Implicit Bias and Emergent Moral Wrongs

Kelsey Vicars
Simon Fraser University

Abstract

My research will explore the normative structure of implicit bias. This project will focus on the philosophical implications of bias in social cognition, and will address the question of moral responsibility for implicit bias. I will argue that individual implicitly biased behaviours constitute moral harms, but are not themselves moral wrongs. This means that the average person is generally not responsible for implicitly biased behaviour. This argument suggests a shift in focus from the current discourse, which has generally assumed that implicit biases always produce wrongs. I will then challenge another assumption in the moral responsibility literature that presupposes that collective level wrongs are always aggregates of individual level wrongs, and will suggest that collective level wrongs can emerge from collections of mere harms.

1 Introduction

Implicit biases are cognitive associations that operate out of reach of a person’s conscious awareness, such that she is generally unaware of their existence and the influence they hold over her behaviour. These biases are importantly different from beliefs: whereas beliefs reflect what an agent explicitly holds, implicit cognition reflects the unconscious connections and associations a person makes between concepts. For example, Alice may believe herself to hold progressive social views and to be a race-egalitarian. Yet she may harbour implicit biases that contradict her explicit beliefs. These biases can influence her behaviour and decisions, which may have detrimental consequences for those who Alice interacts with. For example, suppose that Alice is on a hiring committee in charge of selecting the best candidate from a pool of applicants. Alice’s implicit biases may influence her to overlook a qualified candidate with a typically African-American sounding name in favour of a less qualified applicant with a
European sounding name (Mahzarin and Greenwald, 2013; Moss-Racusin et al., 2012; Saul, 2012). Alice’s biases may cause her, unconsciously, to choose the less qualified candidate and to thus unfairly discriminate against a more qualified candidate. Many people harbour such biases that they are wholly unaware of, and that they would staunchly and sincerely disavow were they to be confronted by their existence.

Implicit biases thus present a unique quandary for moral responsibility: if implicit biases operate out of reach of our cognitive awareness and control, and if they reveal associations that contradict our explicit beliefs, can we really be responsible for them and for the behaviour they influence? This question is at the center of much of the current literature on the matter. In what follows, I will suggest that this focus is misguided. The moral responsibility discourse presupposes that there is something in these cases of individual implicit bias for people to be held responsible for. I will argue that although there are certain cases where implicit biases do produce wrongs at the individual level, individual people’s implicit biases generally do not produce moral wrongs. The responsibility debate has so far failed to consider this point. The literature tends to presuppose that the implicitly biased behaviour of average individuals constitutes wrongs, and therefore mistakenly focuses on the attribution of responsibility in response to wrongs before considering whether individual’s implicit biases are even the kind of things that produce wrongs. I will suggest that the outcomes of such biases are generally harms, not wrongs. I will then challenge another assumption within the literature, which is that if there exists a wrong at the collective level, it can always be causally traced back to individual level contributions that are wrongs in themselves. I argue that this assumption is not true in all cases, and propose the concept of collective emergent wrongs, suggesting that genuine wrongs can emerge out of aggregates of harms which are not themselves wrongs. The notion of a collective emergent wrong is central to my argument, both in addressing the residual sense that people who are
significantly impacted by harmful implicit biases may be owed some form of compensation or redress, and in shifting the focus of responsibility towards an arena where it can be properly addressed.

2 The Current Debate

Much of the responsibility literature that addresses implicit biases focuses on bias within specific types of relationships. The majority of this literature concerns how responsibility should be allocated within institutionalized settings, for example: within hiring committees, admissions boards, to professors in charge of assigning grades to students, or other such gatekeeper roles. Institutionalized settings such as these provide the clearest cases of determining responsibility for implicit biases. In these settings, individuals undertake a particular institutionalized social role that brings with it a clear set of obligations and responsibilities that are more demanding than what would be expected from an average person who did not occupy such a role. People in these roles have amplified responsibility toward fair and equal treatment, because, in virtue of their role, the harm that their biases could produce would be so intensified. For example, consider a professor who is tasked with grading student papers. The professor accepts more responsibility for her implicit biases than would an average person, because the negative effects of any implicit biases she may hold would have a far greater effect. If the professor grades her student unfairly as a result of her implicit biases, then the professor has wronged her student. This is because the student has a claim on his professor that she treats him in an unbiased manner, and he has the reasonable expectation of fair treatment from her. If the professor flouts this claim, she has wronged the student. Cases of implicit bias within institutionalized settings such as these are therefore relatively straightforward: people who undertake special social roles understand that the job title comes with the added responsibility of understanding and addressing one’s implicit biases and implicitly biased actions. The specific responsibilities, and the
wrong, are thus products of one’s specialized role in an institutionalized setting\(^1\).

Yet much of the current literature suggests that this responsibility transcends the undertaking of specialized roles, and extends equally to the average person in ordinary, day-to-day interactions. The theorists contributing to the discourse talk of responsibility\(^2\) in a way that presupposes that responsibility always exists for the average person’s implicit biases, and their projects are largely concerned with determining a set of exculpating or excusing conditions that are satisfied by implicit biases. These conditions are generally taken to be some combination of the following: control, awareness, and/or attributability. In other words, an agent might not be responsible for her implicit biases or implicitly biased behaviour because these biases operate outside her cognitive grasp, and are therefore out of her control. For example, consider a person who unknowingly suffers from sleepwalking. One night while sleeping, this person walks out of his house and into his neighbor’s yard, where he steals a pair of garden shears and brings them back to his own garage. The somnambulant thief may argue that he should not be held responsible for his actions because they operated out of his control: because he was asleep, he had no power to control his behaviour. He might also argue that he was not responsible for the theft because he didn’t even realize he had committed it; he was unaware of what he was doing and therefore could not be to blame. Or he may attempt to excuse himself from responsibility by saying that his actions were not attributable to

---

\(^1\)It will be useful here to bring up two conceptions of responsibility as discussed by Watson (1996) and referenced by Zheng (2016): responsibility as *attributability* and responsibility as *accountability*. The general consensus in the implicit bias literature is that for an agent to be responsible for her implicit biases, she must be attributively responsible for them: they must be attributable to who she is as a moral agent and be consistent with her explicit values and beliefs. However, those who occupy specialized institutional roles may be held accountable for their biases, while not being held attributively responsible for them. In other words, someone in this special social role might be held accountable for the damage caused by her biases, in virtue of her taking on the social role, even though the content of her biases cannot be said to be attributable to who she is as a moral agent.

\(^2\)A paper by Luc Faucher entitled “Revisionism and Moral Responsibility for Implicit Attitudes” does suggest that what we know about implicit biases might make us want to be ‘revisionists’ about moral responsibility in these cases. Faucher suggests that we might revise our notion of moral responsibility in light of implicit biases by dropping one or more of these conditions on responsibility.
who he is as a person; that his actions in no way reflect his value or beliefs. Suppose that the sleepwalker, when he is awake, explicitly condemns thievery, and would never think of taking even a nickel that didn’t belong to him. He may then argue that because his sleepwalking behaviour was so divorced from his identity and the content of his moral character, that he could not be held responsible for his unconscious actions.

The bulk of the responsibility literature is concerned with the application and satisfaction of these three conditions. Robin Zheng, in her article ‘Attributability, Accountability, and Implicit Bias’, argues that responsibility for implicit biases is dependent on what sort of beliefs an agent explicitly holds. For example, Katie may think of herself as very egalitarian and socially liberal. She explicitly supports equal rights for peoples of all ethnicities and would staunchly and sincerely deny harboring any racist attitudes if questioned about her beliefs. Katie takes a race Implicit Association Test and discovers that she holds racist implicit biases. On Zheng’s account, Katie is not responsible for her implicit biases, nor is she responsible for any implicitly biased behaviour they may cause. Zheng draws on Watson’s distinction between responsibility as accountability and responsibility as attributability (Watson, p. 228) and argues that a person such as Katie would not be responsible because her implicit biases are not attributable to who she is a moral agent. Although Katie may be responsible for her actions in an accountability sense, she cannot be held attributively responsible, because Katie’s biases are not dependent on the beliefs she explicitly holds. Katie is thus not responsible for her racist implicit biases if, upon reflection, she does not explicitly endorse them, because they are not attributable to her as a moral agent.

Washington and Kelly, in their work, ‘Who’s Responsible for This?’ focus their account on conditions of awareness and control. In contrast to Zheng, Washington and Kelly argue that people should be held responsible for their
biases, even if the person is question is unaware of her biases (Washington and Kelly, p. 12). The authors are skeptical that a person’s implicit biases are ever truly out of her control, and thus argue that such biases do not exculpate a person from responsibility. Because they do not take the control and awareness conditions to be satisfied, Washington and Kelly argue that people should be held responsible for their implicitly biased actions (Washington and Kelly, p. 24).

Neil Levy, in his article ‘Implicit Bias and Moral Responsibility: Probing the Data’, concentrates on both the control and attributability conditions. He argues that implicit biases almost always block our control over our actions, and contends that we cannot be responsible for biases or behaviours that are simply out of our control (Levy, p. 10). He further argues that we ought to take more seriously the beliefs that a person explicitly believes herself to hold as what is important for determining responsibility. Levy ultimately concludes that those people who believe themselves to be in favour of racial equality but that hold racist implicit attitudes are not morally responsible for these attitudes because they do not satisfy either the control or attributability condition.

Similarly, Jennifer Saul’s account of responsibility focuses on the conditions of awareness and control. In her 2012 work, “Scepticism and Implicit Bias”, Saul proposes that implicit bias should induce in us something called “bias-related doubt”. This doubt is a form of skepticism, that suggests that we have very good reason to think that we are mistaken about some of our beliefs (Saul, p. 243). Saul notes that for an agent to be responsible for her implicit biases and those actions influenced by them, she must have the capacity to “critically and reflectively correct” for her bias (Saul, p. 260). In other words, an agent must be aware of her biases – that they exist, that they are implicit biases, etc. – and she must have the ability to control these biases once she is aware of them in order to be held responsible. Saul thus contends that an agent’s being aware
of her implicit biases is a necessary condition for moral responsibility. Once an agent acquires knowledge about her implicit biases, Saul claims that she does possess the capability to counteract, and thus control, her biases, and thus, at least to some extent, to control her biases. Her responsibility, then, is to self-consciously implement any such counter-biasing strategies as might be available to her. Saul therefore supports the idea that both control and awareness can be exculpating factors for responsibility in these cases.

3 Responsibility

It is clear, then, that in spite of drawing different conclusions about whether people should, in general, be held morally responsible for their implicit biases, most theorists generally agree about what it would take for the concept of responsibility to apply; the main question is whether the subjects displaying implicit biases meet certain excusing conditions. I argue that this focus is misguided. Before we can begin to apply responsibility, we must consider whether a wrong has been committed such that there should be responsibility to follow it. For a moral wrong to have occurred, it is necessary that some right or claim an agent possesses be violated. For example, if Bob kills Sally, he has wronged her in the moral sense because Sally has a right to life, and thus has a claim against Bob and against all other people to not violate her right. Wrongs are thus committed by the flouting of an agent’s rights, and by the violation of a claim she has to that right. When a person has a right to something, they have a claim against others: a claim that dictates what they can reasonably expect of others in regards to some particular transaction. We all have a responsibility not to wrong others. Therefore, responsibility “follows the wrong”, and is attributed in reaction to a wrong having been committed.

Moral wrongs are importantly distinct from harms. A harm, unlike a wrong,

3Saul ultimately proposes a collective solution, as she doubts whether a resolution of the problems caused by implicit bias could be rectified at the individual level. (Saul, p. 260)
need not entail the flouting of a right or claim. We can think of harms as necessary but not sufficient conditions for wrongs: wrongs always entail harms but harms do not always entail wrongs, and a person can be harmed without having being wronged. To harm someone is to cause them to be in a worse situation than they would have been had you not acted. Harms are actions that negatively affect the interests of people, such that they become unable to pursue, obtain, or enjoy interests that they would have otherwise have been able to. A harm can be as simple as a person stubbing their toe, missing out on an opportunity, or spilling hot coffee over their new corduroy pants. To illustrate the concept of a harm, as distinct from a wrong, consider the following example:

1. Ranjit is driving to school to give a presentation. On the way there, she gets stuck in a traffic jam due to major bridge construction. As a result Ranjit is late to her destination, and misses her presentation. Given that this presentation comprised a major component of her class grade, Ranjit will receive a poor grade. Ranjit has therefore suffered a harm but has not suffered a wrong. The reason she has not been wronged is because she does not have a claim to not miss her presentation; she has no right that there not be bridge construction or claim against there being traffic on her way to school.

As I have briefly outlined above, these theorists and the majority of contributions to the current literature have all focused their efforts on determining conditions of responsibility. The discourse has thus been burdened with a significant assumption: namely, that the attribution of responsibility is even at stake in average cases of implicit bias. In what follows I will suggest that the current discourse has missed the crucial point in the responsibility debate, and has mistakenly assumed that there is responsibility to be attributed in these cases. I will suggest that average cases of implicit bias generally do not produce wrongs, and therefore, that there is no need to talk of conditions of responsibility for these cases. The current debate has, I contend, ‘skipped a step’ in its
reasoning, and has jumped to a discussion of how and in what manner responsibility should be attributed without asking whether there is responsibility to be attributed.

Of concern to my project are implicit biases that are negative in valence, and which produce harmful implicitly biased behaviours. These behaviours can manifest in a variety of actions, but I will consider an implicitly biased behaviour to be an action that is directly causally traced to the influence of a person’s implicit bias, such that this action is likely at odds with the person’s explicit beliefs, the actions are likely an automatic response to a perceived signal, and the person is likely unaware of the real motivation behind these actions. As noted, these behaviours need not be negative or harmful, but I will focus on those that are for the purposes of my project. To illustrate such cases, consider the following examples of implicitly biased behaviour:

2. Sarah is walking down a street in her neighborhood. She sees a person of colour, Matthew; approach her on the same sidewalk. Sarah believes herself to be an egalitarian, and would explicitly deny any accusations of racism or racially biased behaviour. Yet as Matthew approaches, Sarah subconsciously associates seeing a person of colour approaching with ‘danger’, and chooses to cross the street to walk on the opposite sidewalk.

3. Kevin is waiting in line at his local Starbucks. When he reaches the front of the queue, the barista who takes his order is a man who Kevin believes to be homosexual. Although he is unaware that he holds any biased beliefs, Kevin finds himself having difficulty looking his barista, Niko, in the eyes. Kevin subconsciously feels uncomfortable interacting with a homosexual person. He avoids eye contact with Niko during the entire transaction.

4. Miriam is riding the bus to school. The bus is quite empty, and she has a row of seats to herself. She leaves her backpack sitting to her left on the chair beside her, with the zipper undone. As the bus starts to fill,
another student, Cody, sits down close to Miriam. Cody is a different ethnicity than Miriam, and while she would be incredibly surprised to discover holding any racist attitudes, Miriam finds herself zipping up her backpack and moving it onto her lap. Despite her explicit beliefs, Miriam has subconsciously associated Cody’s ethnicity with a heightened risk of criminality.

I am sure that we can all agree that these examples and those like them are not desirable situations. Suffice to say, implicit biases can clearly lead to some very unfortunate and unpleasant scenarios. These examples involve some form of harm: each of the recipients in these cases likely feels justifiably harmed by the other person’s actions. However, I argue that it does not follow that the actions described in examples 2-4 are moral wrongs. As I discussed above, harms and wrongs are importantly distinct from one another. Committing a wrong entails that another person’s rights or claims have been violated. So, for these examples to be cases of genuine wrongs, it would require that each of the recipients have had a claim against the other to abstain from the described behaviour. It would require that some person have a claim against another to not do these things, and would require that they have a right against another person to refrain from such. For example, for Matthew to have been wronged by Sarah, it would require that Matthew have a right against Sarah to always walk on the same side of the street as him. It would entail that Matthew have a claim upon Sarah to continue share the same sidewalk with him, and to refrain from crossing the street. I think it is clear that although the behaviour in these examples is undesirable and sometimes unpleasant, it is not the kind of behaviour that anyone has a right against. These kinds of acts are simply not the kind of thing that we can claim of each other: it is simply not reasonable for me to demand that another person always maintain eye contact with me, or always walk on the same side of the sidewalk as me. Given this, I conclude that the implicitly biased behaviour in examples 2-4, while they may cause harms,
do not thereby constitute wrongs. Given that the most common outcome of implicit biases for the average person are similar kinds of microaggressions, I contend that in general, the implicit biases of average people do not produce wrongs at the individual level. For this reason, I suggest that the majority of the implicit bias responsibility discourse has been talking about the wrong thing. Theorists such as Zheng, Washington and Kelly, Saul, and Levy have taken the crucial dilemma of responsibility to be which conditions of responsibility are satisfied by implicit bias, taking for granted that there is responsibility to be attributed because the average person’s biases produce wrongs. I suggest that these approaches miss the point: it is misguided to talk of conditions of responsibility when there is no real responsibility to attribute. And there is no real responsibility to attribute because the cases in question involve harms, not wrongs.

4 Emergent Wrongs

We are now left with a puzzle. If one is to accept my arguments thus far, we are left with the uncomfortable consequence that those people who are the victims of other’s implicit biases have not been wronged by them, and those who act upon their implicit biases are not responsible for these biases or for the consequences of their actions upon others. This would mean that those who experience harmful implicit biases are not owed any compensation or redress for their damages. And this likely will not sit well with us. Many of us are left with the nagging feeling that the recipients of implicitly biased behaviour must be owed some form of compensation or redress. I agree with these contentions. However, I will suggest that the solution to the problems created by implicit biases cannot be found at the individual level: we must look instead to the collective level.

It is important to note that this claim encompasses only the kinds of behaviour I outlined in my examples. There are certainly cases that start out like these examples and progress, or cases that involve but are not limited to such behaviour, where a genuine wrong occurs. For example, if Sarah not only crosses the street but also makes racist comments to Matthew, then she has transgressed the boundaries of my claim. Further, such cases are generally not caused by implicit bias alone.
I would like to draw attention to another assumption that is found in the responsibility literature, namely that if there exists a wrong at the collective level, it is traceable back to individual level contributions, one or more of which are wrongs themselves. For example, consider pollution. Pollution is a wrong that has been contributed to by many different sources. To vastly oversimplify the problem, suppose that pollution was caused by 10 different people, who each littered, left their cars idling for hours, and burned lots of toxic chemicals in their backyards. In this case, each of these 10 people committed an individual wrong, and they each contributed to amass into a bigger collective wrong. It is often assumed that all cases of collective wrongs are like this. I argue that this traditional way of thinking is incorrect. I propose the concept of an emergent wrong. An emergent wrong is a collective moral wrong that is not causally traceable back to a wrong committed by any one individual contribution. I suggest that it is possible for a bona-fide wrong to emerge from an aggregate collection of individual contributions that are not in themselves wrongs. The harms caused by implicitly biased behaviours (outside of contexts where people occupy specific institutional roles) are an excellent example of emergent wrongs. For example, let’s return to Matthew from example 2. I have argued that the implicitly biased behaviour that Matthew was subject to did not constitute a wrong. Sarah may have harmed Matthew, but she did not wrong him. However, suppose that Matthew has experiences just like this one multiple times a day, every day. It would not take long for Matthew to justifiably feel as though he were being discriminated against. And being discriminated against is being wronged. Discrimination is pernicious and damaging, and it also entails the unfulfilled expectation of fair treatment and minimal decency from others. We all have a right to not be discriminated against, and thus a claim on all others to not treat us in a discriminatory way. Matthew has a right to not be discriminated against, and so when he is discriminated against, he has been wronged. Yet as in cases of implicitly biased behaviour such as Sarah expresses in example
Matthew has not been wronged by any one person. I uphold my previous claim that the individual contributions that may have amassed into the wrong of discrimination are not themselves wrongs. Instead, I argue that cases like these are instances of emergent wrongs. While this may sound unorthodox, it is not an entirely new and unusual way of thinking about wrongs and responsibility.

To illustrate this claim, consider the following example:

5. Alex owns an apartment unit. She knows that it will be more profitable for her to rent out her apartment on Air BnB than it will be to rent to long-term tenants, so she lists her unit on Air BnB. Alex has not done anything wrong by doing so. However, imagine that thousands of apartment owners do the same thing, hypothetically in a city called Vancouver. Because all of the available apartments in the city are being rented on Air BnB instead of to long-term renters, a hopeful tenant, Sam, is unable to find a place to live. Sam has a right to a place to live, and this claim has gone unsatisfied. Therefore, Sam has suffered the wrong of being unable to find a living space and to fulfill his claim to such. However, the claim Sam has to find a home is not directed towards Alex, nor is it directed towards any one person at all. Sam has suffered an emergent wrong that has grown from an aggregate of individual contributions that are not each wrongs themselves.

Examples like this show that thinking of collective wrongs as emergent is not entirely unprecedented. The wrongdoing that occurred in this example lies at the level of the institutional structure, and not at the level of the individual. Alex has acted permissibly within the framework of the institutional structure, and as such, has not committed a wrong. I contend that the Air BnB example is much like cases of average people’s implicitly biased behaviour: although wrongs are not committed at an individual level, a genuine moral wrong can emerge from a collection of harms. Incorporating the concept of emergent wrongs into how we think about moral responsibility for implicit bias allows us to deal with
the residual feeling that targets of implicitly biased behaviour are sometimes genuinely wronged, and are thus owed some form of compensation for these damages. We can now acknowledge the significance of the wrongs that sometimes emerge from implicit bias, and recognize the redress owed to those who suffer them, without conceding that the average individual is responsible for her implicit biases.

Thinking of responsibility at the collective level also importantly points us to a place where meaningful action and appropriate redress might be taken. As Saul suggests in her 2012 work, the best way to rectify the damages of implicit bias incurred at the individual level seem to lie at the collective level. Saul suggests that the best way to combat the influence of implicit biases is to re-shape our social world: we need to integrate our neighborhoods, and support varied genders and ethnicities in positions of power. She writes, “The only way to be fully freed from the grip of bias-related doubt is to create a social world where the stereotypes that now warp our judgments no longer hold sway over us” (Saul, p. 260). Saul thus proposes a major restructuring of the social institutions and morals that inform our biases. I agree with Saul that the best way to address the damages caused by individual level implicit biases is found at the collective level. Only many people working together, with the aid of norm-governing institutions can adequately address a problem as detrimental and far reaching as implicit bias. This is further reason for thinking that discussions of individual moral responsibility for implicit biases simply miss their mark.

Because the individual contributions that compose collective wrongs are not wrongs themselves, I have argued that individual people are not responsible in the moral sense for these contributions. Although these contributions may be harms, and although they may amount to a wrong, they are not wrongs themselves. It is uncontroversially accepted that people are held responsible for wrongs they commit. When Mike kills James, Mike is responsible for the
wrong he commits. In such a case, Mike would be held both legally and morally responsible (provided he does not have an exculpating condition he can appeal to, which for the purposes of this example I will assume he has not). He may be responsible in a variety of ways: He may be responsible for the damage done, and may be required to “pay for” his actions. He may make himself liable to punishment and a certain loss of his own rights, for example, by making himself liable to imprisonment and losing his own right to liberty. He may also make himself liable to monetary compensation in the form of fines and reparations.

Cases of emergent wrongs are less clear. Although we traditionally want to say that if there exists a wrong, someone must be responsible for it, I have argued that no one person is responsible for an emergent wrong. This is because no one person’s contribution to the wrong constitutes a wrong. Therefore, I conclude that individual people are not responsible for their contributions to emergent wrongs. Although we have a responsibility to not wrong others, we do not have a responsibility to refrain from crossing streets or moving backpacks. More specifically, we need not hold that people have such responsibilities in order for us nonetheless to maintain that people who are systematically suffering the effects of implicit bias have also suffered a wrong.

5 Conclusion

Implicit biases and the actions they influence pose a unique challenge for responsibility. I have diverged from traditional ways of thinking about responsibility for implicit biases and about moral wrongs, and have suggested a new trajectory for thinking about responsibility. Through analyzing the types of implicitly biased behaviour of an average person in ordinary, everyday interactions, I have concluded that implicit biases generally produce harms and not wrong. Given this, I have suggested that much of the previous literature has been misguided. I have instead proposed the concept of emergent wrongs, and suggested a revisionist approach to how we think about moral responsibility for implicit biases.
References


