness (Kelly 2022). However, depending on the nature of the norm in question, different types of bias can arise. For the current purpose, it is important to distinguish between two notions of bias often encountered in decision-making, but which are sometimes conflated.

The first notion of bias concerns predictions that systematically deviate from the norm of truth. Decision-makers, whether human or machine, primarily aim to generate factually accurate predictions that reflect the actual state of affairs in the world. Truth, therefore, is a norm decision-makers are generally expected to follow. Bias, in this sense, arises when predictions consistently deviate from the truth in a specific direction. Consider, for instance, a decision-making system used to predict the likelihood of recidivism among a group of defendants, where each defendant is classified as either likely to be a recidivist or not likely to be a recidivist.[18] The truth-norm in this

[18] In this paper, for the sake of simplicity, I will focus exclusively on binary classification problems like this one, which predict whether a given object belongs to one of two categories. However, *mutatis mutandis*, the notions of bias introduced here can be extended to multiclass classification problems and even regression problems, where outputs are continuous numeric values rather than categorical memberships.

case would be whether or not the individuals actually re-offend after release. If the system consistently overestimates the likelihood of re-offending—classifying individuals who do not re-offend as recidivists more often than the reverse—it would be exhibiting bias in this sense. I refer to this notion of bias as *epistemic bias*.

However, this is not the notion of bias typically invoked in discussions of implicit bias and algorithmic bias. Implicit bias and algorithmic bias are typically discriminatory, involving violations of the norm of fairness, which may sometimes align with the norm of truth but is typically distinct from it. The norm of fairness is important because, in addition to making factually correct predictions, we also expect decision-makers to treat individuals equally, regardless of their membership in certain social groups—particularly in high-stakes domains such as criminal justice. I call this notion of bias *discriminatory bias*. A decision-making system exhibits discriminatory bias when its predictions result in disparities between different groups based on certain statistical measures.[19] There is a variety of such statistical measures, but two commonly used ones are false-positive and false-negative rates. Discriminatory bias, therefore, arises when a decision-making system generates differing false-positive rates, false-negative rates, or both across different social groups. The COMPAS recidivism prediction system is a paradigmatic example of this form of bias. As the ProPublica study mentioned above demonstrates, COMPAS predictions are more likely to misclassify Black defendants as recidivists (false positives) and White defendants as non-recidivists (false negatives). This disparity in both false-positive and false-negative rates between racial groups constitutes a clear instance of discriminatory bias.

[19] A number of statistical criteria have been proposed to measure the fairness of algorithmic decision-making. For example, Hedden (2021) identified a total of 11 such criteria. However, research by Kleinberg et al. (2016) indicates that, except in marginal cases, it is impossible to satisfy all these fairness criteria simultaneously. It is important to note, though, that this paper does not commit to any specific criterion of fairness. Therefore, the findings of Kleinberg et al. need not pose a challenge to the arguments presented here.

Having explained the difference between the two notions of bias, it is important to note that while a decision-making system can simultaneously exhibit both epistemic and discriminatory bias, it is still crucial to distinguish between them because they do not always occur together. It is possible, and in fact, often happens, that a decision-making system is biased in one sense while not being biased in the other. For example, a recidivism prediction system might consistently misclassify defendants as recidivists, regardless of their race, making it biased in the epistemic sense (due to systematic deviations from the norm of truth) but not in the discriminatory sense (since the misclassification is evenly distributed across groups).[20]

The reverse is also possible, though with certain caveats. The COMPAS system may serve as an example. It produces higher false-positive rates for Black defendants and higher false-negative rates for White defendants. In this case, the system is surely biased in the discriminatory sense because of its unequal treatment of different racial groups. However, we may also say that the system is biased in the epistemic sense with respect to the group of Black defendants, as it systematically overestimates their likelihood of recidivism. Similarly, it also exhibits epistemic bias with respect to White defendants in that it systematically underestimates their likelihood of recidivism. Yet, when considering all defendants as a single group, it is possible that false positives and false negatives across racial groups cancel each other out. In this scenario, the system may be equally likely to overestimate as to underestimate the likelihood of recidivism when averaged over the entire population. In such a case, we could say that although the system is biased in the discriminatory sense, it is not biased in the epistemic sense with respect to the overall population of defendants.

[20] One might argue that the system is still discriminatory in a broader sense—it discriminates against defendants as a whole. However, this form of discrimination differs from the paradigmatic case of discriminatory bias discussed earlier, which systematically favors one social group over another.

The analysis above shows that there is a complex relationship between epistemic bias and discriminatory bias in decision-making. While these two forms of bias may be closely connected, they remain conceptually and practically distinct. As a result, they may have different origins in human and algorithmic decision-making, and efforts to reduce one form of bias may not necessarily address the other. In this paper, I will be primarily concerned with discriminatory bias, as both implicit bias and algorithmic bias are manifestations of this form of bias.[21]

III. Johnson and the kNN model of implicit bias

Since the 1980s, researchers have discovered that decision-making is often biased without the decision-maker realizing the bias. This is the phenomenon known as implicit bias. However, there is much controversy over how this happens. Philosophers have presented various explanations of implicit bias. For instance, Gendler (2008) characterizes implicit bias as *alief*, a mental state that automatically responds to stimuli and may operate independently of—or even in contradiction to—one's conscious beliefs. Levy (2015), on the other hand, describes implicit bias as *patchy endorsement*, a mental state that possesses some propositional structure but lacks the consistent inferential integration characteristic of full-fledged beliefs. Mandelbaum (2016) offers a different perspective, arguing that implicit bias is best understood as *unconscious belief*. Meanwhile, Nanay (2021) conceptualizes implicit bias in terms of *mental imagery*, suggesting that it arises from the way individuals mentally represent social groups or categories. In spite of their significant differences, these accounts all draw on empirical evidence from psychological research

[21] A caveat needs to be noted about the use of the term bias. In this section, I have been speaking of bias as some actually observable pattern of predictions made by a decision-making system. However, in everyday language, we often say that something is biased based solely on our understanding of its internal workings, even without observing any actual decisions or predictions. Thus, it seems plausible to conceive of bias not only as an actual outcome but also as a *disposition* of decision-making systems. In this view, we may say that a decision-making system is biased if it has the disposition to make predictions that result in disparities between different social groups, even if no such predictions have yet been made.

to explain the nature and operations of implicit bias. Johnson (2020a), however, takes a fundamentally different approach. Rather than relying on psychological data, her model of implicit bias is inspired by a widely used algorithm in machine learning known as the k-Nearest Neighbors (kNN) algorithm.

The kNN algorithm is a simple yet powerful method in machine learning and is often used for classification tasks. The algorithm works by organizing data points in a multidimensional feature space, where each data point is described by a set of numerical values that represent its features. For example, in the case of predicting whether a defendant will re-offend, these features might include factors such as race, age, socioeconomic status, prior convictions, and so on. Unlike some other machine learning algorithms, kNN does not involve a separate phase of training where it adjusts model parameters to fit the training data. Instead, it directly stores the pre-labeled training data[22]—where the class of each data point is already known—within the feature space and uses them for classification. Each dimension represents a feature, and the position of each data point is determined by its feature values. When predicting the class of a new data point, the algorithm does so by locating it in the feature space and then comparing it to its closest neighbors in the space. Notably, in making predictions, the algorithm does not rely on explicit rules about these factors but instead makes decisions based on patterns that emerge from the existing data stored in the feature space.

In the kNN algorithm, the parameter k refers to the number of nearest neighbors the algorithm considers when classifying a new data point. For example, if k=5, the algorithm will compute the distance[23] between the new data point and the data points that already exist in the

[22] Although strictly speaking, the kNN algorithm does not go through a training phase like some other machine learning algorithms do, following convention, I still refer to the data it uses for classification as "training data."
[23] In a kNN model, distance is typically calculated using Euclidean distance, though in some cases, it can also be calculated using Manhattan distance. Manhattan distance, also known as taxicab distance, measures the sum of the

feature space, and pick the five of them that are the closest to it. These five data points are called the nearest neighbors. The algorithm then decides the group membership of the new data point through a simple majority vote—if most of the closest neighbors are recidivists, then the new data

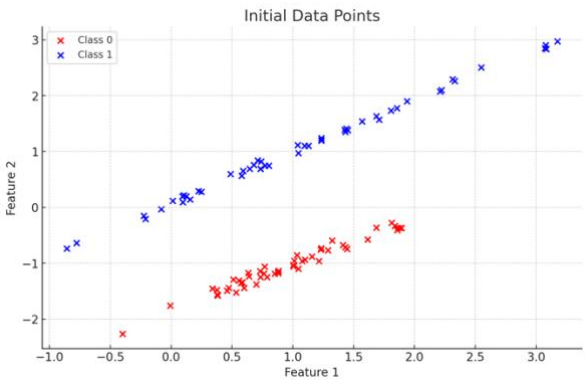

Figure 3.1(a)

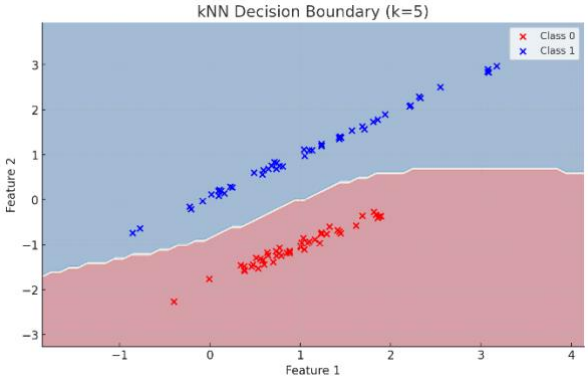

Figure 3.1(b)

Figure 3.1. A simple example of kNN where k=5

point is classified as a recidivist, and vice versa. A simple example of kNN is illustrated in Figure 3.1. Figure 3.1(a) shows the initial dataset, with two features plotted on the x and y axes. The data points are classified into two classes, represented by different colors. In Figure 3.1(b), by computing the distance to the five nearest neighbors, the algorithm creates a decision boundary that divides the feature space into two subspaces. As a result, any new data points that fall within the subspace colored blue will be classified as a member of the blue class and vice versa.

Johnson suggests that "humans can operate with a similar cognitive make-up, one that stores representations of individuals they have encountered" (2020a, p.1209). Over time, we accumulate experiences of interacting with individuals in various contexts. According to Johnson, these experiences function like the training data in a kNN model, forming a repository of past cases

absolute differences between coordinates along each dimension, resembling the path one might take when moving along a grid-like city block.

that inform future decisions. When it comes to making decisions, such as deciding whether to grant parole to a defendant, we may unconsciously evaluate the individual by comparing their characteristics to those stored in memory. If the defendant shares features with a group that has been historically associated with higher recidivism rates, our brain—operating like the kNN algorithm—might automatically categorize the individual based on the most common attributes of their nearest "neighbors" in memory.

Biases can emerge from this epistemically innocuous process. For example, if one's past experiences are dominated by cases where individuals from a certain group are depicted as having high rates of re-offending, these instances form a cluster of similar data points in memory. When encountering a new individual from this group, the brain may unconsciously categorize them as having a high risk of recidivism, even in the absence of explicit stereotypes associating the group with criminality.

This mechanism explains how bias can persist even among decision-makers who consciously—or even unconsciously—hold egalitarian beliefs. If Johnson's account is valid, a judge who firmly reject racial stereotypes may still exhibit biases shaped by the patterns present in the professional experience and social environment. This also explains why implicit biases are so persistent and resistant to change. Like the kNN algorithm that depends on training data to make classifications, a person's judgments are deeply influenced by their accumulated experiences. So correcting implicit bias would thus require changing these experiences, which is often difficult to do.

Johnson's reference to the kNN algorithm provides a novel account of how implicit bias arises in humans. A major contribution of this account is that it bridges the gap between research on implicit bias and algorithmic bias. As I mentioned in the introduction, these two fields have

traditionally developed in isolation, with minimal mutual influence. This separation is not surprising, given the apparent differences between human cognition and machine learning algorithms. But Johnson's kNN model, if valid, shows that these differences may be more superficial than fundamental, and that shared mechanisms underlie both human and algorithmic decision-making processes, leading to biases in both.

Despite this shared mechanism, however, Johnson's kNN model overlooks a crucial characteristic of human implicit bias—its variability. As I will argue in the following section, this variability in human judgment introduces complexities that the kNN model alone cannot fully account for.

IV. The variability of implicit bias

In the last two decades, research on implicit bias has skyrocketed in social psychology, shedding light on the various ways in which unconscious attitudes and stereotypes influence behavior. However, this field of research has faced significant criticism since its emergence. One of the most persistent and widely discussed critiques centers on the *variability* of implicit bias measures—specifically, the fluctuations in individuals' scores on these measures over time or across different situations. As Payne and his colleagues have observed, "the temporal stability of these biases is so low that the same person tested 1 month apart is unlikely to show similar levels of bias" (2017, p. 233). A growing body of research indicates that implicit bias is not a stable, trait-like construct but can vary considerably within the same person.

The Implicit Association Test (IAT), one of the most widely used tools for measuring implicit bias, has been shown to exhibit relatively low test-retest reliability. Test-retest reliability refers to the consistency of a measure over time—that is, whether the same individual produces

similar results when tested on multiple occasions under the same conditions. Although the acceptable test-retest reliability score for a psychological measure to be useful may vary depending on the context, a score below 0.60 is typically considered too low. However, Greenwald and Lai's meta-analyses, aggregating evidence from a large number of empirical studies, have reported a test-retest reliability for IAT measures averaging around only 0.50 (2020, p.424). Even worse, even this low reliability is often reduced in certain research situations, like when researchers have limited data collection time, as is often true of Internet data collection (Greenwald et al. 2022, p.1164). These analyses suggest that an individual's IAT scores can fluctuate significantly even from one testing session to another when no interventions or external factors have been introduced between sessions.

The problem of variability is not unique to the IAT. Other widely used measures of implicit bias, such as the Evaluative Priming Task (EPT) and the Affect Misattribution Procedure (AMP), also exhibit varying degrees of instability across testing sessions (Hu & Hancock, 2024)14. For instance, studies using the EPT have found that implicit bias scores can shift depending on the emotional state of the participant or the specific stimuli used in the task. Similarly, the AMP, which measures implicit attitudes by assessing how individuals misattribute their feelings to neutral stimuli, has been shown to produce inconsistent results when administered under different conditions.

This variability has led some critics to question the validity of implicit bias measures and even the concept of implicit bias itself. If implicit biases are not stable over time or across contexts, can they truly be considered a meaningful predictor of behavior? Some argue that the low reliability of these measures undermines their utility in both research and practical applications, such as diversity training or bias mitigation programs (Machery, 2022). The legitimacy of these

concerns, however, hinges on the assumption that implicit biases are psychological traits or trait-like constructs. This means that, unlike transient psychological states, they are enduring aspects of an individual's personality, and as such, implicit biases are expected to remain relatively stable across time and situations. However, this assumption can be called into question.

As I mentioned in Section 3, there have been competing accounts of the nature of implicit bias. Yet, not all of these accounts entail that bias is a psychological trait or trait-like feature. A notable exception is Nanay's account of implicit bias as mental imagery. Psychological research has established that the formation of mental imagery is heavily influenced by transient mental states, including emotions and desires (Holmes & Matthews, 2010; Kavanagh et al., 2009). For instance, when you try to imagine a dog, the resulting visual imagery often varies considerably depending on your current emotional state: you might visualize a friendly golden retriever with a wagging tail when you are feeling happy and relaxed, whereas you might visualize a large, menacing German shepherd with bared teeth when you are anxious or fearful. This variability indicates that mental imagery functions more as a psychological state rather than a stable trait. Following this line of reasoning, if implicit bias operates as mental imagery as Nanay suggests, then its variations across time and contexts are exactly what we should expect about implicit bias. Therefore, the fluctuations of implicit bias measurements should not be seen as problematic but rather as characteristic of the underlying psychological processes it reflects.

Unlike Nanay's mental imagery account (as well as some other accounts), however, Johnson's kNN-based model conceptualizes implicit bias as a psychological trait that remains relatively consistent across time and situations. This stems from the nature of the kNN algorithm, which classifies new data points based on their similarity to previously stored data points in a defined feature space. That is to say, the algorithm's classification decisions depend solely on the

fixed set of training data stored in memory. As long as the set of training data remains unchanged, the algorithm's predictions are expected to be stable. Under normal circumstances, there should be no significant alterations to the training data. Therefore, if implicit bias emerges through a mechanism akin to kNN, as Johnson proposes, she has to explain the variations observed in empirical studies.

Research on implicit bias reveals numerous forms of psychological and physiological manipulations can lead to variations in implicit bias measures. I classify these factors into three categories: (1) cognitive manipulations; (2) motivational manipulations; and (3) random noise. While the kNN-based model can account for variations in implicit bias resulting from cognitive manipulations, it is less clear how the model, on its own, can accommodate variations arising from the other two categories.

Cognitive manipulations influence implicit bias by manipulating participants' cognitive processes. A prime example of cognitive manipulation involves exposing the participants to examples that contradict existing stereotypes before an implicit bias test, such as presenting them with images of successful professionals from minority groups, women in leadership roles in male-dominated fields, individuals with disabilities excelling in various professions, etc. Such manipulations frequently yield positive results in reducing implicit bias. For instance, a notable study by Dasgupta & Greenwald (2001) demonstrated that presenting participants with images of admired Black individuals and disliked White individuals significantly weakened automatic pro-White and anti-Black attitudes for a long-lasting period of time. Such changes in the participants' attitudes can be easily explained by Johnson's kNN model of implicit bias. According to this model, exposure to counter-stereotypical exemplars can be seen as updating the "training dataset" one uses to classify new information. This update could be done either by adding new data points

to the existing training set or by replacing the old training data set with a new training set. In both cases, the decision boundary of the algorithm will be shifted, altering the classification of new data. Since the classification of new data points in the kNN algorithm depends entirely on the distribution of training data within the feature space, any changes to the training data will consequently change the classification of new data, resulting in a modification of implicit bias—either weakening or, in some cases, strengthening it.

In contrast to cognitive manipulations, motivational manipulations are construed broadly to encompass a wide range of psychological and physiological interventions that do not directly alter participants' cognitive processes.[24] These manipulations include inducing fatigue in participants, priming them with egalitarian norms and imposing time constraints, among others. However, the most extensively studied form of motivational manipulation is affective manipulation, which aims to influence implicit bias by altering participants' moods, feelings, or emotional states.

A notable example of how affective manipulations influence implicit bias comes from Dasgupta and colleagues' studies (Dasgupta et al., 2009). Their studies primarily examined how incidental emotions—emotions that are triggered by events or experiences unrelated to the current situation or task at hand—have an impact on implicit bias. In one of their experiments, participants were randomly placed in one of three emotional conditions (anger, disgust, or neutral). The emotional states were induced by asking the participants to recall and write about their personal experiences related to the assigned emotions and were then reinforced by images (e.g., images of

[24] Some manipulations that appear to be motivational can actually be classified as cognitive manipulations, and as such, variations of implicit bias resulting from them can be accounted for by Johnson's kNN model. A nice example of this comes from Graf and Paolini, who found out that participants' implicit bias was significantly reduced after experiencing positive emotions during intergroup contact. Although the manipulations used in their study involved emotional states, these emotional states were likely the effects of the manipulations, instead of the manipulations themselves. The core manipulation was the intergroup contact itself — a cognitive experience that became a new exemplar learned by the participants, which then updated the "training dataset" stored in their memories

cockroaches to reinforce disgust). The participants were then asked to take a modified version of IAT test involving fictitious social groups. The results showed that participants in both anger and disgust conditions exhibited an increased level of implicit bias against the outgroup compared to the neutral condition. How can Johnson's kNN-based model of implicit bias account for phenomena like this?

Unlike cognitive manipulations, transient emotional states such as anger or disgust have an impact on implicit bias without introducing any new exemplars. In such cases, it is unclear how the influence of affective manipulations, like those in Dasgupta and colleagues' study, can be interpreted as updating the training dataset in the kNN model. As mentioned earlier, in the kNN model, classifications of new data are entirely determined by the training exemplars already stored in memory. Therefore, variations of implicit bias that result from affective manipulations pose a challenge to Johnson's model. Similar difficulties arise with other motivational factors like stress, fatigue, etc. The only plausible way for Johnson to address these challenges, as I see it, is to interpret the manipulations as a mechanism that is extrinsic to the kNN model, yet capable of interfering with its operations. This interference primarily occurs in one of two ways, or through a combination of both.

First, affective manipulations, such as the induction of emotions like anger or disgust, may distort one's access to the training data stored in memory. There are many different ways in which this could be done. For instance, when a person experiences emotions like anger or disgust, emotionally congruent training data may become more salient, and incongruent ones may be suppressed, leading to a distorted representation of the training data in the feature space. Alternatively, emotional states might distort the relative positions of the training data in the feature space. For instance, a data point representing a neutral experience might be "shifted" closer to a

cluster of negative exemplars under the influence of anger, thereby changing the decision boundary and in turn influencing the classification of new data.

Second, rather than altering the training data stored in memory, affective manipulations could alternatively interfere with representations of the new data awaiting classification. Again, there are a number of different ways this could be done. For example, emotional states might distort the perceived features of a new data point, effectively shifting its position in the feature space. Under normal circumstances, the kNN model would identify the nearest neighbors based on the true positions of data points. However, when influenced by emotional states, the representation of the new data point may becomes distorted. This distortion may exaggerate or minimize certain features, relocating the new data point in the feature space. As a result, a different set of stored exemplars may emerge as the nearest neighbors, leading to classifications that differ from those made in an emotionally neutral state.

The mechanisms outlined above offer two distinct pathways through which transient emotional states can influence the classification process of the kNN model. Determining which of these two mechanisms, or what combination thereof, more accurately accounts for the variations in implicit bias caused by affective manipulations is an empirical matter, which can only be resolved through further social psychological research. However, regardless of which account proves to be empirically valid, it is hard to see how either of the two mechanisms can be conceived of as a process intrinsic to the model itself. This is because what these mechanisms do to the model is fundamentally different from cognitive manipulations, which directly update the training dataset with new exemplars. While the latter is part of the regular operations of the model, the former represents an external anomaly. As such, it seems plausible to say that Johnson's kNN model alone

is insufficient in explaining certain forms of variations in implicit bias and must rely on additional mechanisms to do so.

Similar analyses can also be extended to variations due to random noises. In human cognition, random noise refers to the variability and fluctuations in our mental processes and behaviors that occur without any apparent pattern or cause. This noise is an intrinsic part of how our brains function and can influence various aspects of our cognition: your response time when catching a ball or pressing a button often varies each time, even under identical conditions; given the same set of options (e.g., choosing a dish from a familiar menu), you might make different choices on different days without any clear reason. Not surprisingly, implicit bias is also subject to the influence of random noises. This is best evidenced by the relatively low test-retest reliability of IAT test mentioned at the beginning of this section. Greenwald and Lai's meta-analysis reports that the test-retest reliability for IAT measures averages around 0.50, indicating only moderate stability in scores across testing sessions. Since no interventions have been applied to the participants in between test sessions, the variability is most plausibly explained by random noises in human cognition.

So Johnson is once again faced with the challenge of explaining how the observed variations in implicit bias arise from her kNN model. Just as the model cannot fully account for variations caused by affective manipulations, it also struggles to explain variations resulting from random noise. This is because, similar to affective manipulations, random noise does not introduce new exemplars to update the training dataset. As a result, Johnson must once again appeal to the two mechanisms outlined earlier to address this challenge. The key difference, however, is that while affective manipulations can have a more systematic influence on bias, random noise tends

to be more transient and less predictable in its effects. Thus, the influence of random noise introduces another layer of variability that the kNN model alone is ill-equipped to handle.

The analysis above shows that Johnson's kNN model alone cannot fully explain the variability of implicit bias measures and must appeal to additional mechanisms to do so. This highlights a crucial difference between human and algorithmic decision-making.

Humans are highly susceptible to the influences of psychological or physiological factors that are epistemically irrelevant and often transient. The studies described in this section make up only a fraction of the extensive research on how such factors influence human decision-making, even in high-stakes domains. The old adage "justice is what the judge ate for breakfast" may be an exaggeration, but it contains a kernel of truth supported by empirical research. Seemingly unrelated factors, such as the time of day, significantly impact judicial rulings. In a study on Israeli judges, Danziger et al. (2011) found that the likelihood of a favorable ruling was significantly higher earlier in the day or after a break (e.g., a meal), and as time passed, judges were more likely to deny parole. This variability in human cognition has often been overlooked in philosophical discussions, dismissed as merely a contingent feature of it. However, given its pervasiveness in our everyday life and the substantial body of empirical evidence supporting it, there is reason to believe that it is not a contingent but an inherent aspect of human decision-making. This variability in human cognition is also the source of the variability observed in implicit bias.

Machine learning algorithms, on the other hand, are immune from such influences. Different types of machine learning algorithms (kNN, Deep Neural Networks, Naïve Bayesian, etc.) learn the statistical regularity from their training data through different methods and are subject to different constraints. However, a key feature they share is the invariance of their predictions. Once the training of an algorithm is completed, a given input will always produce the

same output, regardless of external circumstances. This is what we expect algorithms to do, and they meet this expectation in most cases. Deviations occur only in exceptional circumstances like software bugs or hardware malfunctions. These cases may lead to problems like training data loss or distortion, computational errors, etc., and ultimately result in inconsistent predictions. These disruptions, however, are extremely rare in algorithmic decision-making, and their influences should hence be considered extra-algorithmic, meaning that they do not originate from the algorithmic process itself.[25]

This inherent invariance of machine learning algorithms stands in sharp contrast to human decision-making processes. Johnson's proposal suggests that human decision-making can be explained, at least in part, through the lens of the kNN algorithm. However, if my analysis in this section holds, then this algorithmic component in humans is constantly subject to extra-algorithmic influences, leading to the observed variability in human cognition. The algorithmic component of human decision-making, as Johnson proposed, explains the emergence of (at least some cases of) implicit bias. But this raises the following important question: how do these extra-algorithmic influences impact the bias emerging from the algorithmic processes? Do they attenuate the bias? Or do they amplify it? These are the questions I will explore in the next section.

V. Variability of human judgments: a bias attenuator or amplifier?

Variability of human judgment has been a well-documented phenomenon in cognitive psychology for decades. However, for a long time, it has been considered merely a contingent property of human cognition rather than a fundamental characteristic. It was not until in their

[25] By claiming that the influences from psychological and physiological states like emotions, fatigue, stress are extra-algorithmic, I do not thereby imply that these influences are not algorithmic. Rather, I am only suggesting that they cannot be accommodated by standard machine learning algorithms. So if these influences can be realized algorithmically, that should be a different sort of algorithm.

recent book *Noise: A Flaw in Human Judgment* that Kahneman, Sibony, and Sunstein argued that variability is so pervasive that it should be considered inherent to human cognition and should be treated with serious attention (2021).[26] Drawing on empirical evidence from various disciplines, the authors show that the variability of human judgments is ubiquitous across domains like medicine, law, hiring, and business decisions. For example, a judge may sentence two defendants found guilty of the same crime to different sentences; a doctor may diagnose or treat two patients with the same condition differently based on individual judgment; and an interviewer might evaluate two equally qualified job candidates differently. Kahneman et al. argue that this variability is not merely a minor inconvenience but a systemic issue that can lead to unfairness, inefficiency, and potentially serious consequences in many areas of social life.

To mitigate these undesirable outcomes, Kahneman and colleagues propose several strategies to reduce the variability in human judgment. One of their key recommendations is to advocate for increased use of algorithms in decision-making processes. As I have explained in the previous section, the outputs of kNN and other machine learning algorithms are determined solely by the input data's relation to training data within a defined feature space, and are not subject to extra-algorithmic influences from fatigue, stress, emotions, random noise, etc. Consequently, once the training of an algorithm is completed, identical inputs should always produce the same outputs. The benefit of using AI algorithms in decision-making is apparent: by leveraging this consistency, institutions and organizations can effectively reduce the variability inherent to human judgment. For instance, in the legal system, algorithms could help ensure more uniform sentencing by

[26] Note that Kahneman et al. primarily focus on the variability of judgment at the group level — that is, how different people with roughly the same level of expertise often make significantly different judgments about the same issue. However, mutatis mutandis, their analysis also applies to variability of judgment at the individual level, where the same person makes different judgments at different times or in different situations. This individual-level variability is the focus of the current paper.

consistently weighing various factors, thus addressing the issue of arbitrariness in judicial decisions highlighted by Kahneman et al; similarly, in financial services, algorithmic credit scoring could provide more consistent loan approvals, reducing the impact of individual loan officer judgments that might be influenced by factors unrelated to creditworthiness. However, Kahneman et al. also made a further claim that in eliminating variability in human judgment, algorithms also tend to reduce bias (Kahneman et al., 2021, Chapters 10, 26). Their claim is also supported by some of the comparative studies of human and algorithmic decision-making, which demonstrate that algorithms are superior to humans in producing less biased outcomes in areas like finance (Gates et al., 2002), facial recognition (Dooley et al., 2021), and jurisdiction (Kleinberg et al., 2018).

However, despite the empirical evidence from these comparative studies, we should approach Kahneman and his colleagues' claim with caution. First, it is essential to acknowledge that these studies are confined to a limited number of specific domains. While the findings within these areas may suggest that algorithms tend to produce less biased outcomes compared to human decision-makers, they do not necessarily justify broad generalizations about the overall superiority of algorithms in reducing bias across all contexts. Decision-making processes vary greatly between fields, and what holds true in one domain may not apply in another. Additionally, even within the domains studied, it remains unclear whether the reduced bias observed in algorithmic decisions is attributable to the elimination of judgment variability. This reduction in bias may stem instead from extensive and diverse datasets on which these algorithms are trained. Such datasets provide algorithms with a more solid foundation for making predictions than the limited experiences typically available to human decision-makers. Consequently, the purported advantage of

algorithms may not lie in their ability to overcome variability in human judgment but in their access to and reliance on richer data.

In the machine learning community, a popular adage states that "garbage in, garbage out." This principle suggests that the quality of an algorithm's outputs is fundamentally determined by the quality of its training data. Take, for example, a kNN algorithm (or any other type of machine learning algorithm) that is designed to predict recidivism. If the training data are biased—for example, including disproportionately more instances of Black defendants recidivating compared to White defendants—then the algorithm's predictions will inevitably reflect this bias. As a result, the algorithm is more likely to classify new Black defendants as recidivists. Assuming the algorithm functions normally and as intended, this outcome is almost unavoidable. This is how implicit bias emerges according to Johnson's model.

To model the variations of implicit bias, suppose then that some external processes interfere with the algorithms' operations. These interferences, as outlined in the previous section, could end up distorting the training data or the new data points to be classified in a number of different ways. But for the sake of illustration, I will only consider a case where the interference distorts the way in which the training data are represented, but it is not hard to see how it applies to other forms of interference.

In this case, some or all of the training data are misrepresented in a way that deviates from their original format. For instance, due to the interference, a non-recidivist who is Black, 54-year-old and has no prior convictions may now be erroneously represented as a recidivist who is Black, 35-year-old and has 3 prior convictions. Such distortions significantly alter the information the kNN algorithm relies on to make its classification. Now, imagine we apply the kNN algorithm with the distorted training data to classify a new group of defendants. Then the difference between

classifications made by the algorithm using distorted training data and those made by the algorithm trained on the original data reflects the impact of the aforementioned interference. Whether this impact results in an amplification or attenuation of the bias present in the original training data depends entirely on the nature and extent of the distortion.

If the training data are distorted in a way that reinforces existing stereotypes, this will likely lead to an amplification of bias. For instance, if one or more Black non-recidivists are misrepresented as recidivists, the algorithm may learn to associate Black individuals with recidivism more strongly than if it is trained on the original and undistorted data. This misrepresentation skews the algorithm's predictions, increasing the likelihood of Black defendants being classified as recidivists. In this case, the interference amplifies the bias present in the original training data. This amplification occurs because the distortion exaggerates the association between Black individuals and the likelihood of recidivism. In the feature space of the kNN algorithm, this distortion changes the positions of the training data, shifting the decision boundary such that Black defendants are more likely to fall on the side of recidivists. This effect can be particularly pronounced if the distorted data points are near the original decision boundary, as even small shifts can significantly impact the algorithm's classifications.

Conversely, distorted training data can also attenuate bias if the distortion turns stereotype-congruent exemplars into incongruent exemplars. For example, if one or more Black recidivists are misrepresented as non-recidivists, then the decision boundary of the algorithm will be shifted such that Black individuals are less likely to be classified as recidivists. In this case, the interference attenuates the bias present in the original training data. This is because the distortion counteracts the existing biased patterns in the original data. By changing stereotype-congruent

exemplars (Black recidivists) into stereotype-incongruent ones (Black non-recidivists), the associations learned by the algorithm are weakened, leading to less biased classifications.

The preceding analysis demonstrates that the impact of external interferences on bias—whether amplification or attenuation—is contingent upon both the nature and extent of the distortion they impose on the training data. The nature and extent of the distortion, in turn, are determined by the source of the interference, with different sources affecting the data in distinct ways. For instance, if the interferences stem from random noise inherent in human cognition, their effect would probably be unpredictable. Such noise might sometimes amplify the bias present in the original training data, while at other times attenuating it.

In contrast, variations in implicit bias stemming from emotions, and other kinds of motivational states tend to exhibit more systematic patterns. The study by Dasgupta et al., discussed earlier, exemplifies this by demonstrating how emotions like anger and disgust can amplify negative implicit bias toward outgroups. According to the account of extra-algorithmic influences outlined above, what happens, in this case, is that these emotional states distort the kNN model's training data in a way that reinforces negative stereotypes about the outgroup. This account can be easily applied to empirical findings regarding how other psychological or physiological factors influence implicit bias.

If my analysis in this section is plausible, it should be clear that in theory, the effects of the extra-algorithmic influences on decision-making can go both ways—sometimes potentially reducing biases learned from training data, while in other instances, amplifying them. This complexity underscores the need for further empirical investigations to evaluate the validity of Kahneman et al.'s suggestion that reducing variability in human judgments leads to decreased bias in decision-making. It is likely that the effectiveness of this approach varies depending on the

specific domain and situational context. For certain tasks under particular conditions, extra-algorithmic influences that increase bias may predominate, resulting in an overall amplification of bias. Conversely, in other scenarios or even the same tasks under different conditions, the reverse could be true. Determining the conditions under which each kind of influence takes place is an empirical question that requires further empirical investigations.

VI. Concluding remarks: the cost of reduced bias

In this paper, I have argued that Johnson's kNN model, all by itself, cannot explain the observed variability of implicit bias, but must appeal to some mechanisms extraneous to the purely algorithmic process. I have also discussed the potential impact of these extra-algorithmic mechanisms on decision-making. Although I have demonstrated that their effects can both attenuate and amplify biases acquired from the purely algorithmic process, in this concluding section, I would like to further explore cases where the overall effects of extra-algorithmic influences result in an attenuation of bias.

In such cases, the variability in human cognition would act as a moderating force, reducing the bias that emerges from algorithmic decision-making processes. These cases would challenge Kahneman et al.'s suggestion, indicating that human decision-makers may, in some instances at least, produce less biased outcomes than AI algorithms. As I will argue in what follows, however, such achievements come with significant trade-offs.

First, as I have explained, the variability in human decision-making, when viewed through the lens of Johnson's kNN model, arises from distortions in the algorithm's training data. These distortions prevent the data from accurately representing the world, which in turn hinders the algorithm's ability to learn real-world patterns, leading to reduced predictive accuracy. This

reduction in accuracy would occur whether the distortion reinforces or disrupts existing stereotypes. Johnson (2020b) has pointed out this trade-off between bias and predictive accuracy in her discussion of algorithmic bias, and it has also received much attention in discussions about human reasoning (Basu, 2019; Gendler, 2011). But what I wish to emphasize here is that compared to human, machine learning algorithms—due to their isolation from extra-algorithmic influences—are generally better at preserving an undistorted set of training data. This also, in part, explains why algorithms often produce more accurate predictions than humans in domains like face recognition and medical diagnosis, in spite of bias (Dooley et al., 2021; Tschandl et al., 2019).

Second, when extra-algorithmic influences attenuate implicit bias, they often do so at the cost of decision-making consistency, which is central to an intuitive principle of fairness: similar individuals should be treated in a similar manner. This notion of fairness is commonly known as *Individual Fairness*, and it is often considered to be a necessary component of a broader definition of fairness, though its sufficiency remains contested (Fleisher, 2021). Given the variability of human judgments and their susceptibility to various internal and external influences, it is not surprising that similar—or even identical—individuals may be evaluated differently by the same decision-maker contingent upon transient factors like emotions, fatigue, and so on. Just consider how often it happens that a professor grades the same paper differently on different occasions. This inconsistency in decision-making apparently violates the principle of Individual Fairness because it is treating similar individuals, or even the same individual, differently based on factors that are irrelevant to the decision at hand.

On the other hand, the concept of bias in current discussions on human implicit bias and algorithmic bias, as explained in Section 2, revolves around the unequal treatment of individuals based on group membership. Reducing or eliminating such biases is often seen as a means to

achieve fair decision-making processes. However, this notion of fairness—commonly referred to as *Group Fairness*—differs from Individual Fairness, as it focuses on ensuring that different social groups are treated equally rather than ensuring consistency in the treatment of similar individuals. In instances where the variability in human cognition results in the reduction of implicit bias, decision-makers may move toward achieving Group Fairness. However, in doing so, they may not necessarily fulfill the requirements of Individual Fairness, and in some instances, may even move further away from it. This highlights a potential tension between these two fairness principles in practice.

Thus, another dilemma arises, in addition to the trade-off between predictive accuracy and bias. While the variability of human judgment can sometimes help mitigate implicit bias, thereby aligning with the principle of Group Fairness, it compromises the principle of Individual Fairness due to the inconsistency of human judgments. On the other hand, the decisions of algorithms are consistent, allowing them to better uphold Individual Fairness. Yet, this consistency also means that algorithms lack flexibility and are consistently biased, which limits their ability to achieve Group Fairness.[27]

The analysis above shows that even in cases where the variability in human cognition, driven by extra-algorithmic influences, reduces implicit bias, it does so at the expense of predictive accuracy and consistency. This highlights the distinct strengths and weaknesses of humans and algorithms in decision-making. While algorithms, shielded from extra-algorithmic influences like emotions, stress, and fatigue, tend to make more accurate and consistent decisions than humans,

[27] It is important to clarify that my claim is not that there is an inherent trade-off between Individual and Group Fairness, making it impossible to meet both fairness criteria simultaneously. While many researchers argue for such an intrinsic conflict, this view has been challenged (Binns, 2020). My claim, rather, is that there is a trade-off in the capacity of two different decision-making systems—humans and algorithms—to satisfy these fairness requirements, regardless of whether the two principles can, in theory, be met together.

their decisions will, under certain circumstances, be more biased. Consequently, a trade-off arises when deciding whether to rely on humans or algorithms for critical decisions. The choice may be context-dependent and there is no universal solution. In contexts where Group Fairness is prioritized, human decision-makers may be better suited to produce less biased and more equitable outcomes than algorithms. Conversely, in contexts that emphasize Individual Fairness and predictive accuracy, algorithms seem to be more appropriate, although they may fall short in maintaining Group Fairness compared to human decision-makers. Identifying which contexts are better suited for human versus algorithmic decisions is a complex issue. It is beyond the scope of this paper and is left for future research. It is also left for future research to explore whether human and algorithmic decision-making can be effectively combined to mitigate the limitations of each.

CHAPTER 4

THE LIMITS OF INDIVIDUAL VIRTUE:

TESTIMONIAL INJUSTICE AS ANOMALOUS AGGREGATE PATTERN[28]

[28] Yang, D. To be submitted to *Disputatio*.

ABSTRACT:

This paper critically examines Miranda Fricker's virtue-theoretic approach to testimonial injustice, which arises when a speaker's credibility is unjustly diminished due to identity prejudice. Fricker proposes that cultivating the virtue of testimonial justice in individual hearers can correct for such biases. However, I argue that this individual-centered approach faces significant epistemological challenges. Specifically, identifying testimonial injustice in singular testimonial exchanges is exceedingly difficult, as neither the presence of credibility deficits nor their causal link to prejudice can be reliably established. Instead, testimonial injustice is best understood as an anomalous aggregate pattern—emerging across clusters of testimonial exchanges rather than in discrete instances. Given this, institutional approaches, rather than individual virtue, offer a more effective means of addressing testimonial injustice. By shifting focus from individual responsibility to structural solutions, this paper highlights the limitations of virtue-based remedies and underscores the necessity of systemic interventions to mitigate epistemic injustice.

I. Introduction

In her seminal book *Epistemic Injustice: Power & the Ethics of Knowing*, Miranda Fricker calls attention to a form of unfairness related to knowledge, understanding, and communicative practices. This form of unfairness occurs when someone is wronged specifically in their capacity as a knower. Since there is a distinctively epistemic dimension to this wrong, it is referred to as epistemic injustice. One form of epistemic injustice, termed testimonial injustice, arises when a speaker's credibility is unjustly deflated due to identity prejudice, such as biases tied to race, gender, or social status. This form of epistemic injustice is particularly pernicious because it not only undermines a speaker's role as a contributor to collective knowledge but also reinforces broader patterns of social inequality.

As a remedy to testimonial injustice, Fricker urges hearers to cultivate the virtue of *testimonial justice*. This virtue equips hearers with the sensitivity to recognize and counteract their biases, enabling them to grant speakers the credibility they deserve. By placing the responsibility on individual hearers, Fricker's approach emphasizes personal accountability in fostering just testimonial exchanges. However, this proposal has been subject to critique. An alternative approach advocates for institutional or structural remedies, arguing that testimonial injustice is a systemic issue arising from entrenched social and institutional biases, rather than merely the biases of individual hearers. Proponents of this view contend that individual virtues are insufficient to address testimonial injustice (Andersen, 2012; Sherman, 2015).

In this paper, I contribute to this debate by proposing a new argument in favor of the institutional approach to addressing testimonial injustice. I argue that Fricker's individual virtue-based approach faces insurmountable epistemological challenges in practice. Specifically, it is extremely difficult to identify testimonial injustice in particular cases of testimonial exchange.

These challenges undermine Fricker's vision of testimonial justice as a virtue that individuals can cultivate to address unfair credibility deficits. By contrast, the institutional approach, which focuses on identifying and addressing aggregate patterns of testimonial injustice at the population level, is better equipped to address this form of epistemic injustice.

The paper is structured as follows. Section 2 introduces Fricker's account of testimonial injustice and her proposal to address it through the cultivation of the virtue of testimonial justice in individual hearers. Section 3 examines the significant epistemological challenges in identifying epistemic injustice within a particular testimonial exchange. I argue that, except in rare cases where credibility deficits can be conclusively linked to a hearer's prejudice, testimonial injustice is better understood as an anomalous aggregate pattern rather than as discrete wrongs in individual testimonial exchanges. Finally, in Section 4, I argue that because testimonial injustice is best understood as an anomalous pattern emerging across clusters of testimonial exchanges, institutional solutions are better suited to address the issue.

II. Testimonial justice as a solution to testimonial injustice

Testimonial injustice, according to Miranda Fricker, is a form of epistemic injustice that occurs when someone is discredited as a knower due to identity prejudice. Specifically, it happens when a speaker's testimony is given less credibility than it deserves, not because of the content of the testimony or the reasoning behind it, but because of the hearer's bias against the speaker's identity. This is what distinguishes testimonial injustice from other forms of credibility deficits.[29]

[29] In her seminal book, Fricker identifies two primary forms of epistemic injustice: testimonial injustice and hermeneutical injustice. A key distinction between the two is that, in cases of hermeneutical injustice, it is often impossible to pinpoint specific hearers as perpetrators. Nevertheless, credibility deficits stemming from identity prejudice still play a role in such cases. Given this overlap, much of my analysis of testimonial injustice in this paper may also extend to hermeneutical injustice.

While testimonial injustice involves credibility deficits, not all such deficits imply testimonial injustice. According to Fricker, what distinguishes testimonial injustice from other forms of credibility deficits is the role of identity prejudice. As she explains, "the speaker sustains such a testimonial injustice if and only if she receives a credibility deficit owing to identity prejudice in the hearer" (2007, p. 28). For Fricker, it is this prejudice—often related to race, gender, class, or other social identities—that turns a mundane credibility deficit into a case of testimonial injustice.

Imagine a female engineer presenting a well-researched, innovative solution to a hard technical problem at a conference. Suppose her proposal is dismissed by a male colleague without any reasonable justification. While this situation clearly involves a credibility deficit, it does not necessarily qualify as testimonial injustice. There could be a number of non-prejudicial reasons for the dismissal. For instance, the male colleague might simply be in a bad mood and instinctively distrust any idea that challenges his beliefs, or he might have missed key parts of the presentation due to inattentiveness. In these scenarios, the female engineer receives less credibility than she deserves. But because the credibility deficit arises from contextual or accidental factors, it does not constitute a case of testimonial injustice.

However, if the male colleague's dismissal of the female engineer's proposal is driven by a stereotype—such as the belief that women are less competent in technical fields like engineering—this would be a clear case of testimonial injustice. The key issue here is that his judgment is the result of his prejudice—his biased view about women's competence leads him to devalue her testimony. This is what Fricker calls the "identity-prejudicial credibility deficit," where the speaker is unfairly discredited because of bias tied to their social identity.

The distinction of identity-prejudicial credibility deficit from other forms of credibility deficit is crucial because it singles out cases of credibility deficits that are not only epistemically wrong in that the hearer fails to conform to evidential norms in making credibility judgments, but are also ethically wrong in that they harm the speaker's epistemic standing. These cases constitute a form of injustice that is uniquely epistemic, supplementing the traditional notion of injustice that focuses on the unfair distribution of material goods.

Since testimonial injustice results from identity prejudice, Fricker claims that addressing it requires us to confront and counteract the identity prejudice that drives these credibility deficits. For that purpose, she proposes cultivating "a corrective anti-prejudicial virtue that is distinctively reflexive in structure" (p. 91). As Fricker sees it, this virtue operates reflexively: a hearer who possesses it would be critically aware of how their own biases might affect their judgment of a speaker's credibility, especially when the speaker belongs to a marginalized or stereotyped group. Recognizing the influence of their prejudice, the hearer would then actively recalibrate their assessments to compensate for credibility deficits and treat the speaker's testimony on its own merits. Fricker refers to this corrective and reflexive virtue as the virtue of testimonial justice.

Fricker's proposal for cultivating testimonial justice has two notable features. First, her approach to addressing testimonial injustice is rooted in virtue theory. Fricker conceptualizes testimonial justice as a virtue that the hearer practices to "reliably neutralize prejudice in her judgments of credibility" (p. 92). This virtue is not merely about having good intentions; it requires continuous self-awareness and critical reflection, enabling hearers to recognize and adjust for biases that might otherwise skew their credibility assessments. Fricker believes that through habitual reflection and correction, individuals would develop a stable disposition to fairly assess a speaker's credibility, even in contexts where prejudices are deeply entrenched in social structures.

This emphasis on self-cultivation leads to the second feature of Fricker's proposal: she envisions testimonial justice as an individual, rather than collective, endeavor. Fricker underscores the responsibility of individuals to cultivate their own moral character and ethical awareness in order to combat the prejudices that give rise to credibility deficits. This individualistic focus implies that each hearer alone can and should develop and exercise this virtue to address testimonial injustice, independently of what other hearers do and what the social and institutional structures are like. Admittedly, Fricker does not deny the value of structural reform, as she acknowledges that "individual virtue is only part of the solution" and that "structural mechanisms also have an essential role in combating epistemic injustice" (Fricker 2010, p. 164). Nonetheless, she maintains that cultivating the virtue of testimonial justice in individual hearers plays an important role in addressing testimonial injustice and epistemic injustice more broadly.

So Fricker's proposal for solving the issue of testimonial injustice is characterized by two features: it is both virtue-theoretic and individual-focused. These two features are closely connected yet distinct. Critics of Fricker's proposal often challenge one (or both) of these features. For instance, Sherman (2015) casts doubt on Fricker's virtue-theoretic approach. He argues that since the prejudices driving testimonial injustice often operate at an implicit level and remain largely inaccessible to conscious reflection, there is no generalizable corrective virtue for rooting out these prejudices in credibility judgments.

On the other hand, Andersen (2012) acknowledges the value of the virtue of testimonial justice, but argues that it should be approached as a virtue of social institutions rather than as an individual responsibility. According to Andersen, epistemic injustice is so deeply embedded within social structures that it cannot be sufficiently addressed by focusing solely on individual character

development. Instead, she advocates for embedding testimonial justice within institutional frameworks to provide structural support for fairer epistemic practices.

In the following sections of this paper, I will, like Anderson, question Fricker's emphasis on individual responsibility. However, while Andersen argues that cultivating the virtue of testimonial justice in individual hearers is ineffective in practice, my focus shifts to the epistemological challenges that make it difficult for hearers to cultivate this virtue in the first place. Specifically, I examine the barriers that prevent individuals from reliably identifying and correcting for testimonial injustice, even when they aim to do so.

III. Testimonial injustice as a systemic pattern

As should be clear from the previous section, testimonial justice consists of credibility deficits resulting from identity prejudice. Thus, to determine that a credibility judgment constitutes testimonial injustice, two facts must be established:

a. There is a deficit in the credibility allocated by the hearer to the speaker.

b. The credibility deficit is caused by identity prejudice the hearer harbors.

However, as I will argue in this section, neither of these two facts can be established for any single credibility judgment, except in rare cases. Instead of manifesting in isolated incidents,

testimonial injustice typically appears as a systemic pattern that can be identified only when observed across multiple judgments over time.

1. Credibility deficit

According to Fricker, a credibility deficit is a situation in which a speaker "receives less than her due credibility" (2007, p.47), meaning the speaker is assigned less credibility than she

actually deserves. To identify a credibility deficit, one therefore has to establish a speaker's due credibility—the amount of credibility they merit—and then compare this with the level of credibility they are actually granted. Unfortunately, however, Fricker does not provide a clear criterion or detailed method for determining a speaker's due credibility.

Consider one of Fricker's paradigmatic examples of testimonial injustice, which comes from Antony Minghella's *The Talented Mr. Ripley*. In this example, Herbert Greenleaf, a wealthy man, dismisses the suspicion of Marge Sherwood, who is engaged to his son, Dickie. Marge suspects that Tom Ripley, a friend of Dickie's, may have harmed or even killed him. However, when she expresses her concerns, Greenleaf silences her by saying, "Marge, there's female intuition, and then there are facts."

Fricker considers this an example of testimonial injustice because Marge does not receive the amount of credibility she merits from Greenleaf, and this deficit is caused by Greenleaf's gender-based stereotype about women. While this interpretation seems plausible, it raises a question: why does Fricker claim there is a credibility deficit in his instance? Fricker does not provide any explanation. Instead, she seems to simply assume it to be the case. This assumption may indeed strike many as intuitively plausible; however, I believe its intuitive appeal likely rests on two considerations.

First, part of the intuitive plausibility of Fricker's assumption seems to stem from the readers' *God's eye view*. For those familiar with *The Talented Mr. Ripley*, it is known from the outset that Ripley did indeed kill Dickie, and thus Marge's suspicion is, in fact, accurate. This knowledge might underpin Fricker's assumption that a deficit exists in Greenleaf's judgment of Marge, and it motivates readers to accept it as a plausible assumption. Measured against the fact that Marge's suspicion reflects the truth, it seems obvious that Marge is granted less credibility

than she merits. However, it is implausible to determine the due credibility of a speaker based solely on the truth of their testimony. Unlike the reader, who has full knowledge of Dickie's fate, Greenleaf lacks this omniscient perspective. Consequently, Greenleaf may not be as epistemically culpable as readers initially take him to be.

This is not to say, however, that Greenleaf bears no epistemic culpability. He certainly does, given his explicit adherence to a gender stereotype that frames women as less rational, as described in the example. This is also another consideration that might have given Fricker's assumption its intuitive appeal. It seems quite reasonable to assume that a hearer with prejudicial views about women is inclined to undervalue the credibility of a female speaker. The presence of prejudice would then serve as an indicator of credibility deficits. This indicator can operate even in the absence of a clear measure of due credibility—the amount of credibility a speaker merits. Thus, it would make sense to infer that since Greenleaf's judgment about Marge's credibility is under the influence of his prejudice, his judgment likely fails to afford her the credibility she deserves. However, as I will argue in Section 3.2, in any single instance of testimonial exchanges, it is often difficult to infer the presence of credibility deficits solely based on the hearer's prejudice against the speaker. If so, then we should conclude that determining whether speakers receive the credibility they deserve cannot rest on the truth of their testimony or solely on the presence of prejudice in credibility judgments.

How, then, should a hearer determine the due credibility of a speaker's testimony? While providing a comprehensive theory of due credibility is beyond the scope of this paper, drawing one Lipton (1998)'s distinction, I will focus on two kinds of evidence relevant to assessing a speaker's credibility: evidence of their sincerity and evidence of their competence. An evaluation of a speaker's due credibility should proceed along these dimensions. However, as I will argue,

significant epistemological challenges arise in interpreting and applying these criteria, making it exceedingly difficult to reliably assess a speaker's due credibility in practice.

According to Searle's canonical view of sincerity, a speaker's assertion is sincere if and only if the speaker believes the proposition they are asserting (1969, p. 66). Sincerity is a crucial component of a speaker's credibility, as it reflects the alignment between what is asserted and what the speaker genuinely believes. While sincerity does not guarantee the truth of a statement, it is, as Williams argues, essential for fostering successful human interaction and building trust between people (Williams, 2002, Chapter 3). Thus, if a hearer has reason to believe that a speaker lacks sincerity in their testimony, the hearer would be justified in granting the speaker a low degree of credibility.

In everyday communication, people frequently rely on verbal and non-verbal cues to gauge sincerity in others' speech. For instance, consistency in statements is often associated with sincerity, so a testimony that contains contradictions may signal dishonesty or insincerity. Similarly, behaviors such as avoiding eye contact or exhibiting darting eyes during a conversation are commonly interpreted as signs of insincerity. A range of such behavioral cues is widely recognized and used in everyday human interactions. These cues provide general rules to aid hearers' evaluation of speakers' sincerity, and in turn, their credibility (Ray Bull et al., 2019).

Despite the widespread reliance on these cues as indicators of sincerity or insincerity, an important question remains: how reliable are these cues in reality? Unfortunately, a significant body of research on lie detection suggests that such cues may be far less dependable than people often believe. In a comprehensive and influential meta-analysis, DePaulo et al. (2003) collected 120 independent samples from previous studies, covering 158 commonly used verbal and non-verbal behavioral cues for lie detection. They found that many behaviors either showed no

discernible links to deceit or, at best, only weak links (p. 74).[30] This finding suggests that most behavioral cues traditionally associated with dishonesty lack consistent empirical support, undermining their usefulness as indicators of sincerity.

Since DePaulo et al.'s meta-analysis was published, lie detection research has advanced with more sophisticated and ecologically valid research paradigms. Some recent research presents a more optimistic view of using behavioral cues to differentiate truth-tellers from liars. For instance, Hartwig and Bond (2014) found that while single behavioral cues may not reliably signal deception, constellations of such cues significantly improve the accuracy of distinguishing between truth-tellers and liars. However, the overall reliability of behavioral cues in detecting deception remains limited. A recent analysis by Patterson et al. (2023) indicates that the connection between behavior and deception was, at best, "faint and unreliable" (p.312).

Furthermore, even if behavioral cues or constellations of such cues do exist that reliably detect lies, it seems that people are not particularly good at using them to tell truth-tellers from liars. It may be that these cues are too subtle or too complex to be noticed by hearers, let alone to use them to judge speakers' sincerity. In a study by ten Brinke et al. (2014), they found that people are very poor lie detectors, performing at about 54% accuracy in traditional lie detection tasks, only marginally above chance. More strikingly, this poor performance does not significantly improve even with specialized training, as Bond and DePaulo (2006)'s study involving professional job interviewers and law enforcement personnel shows.

One might object that this research primarily concerns interactions between strangers, and that familiarity with a speaker's communicative style might improve lie detection accuracy.

[30] DePaulo et al.'s meta-analysis remains one of the most cited studies on deception detection. But for a more recent review of the literature on lie detection and the limitations of behavioral cues, see Vrij, Hartwig, and Granhag (2019), who discuss the challenges of relying on nonverbal cues in high-stakes contexts.

However, empirical evidence on this claim is mixed. For instance, Lee and Welker (2011) found that while "behavioral baselining" (observing a speaker's truth-telling behaviors in controlled settings) improved interviewers' ability to identify truthful behaviors, it did not significantly enhance their ability to detect lies. As Feeley et al. (1995)'s study shows, familiarity improves lie detection accuracy only in cases where a hearer has had longstanding, repeated associations with the speaker, and even then, the improvement is marginal. Since such cases represent only a relatively small subset of all testimonial exchanges, it is reasonable to conclude that, in general, dependable rules for distinguishing sincere speakers from insincere ones are unlikely to exist, let alone for determining a speaker's due credibility based on these cues.

In addition to sincerity, another critical factor in determining a speaker's credibility is evidence concerning the speaker's competence. When two speakers—one competent in a given area and one less so—offer differing testimonies, there is little doubt that the more competent speaker deserves a higher level of credibility, other things being equal. However, the epistemological challenge that complicates judgments about a speaker's sincerity also arises when assessing a speaker's competence. In everyday testimonial exchange activities, people often inevitably make rapid judgments about others' competence based on superficial cues, such as clothing, social demeanor, or other visual indicators.

For example, wearing a white coat is often perceived as a marker of a physician's competence. In a large survey involving 4,062 patients across 10 U.S. academic hospitals, participants were asked to rate physicians based on their perception of how knowledgeable each of them appears. It turned out that doctors wearing formal attire with a white coat received the highest ratings, while those in scrubs with a white coat ranked second (Petrilli et al., 2018). However, in this survey, physicians' clothing was randomly assigned, meaning that attire had no

actual bearing on their medical competence. Studies like this suggest that the visual cues people often rely on in daily life do not provide solid grounds for accurately evaluating a speaker's competence.

Moreover, even professionals trained to assess credibility can struggle to accurately gauge competence based on visual cues. For instance, a study by Gustafsson et al. (2020) found that police detectives with specialized training did not significantly outperform laypeople in judging the accuracy of eyewitness testimony. Findings like this suggest that, just as there are no consistently reliable indicators of sincerity accessible to a hearer, there also seem to be no reliable indicators of competence. Consequently, it is unlikely that any generalizable epistemic rules exist to reliably determine how much credibility a hearer should grant a speaker. Without a plausible method for determining a speaker's due credibility, identifying and the presence of credibility deficits in testimonial exchanges seems impossible.

Fricker has recognized the challenge of identifying credibility deficits in testimonial exchanges, as well as precisely measuring their magnitudes. Nevertheless, she recommends a guiding ideal: adjust credibility judgments upwards for speakers from marginalized groups, even if achieving due credibility is almost always "an imprecise business in practice" (2007, p.170). Fricker believes that this guiding ideal, while imprecise, "makes enough intuitive sense to genuinely guide our practice as hearers" (ibid). However, in Section 4, I will argue that this ideal might be more appropriately understood as a collective goal, rather than one for each individual hearer.

2. Identity prejudice as the cause of credibility deficit

In the previous subsection, I explained why it is difficult to identify the presence of credibility deficits in testimonial exchanges. Here, I will argue that even when a credibility deficit is detected, it is equally—if not more—difficult to establish that it is caused by identity prejudice.

Consider again Fricker's paradigmatic example of testimonial injustice from *The Talented Mr. Ripley,* introduced earlier. As an alleged case of testimonial injustice, it involves a credibility deficit, and moreover, the credibility deficit is caused by Greenleaf's prejudice against Marge. Setting aside the challenge of identifying the presence of credibility deficits which I have discussed in the previous subsection, a further question remains: on what grounds does Fricker claim that Greenleaf's deficient credibility judgment stems from his identity prejudice? In this instance, Fricker's claim is supported by Greenleaf's explicit belittling of women's rational capacities when he remarks, "there's female intuition, and then there are facts." Assuming that a credibility deficit is indeed present in Greenleaf's judgment of Marge, Greenleaf's overt endorsement of gender stereotypes provides *prima facie* evidence that it is his prejudice that leads to the deficit.

However, I doubt that cases like this are common today. As social consensus increasingly condemns racism, sexism, and other forms of discrimination, overt expressions of prejudice, like Greenleaf's, have become less frequent. Instead, implicit biases—subtle, often unconscious forms of prejudice—now become more relevant in social interactions. In the presence of a credibility deficit, what kind of evidence would provide support for its being caused by identity prejudice if the speaker does not explicitly hold a pejorative stereotype about the hearer? In the last three decades, researchers have proposed various research paradigms to experimentally determine and measure one's implicit bias.

For instance, in Affect Misattribution Procedure (AMP), participants are asked to make rapid judgments about ambiguous stimuli (positive or negative) after being briefly exposed to a

prime; in Go/No-Go Association Task (GNAT), participants are exposed to an item from a target category and an attribute, and they must respond quickly if the item and attribute match ("go"), whereas they withhold responses if they do not match("no-go"); in Implicit Association Test (IAT), participants quickly categorize words or images into four categories using two response keys, with category pairings switched across trials. In each of the tasks, participants' response time and/or error rates are recorded to calculate a score indicating the strength of their implicit attitudes.

Administering these tests can, in principle, reveal whether a hearer holds biases against a certain social group, even if they explicitly and sincerely deny being prejudicial. Note, however, that even in the best-case scenario, these tests can only show that a hearer harbors a bias against a group; they cannot establish that this bias has an impact on a testimonial exchange to cause credibility deficits. In cases where a speaker holds explicitly prejudicial views, like the Greenleaf and Marge example, we can identify a clear inferential structure: women are inferior in rational thinking; Marge is a woman; therefore, the credibility of Marge's testimony deserves less credibility. Insofar as we accept a broadly Davidsonian framework (Davidson, 1963) that inferential links between reasons and actions reflect causal connections, the inferential structure identifiable in Greenleaf's assessment of Marge's credibility provides strong evidence that the credibility deficit in Greenleaf's judgment stems from his prejudice against Marge.

This kind of inferential structure, however, is much harder to identify in cases where hearers do not explicitly avow a pejorative stereotype about the speaker. This is because, unlike explicit bias, implicit bias, identified using the kind of implicit attitude tests mentioned earlier, cannot be easily put into a propostional form that allows it to enter into inferential relations with one's reasons and actions. Implicit bias manifests as a statistical pattern emerging from aggregated

data, rather than as a proposition that can support explicit inferences. This poses special challenges to linking implicit bias to credibility deficits in testimonial exchanges.

To understand how implicit bias might be identified in a case like that of Greenleaf and Marge, let us revisit their example. But imagine that, unlike in the original example, Greenleaf does not explicitly express a bias against women and sincerely believes women are not inferior to men in terms of rational thinking. To determine whether Greenleaf harbors implicit bias, suppose he participates in a version of the IAT specifically designed to measure implicit associations between females and rationality.[31]

In this version of the test, Greenleaf is asked to rapidly categorize words or images into paired concepts using two response keys. For instance, he might press one key for "female" names and "rationality" words, and another key for "male" names and "emotionality" words. After completing a certain number of trials (normally 20-40 trials), Greenleaf starts another block of trials in which the pairing is switched (e.g., female + emotionality, male + rationality). These two blocks of trials count as a complete testing session. The IAT measures how quickly and accurately Greenleaf responds in these different pairing conditions. If Greenleaf's responses on average are faster and more accurate when "male" and "rationality" are paired, compared to when "female" and "rationality" are paired, it would indicate that Greenleaf is biased— he has a stronger tendency to associate males with rationality than females. The difference in response times between these pairing conditions is used to calculate Greenleaf's IAT score. The bigger the difference, the more biased Greenleaf is.

[31] In this paper, I use the IAT as the primary example for my analysis. However, my argument extends to other measures of implicit attitudes as well. A common feature of these measures is that they assess downstream behavioral consequences of attitudes rather than the attitudes themselves.

In interpreting Greenleaf's performance on the test, it is important to note that no single IAT trial is sufficient to determine whether he holds an implicit prejudice. Rather, we assess his bias based on the statistical pattern across multiple trials, which generates an overall IAT score. This score reveals, on average, whether Greenleaf shows implicit bias and, if so, the extent to which he is biased. Yet, it does not imply that a particular trial in the test is biased, nor does it imply that Greenleaf's actions beyond that test, such as his judgment about Marge's credibility, is influenced by his bias.

To illustrate, consider a simple analogy. Suppose we take a sample from a larger population of individuals. Suppose also that the average height of individuals in this sample is taller than 5'6''. Now, consider this question: if we randomly pick a person from the population (who is not in the sample), how confident can we be in predicting that the person is taller than 5'6"? To make that prediction with high confidence, two conditions must hold. First, we need to consider the variance of the sample. If the variance of the sample is minimal, meaning that all the individuals in the sample are of roughly the same height, then we can almost be certain that a person randomly picked from the *sample* is taller than 5'6''. I refer to this as the low-variance condition. If this condition is satisfied, we also need to consider the second condition: the representativeness of the sample. If the sample is truly representative of the population, then we can predict with high confidence that a person randomly picked from the *population* is taller than 5'6". I call this the representativeness condition

Likewise, to infer that Greenleaf's judgment about Marge's credibility is influenced by his bias as indicated by his IAT score, the two conditions must also be satisfied. Meeting the low-variance condition requires that Greenleaf's trials in a single IAT session are consistent. On the other hand, the representativeness condition requires that if Greenleaf is asked to re-take the IAT

test, his performance would not be significantly different from that in his first test. However, empirical research indicates that neither of these conditions is reliably satisfied.

The low-variance condition is examined in studies on the internal consistency of IAT measures. Internal consistency refers to how well different trials within a single IAT testing session produce similar and consistent results. Findings regarding the internal consistency of IAT are mixed. Greenwald and Lai (2020)'s meta-analysis, using data from 257 studies, indicates a relatively high aggregate internal consistency ($\alpha = .80$). However, when studies are analyzed individually, the consistency shows significant variations, ranging from moderate ($\alpha = .55$) to high ($\alpha = .88$) (Williams & Steele, 2016). In cases where the internal consistency is only moderate, the condition is not satisfied. In cases where the consistency is high, the condition is satisfied, but then we also need to consider the second condition, namely, the representativeness condition.

The representativeness condition relates to the test-retest reliability of IAT, that is, how consistently an individual's performance holds across different IAT sessions. Research shows that the reliability of IAT test is generally low, with participants' performance varying considerably from one testing session to another. Various factors contribute to these variations. For example, brief exposure to images that reinforce or contradict existing stereotypes (e.g., Black individuals who succeed in traditionally White-dominated fields) can influences participants' performance on IAT, either amplifying or reducing their bias, as demonstrated in studies by Dasgupta and Greenwald (2001). Emotions also have an impact on implicit bias. For instance, Graf and Paolini (2017) found that participants' implicit bias were significantly reduced after experiencing positive emotions during intergroup contact. Furthermore, even emotions triggered by events or experiences that are entirely unrelated to the situation or task at hand can influence implicit bias (Dasgupta et al., 2009). More strikingly, Cochrane et al. (2023) observed order effects in their

studies, indicating that simply changing the sequence in which different pairing conditions are presented is enough to cause variations in IAT scores. These findings suggest that IAT scores are extremely sensitive to external influences. As a result, the IAT test fails to meet the representativeness condition, as participants' performance in one testing session cannot reliably reflect their bias across different contexts or over time.

The empirical findings suggest that neither the low-variance nor the representativeness condition is reliably met in the case of the IAT test. Consequently, even if Greenleaf's performance on an IAT test indicates an implicit bias associating rationality more strongly with men than women, This result offers only limited evidence to conclude that his bias actually influenced his judgment of Marge's credibility. Furthermore, the problem is not unique to IAT. Other measures of implicit attitudes are also subject to this problem to different extents (Hu & Hancock, 2024). Therefore, there seems to be no reliable method to establish that the credibility deficit in Greenleaf's judgment of Marge is causally linked to his prejudice, if he holds this prejudice only implicitly.

3. Testimonial injustice as a pattern of credibility judgments

In the last two subsections, I argued, first, there is no plausible criterion according to which we can establish the presence of credibility deficits in credibility judgments, and second, that even when credibility deficits are present, there is no reliable method for linking them to identity prejudice, except in rare cases where the hearer is explicitly biased against the speaker, as in Fricker's example of Greenleaf and Marge. This uncertainty—both about the presence of credibility deficits and about whether identity prejudice causes them—means that we lack strong evidence to confirm the occurrence of injustice in single instances of testimonial exchange.

However, it is important to emphasize that my claim is epistemological rather than metaphysical. I do not doubt the reality of testimonial injustice. Rather, I am concerned with how we detect injustice in testimonial exchanges. As an epistemological claim, my argument aims to show that there is no reliable method for detecting injustice in a particular testimonial exchange. Yet, this does not prevent us from detecting the presence of testimonial injustice across a broader population, where we can assess patterns of credibility allocation. In this subsection, I will explain how this is possible.

Consider first how credibility deficits in testimonial exchanges can be detected. In Section 3.1, I argued that there are no reliable indicators available to hearers for definitively assessing a speaker's sincerity or competence. As a result, we lack clear criteria for determining the "due credibility" a speaker deserves, making it nearly impossible to confirm the presence of credibility deficits in any individual testimonial exchange.

The core challenge in detecting credibility deficits lies in the difficulty of determining how much credibility a hearer should grant a single speaker. However, this challenge diminishes when we shift our focus to the population level. In a large population, countless testimonial exchanges naturally involve credibility judgments. In such cases, if we observe that a particular social group, as a whole, is granted disproportionately less credibility than another group, this pattern would provide strong evidence that credibility deficits are present within the population. By examining a substantial number of cases across a wide range of interactions, we can identify credibility deficits as an aggregate trend, even if we cannot pinpoint them reliably in isolated exchanges. This population-level approach enables us to detect credibility imbalances without needing to assess each individual case for due credibility, which is exceedingly difficult, if not impossible. Thus, while establishing credibility deficits in a single instance remains problematic, analyzing them at

the population level offers a viable method for identifying and even quantifying imbalances in credibility allocation.

Now, consider the question of establishing identity prejudice as a cause of credibility deficits. In Section 3.2, I argued that in cases where the hearer holds only an implicit prejudice against the speaker, we cannot establish a causal link between their prejudice and credibility deficits because IAT and other implicit attitude measures do not provide strong evidence that the implicit bias affects a particular credibility judgment.

However, much like in the case of the detection of credibility deficits, while it is difficult to assess the influence of implicit prejudice on credibility judgments at the individual level, we may be able to do so at the population level. As the empirical evidence I presented in that section suggests, implicit bias varies significantly within individuals over time. Yet, as Payne et al. (2017) demonstrate, implicit bias consistently emerges in test results across large groups, despite fluctuations in individual scores. When measured at the aggregate level (i.e., liking at average scores across a large group), the presence of bias is highly reliable. Moreover, Payne and his colleagues also found that if different groups or samples are tested repeatedly, similar patterns of implicit bias are found across these groups. This evidence indicates that, despite individual-level variations, implicit bias remains consistent and serves as a reliable predictor of behaviors at the population level.

If we can establish the presence of credibility deficits and implicit prejudice as their cause at the population level, then we can also determine the existence of testimonial injustice at this level. Although we may not be able to identify a particular testimonial exchange as unjust, the widespread credibility imbalance across social groups and the stable presence of implicit bias within the population allow us to assert, with high confidence, that testimonial injustice exists as

an aggregate pattern. This reveals testimonial injustice as a deeply social phenomenon, one that is fundamentally rooted in collective patterns of interaction rather than individual cognitive abilities alone. As I will argue in the next section, this understanding of testimonial injustice has important implications for how we should address it.

IV. Testimonial justice as an institutional virtue

The analysis from the previous section suggests that, epistemologically speaking, testimonial injustice should be better understood as an aggregate pattern rather than as isolated incidents that arise from individual vice or cognitive failings. Given that testimonial injustice manifests as a systemic disparity in credibility allocation between social groups that stems from identity prejudice at the population level, addressing the issue through the cultivation of individual virtues, as Fricker suggests, is unlikely to be effective.

Recall that Fricker's proposal for combatting testimonial injustice, as I explained in Section 2, centers on the virtue of testimonial justice, which is a corrective anti-prejudicial virtue that hearers cultivate to ensure fair credibility judgments. According to Fricker, by practicing self-awareness and critical reflection, hearers can develop a stable disposition to fairly assess testimonies, even in the face of deeply entrenched social structural injustice. Developing such dispositions allows hearers to counteract the influence of their prejudice by appropriately adjusting their credibility judgments to match the credibility a speaker merits.

In a recent paper, Piovarchy (2021) argues that it is impossible for hearers to adjust their credibility judgments due to the involuntariness of belief. For instance, when standing outside on a clear day, one immediately forms the belief that the sky is blue. This belief is automatically formed and beyond conscious control—one cannot simply will oneself to believe that the sky is

green, no matter how hard they try. Similarly, a hearer's belief about a speaker's credibility is also formed involuntarily and out of their conscious control. This means that if Greenleaf has already adopted a belief that Marge is untrustworthy, he cannot revise his judgment about Marge's trustworthiness by simply deciding to believe otherwise. Consequently, it is unreasonable to require individual hearers to adjust their credibility judgments to compensate for deficits.

However, this critique assumes that a hearer's judgment about a speaker's credibility must directly reflect their belief about the speaker's credibility. It presupposes that hearers must first form a belief about a speaker's trustworthiness, and then make a judgment about the speaker's credibility based on their belief. The involuntariness of beliefs thus implies the involuntariness of judgments. But this need not be the case. Rather than basing credibility judgments on beliefs, hearers can base their judgments on *epistemic acceptance* instead. Epistemically accepting a proposition *p*, as Cohen defines it, means "going along with that proposition (either for the long term or for immediate purposes only) as a premiss…Whether or not one assents and whether or not one feels it to be true that *p*" (1989, p. 368).[32] Unlike belief, acceptance is voluntary—one can choose to accept or not accept a testimony regardless of their belief about it, and then act on that acceptance.

For example, a lawyer may accept a client's claim of innocence for the purpose of the trial, even if they privately believe that the client is guilty. Similarly, even if Greenleaf *believes* that Marge's testimony is not credible—feeling that Marge's suspicion of Ripley is unfounded— he can still choose to *accept* Marge as a trustworthy interlocutor and act as if Marge is trustworthy, like listening to Marge's testimony with patience, examining evidence for or against her suspicion,

[32] The distinction between belief and acceptance has been a topic of significant debate in contemporary epistemology. In addition to Cohen (1989), Bratman (1992) also provides an insightful discussion on epistemic acceptance.

etc. If credibility judgments are grounded in epistemic acceptance rather than belief, hearers gain the ability to adjust their judgments at will to counteract the influence of prejudice.

By shifting the cognitive basis of credibility judgment from belief to acceptance, Fricker's proposal to cultivate individual virtue to address testimonial injustice can withstand Piovarchy's critique. However, in light of my arguments from the previous section, while epistemic acceptance enables hearers to adjust their judgments to grant a speaker due credibility, another epistemological challenge, arises for Fricker's proposal: how do the hearers determine the appropriate amount of adjustment needed to achieve testimonial justice?

There are two potential ways for a hearer to determine the amount of adjustment needed in their credibility judgment. First, a hearer could adjust their judgment based on how much credibility a speaker actually deserves. Knowing a speaker's due credibility, the hearer would modify their judgment to align with the level of credibility the speaker merits. Second, since testimonial injustice consists of credibility deficits resulting from identity prejudice, a hearer could adjust their judgment based on the extent to which their prejudice influences their credibility assessments—granting a speaker just enough additional credibility to offset the effects of their bias. However, in light of my analysis from the previous section, neither of these strategies proves workable.

The first strategy fails because, as I argued in Section 3.1, there are no reliable indicators accessible to hearers that allow them to determine how much credibility a speaker actually deserves. Without such indicators, it is unclear how individual hearers can develop a virtue of testimonial justice. As Fricker conceives of it, a virtuous hearer "'just sees' her interlocutor in a certain light, and responds to his word accordingly" (2007, p. 76). However, if there are no characteristics accessible to a hearer that allow them to reliably distinguish a trustworthy speaker from an

untrustworthy one, it is unclear how the hearer could develop such a perceptual-like ability to accurately identify the appropriate level of credibility they should grant a speaker.

The second strategy also fails. As I argued in Section 3.2, in cases where a hearer is unaware of holding prejudicial views against a speaker, they cannot reliably determine whether their credibility judgments are influenced by prejudice. Even if the hearer takes the IAT test multiple times to minimize random error, the test results only indicate that they have a general tendency of being or not being biased. But as I explained, the relatively low internal consistency and reliability of IAT and other implicit attitude measures suggests that this general tendency is not uniform across the hearer's behaviors. Whether a hearer's bias is activated to influence a particular credibility judgment is highly dependent on contextual factors, such as situational cues, emotional states and cognitive load. Thus, a hearer with a high IAT score may not exhibit bias in a specific credibility judgment, while someone with a low score might still make biased judgments in certain contexts.

Therefore, even if IAT reveals that a hearer exhibits significant bias against a speaker's identity, it provides limited evidence about whether or to what extent a particular credibility judgment is influenced by that bias. Consequently, the hearer cannot reliably determine whether or how much they should adjust their judgment to fairly assess the speaker's credibility. In such a case, it is unclear how individual hearers can cultivate reliable, intuitive responsiveness that Fricker's perceptual-like virtue of testimonial justice requires. Given that implicit biases are sporadic and context-dependent, instead of "just seeing" the speaker in a fair and unbiased way, the hearer is left guessing whether their judgment is distorted by prejudice in a particular testimonial exchange.

Unlike implicit bias, explicit bias is more stable and robust across contexts. Yet, the epistemological challenge explained above also applies to cases where a hearer holds explicit prejudice against a speaker. This is because explicit bias is mostly expressed in qualitative rather than quantitative terms. An explicitly biased hearer might believe that women are inferior to men in rational thinking, but it is very rare for someone to quantify such a prejudice by claiming that women are, say, 25% less rational than men. When a hearer is aware of their bias, they have reason to believe that their credibility judgment of the speaker is influenced by the bias and feel compelled to adjust their judgment to compensate for the influence of bias. However, given the qualitative expression of their bias, it is hard for a hearer to determine the amount of adjustment necessary for a fair credibility assessment, and then make the adjustment accordingly.

The failures of both strategies are the result of the epistemological challenge to identifying the existence of injustice in individual testimonial exchanges. Given that neither of the strategies works, it is unclear how individual hearers can cultivate a virtue that allows them to reliably correct for credibility deficits that stem from identity prejudice. As I argued, testimonial injustice is best conceived of as an anomalous aggregate pattern rather than wrongdoings in individual testimonial exchanges. In such a case, addressing the issue of testimonial injustice requires solutions at the population level rather than focusing on individual testimonial exchanges. Individual hearers' capacity is inherently limited in this aspect, whereas institutions are far more effective in tackling the issue.

For example, it can be extremely difficult to determine whether a Black defendant's testimony is discounted in court due to the presiding judge's bias, and thus whether the case constitutes an instance of testimonial injustice. However, if data reveals that disproportionately more Black defendants have their testimonies discounted compared to defendants from other

groups, this aggregate pattern provides strong evidence of testimonial injustice within the legal system. Such a systemic issue demands institutional solutions, such as legal reforms designed to minimize biases and ensure equitable credibility assessments. Addressing testimonial injustice at the institutional level can create structural changes that disrupt these aggregate patterns, thereby fostering a more just and equitable environment for all participants in testimonial exchanges.

CHAPTER 5

Conclusions

This dissertation has delved into three different but interconnected facets of social bias: perceptual bias, implicit bias, and testimonial injustice. Each chapter has examined how bias influences a specific aspect of social life, revealing its underlying mechanisms and evaluating the challenges it presents to traditional interventions. Collectively, these chapters offer a comprehensive understanding of how social bias operates at both individual and systemic levels and provide insights into effective interventions combating social bias.

Chapter 2 investigates how social bias can influence what we see. It argues that bias isn't just a product of our thoughts but can also stem from the way our visual system processes information. Drawing on research in vision science, such as the Müller-Lyer illusion (where lines of the same length appear different) and the carpentered-world hypothesis (which suggests our environment shapes how we perceive shapes), the chapter shows how flawed visual assumptions can lead to distorted views of the world. This challenges traditional efforts to combat bias, as it highlights that some biases are rooted in processes we aren't even consciously aware of. While individuals can try to correct for these biases, their efforts are often limited. Instead, the chapter implies that institutional solutions, such as training programs informed by vision science, to address these issues on a larger scale. For example, fields like law enforcement or healthcare could adopt standardized protocols for evaluating visual evidence, reducing the risk of biased perceptions affecting critical decisions. By embedding these practices into institutional systems, we can create fairer outcomes and reduce the impact of perceptual bias.

Chapter 3 explores the connection between implicit bias in people and bias in artificial intelligence (AI) systems. It critiques models like the k-nearest neighbors (kNN) approach for oversimplifying how implicit bias works in humans. Studies show that implicit bias isn't fixed—it fluctuates depending on factors like emotions, moods or mental fatigue. This variability means that bias isn't just about algorithms; it's also shaped by the broader context in which decisions are made. The chapter argues that purely algorithmic solutions aren't enough to address bias and calls for institutional measures, such as ethical guidelines for AI design and fairness audits, to ensure AI systems are more equitable. For instance, requiring diverse datasets and regular bias checks could help prevent AI from reinforcing existing inequalities. These systemic changes align with the dissertation's broader argument: tackling bias requires more than individual efforts—it demands structural reforms.

Chapter 4 focuses on testimonial injustice, where people are unfairly disbelieved or dismissed because of their identity. The chapter critiques the idea that individual virtues, like being a fair listener, are effective in solving this problem. It argues that such virtues are hard to practice consistently because it is hard to determine the existence of credibility deficits and even if they do exist, it is hard to establish a causal connection between these deficits and identity prejudice. Instead, the chapter advocates for systemic changes, like legal reforms and policies that ensure marginalized groups have a fair voice in decision-making. For example, institutions could implement measures to guarantee diverse representation or create legal pathways for people to challenge unfair treatment. These systemic solutions, the chapter argues, are more effective at addressing testimonial injustice than relying on individual good intentions.

Together, this dissertation highlights the complexity of social bias and how it shows up in various aspects of social life. A recurring theme throughout all three papers is the critical need for

institutional solutions to address social bias. While individual-level interventions, such as cultivating virtues or raising awareness, may be valuable, they often fall short in combating the pervasive and systemic nature of bias. Institutional solutions, by contrast, have the potential to enact structural changes that disrupt patterns of bias and promote equity on a broader scale. This conclusion not only unifies the dissertation but also serves as a call to action for policymakers, educators, and practitioners to prioritize systemic reforms in the fight against social bias.

REFERENCES

Alcoff, Linda (2010). Epistemic Identities. *Episteme*, 7(2), 128-137.

Alkozei, A., Killgore, W. D., Smith, R., Dailey, N. S., Bajaj, S., & Haack, M. (2017). Chronic sleep restriction increases negative implicit attitudes toward Arab Muslims. *Scientific Reports*, 7(1), 4285.

Alpers, G. W., Ruhleder, M., Walz, N., Mühlberger, A., & Pauli, P. (2005). Binocular rivalry between emotional and neutral stimuli: a validation using fear conditioning and EEG. *International Journal of Psychophysiology*, 57(1), 25–32.

Amaral, D. G., Price, J. L., Pitkanen, A. & Carmichael, S. T. (1992). Anatomical organization of the primate amygdaloid complex. In *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction* (pp. 1–66).

American Bar Association. (2022, May 24). Racial Disparities Inherent in America's Fragmented Parole System. *GPSolo eReport*. Retrieved from https://www.americanbar.org/groups/gpsolo/publications/gpsolo_ereport/2022/may-2022/racial-disparities-inherent-americas-fragmented-parole-system/.

Amodio, D., & Hamilton, H. (2012). Intergroup anxiety effects on implicit racial evaluation and stereotyping. *Emotion*, 12(6), 1273–1280.

Anderson, B. (2023). Stop paying attention to "attention". *Wiley Interdisciplinary Reviews: Cognitive Science*, 14(1), e1574.

Anderson, Elizabeth (2012). Epistemic Justice as a Virtue of Social Institutions. *Social Epistemology*, 26(2), 163-173.

Basu, R. (2019). What we epistemically owe to each other. *Philosophical Studies*, 176(4), 915—931.

Berry, J. W. (1971). Müller-Lyer Susceptibility: Culture, Ecology or Race? *International Journal of Psychology*, 6(3), 193–197.

Binns, R. (2019). On the apparent conflict between individual and group fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Blum, L. (2010). Racialized groups: The sociohistorical consensus. *The Monist*, 93(2), 298–320.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of Deception Judgments. *Personality and Social Psychology Review*, 10(3), 214-234.

Bratman, M. E. (1992). Practical Reasoning and Acceptance in a Context. *Mind*, 101(401), 1–15.

Bull, R., van der Burgh, M., Dando, C. (2019). Verbal Cues Fostering Perceptions of Credibility and Truth/Lie Detection. In: Docan-Morgan, T. (eds) *The Palgrave Handbook of Deceptive Communication*. Palgrave Macmillan, Cham.

Carruthers, Peter (2006). *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. New York: Oxford University Press UK.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153-163.

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, 62, 73–101.

Cohen, L. Jonathan (1989). Belief and acceptance. *Mind*, 98(391), 367-389.

Connor, P., & Evers, E. R. K. (2020). The Bias of Individuals (in Crowds): Why Implicit Bias Is Probably a Noisily Measured Individual-Level Construct. *Perspectives on Psychological Science*, 15(6), 1329–1345.

Correll, J., Wittenbrink, B., Crawford, M. T., & Sadler, M. S. (2015). Stereotypic vision: How stereotypes disambiguate visual stimuli. *Journal of Personality and Social Psychology*, 108(2), 219–233.

Crockett, Zachary (2016, September 13). "Gang Member" and "Thug" Roles in Film are Disproportionately Played by Black Actors. *Vox*. Retrieved from https://www.vox.com/platform/amp/2016/9/13/12889478/black-actors-typecasting.

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17), 6889-6892.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, 81(5), 800–814.

Dasgupta, N., DeSteno, D.A., Williams, L., & Hunsinger, M. (2009). Fanning the flames of prejudice: The influence of specific incidental emotions on implicit prejudice. *Emotion*, 9, 585-591.

Davidson, Donald (1963). Actions, Reasons, and Causes. *Journal of Philosophy*, 60(23), 685-700.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, 129, 74–118.

Deręgowski, J. B. (2013). On the Müller-Lyer Illusion in the Carpentered World. *Perception*, 42(7), 790–792.

Dooley, S., Downing, R., Wei, G., Shankar, N., Thymes, B., Thorkelsdottir, G., ... & Goldstein, T. (2021). Comparing human and machine bias in face recognition. *arXiv preprint arXiv:2110.08396*.

Equal Justice Initiative (2021, December 16). Report Documents Racial Bias in Coverage of Crime by Media. Retrieved from https://eji.org/news/report-documents-racial-bias-in-coverage-of-crime-by-media/.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Fazekas, P., & Nanay, B. (2020). Attention Is Amplification, Not Selection. *The British Journal for the Philosophy of Science*, 72, 299–324.

Feeley, T. H., deTurck, M. A., & Young, M. J. (1995). Baseline familiarity in lie detection. *Communication Research Reports*, 12(2), 160–169.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *The Behavioral and Brain Sciences*, 39, e229.

Fiske, S. T. (2017). Prejudices in Cultural Contexts: Shared Stereotypes (Gender, Age) Versus Variable Stereotypes (Race, Ethnicity, Religion). *Perspectives on Psychological Science*, 12(5), 791–799.

Fleisher, Will (2021). What's Fair about Individual Fairness? *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.

Fodor, J. A. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Fricker, M. (2010). Replies to Alcoff, Goldberg, and Hookway on epistemic injustice. *Episteme*, 7(2), 164-178.

Gates, S. W., Perry, V. G., & Zorn, P. M. (2002). Automated underwriting in mortgage lending: Good news for the underserved? *Housing Policy Debate*, 13, 369-391.

Geisler, W. S., & Kersten, D. (2002). Illusions, perception and Bayes. *Nature Neuroscience*, 5(6), 508–510.

Gendler, T. S. (2008). Alief in action (and reaction). *Mind & Language*, 23(5), 552–585.

Gendler, T. S. (2011). On the epistemic costs of implicit bias. *Philosophical Studies*, 156(1), 33—63.

Graf, S., & Paolini, S. (2017). Investigating positive and negative intergroup contact: Rectifying a long-standing positivity bias in the literature. In L. Vezzali & S. Stathi (Eds.), *Intergroup Contact Theory: Recent Developments and Future Directions* (pp. 92-113). Routledge.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1), 4–27.

Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J. F., Friese, M., Hahn, A., Hehman, E., Hofmann, W., Hughes, S., Hussey, I., Jordan, C., Kirby, T. A., Lai, C. K., Lang, J. W. B., Lindgren, K. P., Maison, D., Ostafin, B. D., Rae, J. R., Ratliff, K. A., … Wiers, R. W. (2022). Best research practices for using the Implicit Association Test. *Behavior Research Methods*, 54(3), 1161–1180.

Greenwald, A. G., & Lai, C. K. (2020). Implicit Social Cognition. *Annual Review of Psychology*, 71, 419–445.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74(6), 1464–1480.

Gregory, R. L. (1997). Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 352(1358), 1121–1127.

Hartwig, M., & Bond, C. F. Jr. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, 28, 661–667.

Hayes, M., Fortunato, D., & Hibbing, M. (2021). Race–gender bias in white Americans' preferences for gun availability. *Journal of Public Policy*, 41(4), 818–834.

Hedden, Brian (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs*, 49(2), 209-231.

Helmholtz, H. von. (1925). *Treatise on Physiological Optics* (J. P. C. Southall, Trans.). Optical Society of America. (Original work published 1867)

Hochman, A. (2017). In defense of the metaphysics of race. *Philosophical Studies*, 174, 2709–2729.

Holmes, E. A., & Mathews, A. (2010). Mental imagery in emotion and emotional disorders. *Clinical Psychology Review*, 30(3), 349–362.

Holroyd, J., Scaife, R., & Stafford, T. (2017). What is implicit bias? *Philosophy Compass*, 12(10), e12437.

Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J. H., & Welsh, T. N. (2019). No one knows what attention is. *Attention, Perception, & Psychophysics*, 81, 2288–2303.

Howe, C. Q., & Purves, D. (2005). The Müller-Lyer illusion explained by the statistics of image-source relationships. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4), 1234–1239.

Hu, X., & Hancock, A. M. (2024). State of the science: Introduction to implicit bias review 2018-2020. *The Kirwan Institute for the Study of Race and Ethnicity*. https://kirwaninstitute.osu.edu/research/state-science-introduction-implicit-bias-review-2018-2020

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12), 1489–1506.

Jahoda, G. (1971). Retinal pigmentation, illusion susceptibility and space perception. *International Journal of Psychology*, 6(3), 199–207.

Johnson, Gabbrielle M. (2020a). The Structure of Bias. *Mind*, 129(516), 1193-1236.

Johnson, Gabbrielle M. (2020b). Algorithmic bias: on the implicit biases of social technology. *Synthese*, 198(10), 9941-9961.

Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalization: separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(2), 316–332.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgment*. New York: Little, Brown Spark.

Kanizsa, G. (1985). Seeing and thinking. *Acta Psychologica*, 59(1), 23–33.

Katsuki, F., & Constantinidis, C. (2014). Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist*, 20(5), 509–21.

Kavanagh, D. J., May, J., & Andrade, J. (2009). Tests of the elaborated intrusion theory of craving and desire: Features of alcohol craving during treatment for an alcohol disorder. *The British Journal of Clinical Psychology*, 48(Pt 3), 241–254.

Kelly, Thomas (2022). *Bias: A Philosophical Study*. Oxford, GB: Oxford University Press.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1), 237–293.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Korb, S., Mikus, N., Massaccesi, C., Grey, J., Duggirala, S. X., Kotz, S. A., & Mehu, M. (2022). EmoSex: Emotion prevails over sex in implicit judgments of faces and voices. *Emotion*. Advance online publication.

LeDoux, J. E. (1986). Sensory systems and emotion: A model of affective processing. *Integrative Psychiatry*, 4(4), 237–243.

Lee, C.-C., & Welker, R. B. (2011). Prior exposure to interviewee's truth-telling (baselining) and deception-detection accuracy in interviews. *Behavioral Research in Accounting*, 23(2), 131–146.

Levy, N. (2015). Neither fish nor fowl: implicit attitudes as patchy endorsements. *Nous*, 49(4), 800–823.

Lipton, Peter (1998). The epistemology of testimony. *Studies in History and Philosophy of Science Part A*, 29(1), 1-31.

Macpherson, F. (2012). Cognitive Penetration of Colour Experience: Rethinking the Issue in Light of an Indirect Mechanism. *Philosophy and Phenomenological Research*, 84(1), 24–62.

Macpherson, F. (2017). The relationship between cognitive penetration and predictive coding. *Consciousness and Cognition: An International Journal*, 47, 6–16.

Machery, E. (2022). Anomalies in implicit attitudes research. *Wiley Interdisciplinary Reviews: Cognitive Science*, 13(1), e1569.

Mallon, R. (2006). 'Race': normative, not metaphysical or semantic. *Ethics*, 116(3), 525–551.

Mamassian, P., & Landy, M. S. (1998). Observer biases in the 3D interpretation of line drawings. *Vision Research*, 38(18), 2817–2832.

Mandelbaum, E. (2016). Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Nous*, 50, 629–658.

Mende-Siedlecki, P., Qu-Lee, J., Backer, R., & Van Bavel, J. J. (2019). Perceptual contributions to racial bias in pain recognition. *Journal of Experimental Psychology: General*, 148(5), 863.

Morgenstern, Y., Murray, R. F., & Harris, L. R. (2011). The human visual system's assumption that light comes from above is weak. *Proceedings of the National Academy of Sciences of the United States of America*, 108(30), 12551–12553.

Munton, Jessie (2019). Perceptual Skill and Social Structure. *Philosophy and Phenomenological Research*, 99(1), 131–161.

Nanay, B. (2021). Implicit bias as mental imagery. *Journal of the American Philosophical Association*, 7(3), 329–347.

Neemeh, Z.A. (2020). Bootstrap Hell: Perceptual Racial Biases in a Predictive Processing Framework. *CogSci*.

Nijhawan, R. (1995). 'Reversed' Illusion with Three-Dimensional Müller-Lyer Shapes. *Perception*, 24(11), 1281–1296.

Nour, M. M., & Nour, J. M. (2015). Perception, illusions and Bayesian inference. *Psychopathology*, 48(4), 217–221.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group.

Paolini, S., Harwood, J., Logatchova, A., Rubin, M., & Mackiewicz, M. (2021). Emotions in Intergroup Contact: Incidental and Integral Emotions' Effects on Interethnic Bias Are Moderated by Emotion Applicability and Subjective Agency. *Frontiers in Psychology*, 12, 588944.

Patterson, M. L., Fridlund, A. J., & Crivelli, C. (2023). Four misconceptions about nonverbal communication. *Perspectives on Psychological Science*, 18, 1388–1411.

Payne, B. K., Shimizu, Y., & Jacoby, L. L. (2005). Mental control and visual illusions: Toward explaining race-biased weapon misidentifications. *Journal of Experimental Social Psychology*, 41(1), 36–47.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, 28(4), 233–248.

Petrilli, C. M., Saint, S., Jennings, J. J., Caruso, A., Kuhn, L., Snyder, A., & Chopra, V. (2018). Understanding patient preference for physician attire: a cross-sectional observational study of 10 academic medical centres in the USA. *BMJ Open*, 8(5), e021239.

Pessoa, L., & Ungerleider, L. G. (2004). Neuroimaging studies of attention and the processing of emotion-laden stimuli. *Progress in Brain Research*, 144, 171–182.

Piovarchy, Adam (2021). Responsibility for Testimonial Injustice. *Philosophical Studies*, 178(2), 597–615.

ProPublica. (2016, May 23). How We Analyzed the COMPAS Recidivism Algorithm. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Pylyshyn, Zenon (1999). Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, 22(3), 341–365.

Roberts, Alice (2021). Objectification and vision: how images shape our early visual processes. *Synthese*, 199, 4543–4560.

Scholl, B. J. (2005). Innateness and (Bayesian) Visual Perception: Reconciling Nativism and Development. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Structure and Contents* (pp. 34–52). Oxford University Press.

Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Segall, M. H., Campbell, D. T., & Herskovits, M. J. (1963). Cultural Differences in the Perception of Geometric Illusions. *Science*, 139(3556), 769–771.

Shapiro, J. R., Ackerman, J. M., Neuberg, S. L., Maner, J. K., Vaughn Becker, D., & Kenrick, D. T. (2009). Following in the wake of anger: when not discriminating is discriminating. *Personality & Social Psychology Bulletin*, 35(10), 1356–1367.

Sherman, B. R. (2015). There's No (Testimonial) Justice: Why Pursuit of a Virtue is Not the Solution to Epistemic Injustice. *Social Epistemology*, 30(3), 229–250.

Siegel, Susanna (2020). Bias and Perception. In Erin Beeghly & Alex Madva (Eds.), *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind* (pp. 99–115). Routledge.

Soares, S. C., & Esteves, F. (2013). A glimpse of fear: Fast detection of threatening targets in visual search with brief stimulus durations. *PsyCh Journal*, 2(1), 11–16.

Stokes, D. (2018). Attention and the cognitive penetrability of perception. *Australasian Journal of Philosophy*, 96(2), 303–318.

Stokes, M. B., & Payne, B. K. (2010). Mental Control and Visual Illusions: Errors of Action and Construal in Race-Based Weapon Misidentification. *The Science of Social Vision*, 295–305.

ten Brinke, L., Stimson, D., & Carney, D. R. (2014). Some Evidence for Unconscious Lie Detection. *Psychological Science*, 25(5), 1098-1105.

Trawalter, S., Todd, A. R., Baird, A. A., & Richeson, J. A. (2008). Attending to Threat: Race-based Patterns of Selective Attention. *Journal of Experimental Social Psychology*, 44(5), 1322–1327.

Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., Lallas, A., Lapins, J., Longo, C., Malvehy, J., Marchetti, M. A., Marghoob, A., Menzies, S., Oakley, A., Paoli, J., Puig, S., … Kittler, H. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The Lancet Oncology*, 20(7), 938–947.

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest*, 11(3), 89–121.

Vrij, A., Hartwig, M., & Granhag, P. A. (2019). Reading lies: Nonverbal communication and deception. *Annual Review of Psychology*, 70, 295–317.

Williams, A., & Steele, J. R. (2016). The reliability of child-friendly race-attitude implicit association tests. *Frontiers in Psychology*, 7, 1576.

Wilson, J. P., Hugenberg, K., & Rule, N. O. (2017). Racial bias in judgments of physical size and formidability: From size to threat. *Journal of Personality and Social Psychology*, 113(1), 59–80.