The Imagination Model of Implicit Bias

Anna Welpinghus, TU Dortmund University

Philosophical Studies (2019). https://doi.org/10.1007/s11098-019-01277-1

Abstract

We can understand implicit bias as a person's disposition to evaluate members of a social group in a less (or more) favorable light than members of another social group, without intending to do so. If we understand it this way, we should not presuppose a one-size-fits-all answer to the question of how implicit cognitive states lead to skewed evaluations of other people. The focus of this paper is on implicit bias in considered decisions. It is argued that we have good reasons to assume that imagination plays a vital role in decision making. If this assumption is correct, it offers an explanation for implicit bias in many considered decisions: Human beings who have been frequently exposed to stereotypes have stereotype-congruent expectations as part of their background knowledge. They feed into their imagination, sometimes without their awareness. This model would allow us to explain the key characteristics of implicit bias without recurring to any unconscious attitudes over and above such background knowledge.

Introduction

This paper is about implicit social group-related biases. A person who harbors an implicit bias against members of a social group has the tendency to evaluate, perceive or judge them less favorably than members of another social group (and conversely with biases in favor of a social group). Furthermore, this tendency is to some extent independent of her explicit convictions – biased behavior occurs without intention and sometimes despite egalitarian convictions. Having been socialized in societies structured by gendered and racial hierarchies, many of us are disposed to show such implicitly biased behavior in both considered decisions and spontaneous actions.

I will use the term 'implicit bias' for this very disposition and not for the cognitive states that underlie it. I proceed from the contention that we should not presuppose a one-size-fits-all answer to the question of how implicit cognitive states lead to skewed evaluations of other people. While some recent work on implicit biases has focused on quick, spontaneous, often non-verbal evaluations in social interaction, I focus on implicit biases in considered decisions such as gender bias in the evaluation of CVs.

The model I develop in this paper connects implicit bias to the role of imagination in decision-making. I argue that that we have good reasons to assume that imagination plays a vital role in decision making. If this assumption is correct, it offers an explanation for implicit bias in many considered decisions, for human beings who have been and continue to be frequently exposed to stereotypes.[1] It is not necessary to posit any unconscious, inaccessible mental states to explain why a person is implicitly biased. The model builds on the hypothesis that traces of stereotypes occur in scripts for context-specific social situations. It draws our attention to the automatic and bias-prone nature of imagination.

In Section 1, I characterize the explanandum of models of implicit bias, namely a tendency to evaluate people in a biased way (which is not part of intentional discrimination and often hard to control by the person). I clarify the relation between the project of this paper and existing accounts of implicit bias in Section 2. In Section 3, I argue that we need a specific model of implicit bias for considered decisions. In Section 4, I discuss in which sense imagination contributes to decision-making. Then I show in Section 5 how representations of stereotypes feed into the process of imagination, resulting in implicit bias, before coming to the conclusions.

1. The tendency for skewed evaluations

Models of implicit bias aim to shed light on a disposition of quite many people: namely a tendency to evaluate, perceive or judge members of one or several social groups less (or more) favorably than those of other social groups, which is not part of intentional

1 I use the term 'stereotype' for a cultural representation. This usage deviates from using the term 'stereotypes' only for mental representations such as beliefs. Minds can represent stereotypes but so can artworks and narratives. Stereotypes ascribe a bundle of properties to members of social groups. If they are expressed in propositional form, stereotypes have the form of a generic generalization; this is a generalization that does not include a quantifier. An example is 'women are nurturing'. See Leslie (2017) on the semantics and cognitive functions of generic beliefs and their relevance to stereotyping.

discrimination and often hard to control by the person. This disposition constitutes the explanandum of accounts of implicit bias. In this section, I will describe it in more detail and develop a list of characteristics a model of implicit bias should account for.

To begin, let us consider some examples for implicit bias. First, take a study by Eric Uhlmann and Geoffrey Cohen (2005). They asked participants to assess two (made-up) CVs with regard to how well qualified the candidates from the CVs would be for a job as a police chief. One of the candidates had a lot of practical experience as a police officer but little formal education. The other candidate had quite some relevant formal education but little practical experience. One group of participants got a pair of CVs where the name of the street-smart candidate was male ('Michael') and the name of the educated candidate was female ('Michelle'). A second group of participants was presented with the same pair of CVs but the names were swapped. Most participants in both groups chose the male candidate. When asked for their reasons, members of the first group explained that being street-smart was crucial for doing the job well while members of the second group considered formal education to be crucial. In this study, a certain college degree was interpreted as a stronger proposition for Michael (by group 1) than it was for Michelle (by group 2). There is no reason to assume that people in the two groups differed in their beliefs on the qualifications for police chiefs before the experiment. Since the only difference between the CVs was the candidate's gender, the best explanation is that this information has influenced people's judgment.

Uhlmann & Cohen also calculated how strongly each candidate preferred a male or a female candidate and they let people fill out a questionnaire on their attitudes towards women (item from the Ambivalent Sexism Inventory [Glick & Fiske 1996]). They did not find a correlation between the score on this questionnaire and the strength of preference for a candidate. What predicted the preference was self-perceived objectivity: those who rated their judgment to be very objective tended to favor the male candidate more strongly. I will discuss this gender bias in the evaluation of CVs as an example of implicit bias in considered decisions.

As an example of implicit bias in spontaneous non-verbal behavior, take the way Jules Holroyd (2016) describes a racially biased job interviewer: "his behaviour is more hostile, he reacts with more irritability, he sits marginally further away from the black

interviewees" (p. 9). Holroyd's implicitly biased interviewer reacts with different gestures and postures to the applicant's behavior depending on whether the applicant is taken to be Black or White. Such behavior in a way constitutes a skewed evaluation, too, but it seems to be the result of affective reactions which arise prior to any considered decisions. In theories of emotions, such affective evaluations are called 'appraisals'. We may leave open whether they entail any judgments at all.

A third example is a character from Eric Schwitzgebel (2010): "Juliet, the implicit racist" is a professor who consistently evaluates the contributions of her White students as more insightful than the contributions of her Black students although there *is actually no difference* in overall insightfulness between the comments of the Black and White students.[2]

In all of these examples, we find the same structure: a person's evaluation of other people is skewed as a function of the social group membership this person ascribes to them: she evaluates one and the same qualification, one and the same gesture, differently depending on whether she takes its bearer to be male or female, Black, White or Brown, etc. This structure makes it apt to talk of a bias. It also sheds light on the relation between implicit bias and discrimination (in the sense of discriminating *against* a person): roughly, discrimination entails that a feature like social group membership makes a difference for the way someone is treated where it should not make that difference. If biased evaluations of the kind I have just described guide the way we treat a person, we discriminate against this person. Hence, such skewed evaluations can contribute to perpetuating existing inequalities between different social identity categories such as race, ethnicity or gender.

2 The latter two are fictional cases, while the first one does not include an indirect measure of gender stereotypes or sexist attitudes. In psychological studies, indirect measures (e.g. the Implicit Association Test) are thought to be better indicators for implicit attitudes than a direct measure of attitudes, i.e. openly asking participants about their attitudes (see also Section 2). An anonymous reviewer pointed out that I cannot rely on the assumption that hiring discrimination and hypothetical cases are the result of implicit attitudes, as they are measured by indirect measures. That is true. I do not take implicit bias to refer to the thing measured by indirect measures of attitudes. Neither do I claim that implicit biases are the result of the thing measured by indirect measures. In claiming that these cases are cases of implicit bias, I do not claim that they correlate with indirect measures of stereotypes or attitudes. All that the reader needs to grant me here is that people sometimes behave like it is described in these cases, and often enough for these cases to be of concern to us. There are, however, studies that link hiring decisions to indirect measures of stereotypes: for instance, Agerström & Rooth (2011) found a correlation between indirect measurers of negative stereotypes against obese people and hiring discrimination, and Rooth (2010) found a correlation between indirect measures of negative stereotypes of an immigrant group and hiring discrimination.

So far, I have talked about the sense in which *evaluations* (such as appraisals and judgments) are biased. Yet, we also often use 'biased' for a property of a *person*. A natural way to understand this way of talking is to say that a person who is implicitly biased against members of a social group has a *disposition* to evaluate them in the skewed way I have just described. But not any disposition to exhibit this skewed evaluation is the explanandum for research on *implicit* bias. When a person is implicitly biased, the way stereotypes skew an evaluation is to a certain degree *automatic.* For present purposes, we may understand the automatic nature of the process as follows: a person uses a stereotype for thinking about another person but using this stereotype has not been initiated by any intention or decision to do so.[3] The participants of Uhlmann and Cohen's study were asked whom they considered most qualified for the job. Those who followed these instructions intended to evaluate the CVs of Michael and Michelle for their qualifications, period. But in fact, many participants discounted Michelle's qualifications. Presumably most of them did so unintentionally. This is supported by the fact that they did not mention gender as a reason for their assessment but provided other justifying reasons (and those who were most sure of their objectivity showed the strongest bias). If they had instead reasoned that women should not be police chiefs, this would have counted as an explicit bias.

A further noteworthy characteristic of this disposition is the limited amount of *control* we have over it. The disposition can persist despite conflicting beliefs and desires. Hence, not only people with inegalitarian attitudes, but also professed explicit egalitarians can be implicitly biased. This fact has received quite some attention.[4] Yet, as Jules Holroyd (2016) argues, the focus on the wholehearted explicit egalitarian who is implicitly biased is too narrow: In order to understand the role of implicit attitudes for cognition and action, as well as their contribution to persisting social inequalities, we need to account for more cases. This is the case, according to Holroyd, because an explicit racist is also implicitly biased if he has roughly the same subtle hostile affective avoidance reactions as the explicit egalitarian. Holroyd also describes a third character, the implicitly biased "protocol-

3 Not being initiated by intentions is only one of several regards in which processes operate automatically (Moors & de Houwer 2006). I use this feature here to distinguish implicit from explicit biases: they are automatic in this regard. I do not claim that automaticity in general is reducible to nonintentionality.
4 I use 'egalitarianism' for the conviction that people deserve equal treatment, regardless of their social identity or any other characteristic. This conviction then does not entail a commitment to an equal distribution of material resources.

adhering racist". This person openly admits that he does not like Black people and prefers not to have them as colleagues, but he also wants to hire the most qualified person, no matter what their race is. He fails to do so because of his implicit bias.

I agree with Holroyd on this point. Furthermore, there are even more ways in which implicit bias can be at odds (or in line) with one's explicit attitudes. Bias can persist despite intentions to treat everyone equally and it can persist despite beliefs that are at odds with the bias. The bias of Schwitzgebel's Juliet is at odds with both her explicit beliefs and her intentions. Juliet's tendency to evaluate her Black students as if they were stupid, although she sincerely advances the view that there are no racial differences in intelligence, is an example of a conflict between implicit bias and descriptive (explicit) beliefs. Juliet also intends to be fair, but she is not. The behavior of Holroyd's protocol-adhering racist conflicts with his intentions to be fair, but not with his evaluative beliefs about people of color. Or consider a hypothetical participant of Uhlmann and Cohen's study who tries to be not influenced by gender stereotypes when evaluating CVs, and yet explicitly endorses several stereotypes about men and women. In this case, there is a conflict between intentions and actual behavior but not between beliefs and behavior.

Characters like Holroyd's explicit racist are not the only ones whose implicit biases are not in conflict with their explicit beliefs. Consider June, a colleague of Juliet who has the same racial bias as Juliet. June also does not intend to evaluate her Black students any differently from her White students. This distinguishes her from Holroyd's explicit racist. But unlike Juliet, June does not care about the fact that she might be unfair.

We can summarize the characterization so far as follows: a person harbors an implicit bias against (in favor of) members of a social group only if

(1) she is disposed to evaluate members of this group in a less (more) favorable light than members of other social groups;

(2) she has this disposition also when she does not intend to favor or disfavor human beings based on their group membership.

Furthermore, I have noted some other characteristics of the disposition described by (1) and (2): first, the way in which an evaluation is biased is in line with culturally shared

stereotypes. Second, the disposition can persist if a person also harbors egalitarian beliefs and/or intends to be unbiased.

2. Explaining the tendency for skewed evaluations

How do we explain what is going on in implicitly biased individuals in a way that accounts for these characteristics? A rough and ready description which I take to be widely acceptable is this: stereotypes leave traces in memory.[5] These traces get regularly activated and influence our judgment without our intention. The amount of control we have over this process is limited. This leads to a disposition to evaluate members of this group in a less (or more) favorable light than members of other social groups, also if one does not intend to favor or disfavor human beings based on their group membership. Models of implicit bias put flesh on this rough description in different ways. In order to clarify the contribution of the imagination model I will develop later in this paper, let me make a few remarks.

First, a note on terminology: the term 'implicit bias' can be used for different elements of this description. It can be used for the explanandum, this is, for the very disposition described in Section 1. It can also be used for the explanans, this is, for the traces of stereotypes in memory. These are then often called 'implicit attitudes'. Implicit attitudes are either understood as *implicit mental representations* or as overall (implicit) likings or dislikings.[6] In each case, there are different ways of understanding what is implicit about them. I will use the term 'implicit bias' for the explanandum, this is, for the disposition

5 This formulation deliberately echoes Greenwald and Banaji's (1995) seminal formulation of implicit attitudes as "traces of past experience" which influence responses (p. 5).

6 Attitudes in the sense used in social psychology are not directed at propositions. They are directed at objects or categories and they constitute likings or dislikings of their objects. Edouard Machery (2016) argues that implicit attitudes are not mental representations. Referring to Fazio and Olson (2007), Alessandra Tanesini describes attitudes as "cognitive shortcuts, based on experience, that summarize one's overall evaluation of an object", which have a cognitive base consisting of representational states (Tanesini 2018, p. 410). According to her conception, attitudes are not implicit representations of stereotypes but these implicit representations may constitute the cognitive base for group-directed attitudes. Note, however, that attitudes, as Tanesini and Machery conceive of them, are not identical to the disposition I call an implicit bias. The difference between the disposition to evaluate others in a more or less favorable light and attitudes becomes clearer when we consider how positive and negative attitudes relate to each other, and how the dispositions to evaluate others in a favorable or unfavorable light relate to each other: every disposition to evaluate members of Group A favorably comes with a disposition to evaluate non-As unfavorably, and vice versa. This is simply the result of the fact that this disposition concerns the evaluation of a group relative to others. In contrast, a negative implicit attitude towards non-As does not come with a positive implicit attitude towards As: a person can be fond of As and have a neutral attitude towards all others. Or she might have a hostile attitude towards all non-As without being particularly fond of As.

described in Section 1. A prima facie reason for this terminological choice is that I do not think that we will find an informative characterization of underlying mental representations that fits all variants of the disposition (more on this in Section 3). The point of this paper, however, is not to argue about words. Its main arguments are compatible with using the term for an implicit mental representation or overall liking or disliking.

Second, a note on measurement: The most popular instrument to measure implicit attitudes is the Implicit Association Test, or IAT. When taking this test, participants are instructed to press buttons when pictures or words show up on a computer screen instead of being asked about their attitudes directly. For instance, in an IAT on racial attitudes, subjects are first instructed to press the same button whenever they see a face of a Black person and a positive word, and another button, whenever they see a White face or a negative word. In subsequent rounds, pairings are switched. The subjects' performance indicates how easy it is for them to pair pictures of members of a social group with a family of concepts. The IAT measures semantic associations and/or evaluative attitudes. How well the IAT also predicts implicit biases, understood as dispositions to treat members of different groups differently, is still debated.[7] Also, there is no reason to assume that the IAT measures all of the mental representations that contribute to implicit biases.

A final point of clarification: the rough-and-ready description raises at least two questions: first, how are traces of stereotypes represented in the mind? Second, how do they come to skew a person's judgments so that this person harbors the above described disposition? Several recent papers in philosophy have focused on the former question, with special emphasis on the issues of whether traces of stereotypes are propositionally structured or mere associations between concepts, and of whether they qualify as beliefs. Mandelbaum (2016) takes them to be unconscious beliefs, while Madva (2016b), Levy (2015), Brownstein (2018) and Toribio (2018), for instance, argue that implicit representations are not beliefs because they do not update in the same way as beliefs do. My focus, in contrast, is on the second question – although I will say a bit on the first question, too. As l will show in Section 5, focusing on the second question may allow us to provide a role for both

7 E.g. Oswald et al. (2013); Greenwald, Banaji & Nosek (2014); Lai et al. (2017); Carlsson & Agerström (2016); Singal (2017); Kurdi et al. (2018).

propositionally structured representations and associations. I will take a neutral stance towards the trickier issues pertaining to the nature of beliefs.

Nonetheless, the debate on whether implicit attitudes are propositionally structured tackles some tentative empirical results that constrain answers to the main question of this paper. These are results on the malleability of IAT scores. Mandelbaum, in particular, highlights empirical evidence that IAT scores are influenced not just by conditioning, but also by interventions that, according to Mandelbaum, presuppose inferential belief updating. In the study by Gregg, Seibt and Banaji (2006), participants were introduced to two fictional tribes. One was described as benevolent, the other as belligerent. The IAT scores of the participants indicated negative and positive attitudes towards these tribes. IAT scores were significantly influenced when participants were asked to now imagine that the first tribe was belligerent, and the second tribe was benevolent. In a study by Briñol, Petty and McCaslin (2008), participants read good and bad arguments for hiring more Black professors. This influenced their IAT scores; they were more successful in pairing positive words and Black faces after having read the strong arguments. The results are to be taken with a grain of salt: the experiment by Gregg et al. had a small sample size, while Briñol et al. do not report the sample size of their study. Still it would speak for the account developed in this paper if it was compatible with these results. As I will show in Section 5, the imagination model is quite capable of integrating them. The studies Mandelbaum discusses can be described as interventions on imagistic processes. But before getting there, in the next section I provide reasons for this paper's focus on bias in considered decisions.

3. Variants of implicit bias

So far, we have been concerned with the question what different examples of implicit bias (bias in evaluating CVs, in spontaneous subtle, non-verbal reactions during a job interview, in grading papers and in evaluating the insightfulness of a student's comment) have in common. In this section, we will have a look at some differences between them. As I will argue, because of these differences, we might need different models for different variants of implicit bias. At least, we cannot start from the assumption that there is a one-size-fits-all model. Showing that there is such a model would be a substantial theoretical achievement, and as long as we have not shown this, we should instead acknowledge that

we may need different models for these variants. Afterwards I show that recent models of implicit bias which focus on bias in spontaneous, often automatic and nonverbal evaluations, do not adequately explain how bias in decisions that involve some deliberation occurs. I call these decisions 'considered decisions'. Hence the need to fill this gap.

This point does not hinge on the dispositional understanding of the term 'implicit bias' I am using in this paper. It does depend on the following claim: in considered decisions, just as in spontaneous behavior, traces of stereotypes influence judgments without intention, while this often escapes our awareness and can happen despite conflicting beliefs and desires. Considered decisions are like the ones that participants in Uhlmann's and Cohen's study are asked to make. Examples of biases in spontaneous affective evaluations are found in Holroyd's biased interviewers who show subtle avoidance behavior towards Black interviewees.

There is ample evidence that considered decisions can be biased in the sense above (see e.g. Rooth (2010) and Rooth & Agerström (2011) for hiring decisions; Croskerry, Singhal & Mamede (2013) and Croskerry (2003) for medical diagnoses, Kurdi et al. (2018) for a range of spontaneous and considered behaviors). Many theoretical models also predict that implicit processes influence considered decisions. For instance, Gawronski and Bodenhausen's (2006) Associative-Propositional Evaluation (APE) model describes several interactions between quick, automatic assessments and slow, deliberate propositional reasoning.[8] Greenwald and Banaji (2017) have recently presented an understanding of the relationship between indirect and direct measures that does not presuppose two cognitive systems. They emphasize how strongly our conscious experiences and judgments result from unconscious processes. Unconscious stereotypes may skew conscious judgments.

[8] Gawronski & Bodenhausen's (2006, 2011) APE model is a dual process model of implicit social cognition which locates the source of implicit biases primarily among the type-I processes. According to dual-process models of the mind, cognitive processes can be divided into two types: type-I-processes provide a quick assessment of inputs. They require relatively few cognitive resources but are fairly error-prone. Type-II processes, in contrast, are slower, cognitively more demanding but less error-prone. APE's basic claim is that indirect measures of group attitudes like the IAT measure quickly and automatically activated associations, while questionnaires on one's attitudes towards social groups measure the outcome of propositional processes. Propositional processes are concerned with validation of information. A precursor of APE is Strack & Deutsch's (2004) reflective-impulsive model of social behavior.

The following differences between considered decision and spontaneous affective evaluations are prima facie reasons for considering different models for them: In considered decisions, we deliberately evaluate the other person and we know what our evaluation is. We also have sufficient time to do so. Thinking about what to do is furthermore unhooked from sensory input and action guiding mechanisms. Spontaneous interactions leave a person little, if any, time to think. The cognitive processes that are giving rise to implicit bias are not unhooked from sensory input and action. They are non-verbal and closely connected to affective reactions. While we can speak of a non-verbal evaluation of the interviewee's likability or trustworthiness, it is not clear whether the interviewers deliberately evaluate the interviewees in this regard at this moment. At the very least, it is possible that they are not aware of the fact that they are evaluating the interviewee in this regard.

As far as they have considered the question how traces of stereotypes come to skew evaluations, recent philosophical models of implicit bias tend to focus on spontaneous affective evaluations. An example for this is Alex Madva's and Michael Brownstein's (2018) model of implicit biases, as well as Brownstein's own take on spontaneous inclinations in his book on "The Implicit Mind" (2018). Another example is Jules Holroyd's (2016) sketch of a model. Madva and Brownstein argue that, in implicit cognition, putative purely semantic associations entail affective and motivational elements, while putative implicit prejudices include semantic associations. Brownstein (2018) describes implicit attitudes as integrated bundle states with representational, affective, and behavioral content as well as a tendency for alleviation. Holroyd (2016) sketches the idea that implicit bias can be explained as a result of a bundle of co-activated contents, representational, behavioral, affective and evaluative.[9] While Brownstein (2018) and Madva & Brownstein (2018) argue that we should think of this bundle as one integrated state with evaluative, affective and behavior-guiding contents, Holroyd thinks of the bundle as several cognitive states that are often activated together but can also come apart.

9 Holroyd credits Currie and Ichino (2012) but they are actually proposing something different than Holroyd does. Their paper is a critical comment on Tamar Gendler's (2010) concept of aliefs. Their proposal is quite in line with the point of this section: we need different cognitive explanations for the different behaviors Gendler explains through aliefs. According to Currie and Ichino, racist hiring behavior could be explained through co-activated beliefs and desires, while we might have to recur to automatic associations (a "poor cousin" of aliefs) for racist subtle affective reactions.

Both models can with some plausibility explain unintentional, spontaneous, non-verbal biases in interaction: working to some degree independently of explicit judgment, subtle affective reactions entail motor routines and hence influence how friendly (for instance) a person is – without any intermediate deliberative steps. But in considered decisions, people do execute such intermediate deliberate steps, both in order to come to an overall judgment about, say, a person's qualifications and in order to act in line with it. Brownstein (2018) applies his bundle account of implicit attitudes to a hiring discrimination case. Someone is reviewing CVs. When she perceives a female name on a CV, she feels a tension Brownstein describes as "risky hire!" and a corresponding behavioral tendency is activated, namely: "place in low-quality pile" (p. 60). However, it is highly unlikely that this specific action tendency is intimately tied to the perception of female names on CVs. If the same person was not engaged in the task of piling CVs but of formulating her impressions orally, she would not have a behavioral tendency to pile anything. It is highly questionable whether there is a motor routine specifically tied to seeing female names on CVs. Rather, because of implicit attitudes, the person in Brownstein's example comes to view women more easily as risky hires. What she does with this assessment depends on her further beliefs and desires.

The point applies to the other bundle accounts as well. Hence, recent philosophical accounts of implicit bias say interesting things on the question 'how do representations of stereotypes come to skew an evaluation?' for spontaneous affective evaluations, but these answers cannot be simply transferred to considered decisions.[10]

Note that I am not claiming that spontaneous, affective evaluations and deliberation operate completely independently of each other. Of course, it is possible and plausible that something like the semantic/affective/motivational states that Madva and Brownstein describe also play a crucial role for biased considered decisions. But it is also possible that biases in considered decisions do not, or not exclusively, result from such semantic/affective/motivational states.

10 My argument is more far-reaching than the arguments presented by Holroyd and Sweetman (2016) against a one-size-fits-all-model. Holroyd and Sweetman are primarily concerned with the way in which implicit biases are represented in the mind, while my focus lies on the question how traces of stereotypes come to skew a judgment. Holroyd and Sweetman do not distinguish between biases in spontaneous evaluations and in considered decisions. Furthermore, they still assume that automatic associations are underlying all forms of implicit bias, while I do not assume this.

4. Imagination in decision-making

In the next two sections, I will argue for the following claim: we have good reasons to assume that imagination plays a vital role in decision making. If this assumption is correct, it offers an explanation of implicit bias in many considered decisions. Roughly, the idea is this: when you sit at your desk with CVs in front of you and choose whom to invite for a job interview, you will imagine the reasonably qualified candidates in the job to be given. You imagine Michael having the job: which challenges would he encounter? How would he deal with them? How would colleagues and stakeholders react to the way he deals with them? You do the same for Michelle. Based on the things you imagine, you will judge how the candidates will likely fare in this job. This process integrates a large body of social knowledge, among them representations of stereotypes. They can bias your judgments.

The most compelling reasons for this model are of a theoretical nature: the focus on imagination allows us to explain how implicit bias occurs by referring to concepts that are already in use in philosophy of mind and cognition. Together with some plausible assumptions about the way human beings pick up and store stereotypes, we can then develop a neat model of implicit bias in decision making. However, imagination is not the only route to implicit bias in considered decisions, and the question when imagination is at play in considered decisions will remain open in this paper. I will present some theoretical reasons for the hypothesis that situations in which we have to recur to imagination make us particularly prone to biases that are difficult to control and easily escape awareness. This hypothesis is compatible with the existence of other pathways to implicit bias.

In this section, I tackle the question what it entails to *imagine* that, for instance, Michael gets the job. While it is beyond the scope of this paper to propose a definition of imagination as such, some clarifications will help us to get a grip on the characteristics of imagination as it figures in decision-making. In Section 5, I lay out how the role of imagination in decision-making explains persisting dispositions to exhibit biased judgments.

Activity, products and attitude

According to one central meaning of 'imagination', imagination is a mental activity. It is a "temporally-extended constructive process of assembling mental representations" (Van Leeuwen 2013, p. 221). Van Leeuwen calls this process 'constructive imagination'. The

mental representations a person comes up with during constructive imagination are the *products* of her imagination. Let us call them 'imaginings'. A mental image, for example, is an imagining. Mental images are formatted like perceptions but they are not caused by sensory stimulation in the same way as perceptions: if I experience an image of an apple but no light from an apple-shaped object has stimulated my retina, that mental image is a product of my imagination (Nanay 2016a). Furthermore, the verb 'to imagine' can describe a *propositional attitude*. In the sentence 'you imagine that Michelle has the job', 'imagine' refers to the attitude you take towards the proposition that Michelle has the job. Imagining that p entails roughly that you take p to be true in some fictional context (Van Leeuwen 2016a). Since imagining that p is different from believing that p, it is perfectly compatible to imagine that Michelle gets the job while believing that she did not.[11] While it is comparatively straightforward to understand what we do when we imagine an object – we construe a mental image – it is much less clear what we do when we imagine a proposition.

What does it mean to imagine a proposition?

This paper's concern is to clarify how imagination in the context of decision-making leaves us vulnerable to implicit bias. Therefore, we need to describe the imagination of propositions, as it plays a role in decision-making. We do not have to clarify what imagination in general is and in which sense it is different from other mental operations. Here is the suggestion: Imagining that Michelle has gotten the job is a way to reason about counterfactual scenarios. However, it does not just consist in drawing conclusions from a limited set of premises by applying a limited set of rules of inference. Rather, imagining that p can (at least in this context) be described as *mental simulation* of a counterfactual scenario. In taking the attitude of imagination towards p, we engage in constructive imagination – we mentally simulate what would be the case if p were the case.

By simulating what would happen if p, a person integrates information in a way that is different from only applying rules of inference to a limited set of premises. We draw

11 This distinction is inspired by Neil Van Leeuwen's (2013) distinction between constructive imagination, attitude imagination and mental imagery. Van Leeuwen sometimes uses 'constructive imagination' for the activity, as I do here, too (2013). But sometimes he uses this expression for the *capacity* to imagine different scenarios that are to be considered in rational choice (2016a). Another difference to Van Leeuwen's taxonomy is that I use the term 'imaginings' for the products of imagination. This allows that some products of imagination are not formatted like perception.

instead from a wide range of general knowledge about the world. The sources we draw on in imagination might include tacit background beliefs about men and women in general. They may also be guided by semantic associations. And, as Nichols and Stich (2000) put it in an influential paper, imagination can be described as guided by *cognitive scripts*. Nichols and Stich draw from an established framework in AI, according to which scripts are bundles of expectations about the way events typically unfold in a particular setting (Schank & Abelson 1977). Schank and Abelson's example is a restaurant script. To illustrate their idea, consider the following story: 'A man goes into a restaurant. He orders a dish. After a while he pays and leaves.' If you hear that story, you will assume that he has received his dish and eaten it. However, the fact that he has eaten his food does not follow from any of the sentences in the story. You assume so because these expectations belong to your restaurant script. We have cognitive scripts for a lot of specific places and social situations. We use scripts to navigate within the social world. Scripts are exactly what we use in order to run a simulation in our head.

Simulation also calls for affective responses (see Van Leeuwen (2016c) and Gendler & Kovakovich (2005) for helpful models). The person might like, respect or belittle the way in which she imagines Michelle would solve a challenge. Later I will discuss how the connection to emotions contributes to implicit bias.

Epistemic function

In order to say that imagination contributes to decision-making, we need to clarify how it contributes to making *adequate* decisions. It seems to be the case that imagination sometimes provides us with information that we cannot get from other sources, or at least, that would be harder for us to obtain otherwise. Bence Nanay (2016b) makes this point for assessing options regarding the course of our own life. He argues that the rational choice model of decision-making leaves unexplored why we assign specific values to the options we have. For this, we do and must rely on imagination.

Nanay does not define imagination in his paper. However, my characterization of imagination as a form of mental simulation provides us with a better understanding of its epistemic role: precisely because we integrate a large array of knowledge (explicit beliefs as well as scripts) into a coherent simulation of events, it provides us with ideas about the way

events could likely unfold. We would not have arrived at these ideas if we had merely reasoned over a fixed set of premises by using a fixed set of inferences.

Peter Langland-Hassan (2016) makes a related point when he describes imagination as guided by both the subject's intentions and by lateral constraints. The latter ones are background knowledge and scripts which enable a simulation to simply unfold in our mind, but we can and sometimes must decide how to elaborate on these scripts. Guidance by intentions allows us to explore different likely or unlikely consequences. Langland-Hassan describes this as a feedback loop between intentionally chosen contents and lateral constraints. With this suggestion, he aims to solve the following puzzle: on the one hand, imagination must be constrained by beliefs about the subject matter in question. Of course, we can mentally simulate very unlikely, widely implausible courses of events. However, when we use imagination in decision-making, we aim at a realistic assessment of what would happen if we chose one option or the other. On the other hand, the point of imagination is that we come up with new ideas and can explore possibilities at will. In line with Gendler and Kovakovich (2005), we might add that one way in which imagination provides us with information is via emotions: we respond with real emotions to imagined scenarios and these emotions can give us a sense of what we would like or not.

However, imagination is also subject to systematic biases. Nanay emphasizes that we underestimate how much our preferences change over time. The description of imagination as mental simulation also points towards sources of error: given that we draw on a wide range of representations without having to choose them beforehand, we might use false or irrelevant ones. If, as Nanay argues, imagination is nonetheless the best thing we have for making certain complex decisions, this would explain why these decisions are vulnerable to systematic biases.

Elaboration

Imagination can take place at different degrees of elaboration both in terms of detail and in terms of the range of options that are imagined. I can imagine quite elaborate narratives (when I am exploring at length about how Michael and Michelle would fare in the job). But I can also imagine simple events: I might mentally complete a gesture by the job candidate in

front of me. Imagination in the context of decision-making can consist of sketches and glimpses.

(Non-)deliberateness

We do not always explicitly *choose to rely on imagination* during decision-making. Accepting Langland-Hassan's point that what we imagine is set by intentions and lateral constraints does not commit us to the point that we have intended to run a simulation at the onset of an imaginative episode. Many of us will use imagination as a natural part of thinking through counterfactual scenarios without reflectively choosing this method.

Awareness

It seems that the function of imagination can be best fulfilled if the simulation is conscious: through simulating counterfactual scenarios we make our expectations about what would happen available for reasoning about what to do. This does not rule out the possibility of unconscious simulation. We do not have to settle the issue for present purposes. In any case, it is certainly not necessary that we are aware *that* we are engaging in imagination when we do so if it is part of our ordinary way of counterfactual reasoning.

Prevalence

If elaborate daydream-like simulations of events are not a mark of imagination, but imagination can consist of sketches and glimpses, and we are not always aware that we are engaging in imagination, imagination may be more common than we think. Yet, how often do we actually use imagination in decision making? Under which conditions do we do so? This is important for the question how many implicit biases in considered decisions are in fact explained by the imagination model. Certainly, we do not recur to imagination in all decisions. Sometimes we follow simple rules. But what about the cases where we think through the consequences of our options?

I cannot fully answer these questions in this paper. There are some open empirical and conceptual issues at stake here. The open conceptual issue is what counts as imagination. I do not claim that thinking through hypothetical consequences already constitutes imagination. Earlier I said that imagination means to mentally simulate what would happen if p was the case and contrasted this to reasoning from a limited set of premises. But that

does not help much, since it remains an open question when one ends and the other begins. This question has evoked a debate that cannot be settled in this paper. Liao & Gendler (2019), especially the section on supposition, contains a summary of the issues.

Complex decisions about the future like 'what will happen if Michael gets the job?' are the most obvious candidate for cases where we recur to imagination. It seems to me that when we need to assess complex abilities or character traits of people, simulating how they would act in different relevant scenarios is what we usually do. Recall the proposed function of imagination: to integrate a large array of knowledge (beliefs, scripts, how to react emotionally). This seems what we must do when assessing complex abilities or traits.

We might also recur to imagination when it is less obvious that we assess character traits or abilities. When grading papers, Juliet is assessing the quality of the papers in front of her. But she sometimes has to interpret a somewhat confused remark as either a clumsy formulation of an original thought or as words the student put together without an understanding of the subject matter. In order to decide this, Juliet might imagine how the student is thinking and writing. This is admittedly speculative. Maybe imagination is not an important source of bias in grading, but it could well play a role.

To conclude this section, imagination, understood as mental simulation, is a valuable tool, maybe an indispensable tool, for decision-making. This is why we frequently use it and also why we should use it to make adequate decisions. However, it is prone to errors. As we will see now, this combination of characteristics leaves us vulnerable to implicit bias.

5. Implicit bias as a product of imagination

In this section, I will show that, implicit biases are to be expected in decisions that involve imagination. While I do not think that this is the only route through which decisions can become biased, it is a route that allows us to explain the intriguing features of implicit bias quite well. I first discuss how traces of stereotypes are represented in the mind. This allows showing how they feed into imaginative processes and thereby bias judgments. We will see that some features of this process explain why a person's disposition for biased judgments can persist despite conflicting intentions and beliefs, and why it can escape a person's awareness. Finally, I point out which tentative empirical support the model enjoys, and which questions are still open.

Traces of stereotypes in the mind

When you imagine how Michelle and Michael would fare as police chiefs, your imagination is constrained by the things you believe about Michelle, Michael and the job. In addition, you draw on a wide array of representations; for instance, beliefs about men and women in general or police work. Some of these representations likely contain traces of culturally prevalent stereotypes. In addition, some stereotypical assumptions underlie our everyday interactions as well as shared narratives – and it is difficult, if not impossible, to successfully interact with others without presupposing these stereotypes. This does not mean that by presupposing them a person already uncritically accepts them. Nor does it mean that all members of one society have the same stereotypes.

In the imagination model, traces of stereotypes can be represented in different formats and all of them can feed into an imaginative episode and bias a judgment. I will discuss traces in the form of generic beliefs, of context-specific expectations, and of semantic associations, without assuming that this list is exhaustive. As I mentioned in Section 2, generic beliefs and associations have been discussed in the literature (usually as competing models of implicit attitudes). Context-specific expectations play a role in imagination and paying attention to them helps us account for some of the seemingly puzzling characteristics of implicit bias.

First, stereotypes are represented in the form of generic beliefs about members of social groups. Generics have the form 'A's are B'. Not all generic beliefs about the members of a social group are malevolent. However, benevolent beliefs about a social group can also influence a decision to its members' disadvantage. Classic examples are benevolent stereotypes that present women as high in warmth but low in competence (Fiske et al. 2002). If a person imagines Michelle as friendly but not assertive because she believes that women are friendly but not assertive, this will work to Michelle's disadvantage whenever assertiveness is an asset.

However, such global generic beliefs are not the only way in which stereotypes are represented. Stereotype-congruent expectations, as they occur in cognitive scripts for specific social situations, are another way in which they can be represented (Casper, Rothermund & Wentura 2010). They deserve our attention because of their central role in

guiding imagination. For example, your restaurant script might include the expectation that the man takes the bill if he is dining with a woman. I remain neutral about the question whether expectations, as they occur in scripts, are beliefs. We might formulate the expectations I have just mentioned as generic beliefs like, 'in a restaurant, if a man and a woman dine together, the man takes the bill'. However, scripts concern the way people behave in a specific context. They do not concern *global* ascriptions of character traits or abilities to members of a social group independently from a specific social context. We may understand an explicit inegalitarian as a person who holds true some sexist stereotypes such as 'women cannot provide for themselves, so men have to do it' and who is prepared to use this belief as a premise in reasoning; she would not learn anything new about herself if she was told that she does so. An explicit egalitarian would seriously disavow such a stereotype. But she may nonetheless have the expectation that the man takes the bill. In this sense the stereotype leaves a trace in her mind.

Although it is possible to understand my model as a belief model of the cognitive underpinning for implicit bias, there is a significant difference to Mandelbaum's (2016) claim that implicit bias derives from reasoning over unconscious beliefs. Mandelbaum seems to suggest that the difference between explicit stereotyping and implicit bias is that explicit stereotyping is guided by a conscious belief and implicit bias by an unconscious belief with the same content. In other words: a person who explicitly argues that Michelle should not get the job as a police chief because women are nurturing rather than assertive, employs a conscious belief. A person who discounts a woman's qualifications for jobs that seem to require assertiveness and no nurturing attitude, has an unconscious belief that women are nurturing rather than assertive.[12] In my model, however, the expectations that guide implicit bias are different from those that guide explicit stereotyping. They are expectations tied to a specific context. Such an expectation is not always easily identified as a trace of a particular stereotype. Thus, a person will often not notice an incongruency between these expectations and her more generic beliefs about groups. And this would

12 A note on the elephant in the room: what you imagine also rests on your scripts about police work and norms about good police work. The judgment about Michelle as unfit for the job is the result of a mismatch between stereotypes about women and a model of good police work. If, however, dominant cultural models of good police work are inadequate in the sense that a police officer who fulfills the model does not act particularly effectively and fairly, we have good reasons for changing our model of good police work, over and above the model's potential contribution to discrimination against female job candidates.

explain how a person can retain both a trace of a stereotype and believe that this stereotype is wrong at the same time.

Mandelbaum must assume that a global generic belief is driven into unconsciousness in people who sincerely disavow that very belief, in order to explain how a person might retain both – obviously contradictory – beliefs. But if the context-specific expectation does not obviously contradict a person's egalitarian beliefs, this would explain how both beliefs can be retained even though they are contradictory (or at least in some other sense incongruent). We do not have to assume that a person's remaining representations of stereotypes are driven into unconsciousness as a consequence of disavowing the generic belief. An expectation may be as accessible as an ordinary belief. We nonetheless may call such expectations 'implicit' if that means that they are not easily identified as representations of stereotypes.

Integrating expectations into the model also accounts for Alex Madva's (2016a) point that not all representations of stereotypes cause implicit bias: I can represent a stereotype about group A without expecting that members of group A conform to the stereotype. In other words, I do not have any expectations that partially represent that stereotype. Hence, it does not feed into my imagination about members of group A.

The imagination model also allows for the influence of *semantic associations* on imagination. The underlying assumption of models that explain implicit bias as a result of semantic associations is that semantic memory is associatively structured (Gawronski & Bodenhausen 2006): if one concept is activated, associated concepts are more easily activated than concepts that are not associated with the activated concept. Say, because of the associative structure of her semantic memory, it is easier for a person to think about musical activities than it is to think about intellectual activities when thinking about a Black colleague. Hence, her imagination might be driven this way. Note that scripts explain this tendency, too. We can leave it open whether we need semantic associations *in addition* to scripts for explaining how imagination contributes to group-related biases in decision-making.

So far, I have argued that exposure to stereotypes shapes how Michael and Michelle would behave in the job and how others react to them. Another source of bias concerns the way

we *evaluate* Michael's and Michelle's imagined performance. Our evaluation is influenced by the affective responses the imagination provokes. Say, assertive women have been presented to me as scary when growing up, so I have learned to respond to them with some anxiety. Now I imagine Michelle doing her job as a police chief in an assertive manner and I feel uneasy about her.[13] Imagining assertive Michael does not lead to feelings of discomfort; to the contrary, I trust him. I use these feelings as an indicator of my overall judgment about who should get the job: Michael is the man for the job. Michelle? Not sure if she is trustworthy.

Characteristics of implicit bias

Now we have the ingredients to account for the characteristics of implicit bias identified in Section 1: a disposition for skewed evaluations, without intention, which may persist despite conflicting intentions or beliefs.

We can now answer the question how representations of stereotypes come to skew a judgment about a person if imagination is at play. There are several ways in which imagination can lead to implicit bias: first, I might base my imagination on false beliefs or inaccurate scripts. Because I aim for the scenarios to be realistic, I do not imagine any scenarios that contradict my beliefs and I aim to imagine the consequences of relevant beliefs. If some of these beliefs are false, I might imagine the consequences of false beliefs. These can be very unlikely to occur in fact. Second, and maybe more important, while the scenarios I imagine are all reasonably realistic and none of them is based on blatantly false beliefs, the range of scenarios I imagine can be one-sided. Scripts, generic beliefs and maybe associations will leave some realistic scenarios unexplored: I did not imagine X although X was just as likely as the scenarios Y and Z that I have imagined – given my justified true beliefs. On the other hand, maybe I imagine some rather unlikely outcomes and they seem extremely plausible to me. In both cases, the scenarios I explore are one-

13 Heilman et al. (2004) found evidence that many people like successful women in male-dominant fields less than equally successful male colleagues. Williams and Tiedens (2016) conclude their review with the finding that women who behave in an explicitly dominant manner are considered less likable and less hirable than non-dominant women, while those whose dominant performance is implicit do not face these negative reactions.

sided. Third, affective reactions show how I evaluate the different scenarios I have imagined, but these affective reactions may get it wrong, too.[14]

It is possible that imagination goes awry in several of these ways at the same time but one of them can already skew an overall judgment: in the mind of a biased person, the same qualification (a particular degree, say) counts more for Michael than it would have counted for Michelle because this person imagines Michael with this qualification as more competent that she would have imagined Michelle with the same qualification. This is the structure that is characteristic of implicit bias in general – a property of a person is evaluated differently as a function of her (ascribed) social group membership.

In Section 1 I said that implicitly biased judgments are influenced by ascriptions of group membership even if a person does not *intend* to favor or disfavor members of the respective social groups. The imagination model accounts for this because it entails that we do not intentionally choose all sources we use during imagination. We let the simulation run, guided by lateral constraints, as Langland-Hassan (2016) would put it. The onset may be intentionally chosen, but that stereotypical expectations skew the possibilities we explore and how we evaluate them is not chosen.

Furthermore, as I pointed out in Section 1, it is a characteristic of implicit bias that the disposition can persist despite conflicts with other attitudes. First, the disposition sometimes persists despite intentions to the contrary; second, it sometimes persists although a person is holding descriptive beliefs which contradict the stereotypes that apparently influence her judgments.

The first sort of conflict occurs when, for instance, I intend that a job candidate's gender does not influence my assessment of the candidate's suitability for the job. Why is this intention, however sincere, not sufficient for eliminating implicit bias? It is due to a combination of two factors: first, as before, it is due to the fact that we let the simulation run without intentionally choosing each and every premise. Again, what is a strength of the process also results in a weakness, in this case limited control when it would be desirable.

[14] Yet another route is proposed by Strack & Deutsch (2004): imaginings may activate behavioral schemata via purely associative processes, and by this influence our actions. I thank an anonymous reviewer for bringing this route to my attention.

Yet, as long as imagination plays a valuable epistemic role, we will not refrain from imagination altogether.

The second sort of conflict occurs when implicit bias persists although the person harbors *egalitarian beliefs* with regard to the social groups she treats or views differently. How does my model explain this? As I have argued earlier in this section, because stereotypes are represented in many different ways in memory, it is not sufficient to reject some sexist assumptions in order to eliminate all those expectations that represent sexist stereotypes (for instance). I argued that it is sometimes difficult to understand that a context-specific expectation is in tension with one's egalitarian beliefs. Another reason is that it might be epistemically and otherwise costly to give them up. In addition, since we only have to let the simulation run instead of consciously choosing our premises, we are not aware of all the sources that feed into a particular imaginative episode. Hence, to sum up, the imagination model naturally explains why implicit biases persist despite conflicting intentions or conflicting beliefs.

Support and open questions

Over and above the theoretical considerations, what reasons do we have to assume that imagination is indeed a route to implicit bias in considered decisions? There is some tentative support that imagination is at play in implicit bias. Somewhat indirect support stems from empirical studies on interventions that impact IAT scores. Some of them apparently manipulate how we imagine people, for instance by exposure to pictures of counter-stereotypical role models (Dasgupta & Greenwald 2001), or by encouraging mental imagery (Blair, Ma & Lenton 2001). In a replication study, Lai et al. (2014) found that exposure to counter-stereotypical role models was among the effective interventions for reducing implicit bias. The intervention that was most effective in their study was imagining a vivid counterfactual scenario with counter-stereotypical characters. A study by Joy-Gaba and Nosek (2010), however, found only small effects of exposure to counter-stereotypical exemplars on IAT scores. Furthermore, these effects were not significant in all experimental conditions. Both Lai et al. and Joy-Gaba & Nosek concluded that it seems important to not only include positive exemplars of Black people, but also to include negative exemplars of White people. Contrary to what one might expect if the imagination model holds,

interventions that included imagining oneself in the shoes of a Black person did not reduce implicit preferences for Whites in the study by Lai et al. (2014). Maybe these interventions did not prompt the participants to expect different things of Black and White people than they did before.

The studies that Mandelbaum has put forward as evidence for his unconscious belief model also manipulate imagination. In the study by Gregg, Seibt and Banaji (2006), participants were *asked to imagine* one tribe as belligerent and the other as benevolent. This instruction affected their subsequent IAT scores. So here is an explanation in line with the imagination model: The descriptions of the 'tribes' (which were depicted by abstract shapes) is apt to evoke narratives and stereotypes of friendly and unfriendly human tribes which are already part of Western culture – and, by the way, entail colonial overtones. Hence participants could rely on their cognitive scripts in order to imagine these tribes. In the study by Briñol, Petty and McCaslin (2008), different scenarios were made available to the participants in both conditions ('we recruit new excellent academics' vs. 'search for excellence is threated by Affirmative Action imperatives'), leading to different subtle affective reactions towards Black faces, and hence IAT scores.

This evidence is indirect because it took IAT scores and not decisions as their dependent variable. It is in line with the imagination model that interventions on imagination make it easier to see White people in a negative light and Black people in a positive light, especially right after the intervention. But the imagination model predicts that interventions on imagination can also reduce bias in decisions like hiring, grading, making judgments on trustworthiness, etc. It would be interesting to test how such interventions on imagination influenced decisions.

The imagination model of implicit bias offers a neat way of explaining why discrimination can occur unintentionally, is hard to control and may escape awareness in considered decisions. While this is compatible with there being other pathways to biased decisions, the model suggests that we might be particularly bias-prone in decisions for which we (have to) rely on imagination. It might be more difficult for subjects to identify the stereotype-congruent expectations that fed into their mental simulation than it is to identify those they used in reasoning. It may also be more difficult to choose not to reason from particular

expectations. After all, if the hypothesis presented in this paper is correct, the advantage of simulation is exactly that we can integrate a wide range of knowledge without having to choose what to integrate.

Imagination may also be more strongly moderated by the temporal availability of certain contents. This can work to temporally strengthen or weaken biases. For permanent reductions of biases, it would be more effective to change situation-specific expectations, as they occur in scripts.

It would be interesting to pursue these issues further and develop them into empirically testable predictions because this will provide us with a better understanding of the influences on actual discriminatory behavior and not just IAT scores.

6. Conclusions

Using a dispositional understanding of implicit bias, I provided a model of the cognitive underpinnings of implicit bias in some considered decisions – the imagination model. The model draws attention to the role of group-specific *expectations* for implicit bias, which are part of scripts for social contexts. I have argued that it is to be expected that people who would honestly disavow certain stereotypes when they are expressed as generic context-independent generalizations, can still hold some more specific stereotype-congruent expectations. We do not have to postulate that these expectations are inaccessible. It is sufficient to say that often we have simply not understood that some of our expectations perpetuate culturally transmitted stereotypes. This part of the model is independent from the particular causal contribution that imagination plays for biased decisions. Furthermore, I have also argued that traces of stereotypes can feed into imagination although we did not choose this to happen. Therefore, so a central hypothesis of this paper, we might be particularly vulnerable to biases when we (have to) rely on imagination.

Another line of future research is to explore agent control of imagination. To a certain degree, I can train myself to imagine counter-stereotypical persons. Hence, imagination is not only a source of bias, it could also be a source for combating bias. Behind these considerations lies the question to which degree we rationally ought to control imagination and when it is better to let the simulation run in the context of decision-making. How this

issue intersects with attempts to control and reduce implicit bias deserves further attention.

Acknowledgements

Earlier versions of this paper were presented at a workshop on implicit attitudes at KWI Essen, at SWIP Germany's jour fixe at HU Berlin, at the 4th mental fragmentation workshop at Graz University, and at the ECAP9 at LMU Munich. I thank all audiences for helpful discussions. I also thank Christine Bratu, Katja Crone, Lena Kästner, Andrea Lailach, Francesco Marchi, Nora Olbrisch, and last, but not least, an anonymous reviewer, for helpful comments on earlier versions of this paper.

References

Agerström, J., & Rooth, D.-O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, 96(4), 790–805. DOI: 10.1037/a0021594

Blair, I. V., Ma, J., & Lenton, A. (2001). Imagining Stereotypes Away: The Moderation of Implicit Stereotypes through Mental Imagery. *Journal of Personality and Social Psychology*, 81, 828-841.

Briñol, P., Petty, R., & McCaslin, M. (2008). Changing Attitudes on Implicit versus Explicit Measures: What is the Difference? In R. Petty, R. Fazio, and P. Briñol (Eds.), *Attitudes: Insights from the New Implicit Measures* (pp. 285-326). New York: Psychology Press.

Brownstein, M. (2018). *The Implicit Mind.* Oxford: Oxford University Press.

Carlsson, R., & Agerström, J. (2016). A Closer Look at the Discrimination Outcomes in the IAT Literature. *Scandinavian Journal of Psychology,* 57, 278–287.

Casper, C., Rothermund, K., & Wentura, D. (2010). Automatic stereotype activation is context dependent. *Social Psychology*, 41(3), 131-136.

Croskerry, P. (2003). The Importance of Cognitive Errors in Diagnosis and Strategies to Minimize Them. *Academic Medicine*, 78(8), 775–780.

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Quality & Safety*, 22. DOI: 10.1136/bmjqs-2012-001712

Currie, G., & Ichino, A. (2012). Aliefs Don't Exist, though Some of Their Relatives Do. *Analysis Reviews,* 72(4), 788–798. DOI:10.1093/analys/ans088.

Dasgupta, N., & Greenwald, A. G. (2001). On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals. *Journal of Personality and Social Psychology,* 81, 800–814.

Gawronski, B., & Bodenhausen, G. (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological Bulletin*, 132(5), 692–731.

Gawronski, B., & Bodenhausen, G. (2011). Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model. Theory, evidence, and open questions. *Advances in Experimental Social Psychology*, *44*, 59-127. https://doi.org/10.1016/B978-0-12-385522-0.00002-0

Gendler, T. (2010). *Intuition, Imagination, and Philosophical Methodology*. Oxford: Oxford University Press.

Gendler, T., & Kovakovich, K. (2005). Genuine Rational Fictional Emotions. In M. Kieran (ed.), Contemporary Debates in Aesthetics and the Philosophy of Art (pp. 241–253). Malden, MA: Blackwell.

Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70, 491–512.

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

Greenwald, A. G., & Banaji, M. R. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *The American Psychologist*, 72(9), 861-871. DOI: 10.1037/amp0000238

Greenwald, A. G., Banaji, M. R. & Nosek, B. (2014). Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects. *Journal of Personality and Social Psychology*. DOI:10.1037/pspa0000016.

Gregg A., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology,* 90, 1–20.

Fazio, R. H., & Olson, M. A. (2007). Attitudes: Foundations, Functions and Consequences. In M.A. Hogg & J. Cooper (Eds.), *The Sage Handbook of Social Psychology* (pp. 139–60). London: SAGE.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology,* 82(6), 878–902.

Heilman, M. E., Wallen, A. S., Fuchs, D., & Tamkins, M. M. (2004). Penalties for Success: Reactions to Women Who Succeed at Male Gender-Typed Tasks. *Journal of Applied Psychology,* 89(3), 416–427. DOI: 10.1037/0021-9010.89.3.416.

Holroyd, J. (2016). What Do We Want from a Model of Implicit Cognition? (digital preprint / draft). *Proceedings of the Aristotelian Society,* 116(2). https://www.aristoteliansociety.org.uk/pdf/holroyd.pdf. Accessed 20 June 2016.

Holroyd, J., & Sweetman, J. (2016). The Heterogeneity of Implicit Bias. In Michael Brownstein and Jennifer Saul (Eds.), *Implicit Bias and Philosophy,* Volume 1 *(pp. 80-103).* Oxford: Oxford University Press.

Joy-Gaba, J. A., & Nosek, B. A. (2010). The Surprisingly Limited Malleability of Implicit Racial Evaluations. Social Psychology, 41(3), 137-146. DOI: 10.1027/1864-9335/a000020

Kurdi, B., Seitchik, A., Axt, J., Carroll, T., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2018). Relationship between the Implicit Association Test and Intergroup Behavior: A Meta-Analysis. *Open Science Framework*. June 20. osf.io/ryjva. Accessed 24 January 2019.

Lai, C.K., Forscher, P. S., Axt, J., Ebersole, C. R., Herman, M., & Nosek, B. A. (2017). A Meta-Analysis of Change in Implicit Bias. *Open Science Framework*. February 17. osf.io/awz2p. Accessed 24 January 2019.

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. http://dx.doi.org/10.1037/a0036260

Langland-Hassan, P. (2016). On Choosing What to Imagine. In Amy Kind and Peter Kung (Eds.), *Knowledge through Imagination* (pp.61-84), Oxford: Oxford University Press.

Leslie, S.-J. (2017). The Original Sin of Cognition: Fear, Prejudice, And Generalization. *Journal of Philosophy,* 114(8), 393-421.

Levy, N. (2015). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs,* 49(4), 800–823. DOI:10.1111/nous.12074.

Liao, S., & Gendler, T. (2019). Imagination. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), forthcoming. https://plato.stanford.edu/archives/spr2019/entries/imagination/. Accessed 24 January 2019.

Machery, E. (2016). De-Freuding Implicit Attitudes. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy,* Volume 1 (pp. 104-129). Oxford: Oxford University Press.

Mandelbaum, E. (2016). Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Noûs*, 50, 629-658. DOI:10.1111/nous.12089

Madva, A. (2016a). Virtue, Social Knowledge, and Implicit Bias. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy,* Volume 1 (pp. 191-215). Oxford: Oxford University Press.

Madva, A. (2016b). Why Implicit Attitudes Are (Probably) not Beliefs. *Synthese,* 193, 2659–2684.

Madva, A., & Brownstein, M. (2018). Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind. *Noûs*, 52, 611-644. DOI:10.1111/nous.12182

Moors, A., & de Houwer, J. (2006). Automaticity: A Theoretical and Conceptual Analysis. *Psychological Bulletin,* 132(2), 297–326.

Nanay, B. (2016a). Mental Imagery. Video blog hosted by The Brains Blog. First video. http://philosophyofbrains.com/2016/05/02/how-should-we-use-the-concept-of-mental-imagery.aspx. Accessed 22 February 2019.

Nanay, B. (2016b). The Role of Imagination in Decision-Making. *Mind and Language,* 31(1), 127–143. DOI: 10.1111/mila.12097/full.

Nichols, S., & Stich, S. (2000). A cognitive theory of pretense. *Cognition,* 74, 115-147.

Oswald, F., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. (2013). Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies. *Journal of Personality and Social Psychology*, 105(2), 171–192. DOI: 10.1037/a0032734.

Rooth, D.-O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, 17(3), 523-534.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding*, Hillsdale: Lawrence Erlbaum.

Schwitzgebel, E. (2010). Acting Contrary to Our Professed Beliefs, or the Gulf between Occurrent Judgment and Dispositional Belief. *Pacific Philosophical Quarterly,* 91, 531–553.

Singal, J. (2017). Psychology's Favorite Tool for Measuring Racism Isn't up to the Job. *New York Magazine / Science of Us.* Published January 11, 2017. URL: http://nymag.com/scienceofus/2017/01/psychologys-racism-measuring-tool-isnt-up-to-the-job.html. Accessed 25 January 2017.

Strack, F., & Deutsch, R. (2004). Reflective and Impulsive Determinants of Social Behavior. Personality and Social Psychology Review, 8(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

Tanesini, A. (2018). Intellectual Humility as Attitude. *Philosophy and Phenomenological Research*, 96, 399-420. DOI: 10.1111/phpr.12326.

Toribio, J. (2018). Implicit Bias: From Social Structure to Representational Format. *THEORIA,* 33(1), 41-60.

Uhlmann. E. L., & Cohen, G. (2005). Constructed Criteria. Redefining Merit to Justify Discrimination. *Psychological Science,* 16(6), 474-480.

Van Leeuwen, N. (2013). The Meanings of 'Imagine'. Part I: Constructive Imagination. *Philosophy Compass,* 8(3), 220-230.

Van Leeuwen, N. (2016a). Imagination and Action. In Amy Kind (Ed.), *The Routledge Handbook of Philosophy of Imagination* (pp.286-299). London/New York: Routledge.

Van Leeuwen, N. (2016b). The Imaginative Agent. In Amy Kind and Peter Kung (Eds.), *Knowledge through Imagination* (pp.85-109). Oxford: Oxford University Press.

Williams, M.J., & Tiedens, L.Z. (2016). The Subtle Suspension of Backlash: A Meta-Analysis of Penalties for Women's Implicit and Explicit Dominance Behavior. *Psychological Bulletin,* 142(2), 165-97. DOI: 10.1037/bul0000039.