To cite this chapter:

Goffin, K., & Moors, A. (2023). Implicit bias, fiction, and belief. In A. James, A. Kubo, & F. Lavocat (Eds.), *The Routledge handbook of fiction and belief* (pp. 173-184). Routledge.

13

IMPLICIT BIAS, FICTION, AND BELIEF

Kris Goffin and Agnes Moors

Introduction

In this chapter we discuss, from a psychological point of view, how fiction can influence implicit biases, such as unconscious racist and sexist attitudes and stereotypes. It might be argued that we learn these attitudes and stereotypes from fiction.

It is well known that many fictional narratives contain stereotypes. Today, many fictions no longer display blatant racist or sexist stereotypes, but several of them still fall prey to more subtle stereotypes. For example, in many contemporary romantic comedies, women are no longer explicitly portrayed as helpless, humble, and subservient, but they are still pictured as overly emotional and primarily dependent on romantic relationships for their happiness (e.g., see the film *Isn't It Romantic*, 2019). For another example, Asian women are often sexualized or exoticized in movies (see the film *Memoirs of a Geisha*, 2005) and Asian men often have an accent and lack emotional depth (see the "Karate Kid" film franchise, 1984–2010). Finally, Black men are still all too often presented as criminals in crime TV shows (e.g., *Law & Order: Special Victim's Unit*, 1999). Indeed, Black men are disproportionally portrayed as wrongdoers and criminals, while the crime solvers and law enforcers are disproportionally White. More indirectly, these Crime TV shows often idealize the justice system and normalize various inequalities and injustices (for an overview, see, Color of Change, 2017).[1] This has come to the foreground, especially since recent problems of police violence and the related unresolved legacies of slavery and segregation in the United States have been the focus of many debates.

Given that many fictions contain stereotypes, the first question we will address in the present chapter is: does fiction teach us implicit bias? A second question we will tackle is whether fiction can help in mitigating these implicit attitudes. Before doing so, we will clarify what implicit biases are and how they are measured, and we will consider mechanisms that plausibly install implicit bias in daily life.

What Is Implicit Bias and How Is It Measured?

Much recent research in (social) psychology has been devoted to how people often unconsciously endorse stereotypes. For instance, one can believe that it would be racist to think that Black men are likely to be criminals, and that racism is horrible. Yet, research indicates that even if one is anti-racist, one can behave in ways that seem to indicate that one does (unconsciously or implicitly) endorse such stereotypes (Banaji and Greenwald, 2013).

DOI: 10.4324/9781003119456-16

In this chapter, we are concerned with implicit social biases. A social bias can be understood as a bias in cognitions about a social group (Banaji and Greenwald, 2013; Saul, 2013). While interacting with members of particular social groups, one's mental processes can be biased. Examples of targeted social groups are race and gender groups, such as Black people and women. There are other cases of bias, which are not directed toward social groups, such as purely cognitive biases (e.g., biases in attention or reasoning), but in this chapter, we only discuss social bias. What makes a social bias implicit is that the individual is not typically aware of the bias. Sometimes it even contradicts their explicit mental states.

The mental states that are involved in social bias can be quite diverse (Holroyd and Sweetman, 2016). Many distinguish between an attitude and a stereotype. An attitude specifies a relation between a social group and a valence attribute (i.e., good or bad, or positive or negative), for instance, a relation between "Black people" and "bad." A prejudice is a negative attitude. A stereotype, on the other hand, specifies a relation between a social group and a non-valence attribute, for instance, a relation between "Black people" and "athletic" or between "Black people" and "criminal." The attributes in stereotypes often have a valence, but they do not simply convey valence like attitudes do. Both negative attitudes toward social groups (e.g., racialized attitudes) and stereotypes can be implicit.

To assess people's social biases, researchers have developed so-called implicit measures. These are contrasted with traditional, explicit measures. Explicit measures directly ask a person, for instance, whether Black people are more dangerous than White people. In implicit measures, on the other hand, a person's mental content is inferred in an indirect way from their behavior.

A famous example of an implicit measure is the Implicit Association Task (IAT), which is a task designed to measure people's attitudes toward social groups (Banaji and Greenwald, 2013; Greenwald and Banaji, 1995; Greenwald, McGhee, and Schwartz, 1998; Greenwald et al., 2009). One example is an IAT which tests whether individuals relate Black and White people with positive or negative valence. The task requires participants to categorize pictures of Black and White people and positive and negative words using the same or a different response. In one phase of the task, participants are asked to press left if they see a Black person or a positive word and to press right if they see a White person or a negative word. In another phase of the task, participants receive the opposite mapping between Black and White persons and positive and negative words. Then words and pictures are shown. What is measured is the speed by which participants resolve this task and the number of errors they make in both phases. If results show that a participant is slower and makes more errors in the first phase, when they have to combine Black faces and positive words, it is inferred that they have an implicit bias toward Black people.

Other IATs measure stereotypes, understood as relations between social groups and other concepts than positive or negative. An example is the gender-career IAT (Banaji and Greenwald, 2013).[2] Here, participants are asked to categorize words referring to male and female categories and words referring to career and family using the same or a different response. In one phase of the task, they have to press left for words referring to female and career and right for words referring to male and family. In another phase, they have to press left for words referring to male and career and right for words referring to female and family. The majority of people tend to be quicker and more accurate in combining male with career and female with family, which suggests that they hold an implicit gender-career stereotype. This is also true of people who explicitly reject these stereotypes (Banaji and Greenwald, 2013; Greenwald and Banaji, 1995; Greenwald et al., 2009).

Another example of an implicit measure of stereotypes is the first-person shooter task (Correll, Park, Judd, and Wittenbrink, 2002). This task is designed to measure whether people implicitly endorse the stereotype that Black men are likely to be criminals and tend to behave according to this stereotype. In this task, participants play a sort of first-person shooter game, in which they have to shoot armed people. Pictures are shown of Black and White people holding objects: some hold a gun, some

do not. The task goes quite fast, and it is not always clear which objects the people are holding. The task is, however, to only shoot the people who are holding guns. It turns out that most participants, even those who explicitly reject the stereotype, have a bias in the mistakes they make (i.e., the so-called shooter bias, Payne, 2001): unarmed Black men are more often shot than unarmed White men.

Implicit measures have been criticized. It has been argued, for instance, that it is not clear what these measures actually measure (for an overview and responses to the criticism, see Brownstein, Madva, and Gawronski, 2020) and whether they can meaningfully be called implicit (Corneille and Hütter, 2020). It is also debated what constitutes these biases: unconscious beliefs (Mandelbaum, 2016) or unconscious associations (Gawronski and Bodenhausen, 2011; Strack and Deutsch, 2004). For instance, is it the case that people unconsciously believe that "Black men are dangerous" or do they simply associate Black men with danger? A belief can be understood as a qualified relation between two concepts that is held to be true. An association, on the other hand, points at the fact that two concepts are related without specifying the nature of the relationship. Concepts simply conjure each other up (see De Houwer, 2014), as we will address further later.

A final line of criticism denounces the fact that implicit measures of attitudes toward groups are poor predictors of discriminative behavior (e.g., Kurdi et al., 2019; Oswald et al., 2013). This attitude-behavior gap has led some authors to search for moderators of attitude-behavior relations (e.g., Brownstein et al., 2020; Friese et al., 2008) and others to argue for the measurement of other (more proximal) causes of behavior such as the expected utility of behavior (Moors and Köster, 2022).

How Do We Acquire Implicit Biases?

It has often been suggested that the learning of the attitudes and stereotypes responsible for implicit biases can be framed as a case of evaluative learning or conditioning (De Houwer, 2007; Olson and Fazio, 2001, 2006). Evaluative learning is the phenomenon in which the pairing of an initially neutral stimulus with another, positive or negative stimulus results in a change in the perceived valence (i.e., goodness/badness) of the former stimulus (Baeyens et al., 1992; De Houwer, 2007; De Houwer et al., 2001; Hofmann et al., 2010; Mitchell et al., 2009).[3] For instance, take the stimulus "men with ginger hair." Suppose you initially have a neutral attitude toward ginger men: you do not think ginger men are particularly good or bad. As a kid, however, you were bullied by an irritating ginger boy. In high school, many other people spoke of ginger men as undesirable and ugly losers and treated them accordingly. In your youth, the stimulus "ginger men" is thus repeatedly coupled with other, negative stimuli such as being undesirable, being ugly, and so on. As a result, you may learn that "ginger men" are associated with negative valence. So the initially neutral valence of ginger men shifts toward negative valence because the stimulus is coupled with negatively valenced stimuli. In another example, the initially neutral valence may shift toward positive valence because one sees ginger men, such as Ed Sheeran, being coupled with positively valenced stimuli, such as "nice music" and "likeable guy." Of course, this example does not hold for everyone. If you do not like Ed Sheeran, the effect will not occur.

Evaluative learning is a somewhat different kind of learning than studying for your statistics exam. Evaluative conditioning is something that happens to you, rather than something that you do. Learning that ginger men are positive or negative is the result of being passively exposed to pairings between ginger men and positive or negative other stimuli, not something that you actively try to memorize. Evaluative conditioning is a useful tool to navigate the world, as we automatically learn which things are good and bad for us. The dark side of evaluative conditioning, however, is that it also applies to social groups.[4]

Let us now unpack the mechanisms underlying evaluative learning. As we said before, evaluative conditioning is the pairing of an initially neutral stimulus with another positive or negative stimulus,

which results in a change in the perceived valence of the former stimulus. A mechanistic explanation of evaluative conditioning is an explanation of how this change in perceived valence comes about. In line with the aforementioned debate between theorists who believe that implicit attitudes are associations versus theorists who think they are beliefs, there are theorists who think evaluative learning is associative and theorists who think that it is more like learning a belief. The traditional way of understanding evaluative learning is the associative model. When two concepts are paired: "ginger men" and "negative," one learns to associate "ginger men" with bad valence: one learns "ginger men—bad." When one thinks of ginger men, the next thought is "bad." Recently, however, evidence is accumulating that evaluative conditioning might not be merely associative, but propositional (for an overview see Madva, 2016). Instead of learning that two concepts are related in some way, as is the case in a mere association, one learns something about the nature or the quality of the relation between these concepts. One learns, for instance, that "ginger people are bad." Mitchell and De Houwer et al. (2009) argue that all learning is propositional, that is, based on knowledge about propositions understood as qualified relationships rather than unqualified associations. A similar statement has been made by philosophers such as Mandelbaum (2016). We wish to point out that despite disagreement about the details of the various processes underlying the evaluative conditioning effect, there is consensus that the effect itself—a change in liking of an initially neutral stimulus resulting from the pairing of this stimulus with another positive or negative stimulus—is robust (De Houwer, 2007; De Houwer et al., 2001).

Note that evaluative learning is about goodness or badness. In line with the aforementioned distinction between attitudes and stereotypes, we could make a distinction between evaluative learning and other types of relationship learning. Instead of learning the relation between an object and good or bad, we can (by means of the same procedure learn the relation between an object and another concept. Thus, we can learn a relation between "Black person" and "bad," but we can also learn the relation between "Black person" and "dangerous."

Now that we understand the processes involved in the formation of attitudes and stereotypes in real life, we will focus on how this applies to learning attitudes and stereotypes from fiction.

Fiction as a Bias Teacher

Many fictional narratives, such as TV shows, contain stereotypes, often in subtle sometimes in less subtle ways. In the introduction we mentioned some examples: Romantic comedies reinforce gender stereotypes such as women being dependent on romantic relationships for their happiness. Asian women are often objectified and overly sexualized; Asian men often appear as one-dimensional skilled martial artists. Black men often are presented as criminals, while the heroes of the story are often White people.

A good case could be made that we learn implicit attitudes and stereotype from fiction. Indeed, in many articles on implicit bias, it is presumed that popular media are important sources of bias learning and "unlearning" (for an overview, see Dasgupta, 2013). These media not only include fictional TV shows but also news shows and non-fictional programs.[5] In fact, the influence of non-fictional media is more widely studied than that of fictional media (see e.g., Entman, 1994; Payne and Dal Cin, 2015; Roos et al., 2013). The idea that fiction is a bias teacher is often presupposed but not very thoroughly empirically investigated. Nevertheless, it makes sense to think that engaging with fiction is sufficient to learn implicit attitudes and stereotypes. We will discuss this hypothesis.[6]

It has been shown that people can learn relationships between things by merely observing other people. In the evaluative conditioning literature, this phenomenon is described as "observational evaluative conditioning" (Baeyens and Vansteenwegen et al., 1996). Here, learning does not require directly experiencing the stimuli being paired, but occurs when seeing other people experiencing

those stimuli. One might understand attitude learning from fiction as a case of observational evaluative conditioning. Maybe our negative attitudes toward a certain social group do not reflect our own negative experiences from interacting with members of this group, but rather stem from the experiences of the fictional characters that we observe. If we are exposed to many fictional Black men holding guns on television, we might learn to associate black men with guns. If we see or read that most fictional scientists are old White men with beards, we might learn to associate the profession of scientist with men and not with women.

Goffin and Friend (2022) argue in their paper for the pessimistic stance that many works of fiction install harmful biases in people. As an example, they discuss the "glass slipper effect" (Rudman and Heppen, 2003), which is an effect that indicates that implicit gender stereotypes as displayed in "Prince Charming" narratives, in which the male love interest is a brave hero who saves the helpless woman, influence one's career aspirations. Rudman and Heppen (2003) argue that these implicit stereotypes are correlated with women being less ambitious.

Goffin and Friend (2022) also referred to the "bias contagion" phenomenon, which is a form of observational learning in which individuals copy biased behavior as seen by others (Weisbuch, Pauker, and Ambadi, 2009). The behavior in question can be extremely subtle, such as standing further away from people of color than from White people in elevators. A number of studies have indicated that bias contagion also works by watching television series (Weisbuch, Pauker, and Ambadi, 2009; Willard, Isaac, and Carney, 2015). By watching characters on TV displaying subtle signs of aversion, such as smiling less toward characters from a specific social group, spectators can copy this subtle aversive behavior. Tragically, however, bias contagion even occurs when TV shows, such as "CSI" and "Grey's Anatomy," actively contra-stereotype their characters and portray women and people of color as having high-level jobs (Weisbuch, Pauker, and Ambadi, 2009). When the actors interacting with these characters display subtle behavior vis-à-vis these characters, such as standing a little bit further away from them, spectators still pick this up, even if only implicitly.

There are also studies on the influence of video games on implicit bias. Saleem and Anderson (2013), for instance, showed that violent videogames with Arab stereotypes and Arab terrorists increased anti-Arab attitudes as measured by an IAT. Moreover, a great deal of research examines stereotypes presented in children's fiction (e.g., Kneeskern and Reeder, 2020; Peterson and Lach, 1990). It is often thought that many social biases are learned at a very young age, and then maintained throughout the rest of life (for an overview, see Dasgupta, 2013).

Another mechanism that may be involved in the learning of biases from fiction is known as Spinozan Belief Formation. Dan Gilbert (1991) distinguishes between two accounts of the acquirement of new beliefs: the Cartesian model and the Spinozan model (referring to the philosophers Descartes and Spinoza). The Cartesian model states that when a person is confronted with a statement or proposition, a first step is to process the proposition or to consider it. A second step is to evaluate the proposition as true or false after evidence in favor or against the proposition is gathered. A person may also suspend judgment if the evidence is deemed insufficient to hold something true or false. The Spinozan view, which Gilbert (1991) himself defends, states that belief formation happens instantly. When a person is confronted with a proposition, the first step is to process and believe it. In a second step, however, one can re-evaluate this belief and then reject it or conclude that it was correct. Here, believing a proposition is the default, which may or may not be overruled in a later stage. On the Cartesian model it takes effort to believe or reject something. On the Spinozan model, on the other hand, it only takes effort to reject a belief. Because people do not always have sufficient opportunity or motivation, they will often hold beliefs they have not critically examined. Empirical support for the Spinozan belief formation hypothesis comes from studies showing that participants seem to immediately believe things they read, even if the text is explicitly presented as fiction (Gilbert et al., 1993). It would take effort to critically examine what we read or to stop believing the things that

are offered to us while reading. If this is true, then fiction should not only provide us with relational information, in the form of associations or propositions, but also generate instant beliefs, including those that constitute implicit bias.[7]

One might raise skepticism against the idea that it is fiction that teaches us these biases. One might think that biases live in a society, and that fiction presents a reflection of what lives in that society (see Payne et al., 2017, for a structural account of bias). One could argue that we pick up biases from real-life interactions and pass them on through fiction. Even if it seems plausible to assume that real-life biases find their way into fiction, fiction may in turn select and exaggerate certain stereotypes, thereby creating novel biases in people who never had them in the first place and sustain or enlarge them in people who did. In other words, it may be a two-way street or even a vicious cycle.

Given that we can learn biases from real-life experiences as well as from fiction, the question arises whether fiction is special. Is fiction especially good or bad at teaching biases? Some argue that narratives in particular are better in persuading us to believe something. This idea is referred to as "narrative persuasion" (Busselle and Bilandzic, 2008; Hamby, Brinberg, and Jaccard, 2016). Two factors may contribute to this narrative persuasion. The first is that fiction is often designed to create strong emotional experiences (Busselle and Bilandzic, 2008). The second is that people are less likely to check for accuracy in fiction than in real life (Dal Cin et al., 2004; Prentice and Gerrig, 1991). We discuss each of these two factors in more detail.

Regarding the first point, some authors claim that fiction is persuasive because it elicits strong emotions (Busselle and Bilandzic, 2008; Hamby, Brinberg, and Jaccard, 2016). This should create an ideal circumstance for observational evaluative learning, as positive and negative experiences are essential ingredients of evaluative learning. One might object, however, that fiction produces less intense emotions precisely because we know we are dealing with fiction. Witnessing a crime in real life will evoke much stronger emotions than seeing a crime in a movie. However, the knowledge that what is happening in fiction is not real is often compensated by the use of several strategies such as the use of extreme stimuli, the use of superstimuli, and the use of techniques that foster "immersion." To create an equally strong affective experience in fiction as in real life, the events need to be far more intense or extreme. That is why in movies there is so much horror, tragedy, and violence. You need to present more extreme events than in real life to obtain an emotional reaction of some sort. The same goes for psychological experiments: the more you simulate real life (e.g., in VR) the less intense your stimulus must be. The two positions discussed here—that fiction elicits strong versus weak emotions—focused mainly on the intensity of emotions. It could also be argued that the emotions generated by fiction are of a different kind than the ones generated in real life (Walton, 1990).

In addition to more extreme stimuli, certain forms of fiction also make use of "superstimuli" (Ramachandran and Hirstein, 1999; but see criticism by Gombrich, 2000). Superstimuli are not necessarily more extreme, but they have exaggerated features, which may create affective experiences in the audience. For instance, in cartoons, characters have large eyes and big heads, like babies do, to make sure they are perceived as cuter and elicit more empathy. Finally, works of fiction employ a range of techniques to make sure the audience becomes immersed in the fictional world and empathizes with the fictional characters living in it. This too may contribute to the emotional power of fiction (Visch, Tan, and Molenaar, 2010). We may conclude that depending on the successful use of the discussed strategies, fiction may be more or less effective in eliciting affective experiences that contribute to observational evaluative learning.

Regarding the second factor, some researchers argue that in real life and non-fictional stories, people tend to check whether the facts that are presented are actually true (Dal Cin et al., 2004; Prentice and Gerrig, 1991). In fiction, they are less likely to do this because it is clear from the start that they are dealing with fiction. Paradoxically, however, the fact that people do not engage in fact-checking

also makes them more susceptible to uncritically adopt the stereotypes presented in fiction and develop biased attitudes based on them. In other words, they do not necessarily consider the presented stereotypes to be fictional. The argument to support the claim that we gather information without checking is based on the ideas of the aforementioned Spinozan belief formation framework (Hamby, Brinberg, and Jaccard, 2016).

In addition to the factors involved in narrative persuasion, we may speculate that fiction can be especially powerful in generating stereotypes. Creators of fiction make use of various "tropes" or predictable elements in crafting their stories. Fiction presents us with archetypes or caricatures of people such as "the foreigner" or "the professor." Instead of presenting us with real people, we see fictional characters who are one-dimensional. Often these caricatures are problematic because they reflect harmful stereotypes. For instance, as mentioned before, the "Karate Kid" film franchise (1984–2010) portrays Asian characters as martial arts fighters with a thick accent, who also lack emotional depth. Thus, fiction may facilitate observational learning of relational information involved in stereotypes.[8]

In this section, we have presented various ways in which we can conceptualize fiction as a bias teacher. In the final section, we will examine whether fiction can also be a bias remover.

Fiction as Bias Remover

Some argue that fiction can help people to get rid of racist, sexist, and other social biases. If it is true that we can learn biases from fiction, then fiction can be used strategically to unlearn biases.

In the previous section, we have mentioned "narrative persuasion." The idea that fictional narratives are supposedly really good at making someone believe something can be used for the better. One application might be to narratively persuade people to stop (implicitly) endorsing harmful attitudes and stereotypes (Hamby, Brinberg, and Jaccard, 2016).

In fact, this is a very old idea. Psychological researchers have experimented with using fictional stories to re-educate people and save them from their vices and addictions. For instance, Cautela (1971) discussed a story that is supposed to mitigate alcohol addiction. It is a story that makes you imagine in detail a case in which drinking alcohol makes you vomit. It also describes how you feel wonderful if you reject drinking more alcohol. This story is supposed to teach you a negative attitude toward alcoholic beverages and help with alcohol addiction.

Let us focus on the mechanism and procedure that may be behind these kinds of unlearning from fiction. Some have argued that by means of evaluative conditioning, which we have discussed in the beginning of the chapter, we can reduce biases (Olson and Fazio, 2006). Applying this to fiction, we could then argue that by coupling members of certain social groups to positively valenced objects, we can reduce implicit prejudice.

The procedure or phenomenon at stake here is called "counterconditioning." A researcher tries to reverse the outcome of a previous conditioning procedure that led to the initial attitude by using another conditioning procedure that would create the opposite attitude. If the initial conditioning episode paired "ginger men" with negatively valenced stimuli, counterconditioning will entail the pairing of "ginger men" with positively valenced stimuli, with the aim of creating a positive attitude toward ginger men.

There is fair consensus that counterconditioning does not consist of the "unlearning" or updating of a previous attitude, but rather consists of the learning of a new attitude that sits next to the old one (Baeyens, 1988; for an overview see Bouton, 2002). Research thus suggests that the best one can do is to learn new attitudes that contradict the old ones, but this does not mean that the old ones disappear. The old attitudes may re-emerge when the right contextual cues are presented. This would imply that fiction cannot really erase biases through counterconditioning. This thus means that we cannot "unlearn" bias from fiction.[9]

So far, counterconditioning has been targeted at implicit attitudes or prejudices toward certain social groups. One can build a similar case for implicit stereotypes. Counter-stereotype exposure (Dasgupta and Greenwald, 2001) is meant to diminish people's implicitly endorsed stereotypes by confronting them with images and scenarios that reject these stereotypes. We believe fiction can play a role in this. If people are confronted with images or scenarios of scientists who are not old white men with beards, existing gendered and racialized stereotypes of a scientist may be counteracted. However, as is the case in evaluative counterconditioning, it is likely that counter-stereotype exposure will also not automatically lead to the "unlearning" of the old stereotypes (see also Madva, 2017).

Some researchers have argued that intergroup contact is an effective way to reduce prejudice (see Dasgupta and Rivera, 2008). The idea is simple: the more you interact with people from a particular social group, the less biased you become toward members of this social group because intergroup contact creates opportunities for counterconditioning. However, not all interactions turn out to have this effect. Research has shown, for instance, that prejudice is not reduced if there is a power imbalance, for instance, if a person of one social group is serving or obeying a person of another social group (see Anderson, 2010; Dasgupta and Rivera, 2008). The two people interacting have to be each other's equals to some extent.

Fiction could be designed to simulate intergroup interaction among equals: by reading about or viewing screenings with people from different social groups, or by reading about interactions between people from social groups, one can create a fictional intergroup contact. There are some psychological studies that have investigated fictional intergroup contact. In an overview article about experiments testing various prejudice mitigation strategies (Paluck and Green, 2009), the fictional intergroup contact strategy seems to be quite successful. In a number of studies with school children as participants (Cameron and Rutland, 2006; Cameron et al., 2006; Liebkind and McAlister, 1999; Slone et al., 2000), the children were presented with fictional stories in which there was a positive contact between them and children from the same and from other social groups (race, gender, disability). Results showed a positive effect in prejudice reduction. However, this positive effect is only measured by means of self-reports and not by means of implicit measures (Paluck and Green, 2009).[10] Thus, it is unclear whether the results of these studies also apply to implicit prejudices. This could be a topic for further research.[11]

It might also be the case that the true unlearning of biases from fiction is not done by a passive viewing or reading, but by means of reflection and discussion on the fictional narrative. Fiction might stimulate you to reflect on your biases and make you aware that you have these biases (see Fisher, 1998).

These are various ways in which we might mitigate biases through fiction. Some researchers are, however, skeptical of the idea that fiction would be helpful in combatting biases. Goffin and Friend (2022), for instance, argue that the initial conditioning of "bad" attitudes is more likely to happen than the possibility of counterconditioning of these "bad" attitudes into "good" ones.

However, instead of making big optimistic claims like "fiction can help us to unlearn bias" or alternative pessimistic claims like "fiction is more likely to create than reduce bias," it would be better to further investigate the phenomenon of learning implicit bias from fiction through rigorous psychological research. A promising avenue for future research would be to investigate the conditions under which fiction leads to the creation and reduction of biases. More broadly speaking, the question can be raised of what the moderators are that influence whether fiction is better able to teach and/or to reduce biases? Examples of moderators could be the artistic quality of the fictional work. Is good fiction better at reducing biases than bad fiction?[12] Maybe when consumers of fiction think that it is "too obvious" that the author has as an aim to reduce bias, or the creator is "too manipulative," the effect might be absent or opposite.

It is often said that fiction can teach biases as well as remove them. This chapter presented an overview, from a psychological perspective, of several ways that both phenomena might occur.

Notes

1 It is worth noting that in our examples, we mainly focus on the U.S. context, as U.S. visual culture has arguably one of the biggest impacts on Western culture. The psychological claims we make, however, are taken to be universal.
2 The test may be taken online for free here: https://implicit.harvard.edu/implicit/user/agg/blindspot/indexgc.htm.
3 For a good introduction to the psychology of learning, see De Houwer and Hughes (2020).
4 There is debate about whether evaluative learning really is automatic, or whether it requires awareness of the pairings (for a review, see Corneille and Stahl, 2019).
5 Interestingly, many stimuli in an experimental setting are in fact "fictional." In the lab pictures or videos are shown; or one needs to read a vignette. Experiments "in the wild" are rarely done. So, in fact many psychological studies engage with fiction, without explicitly mentioning it.
6 We would like to point out that we are neutral on theories of fiction. The claims of this chapter are meant to be consistent with most theories of fiction.
7 One might think that the Spinozan theory of the mind contradicts the idea that fiction involves "make-believe" (Walton, 1990). This is not necessarily the case. Walton is neutral about the specific subpersonal mental processes that give rise to his theory of make-believe. The make-believe hypothesis states that while engaging with fiction you exercise your imagination. The Spinozan theory would state that if you imagine something, you then consequently also believe that this is the case, but you can always reject these beliefs right after you acquire them. Spinozan theory and make-belief are not necessarily contradictory, but one might think that they form a bit of an awkward pair. However, the Spinozan theory is indeed controversial.
8 Of course, not all fiction has these one-dimensional stereotypes. Fiction can also focus on the complexity of the characters.
9 The fact that counterconditioning does not lead to erasing attitudes but rather to the formation of new attitudes that contradict the old ones can be seen as evidence for the fragmentation of the mind hypothesis. This hypothesis states that people often have contradictory beliefs and attitudes, and in different contexts our behavior is guided by different (and possibly contradicting) attitudes and beliefs (Mandelbaum, 2014).
10 Paluck and Green (2009) mentioned that similar studies done with reading about other cultures or about people from different social groups have no, or a less overwhelmingly positive effect than the studies just mentioned. So the effect is stronger if the fictional narrative describes an interaction between members of different social groups.
11 It might be remarked that although fiction is used in this way in many U.S. schools, the question remains whether this is effective to address deep social inequities.
12 See also Nussbaum (1996).

Works Cited

Baeyens, Frank, Geert Crombez, Omer Van Den Bergh, and Paul Eelen. 1988. "Once in Contact Always in Contact: Evaluative Conditioning is Resistant to Extinction." *Advances in Behavior Research and Therapy* 10, no. 4: 179–99.

Baeyens, Frank, Paul Eelen, Geert Crombez, and Omer Van Den Bergh. 1992. "Human Evaluative Conditioning: Acquisition Trials, Presentation Schedule, Evaluative Style and Contingency Awareness." *Behaviour Research and Therapy* 30, no. 2 (March): 133–42.

Baeyens, Frank, D. E. B. Vansteenwegen, J. A. N. De Houwer, and Geert Crombez. 1996. "Observational Conditioning of Food Valence in Humans." *Appetite* 27, no. 3: 235–50.

Banaji, Mahzarin R., and Anthony G. Greenwald. 2013. *Blindspot: Hidden Biases of Good People*. New York: Bantam.

Bouton, Mark E. 2002. "Context, Ambiguity, and Unlearning: Sources of Relapse After Behavioral Extinction." *Biological Psychiatry* 52, no. 10 (November): 976–86.

Brownstein, Michael, Alex Madva, and Bertram Gawronski. 2020. "Understanding Implicit Bias: Putting the Criticism Into Perspective." *Pacific Philosophical Quarterly* 101, no. 2 (June): 276–307.

Busselle, Rick, and Helena Bilandzic. 2008. "Fictionality and Perceived Realism in Experiencing Stories: A Model of Narrative Comprehension and Engagement." *Communication Theory* 18, no. 2 (May): 255–80.

Cameron, Lindsey, and Adam Rutland. 2006. "Extended Contact through Story Reading in School: Reducing Children's Prejudice Toward the Disabled." *Journal of Social Issues* 62, no. 3 (September): 469–88.

Cameron, Lindsey, Adam Rutland, Rupert Brown, and Rebecca Douch. 2006. "Changing Children's Intergroup Attitudes toward Refugees: Testing Different Models of Extended Contact." *Child Development* 77, no. 5 (September–October): 1208–19.

Cautela, Joseph R. 1971. "Covert Conditioning." In *The Psychology of Private Events: Perspectives on Covert Response Systems*, edited By Alfred Jacobs and Lewis B. Sachs, 109–30. New York: Academic Press.

Color of Change. 2017. "Normalizing Injustice: The Dangerous Misrepresentations that Define Television's Scripted Crime Genre." https://colorofchange.org/press_release/normalizing-injustice-new-landmark-study-by-color-of-change-reveals-how-crime-tv-shows-distort-understanding-of-race-and-the-criminal-justice-system/

Corneille, Olivier, and Christoph Stahl, C. 2019. "Associative Attitude Learning: A Closer Look at Evidence and How It Relates to Attitude Models." *Personality and Social Psychology Review* 23, no. 2: 161–189. https://doi.org/10.1177/1088868318763261

Corneille, Olivier, and Mandy Hütter. 2020. "Implicit? What Do You Mean? A Comprehensive Review of The Delusive Implicitness Construct in Attitude Research." *Personality and Social Psychology Review* 24, no. 3 (August): 212–32.

Correll, Joshua, Bernadette Park, Charles M. Judd, and Bernd Wittenbrink. 2002. "The Police Officer's Dilemma: Using Ethnicity To Disambiguate Potentially Threatening Individuals." *Journal of Personality and Social Psychology* 83, no. 6: 1314–29.

Dal Cin, S., M. P. Zanna, and G. T. Fong. 2004. "Narrative Persuasion and Overcoming Resistance." In *Resistance and Persuasion*, edited by Eric S. Knowles and Jay A. Linn, 175–91. New York: Psychology Press.

Dasgupta, Nilanjana. 2013. "Implicit Attitudes and Beliefs Adapt To Situations: A Decade of Research on the Malleability of Implicit Prejudice, Stereotypes, and the Self-Concept." *Advances In Experimental Social Psychology* 47: 233–79.

Dasgupta, Nilanjana, and Anthony G. Greenwald. 2001. "On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals." *Journal of Personality and Social Psychology* 81, no. 5 (November): 800–14.

Dasgupta, Nilanjana, and Luis M. Rivera. 2008. "When Social Context Matters: The Influence of Long–Term Contact and Short–Term Exposure to Admired Outgroup Members on Implicit Attitudes and Behavioral Intentions." *Social Cognition* 26, no. 1 (February): 112–23.

De Houwer, Jan. 2007. "A Conceptual and Theoretical Analysis of Evaluative Conditioning." *Spanish Journal of Psychology* 10, no. 2 (November): 230–41.

De Houwer, Jan. 2014. "A Propositional Model of Implicit Evaluation." *Social and Personality Psychology Compass* 8, no. 7: 342–53.

De Houwer, Jan. 2019. "Implicit Bias is Behavior: A Functional-Cognitive Perspective on Implicit Bias." *Perspectives on Psychological Science* 14, no. 5 (September): 835–40.

De Houwer, Jan, and Sean Hughes. 2020. *The Psychology of Learning: An Introduction from A Functional-Cognitive Perspective*. Cambridge, MA: MIT Press.

De Houwer, Jan, Sarah Thomas, and Frank Baeyens. 2001. "Association Learning of Likes and Dislikes: A Review of 25 Years of Research on Human Evaluative Conditioning." *Psychological Bulletin* 127, no. 6 (November): 853–69.

Entman, Robert M. 1994. "Representation and Reality in the Portrayal of Blacks on Network Television News." *Journalism Quarterly* 71, no. 3 (Fall): 509–20.

Fisher, Robert. 1998. "Stories for Thinking: Developing Critical Literacy Through the Use of Narrative." *Analytic Teaching* 18, no. 1: 16–27.

Friese, Malte, Wilhelm Hofmann, and Manfred Schmitt. 2008. "When and Why Do Implicit Measures Predict Behaviour? Empirical Evidence for the Moderating Role of Opportunity, Motivation, and Process Reliance." *European Review of Social Psychology* 19: 285–338.

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

Gerrig, Richard J., and Deborah A. Prentice. 1991. "The Representation of Fictional Information." *Psychological Science* 2, no. 5 (September): 336–40.

Gilbert, Daniel T. 1991. "How Mental Systems Believe." *American Psychologist* 46, no. 2 (February): 107–19.

Gilbert, Daniel T., R. Tafarodi, and P. Malone. 1993. "You Can't Not Believe Everything You Read." *Journal of Personality and Social Psychology* 65, no. 2 (August): 221–33.

Goffin, Kris, and Stacie Friend. 2022. "Learning Implicit Biases from Fiction." *The Journal of Aesthetics and Art Criticism* 80, no. 2 (Spring): 129–39.

Gombrich, Ernst H. 2000. "Concerning 'The Science of Art': Commentary on Ramachandran and Hirstein." *Journal of Consciousness Studies* 7, nos. 8–9 (August): 17.

Greenwald, Anthony G., T. Andrew Poehlman, Eric Luis Uhlmann, and Mahzarin R. Banaji. 2009. "Understanding and Using The Implicit Association Test: III. Meta-Analysis of Predictive Validity." *Journal of Personality and Social Psychology* 97, no. 1 (July): 17–41.

Greenwald, Anthony G., and Mahzarin R. Banaji. 1995. "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes." *Psychological Review* 102, no. 1 (January): 4–27.

Greenwald, Anthony G., Debbie E. Mcghee, and Jordan L. K. Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74, no. 6 (June): 1464–80.

Hamby, Anne, David Brinberg, and James Jaccard. 2016. "A Conceptual Framework of Narrative Persuasion." *Journal of Media Psychology* 30: 113–124. https://Doi.Org/10.1027/1864-1105/A000187.

Hofmann, Wilhelm, Jan De Houwer, Marco Perugini, Frank Baeyens, and Geert Crombez. 2010. "Evaluative Conditioning in Humans: A Meta-Analysis. *Psychological Bulletin* 136, no. 3 (May): 390–421.

Holroyd, Jules, and Joseph Sweetman. 2016. "The Heterogeneity of Implicit Bias." In *Implicit Bias and Philosophy*, edited By Michael Brownstein and Jennifer Saul, 1: 80–105. New York: Oxford University Press.

Kneeskern, Ellen E. and Patricia Reeder. 2020. "Examining the Impact of Fiction Literature on Children's Gender Stereotypes." *Current Psychology* 41: 1472–1485.

Kurdi, Benedek, Allison E. Seitchik, Jordan R. Axt, Timothy J. Carroll, Arpi Karapetyan, Neela Kaushik, Diana Tomezsko, Anthony G. Greenwald, and Mahzarin R. Banaji. 2019. "Relationship Between the Implicit Association Test and Intergroup Behavior: A Meta-Analysis." *American Psychologist* 74, no. 5 (July-August): 569–86.

Liebkind, Karmela, and Alfred L. Mcalister. 1999. "Extended Contact Through Peer Modelling to Promote Tolerance In Finland." *European Journal of Social Psychology* 29, no. 5–6 (August–September): 765–80.

Madva, Alex. 2016. "Why Implicit Attitudes Are (Probably) Not Beliefs." *Synthese* 193, no. 8 (August): 2659–84.

Madva, Alex. 2017. "Biased Against Debiasing: On the Role of (Institutionally Sponsored) Self-transformation in the Struggle Against Prejudice." *Ergo: An Open Access Journal of Philosophy* 4, no. 6: 145–179. https://Doi.Org/10.3998/Ergo.12405314.0004.006.

Mandelbaum, Eric. 2014. "Thinking Is Believing." *Inquiry* 57, no. 1: 55–96.

Mandelbaum, Eric. 2016. "Attitude, Inference, Association: On the Propositional Structure of Implicit Bias." *Noûs* 50, no. 3 (September): 629–58.

Mitchell, Chris J., Jan De Houwer, and Peter F. Lovibond. 2009. "The Propositional Nature of Human Associative Learning." *Behavioural and Brain Sciences* 32, no. 2 (April): 183–246.

Moors, Agnes, and Massimo Köster. 2022. "Behavior Prediction Requires Implicit Measures of Stimulus-Goal Discrepancies and Expected Utilities of Behavior Options Rather Than of Attitudes Toward Objects." *Wiley Interdisciplinary Reviews: Cognitive Science* 13, no. 5: E1611.

Nussbaum, Martha. 1996. *Poetic Justice: The Literary Imagination and Public Life.* Boston, MA: Beacon Press.

Olson, Michael A., and Russell H. Fazio. 2001. "Implicit Attitude Formation Through Classical Conditioning." *Psychological Science* 12, no. 5 (September): 413–17.

———. 2006. "Reducing Automatically Activated Racial Prejudice Through Implicit Evaluative Conditioning." *Personality and Social Psychology Bulletin* 32, no. 4 (April): 421–33.

Oswald, Frederick L., Gregory Mitchell, Hart Blanton, James Jaccard, and Philip E. Tetlock. 2013. "Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies." *Journal of Personality and Social Psychology* 105, no. 2 (August): 171–92.

Paluck, Elizabeth Levy, and Donald P. Green. 2009. "Prejudice Reduction: What Works? A Review and Assessment of Research and Practice." *Annual Review of Psychology* 60 (January): 339–67.

Payne, B. Keith. 2001. "Prejudice and Perception: The Role of Automatic and Controlled Processes in Misperceiving a Weapon." *Journal of Personality and Social Psychology* 81, no. 2 (August): 181–92.

Payne, B. Keith, and Sonya Dal Cin. 2015. "Implicit Attitudes in Media Psychology." *Media Psychology* 18, no. 3: 292–311.

Payne, B. Keith, Heidi A. Vuletich, and Kristjen B. Lundberg. 2017. "The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice." *Psychological Inquiry* 28, no. 4: 233–48.

Peterson, S. Bender and Mary Alyce Lach. 1990. "Gender Stereotypes in Children's Books: Their Prevalence and Influence on Cognitive and Affective Development." *Gender and Education* 2: 185–197.

Ramachandran, Vilayanur S., and William Hirstein. 1999. "The Science of Art: A Neurological Theory of Aesthetic Experience." *Journal of Consciousness Studies* 6, nos. 6–7 (June–July): 15–51.

Roos, Leslie E., Sophie Lebrecht, James W. Tanaka, and Michael J. Tarr. 2013. "Can Singular Examples Change Implicit Attitudes in the Real-World?" *Frontiers In Psychology* 4: 594.

Rudman, Laurie A., and Jessica B. Heppen. 2003. "Implicit Romantic Fantasies and Women's Interest in Personal Power: A Glass Slipper Effect?" *Personality and Social Psychology Bulletin* 29, no. 11 (November): 1357–70.

Saleem, Muniba, and Craig A. Anderson. 2013. "Arabs as Terrorists: Effects of Stereotypes Within Violent Contexts on Attitudes, Perceptions, and Affect." *Psychology of Violence* 3, no. 1 (January): 84–99.

Saul, Jennifer. 2013. "Implicit Bias, Stereotype Threat, and Women in Philosophy." In *Women in Philosophy: What Needs To Change*, edited By Katrina Hutchinson and Fiona Jenkins, 39–60. Oxford: Oxford University Press.

Slone, Michelle, Ricardo Tarrasch, and Dana Hallis. 2000. "Ethnic Stereotypic Attitudes Among Israeli Children: Two Intervention Programs." *Merrill-Palmer Quarterly (1982-)* 46, no. 2 (April): 370–89.

Strack, Fritz, and Roland Deutsch. 2004. "Reflective and Impulsive Determinants of Social Behavior." *Personality and Social Psychology Review* 8, no. 3 (August): 220–47.

Visch, Valentijn T., Ed S. Tan, and Dylan Molenaar. 2010. "The Emotional and Cognitive Effect of Immersion in Film Viewing." *Cognition and Emotion* 24, no. 8: 1439–45.

Walton, Kendall. 1990. *Mimesis as Make-Believe: On the Foundations of the Representational Arts*. Cambridge, MA: Harvard University Press.

Weisbuch, Max, Kristen Pauker, and Nalini Ambady. 2009. "The Subtle Transmission of Race Bias via Televised Nonverbal Behavior." *Science* 326, no. 5960 (December): 1711–14.

Willard, Greg, Kyonne-Joy Isaac, and Dana R. Carney. 2015. "Some Evidence for the Nonverbal Contagion of Racial Bias." *Organizational Behavior and Human Decision Processes* 128: 96–107.