Learning Implicit Biases from Fiction

Kris Goffin & Stacie Friend

Abstract: *Philosophers and psychologists have argued that fiction can ethically educate us: fiction supposedly can make us better people. This view has been contested. It is, however, rarely argued that fiction can morally "corrupt" us. In this paper, we focus on the alleged power of fiction to decrease one's prejudices and biases. We argue that if fiction has the power to change prejudices and biases for the better, then it can also have the opposite effect. We further argue that fictions are more likely to be a bad influence than a good one.* [6431 words]

Keywords: fiction and ethics; implicit bias; moral education

1. Introduction

Many philosophers and other scholars in the humanities have maintained that works of fiction are ethically educational, for instance because they refine our moral perception (Beardsmore 1971; Nussbaum 1990); promote ethical thinking (Diamond 1991); enhance empathy and perspective-taking (Nussbaum 1996); or clarify our understanding of moral principles (Carroll 1998).[1] Recently, a number of psychologists have offered evidence suggesting that reading literary fiction increases empathy, compassion, and prosocial behavior.[2] Such studies have prompted media speculation that engaging with fiction might "make us better people."[3]

[1] For an overview see Gaut (2013).

[2] Meta-analyses (Mumper & Gerrig 2017; Dodell-Ferrer & Tamir 2018) show weak positive effects.

[3] The phrase appears in headlines in (e.g.) the *Washington Post* and *BBC Future*, referring to Kidd & Castano (2013).

Although there is good reason for skepticism about these claims, for present purposes we do not deny that moral improvement through fiction is possible.[4] Instead, we argue that if fictions can have this effect, they can just as easily—and via the same mechanisms—have the opposite effect, corrupting rather than bettering us. And we suggest that at least in certain respects, fictions are more likely to be a bad influence than a good one.

In this paper we focus on one kind of potential moral improvement which has been claimed for fiction. Both philosophers and psychologists have suggested that engaging with fiction may reduce prejudice or bias against social groups, citing a variety of potential mechanisms. We take for granted here that these mechanisms can decrease harmful biases. We argue, however, that they are far more likely to increase them. In other words, there is a dark side to learning from fiction, which has not received the attention it should. We conclude, though, on a brighter note: there is little reason to think that fictions cause significant harm.

Before turning to that argument, one clarification: Because the philosophical arguments and psychological studies we discuss presuppose a Western, and especially Anglo-American, perspective, our examples are consistent with that approach. However, we take our broader conclusions to apply to fiction produced in any culture.

2. Implicit Bias

There are a variety of ways in which fictions might reduce prejudice toward disadvantaged social groups, by exploiting mechanisms that have been studied by

[4] Skepticism about the evidence arises because some studies have not been replicated (see e.g. Panero et al. 2016; Samur, Tops, & Koole, 2018) and more recent studies have found no effects (Wimmer et al. 2021, n.d.). There are also philosophical reasons for skepticism (see, e.g., Harold 2005; Landy 2008; Currie 2020).

psychologists in standard anti-bias interventions. These include portraying people who are counter-stereotypical exemplars of a social group; providing relevant information about a group or culture; depicting positive intergroup interactions; prompting perspective-taking or empathy; and evaluative conditioning, in which a group is associated with positive emotional stimuli.[5] Such interventions often involve an imaginative component—for instance, participants are asked to *imagine* counter-stereotypical exemplars or positive intergroup interactions—and some use stories to guide such imaginings.[6]

A few psychological studies have more closely examined the role of fictional narratives, suggesting that effects on prejudice might be mediated by features they take to be characteristically associated with fiction as opposed to nonfiction, such as role-taking (Hakemulder 2001; 2006); the degree of transportation or immersion in the world of the story (Johnson 2013); or the extent of identification with characters (Moyer-Gusé et al. 2019; Vezzali et al. 2014). Johnson (2013), for example, appeals to Mar and Oatley's (2008) argument that narrative fiction prompts imaginative simulations, so that "readers engage in a special form of perspective-taking that should foster empathic growth and prejudice reduction" (78).

Philosophers have advocated a role for some of these same mechanisms. For example, Martha Nussbaum (2013) argues that literary narratives are essential to a just society insofar as they enhance empathy, which in turn renders us more compassionate toward others who are unlike ourselves—including those unjustly oppressed by societies. Nussbaum defines empathy "as the ability to imagine the situation of the

[5] For overviews of empirical studies of anti-bias interventions, see FitzGerald et al. 2019; Lai et al. 2014; Paluck & Green 2008.
[6] See, e.g., Intervention 4 reported in Lai et al. 2014.

other, taking the other's perspective" (145). Empathy need not produce compassion, as illustrated by the empathetic sadist. Nonetheless, Nussbaum claims that there is a connection, relying on experiments by the psychologist Daniel Batson (2011), which showed that that participants who were instructed to vividly imagine the plight of a character later exhibited a greater tendency help another person (e.g., picking up a pencil when someone drops it).

Noël Carroll (2014) describes a process very much like evaluative conditioning. He proposes that fictions can "recalibrate" our attitudes "by reorienting our emotional responses, by associating a previously indifferent or aversive stimulus with a different, preexisting paradigm scenario" or "existing set of emotion elicitors" (47).[7] Carroll illustrates the proposal with Harriet Beecher Stowe's novel *Uncle Tom's Cabin* and Jonathan Demme's film *Philadelphia*. In these fictions, he suggests, Black or homosexual people (respectively) are "linked" with the familiar scenario of a loving family. By associating members of the maligned group with a stimulus that typically elicits positive emotions, the reader learns to see them as elicitors of positive emotions. The result of changing the valence of emotion in this way is that "persons regarded previously as objects of disgust within a culture, are transformed into objects of moral respect" (Carroll 2014, 44).

Carroll's proposal is particularly interesting because evaluative conditioning is a form of implicit learning—that is, learning without awareness—which is widely agreed to be effective for reducing *implicit bias*.[8] We follow Jennifer Saul in defining implicit bias as "unconscious biases that affect the way we perceive, evaluate, or interact with

[7] The concept of a *paradigm scenario* is from de Sousa (1987).

[8] Harold (2005) also discusses "automatic processes" which may be morally harmful, though not the processes we consider in this paper.

people from the groups that our biases 'target'" (Saul 2013, 40). It turns out that many people who explicitly reject prejudicial attitudes or stereotypes nonetheless display such unconscious biases.

Much of the evidence for this conclusion comes from Implicit Association Tests (IATs) (see, e.g., Banaji & Greenwald 2013; Greenwald & Banaji 1995; Greenwald et al. 2009). In an IAT, participants are asked to quickly categorize negatively and positively valenced words with other words or images, for example pictures of white and Black faces.[9] Measuring the speed and accuracy of one's performance is taken to measure implicitly endorsed associations.[10] The majority of people seem to associate white people with positive words and Black people with negative words (see, e.g., Xu et al. 2014). This effect is also true of people who explicitly reject racist claims when asked.

In the past few decades, social psychology has also produced behavioral studies that indicate widespread implicit bias. For example, several studies have examined the "Shooter Bias" (Correll et al. 2002; Mekawi & Bresin 2015). In a computer simulation task, participants are asked to "shoot" when they encounter a person with a gun. The people they see in the simulation either hold a gun or an object which resembles a gun. Participants are significantly more likely to shoot unarmed Black men than unarmed white men, and to fail to shoot armed white men than armed Black men.

It is important to note, however, that implicit bias is a heterogeneous phenomenon, which may involve a variety of different psychological features and processes, prompting distinct behaviors and requiring distinct mitigation strategies (Holroyd and Sweetman 2016). For example, implicit stereotypes—such as associating

[9] This test is available at https://implicit.harvard.edu/ .

[10] The use of IATs to uncover biases has been subject to widespread criticism in philosophy and psychology. However, there is reason to think that much of the criticism is misguided. See Brownstein, Madva and Gawronski (2020) for thorough discussion. See also FitzGerald et al. (2019), 2-3.

Black people with athletic ability or Asians with mathematical aptitude—can come apart from implicit attitudes, that is, positive or negative feelings toward a social group (Amodio and Devine 2006; see Holroyd and Sweetman 2016 for discussion).

Because implicit biases of these kinds have pernicious social effects, efforts to reduce them are widely considered a moral and political imperative. We will therefore focus on the potential for fiction to influence implicit biases in what follows.

3. Moral Impartiality

In this section we argue that features of fiction thought to reduce bias are just as likely to increase it, and that different features of a work may have opposite effects. The claim that fictions can have both good and bad effects, and via the same mechanisms, might seem too obvious to need stating; however, it is rarely acknowledged in discussions of fiction's impact on prejudice that the relevant mechanisms are at best morally impartial.

Consider, Carroll's proposal concerning evaluative conditioning. *Evaluative conditioning* or *evaluative learning* refers to a process whereby one's positive or negative affective evaluation—one's liking or disliking—for a stimulus is changed by its co-occurrence with another stimulus which one perceives as good or bad. The perceived valence (i.e., perceived goodness/badness) of a particular stimulus changes because it is paired with another positive or negative stimulus.

An extreme example comes from Anthony Burgess's 1962 novel *A Clockwork Orange* (or the Stanley Kubrick 1971 film adaptation), in which the protagonist Alex's positive attitudes toward violence (and Beethoven) are altered by exposing him to depictions of violence with a Beethoven soundtrack, while a drug induces disgust and extreme discomfort. A more familiar example is a child's learning to be afraid of touching a flame, after the co-occurrence of touching with the sensation of pain. In both cases the valence of a previously positive stimulus changes because it is paired with a negative stimulus.

There is some disagreement about the cognitive architecture that underpins evaluative conditioning. According to some, the process is purely associative; according to others, genuine representations are involved (see Mitchell et al. 2009). However the underlying process is described, there is no doubt that evaluative conditioning is a genuine phenomenon (see, e.g., De Houwer, 2007; De Houwer et al. 2001.) There is equally widespread agreement that it is automatic: fast, unconscious and beyond our control. The (repeated) co-occurrence of the stimuli automatically leads to evaluative conditioning.

It is plausible that fictions can reduce biases via evaluative conditioning, because it makes no difference to this process whether the stimuli are real or fictional. The process is automatic, generated by repeated co-occurrence. In fact, studies in experimental settings typically deploy depictions of stimuli, rather than real-world interactions. If fictions similarly prompt readers to associate certain social groups with positive emotional elicitors, this may reduce bias.

However, the same mechanism can easily have the opposite effect. One reason for investigating the use of evaluative conditioning as a tool for mitigating harmful biases is that it seems to be a primary mechanism by which we acquire such biases in the first place (see Olson & Fazio 2001; 2006). Not all fictions associate certain social groups with positive emotional stimuli. This is obvious when the fiction is not merely implicitly but also explicitly racist. For example, in *Birth of a Nation,* D.W. Griffiths cuts together alternating images of white Klansmen behaving heroically and Black freedmen attacking people. These imagistic associations plausibly prompt automatic associations that reinforce other racist elements, such as the stereotypical portrayal of Black people as corrupt and violent, or the use of close-ups to invite empathy with white characters.

Birth of a Nation is a good example of a fiction in which multiple features work together to influence bias, in this case negatively. *Philadelphia* provides another example, with the opposite aim. Demme does not just associate the homosexual characters with the image of a loving family; he also purposely portrays these characters in a counter-stereotypical way. Most fictional narratives exploit a variety of different features which can have an impact on implicit bias. Although these typically work in the same direction, they need not.

Carroll's example of *Uncle Tom's Cabin* illustrates the point. Because the novel played a role in motivating abolitionism in the American North, it is often held up as an example of fiction's capacity to promote positive moral change. However, it has also been widely condemned for propagating harmful stereotypes. For instance, James Baldwin argued in his famous essay "Everybody's Protest Novel" (1998) that the narrative is a cowardly, sentimentalist condemnation of racial oppression. The character of Uncle Tom, in particular, is depicted as a good person in virtue of his humility toward white people, generating the familiar stereotype of an "Uncle Tom," a Black person who is overly subservient toward whites. Baldwin claims that Tom's humility functions as a "redemption" of his blackness.

In different terms, it is not the association of Tom with a loving family, but with *humility*, that creates the positive attitude toward him. This link impacts the affective dimension of implicit bias, transforming a negative feeling about Black people into a positive one. But because the attribution of humility reinforces the power dynamics between white and Black people, readers are implicitly learning a harmful stereotype. The link created by *Uncle Tom's Cabin* constitutes a racialized attitude.[11]

We therefore maintain that the very same psychological mechanism that brings about a positive moral change can also reinforce harmful biases.[12]

4. The Dark Side

So far, we have argued that the mechanisms by which fictions might influence bias are morally impartial. And we think this is right; taken by themselves, they can be used for good or bad. However, there are several reasons to think that overall, the changes wrought by engaging with works of fiction are more likely to be negative.

Many arguments for the ethical value of engaging with fiction cite classic works of literature. Although these are less likely to traffic in crude stereotypes of the sort Stowe creates—with some notable exceptions, such as Fagin in Dickens's *Oliver Twist*—they are rarely free of more subtle biases. As feminist and postcolonial critics have highlighted, most fictions in the Western literary canon reflect the perspective of white middle-class men, the dominant group (see e.g., Gilbert & Gubar 1979; Said 1993; Spivak 1985; Wekker 2016). The exceptions (such as Austen, Eliot, or the Brontë sisters) are middle-class white women. The chances are therefore high that canonical works of literature will reinforce harmful implicit attitudes towards less privileged social groups. After all, it is only lately that awareness of implicit bias, along with efforts to promote fictions by women and ethnic minority authors, have become prevalent.

[11] To be fair, Carroll acknowledges the possibility that the mechanism he describes can be used for bad as much as good; we return to this in §5.

[12] Or it can be the bias itself which can have both positive and negative effects, such as the Uncle Tom humility-related bias.

For example, until relatively recently there was little recognition of the racist/colonialist biases in Charlotte Brontë's portrayal of Bertha Mason, the "madwoman in the attic," in *Jane Eyre*.[13] Mason is a Creole from Jamaica, the daughter of European settlers, but with ambiguous racial features; her darkness of hair and skin tone are highlighted. Rochester describes her as a sensual, savage seductress, and himself as her innocent victim. Because many (white) readers today remain ignorant of the critical discussion inspired by Jean Rhys's groundbreaking postcolonial novel *The Wide Sargasso Sea*—which creates a backstory for Mason—they probably still fail to pick up on the subtle associations between her racially mixed heritage on the one hand, and sensuality and savagery on the other (Spivak 1985). Nothing so egregious appears in (say) Austen, who includes condemnations of slavery in *Mansfield Park* and *Emma*; or in Eliot, who uses *Daniel Deronda* to denounce her contemporaries' anti-Semitism. Still, both authors have been subject to critique by post-colonialists who detect implicit prejudice, for example in Austen's relatively sympathetic treatment of the slaveholder Sir Thomas Bertram.[14]

Similarly, critics have pointed out that people with Autistic Spectrum Disorder (ASD) are often portrayed in a stereotypical way in popular fiction, creating stigma and negative attitudes (van Goidsenhoven, 2020; Murray 2008; Osteen 2008). Snyder and Mitchell (2000) describe how fiction shapes the way society sees disabled and neurodiverse people. The movie *Rainman*, for instance, helped to create a stereotype of how an autistic person behaves, by muttering constantly, avoiding eye contact, and displaying unusual mathematical and memory skills. Only recently have works like *The Curious Incident of the Dog in the Night-Time* provided an alternative narrative. Fictions

[13] The phrase is from Gilbert & Gubar (1979).

[14] See, e.g., Ferguson (1991) on Austen and Meyer (1993) on Eliot.

that portray autistic people are more likely to reinforce rather than reduce prejudice against those with ASD.

Or consider the "Prince Charming" trope, common in fairy tales and Disney movies: the prince is heroic and chivalrous and typically rescues a helpless female. Prince Charming is clearly a fictional construct, not to be met with in real life. However, Rudman and Heppen (2003) found that women whose implicit romantic fantasies involved a Prince Charming displayed systematically less ambitious career aspirations, in line with stereotypes of women as less successful professionally. This was so even when their implicit romantic fantasies, as measured by an IAT, diverged from their conscious beliefs about what they wanted in a romance, as measured by self-report. Rudman and Heppen call this the "glass slipper" effect.

In fact, it is unnecessary to cite works which contain explicit stereotypes or are written from privileged perspectives to make the point. As we have seen, empirical research suggests that even those who consciously reject such stereotypes are subject to implicit biases. This includes authors and fiction filmmakers. It is therefore more likely than not that implicit attitudes will influence their creative processes. And there is evidence that this is exactly what happens.

For instance, a study by Weisbuch, Pauker and Ambady (2009) demonstrated effects of *bias contagion* in audiences of popular television series. Bias contagion is the process by which we unconsciously adopt biases as a result of observing the nonverbal behavior of others, including very subtle behavior of which we are not aware (Willard et al. 2015; Wiesbuch & Pauker 2011). What is striking is that the study found bias contagion even in shows created with the intention of reducing harmful stereotypes. The researchers first analyzed eleven different shows for negative, nonverbal behavior toward Black characters. They found that in series like *CSI* and *Grey's Anatomy*, in which women and people of color are portrayed as agents, scientists, and doctors—that is, counter-stereotypically—the body language of the white actors often betrayed implicit negative attitudes. For example, they subtly shifted away from, or squinted slightly more at, Black actors. The researchers showed participants several scenes in which such biased behavior was displayed. Changes in attitude were measured by an IAT as well as by asking participants how much they "liked" a particular character. The results indicate that bias contagion occurs even in response to these well-meaning fictions.

Similar points can be made about the proposal that fictions, such as great works of literature, might reduce prejudice by enhancing empathy or perspective-taking (Nussbaum 1990, 1996, 2013). The claim that fiction is morally improving in virtue of generating empathy has been contested.[15] But even if it is true that engaging with fictions renders people more empathetic, this does not necessarily imply a positive moral change. The reason is that empathy itself is biased.[16] One is more likely to feel empathy toward ingroup members than outgroup members (Cikara et. al. 2014, 120). Moreover, there is a widespread empathy bias in favor of white people, including among people of color (Xu et al. 2009). Empathy biases seem to favor already dominant groups in society, in the same way as biases that rely on stereotypes. As we have seen, canonical literature is typically written from socially dominant perspectives. If reading such literature enhances empathy, then it is more likely to foster than weaken the existing biases toward ingroup members and the socially advantaged.

Biased empathy is problematic because it encourages what Merton (1968) calls the "Matthew Effect": privileged people become more privileged by acts of benevolence.

[15] Recent studies that explored this question in depth found no effects on either cognitive or affective empathy even for lifetime readers of fiction. See Wimmer et al. 2021, n.d.

[16] For discussions of the empirical literature that support this claim, as well as arguments about the ethical implications, see Bloom 2016; Prinz 2011.

Social science research suggests that social inequality is mainly enforced, not by acts of explicit discrimination, but by acts of benevolence (Banaji, Bazerman & Chugh 2003). If a person in power is more apt to help people who are like them, then even if these acts are altruistic, the socially unequal power structure remains in place. For instance, nepotism in job hiring is typically an act of benevolence; one wants to help one's friends. However, this excludes people who are not part of the right network, most often minorities and those from different social classes.

We have already seen that works of fiction are more likely to reinforce biases, if only implicitly. Given that our empathetic capacities are already biased, increasing the tendency to greater empathy is more likely to do harm than good. In this respect, most works of fiction, from novels in the literary canon to popular television series, are probably morally harmful rather than morally improving.

5. Resistance

Are we being too pessimistic? Carroll acknowledges the possibility that the mechanism he describes can be used for bad as much as good. However, he suggests that people are capable of resisting negative influences:

> Recalibration can be employed by angels or demons. Leni Riefenstahl's invocation of community in *Triumph of the Will* is perhaps the most notorious instance of the latter. However, that does not entail that audiences are at the mercy of just any invocation of a positive paradigm scenario. For, once the emotion process is set in motion by the invocation of a paradigm scenario, it is still open to deliberative monitoring in light of whether or not the scenario fits

> with or is coherent with our preexisting cognitive stock. Paradigm scenarios can be resisted. (Carroll 2014, 55)

That is, we are in little moral danger from works like *Triumph of the Will* because, given our existing views, we will reject Riefenstahl's attempt to link the Nazis with a positive image of community.

Carroll's suggestion calls to mind Tamar Szabó Gendler's (2000) account of imaginative resistance. *Imaginative resistance* refers to the psychological obstacle people face when trying to imagine certain events in fiction. Confronted with a story which includes the sentence, "In killing her baby, Giselda did the right thing; after all, it was a girl" (Walton 1994, 37), we find it difficult to accept the moral judgment or imagine a world in which it is true. According to Gendler, this is not because we are unable to imagine as prescribed, but instead because we refuse to do so. And we refuse to do so whenever we assume that the ethical perspectives manifested in the work are meant to be "exported" to the actual world. Therefore, if a work manifests racist, sexist, or other prejudiced attitudes, we are likely to refuse to go along. In this way, imaginative resistance can protect us from the biasing effects of fiction.

Setting aside debates over imaginative resistance as a phenomenon (see Tuna 2020), it is fair to say that Carroll overestimates the capacity of audiences of fiction to overcome the effects of implicit learning through resistance. This kind of learning is typically unconscious and automatic, often responding to very subtle features. It is unlikely, for example, that viewers of *CSI* are in a position to resist bias contagion, since they are unaware of the biased nonverbal behavior influencing their attitudes.

Moreover, imaginative resistance itself can be biased. Adriana Clavel-Vazquez (2018) argues that we experience more imaginative resistance to the violent, immoral actions of “rough heroines” than we do in response to “rough heroes.”[17] Creators of fictions about immoral protagonists use a variety of techniques to weaken our resistance; their skill in overcoming this resistance and eliciting sympathy for the perpetrators of horrendously unethical acts is arguably an artistic achievement (Eaton 2012). However, the standard examples all involve male characters, such as Humbert Humbert, Tony Soprano, or Omar Little. Few readers or audiences are as sympathetic to Amy Dunne in David Fincher’s film *Gone Girl* or Cersei Lannister in *Game of Thrones*. Clavel-Vazquez (2018, 207) argues that “we resist allying with rough heroines because in being morally transgressive, they break with gender norms and expectations.” In other words, the explanation for the asymmetry is implicit sexism: we do not like characters who violate our gender-based assumptions.

The contrast highlighted by Clavel-Vazquez is not limited to fiction. As Kate Manne (2017) has argued, there is a widespread tendency to feel disproportionate sympathy toward powerful men. Manne calls this *himpathy*. For instance, in cases of sexual assault people often display greater sympathy for perpetrators, especially if they are men in positions of authority or power—such as Bill Clinton or Brett Kavanaugh—than for the victims, whose accusations are frequently greeted with skepticism. Although sympathy is typically construed as morally positive, Manne argues that sympathy is socially biased in ways that are ultimately harmful. Fictions which portray rough heroes might therefore tend to reinforce our existing tendencies toward himpathy.

In summary: If imaginative resistance is itself biased, then it is not a reliable tool for evading negative influences from fiction. Imaginative resistance works only if the

[17] The phrase *rough hero* is originally Hume’s in “The Standard of Taste.”

biases manifested in the fiction are in conflict with one's pre-existing biases. And even then, it will not have an effect when the acquisition of bias is due to implicit learning.

A different way to "resist" bias is to widen one's reading, consciously trying to engage with fiction from other parts of the world or from underrepresented groups in one's own culture. Exposing oneself to other perspectives is likely to be cognitively beneficial. However, it will not erase the potential for increased bias. First, biases are widespread; though the specific biases may differ from culture to culture and group to group, they exist nonetheless. Second, Western biases are dominant globally, reflecting colonial power imbalances. Western fiction, in the form of Hollywood movies and curricula skewed to the Western canon, is similarly dominant.[18]

Engaging with fictions produced by non-Western or oppressed groups is a positive development. However, it would be naive to think that a few works of fiction could destroy the dominance of Western biases, let alone biases toward any and all social groups.

6. The Bright Side

If we are right, engaging with fiction is likely to have a negative influence with respect to the reinforcement of harmful biases. Does this mean we should stop reading novels or watching movies and television—or even more radically, censor their production, in a modern-day equivalent of Plato's banishment of the tragic poets?

The answer to this question is (with all due respect to Plato) certainly no. First, the fact that works of fiction are likely to reinforce implicit biases does not mean that they cannot be ethically valuable in other respects. None of our arguments undermines

[18] Mudimbe (1988) uses the concept of the "colonial library" to refer to the cultural and epistemological aspects of colonization, of which the ubiquity of Western fiction is one dimension.

the more traditional claims for fiction's ethical value mentioned in §1, such as that fictions refine moral perception, promote ethical thinking, clarify moral principles, or increase prosocial behavior. Furthermore, we have already drawn attention to works which may influence implicit bias in different directions, such as *CSI* or *Gray's Anatomy,* which offer counter-stereotypical portrayals at the same time as they increase biased attitudes. Nothing we have said excludes the possibility that fictions do more good than bad, all things considered.

This observation raises the question of how we make such "all things considered judgments": how do we weigh different ethical benefits and deficiencies against each other in evaluating a particular work?[19] We do not think there are any general rules; we make these judgments on a case-by-case basis. And we expect that most artworks, like most people, are ethically complex: good in some ways and bad in others. What we want to urge here is that some ethical impacts may not be apparent at first glance. Without empirical investigation, we would not know that shows like *CSI* transmit implicitly biased attitudes despite other positive ethical features.

Second, ethical value is not the only kind of value. Many works of fiction are aesthetically or artistically valuable, and our lives would be impoverished without the experiences afforded by literature, cinema, and so on. To say this is not to assume that the aesthetic is autonomous from the ethical.[20] Even if ethical defects (sometimes or always) constituted aesthetic defects, it would not follow that an ethical defect—such as manifesting bias—outweighs other meritorious features of a work (Gaut 1998). A useful analogy might be the value of power or money. Having either is not intrinsically morally

[19] Thanks to an Editor for pressing us on this question.

[20] For an overview of the debate over autonomism, see Gaut (2007).

bad, but it will not make you a better person. It will probably make you worse. The same goes for fiction.

Third, our claim that fictions can reinforce negative biases is a contingent one. It turns on the fact that most of the works of fiction we consume manifest at least some degree of bias. Is there anything distinctive about fiction that makes it more likely that it will reinforce implicit biases?[21] There are features standard for fiction that could have this effect.[22] For example, most works of fiction are narratives designed to transport readers or audiences to imaginary worlds, to engage them affectively and attentionally, and so on. There is empirical evidence that the persuasive effects of narrative are increased to the extent that we are transported into the story (see, e.g., Green and Brock 2000). Some psychologists also have suggested we are less likely to scrutinize fiction than nonfiction, thereby rendering us more susceptible to persuasive effects (see, e.g., Prentice and Gerrig 1991). These differences are themselves contingent, however; they are a function of conventional features of fiction and our ways of engaging with it. Many works of nonfiction are transporting narratives, and thus can be expected to have similar effects.[23]

Finally, even if fiction were inherently more bias-prone than nonfiction, giving up fiction just to avoid bias would be an enormous sacrifice for a questionable gain. After all, fiction is not the only medium which reinforces bias. So a fiction-free life would not guarantee a bias-free life. We are socialized into different attitudes from a very young age, through upbringing, education, and interactions with other people. Biases against socially disadvantaged groups are so widespread that reading a novel or watching a film

[21] Thanks to an anonymous reviewer for encouraging us to address this question explicitly.

[22] On the concept of 'standard features' of fiction, see Friend (2012).

[23] In fact Green and Brock (2000) found no difference between fiction and nonfiction as far as transportation and persuasive effects.

will make very little difference. In fact, it is likely the causal direction is the other way around: fictions manifest and reinforce biases because creators and audiences are already subject to those biases.

We therefore do not condemn fiction. The main claim of this paper is conditional: If fiction can effect positive moral change by affecting biases, it is more likely to be a bad influence than a good one. But given the many other sources of bias, fictions probably do very little ethical harm. Implicit bias is a pervasive phenomenon, deeply rooted in society. Getting rid of bias requires fundamental structural change which goes well beyond anything that fiction can provide.

Acknowledgements

This research was funded by the British Society of Aesthetics Postdoctoral Fellowship (Kris Goffin) and the Leverhulme Research Project Grant RPG-2017-365 (Stacie Friend). Kris Goffin is an FWO (Fonds Wetenschappelijk Onderzoek Vlaanderen/ Research Foundation - Flanders) Junior Postdoctoral Fellow.

We want to thank Katrien Schaubroeck and the members of the UAntwerp Moral Psychology Seminar, as well as the anonymous referees, for their valuable feedback.

REFERENCES

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: evidence for independent constructs and unique effects on behavior. Journal of personality and social psychology, 91(4), 652.

Banaji, M. R., Bazerman, M., & Chugh, D. (2003). How (un)ethical are you? *Harvard Business Review, 81*, 56–64.

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

Baldwin, J. (1998). Everybody's Protest Novel. In *Collected Essays*. New York: Library of America, 11-18.

Batson, C. D. (2011). *Altruism in humans*. New York: Oxford University Press.

Bloom, P. (2016). *Against Empathy: The Case for Rational Compassion*. New York: Ecco.

Brownstein, M., Madva, A. & Gawronski, B. (2020). Understanding Implicit Bias: Putting the Criticism into Perspective, *Pacific Philosophical Quarterly,* 101: 267-307.

Carroll, N. (1998). Art, narrative, and moral understanding. In J. Levinson (ed.), *Aesthetics and Ethics: Essays at the Intersection*. Cambridge: Cambridge University Press, 126-60.

––– (2014). Moral Change: Fiction, Film & Family. In Choi, J., & Frey, M. (Eds.). *Cine-Ethics: Ethical Dimensions of Film Theory, Practice, and Spectatorship*. New York: Routledge, 43-56.

Cikara, M., Bruneau, E., Van Bavel, J.J. & Saxe, R. (2014). Their Pain Gives Us Pleasure: How Intergroup Dynamics Shape Empathic Failures and Counter–Empathic Responses, *Journal of Experimental Social Psychology*, 55: 110–115.

Clavel-Vazquez, A. (2018). Sugar and spice, and everything nice: what rough heroines tell us about imaginative resistance. *The Journal of Aesthetics and Art Criticism, 76*(2), 201-212.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of personality and social psychology, 83*(6), 1314-1329.

Currie, G. (2020). *Imagining and knowing: The shape of fiction*. New York: Oxford University Press.

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish journal of psychology, 10(*2), 230-241.

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological bulletin, 127*(6), 853-869.

de Sousa, R. (1987). *The Rationality of Emotion*. Cambridge, MA: MIT Press.

Diamond, C. (1991). *The Realistic Spirit: Wittgenstein, philosophy, and the mind*. Cambridge, MA: MIT Press.

Dodell-Feder, D., & Tamir, D. I. (2018). Fiction reading has a small positive impact on social cognition: A meta-analysis. *Journal of Experimental Psychology: General, 147*(11), 1713-1727.

Eaton, A. W. (2012). Robust Immoralism. *Journal of Aesthetics and Art Criticism, 70* (3), 281-292.

Ferguson, M. (1991). *Mansfield Park:* Slavery, colonialism, and gender. *Oxford Literary Review, 13*(1/2), 118-139.

FitzGerald, C., Martin, A., Berner, D., & Hurst, S. (2019). Interventions Designed to Reduce Implicit Prejudices and Implicit Stereotypes in Real World Contexts: A Systematic Review. *BMC Psychology, 7*(1)29, 1-12.

Friend, S. (2012). Fiction as a genre. *Proceedings of the Aristotelian Society, 112*(2), 179-209.

Gaut, B. (1998). The ethical criticism of art. In J. Levinson (ed.), *Aesthetics and Ethics: Essays at the Intersection*. Cambridge: Cambridge University Press, 182-203.

--- (2007). *Art, Emotion and Ethics*. Oxford: Oxford University Press.

--- (2013). Art and Ethics. In Gaut, B. & Lopes, D. (eds.), *The Routledge Companion to Aesthetics*, 451-464.

Gendler, T. S. (2000). The Puzzle of Imaginative Resistance. *The Journal of Philosophy, 97*, 55–81.

Gilbert, S. & Gubar, S. (1979). *The Madwoman in the Attic: The Woman Writer and the Nineteenth-Century Literary Imagination*. Yale: Yale University Press.

Green, M. & Brock, T. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology, 79*(5), 701-721.

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

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

Hakemulder, J. (2001). How to make alle Menschen Brüder: Literature in a multicultural and multiform society. *The psychology and sociology of literature*, 225-42.

Hakemulder, J. (2006). Imagining what could happen: Effects of taking the role of a character on social cognition. In S. Zyngier et al. (eds.), *Directions in Empirical Literary Studies: In Honor of Willie van Peer*. Amsterdam: John Benjamins, 139-153.

Harold, J. (2005). Infected by evil. *Philosophical Explorations, 8(*2), 173-187.

Holroyd, J. & Sweetman, J. (2016). The Heterogeneity of Implicit Bias. In M. Brownstein & J. Saul (eds.), *Implicit Bias and Philosophy*. New York: Oxford University Press, 80-103.

Johnson, D. R. (2013). Transportation into literary fiction reduces prejudice against and increases empathy for Arab-Muslims. *Scientific Study of Literature, 3*(1), 77-92.

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. *Science, 342*(6156), 377-380.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. E. L., Joy-Gaba, J. A., ... & Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*(4), 1765-1785.

Landy, J. (2008). A nation of Madame Bovarys: on the possibility and desirability of moral improvement through fiction. In G. Hagberg (ed.), *Art and Ethical Criticism*. Blackwell, 63-94.

Manne, K. (2017). *Down Girl: The Logic of Misogyny*. Oxford: Oxford University Press.

Mekawi, Y., & Bresin, K. (2015). Is the evidence from racial bias shooting task studies a smoking gun? Results from a meta-analysis. *Journal of Experimental Social Psychology, 61*, 120-130.

Merton, R. K. (1968). The Matthew effect in science. *Science, 159*, 56–63.

Meyer, S. (1993). “Safely to their own borders”: Proto-Zionism, feminism and nationalism in *Daniel Deronda*. *ELH, 60*(3), 733-758.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32*(2), 183-246.

Moyer-Gusé, E., Dale, K. R., Ortiz, M. (2019). Reducing prejudice through narratives: An examination of the mechanisms of vicarious intergroup contact. *Journal of Media Psychology, 31*, 185–195.

Mudimbe, V. Y. (1988) *The Invention of Africa: Gnosis, Philosophy, and the Order of Knowledge*. Bloomington: Indiana University Press.

Mumper, M. L., & Gerrig, R. (2017). Leisure reading and social cognition: A meta-analysis. *Psychology of Aesthetics, Creativity, and the Arts, 1*1(1), 109-120.

Murray, S. (2008). *Representing autism: Culture, narrative, fascination*. Liverpool: Liverpool University Press.

Nussbaum, M. C. (1990). *Love's knowledge: Essays on philosophy and literature*. New York: Oxford University Press.

––– (1996). *Poetic Justice: The Literary Imagination and Public Life.* Boston: Beacon Press.

––– (2013). *Political Emotions.* Harvard: Harvard University Press*.*

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12*(5), 413-417.

––– (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, *32*(4), 421-433.

Osteen, M. (Ed.). (2010). *Autism and representation*. New York: Routledge.

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual review of psychology, 60*, 339-367.

Panero, M. E., Weisberg, D. S., Black, J., Goldstein, T. R., Barnes, J. L., Brownell, H., & Winner, E. (2016). Does reading a single passage of literary fiction really improve theory of mind? An attempt at replication. *Journal of Personality and Social Psychology, 111(*5), e46.

Prentice, D. A. & Gerrig, R. J. (1999). Exploring the boundary between fiction and reality. In Chaiken, S. & Trope, Y. (eds.), *Dual-Process Theories in Social Psychology*. New York: Guilford Press, 529-546.

Prinz, J. (2011). Against Empathy. *The Southern Journal of Philosophy*, *49*, 214–233.

Project Implicit. (2011). *Take a Test.* https://implicit.harvard.edu/implicit/takeatest.html

Rudman, L. A., & Heppen, J. B. (2003). Implicit romantic fantasies and women's interest in personal power: A glass slipper effect?. *Personality and Social Psychology Bulletin, 29*(11), 1357-1370.

Said, E. (1993). *Culture and Imperialism.* New York: Knopf.

Samur, D., Tops, M., & Koole, S. L. (2018). Does a single session of reading literary fiction prime enhanced mentalising performance? Four replication experiments of Kidd and Castano (2013). *Cognition and Emotion, 32*(1), 130-144.

Saul, J. (2013). Implicit bias, stereotype threat, and women in philosophy. In F. Jenkins & K. Hutchison (eds.), *Women in philosophy: What needs to change*. Oxford University Press, 39-60.

Snyder, S. L., & Mitchell, D. T. (2010). *Cultural locations of disability*. Chicago: University of Chicago Press.

Spivak, G. (1985). Three Women's Texts and a Critique of Imperialism. *Critical Inquiry, 12*(1), 235-61.

Tuna, E. H. (2020). Imaginative Resistance. *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2020/entries/imaginative-resistance/>.

van Goidsenhoven, Leni. (2020) *Autisme in Meervoud*. Antwerp: Garant.

Vezzali, L., Stathi, S., Giovannini, D., Capozza, D., & Trifiletti, E. (2015). The greatest magic of Harry Potter: Reducing prejudice. *Journal of Applied Social Psychology, 45*(2), 105-121.

Walton, K. L. (1994). Morals in fiction and fictional morality I. *Proceedings of the Aristotelian Society, Supplementary Volume*, *68*, 27-50.

Weisbuch, M., & Pauker, K. (2011). The nonverbal transmission of intergroup bias: A model of bias contagion with implications for social policy. *Social issues and policy review*, *5*(1), 257-291.

Weisbuch, M., Pauker, K. & Nalini A. (2009). The subtle transmission of race bias via televised nonverbal behavior. *Science, 326*(5960), 1711-1714.

Wekker, G. (2016). *White Innocence: Paradoxes of Colonialism and Race*. Durham, NC: Duke University Press.

Willard, G., Isaac, K. J., & Carney, D. R. (2015). Some evidence for the nonverbal contagion of racial bias. *Organizational Behavior and Human Decision Processes*, *128*, 96-107.

Wimmer, L., Currie, G., Friend, S. & Ferguson, H. (2021). Testing Correlates of Lifetime Exposure to Print Fiction Following a Multi-Method Approach: Evidence from Young and Older Readers. *Imagination, Cognition and Personality*. Published online: https://journals.sagepub.com/eprint/GPWRFR4WYGFS4SCU6G6Z/full

––– (n.d.). The Effects of Reading Narrative Fiction on Social and Moral Cognition: Two Pre-Registered Experiments Following a Multi-Method Approach.

Xu, K., Nosek, B., & Greenwald, A. (2014). Psychology data from the race implicit association test on the project implicit demo website. *Journal of Open Psychology Data*, *2*(1), e3.

Xu, X., Zuo, X., Wang, X. , & Han, S. (2009). Do You Feel My Pain? Racial Group Membership Modulates Empathic Neural Responses, *Journal of Neuroscience*, *29*, 8525–8529.