Katharina Berndt Rasmussen
katharina.berndt.rasmussen@iffs.se

Implicit Bias and Discrimination

1 Introduction

Imagine the following two people:

> *Alex*, a tech company CEO, is an outspoken defender of everyone's equal rights. Still, whenever employees are considered for promotion, Alex ranks female employees' achievements lower than the equivalent achievements of their male counterparts.
>
> Even *Billie*, one of the company's employees, is an outspoken defender of everyone's equal rights. Yet, whenever her company hires a black person, she fails to be as nice – as welcoming, helpful and curious – toward them as she is with newly hired white people.
>
> Both Alex and Billie are unaware of these subtle, but noticeable differences.

What could explain Alex's and Billie's differential treatment of their employees and co-workers, respectively? Simply asking them would clearly be to no avail, as they are unaware of their behaviour. Here is another idea for how to test for a possible explanation, which originates in recent social-psychological research: we could ask them to take an Implicit Association Test (IAT). The Race-IAT might then reveal that Billie harbours an automatic preference for white people over black people. And the Gender–Science IAT might reveal that Alex associates liberal arts with females and STEM-fields with males. This could help explain their differential treatment of male and female employees, or black and white co-workers.

Simply put, the Race-IAT is a web-based application that is designed to test the strength of one's racial prejudices.[1] It presents the user with either pictures of black or white faces, or with positively or negatively valenced words ('smile' or 'peace', 'rotten' or 'agony'). The user is required to sort these items, as fast and accurately as possible, into categories. These categories are disjunctive: e.g., 'African-American *or* unpleasant' and 'European-American *or* pleasant'. It turns out that a clear majority of users is faster, and makes fewer mistakes, when they sort the items into these exact categories, as compared to when they sort the items into the reverse categories 'African-American *or* pleasant' and 'European-American *or* unpleasant'. This greater ease of association is interpreted as an automatic preference for white people over black people.

For the Gender–Science IAT, users are required to sort either male or female coded terms, ('grandpa', 'wife'), or science or liberal arts subfields ('engineering', 'literature'). Into similarly disjunctive categories. A clear majority of users is faster, and makes fewer mistakes, when they sort these items into the categories 'Male *or* Science' and 'Female or Liberal Arts', as compared to when they sort the items into the reverse categories. This is interpreted as an automatic association for Male with Science and Female with Liberal Arts.

[1] These tests are available at https://implicit.harvard.edu/implicit/takeatest.html.

Just as in the hypothetical case of Alex and Billie, a majority of the real-world users score such a (slight, moderate or strong) preference or association – an implicit bias – yet most of them endorse egalitarian views, when asked explicitly. One hypothesis is that a significant amount of differential treatment along the lines of race, gender (as well as other socially salient categories), is caused by the implicit biases of outwardly egalitarian-minded people.

How can we analyse differential treatment from implicit bias? This is the guiding question for this article. My aim is twofold. First, I seek to improve our understanding of the phenomenon implicit bias, including its moral status, by examining it through the lens of a theory of discrimination. Second, I simultaneously seek to improve this theory of discrimination by making conceptual space for implicit bias discrimination. My focus in this article is mainly taxonomical, though I will sketch some of the moral implications of my taxonomy along the way.

In section 2, I introduce the phenomenon of implicit bias more thoroughly. Section 3 gives a brief overview of the theory of discrimination I work with and explores two ways of distinguishing direct and indirect discrimination. In the light of this pair of distinctions, section 4 spells out four different forms of discrimination and locates implicit bias discrimination in the resulting conceptual space. This section also deals with four challenges to my proposal of capturing implicit bias within my discrimination framework: the metaphysical challenge, the moral insignificance challenge, the causal connection challenge, and the challenge from irreducibly collective bias. Section 5 concludes.

2 Implicit bias

One way to interpret the discrepancy between Alex's explicit beliefs and attitudes and her IAT-score would be to deny the credibility of her explicit statements. Behind a thin varnish of civilisation, the idea goes, deep down Alex really *is* a racist. Corroborating this interpretation, empirical studies show that social desirability issues affect attitudes measured in surveys: many people hesitate, even knowing that their anonymity is protected, to answer in a way that deviates from the established norm on politically or morally sensitive questions.[2] So under normal circumstances, Alex just says what everyone expects to hear, covering up her secretly held convictions – possibly even to herself. According to this interpretation, the IAT just reveals these "true" convictions, by forcing Alex to answer quickly, stripping her of the possibility to cover them up.

This interpretation is surely plausible in some cases. Yet there is an alternative interpretation, which has the strength of granting the possibility that people like Alex and Billie are truthful in reporting their convictions. According to this alternative interpretation, they *also* harbour mental processes which contradict these convictions and influence their behaviour – such as the behaviour measured by the IAT – and of which they are unaware. A common, functional definition of implicit bias captures this thought: "implicit biases are whatever unconscious processes influence our perceptions, judgements and actions—in this context, in relation to social category members (women, blacks, gays, for example)".[3]

[2] E.g. (Greenwald and Banaji, 1995).
[3] Cf. (Holroyd and Sweetman, 2016, p. 81), (Saul, 2013, p. 40).

For an illuminating analogy, consider the Dark Matter of astrophysics. In the 1970's, Vera Rubin and colleagues discovered that the outer stars in a number of distant spiral galaxies rotated at the same speed as the stars closer to their galaxy's centre. These observations were inconsistent with the Newtonian theory of gravity, given other observations of the galaxies, concerning e.g. the number and masses of its stars. According to these factors, the outer stars should move much more slowly around their galaxy's centre than the inner ones. One possible explanation of the discrepancy was of course that there is something wrong with the Newtonian theory of gravity. A more conservative, and according to the scientists overall more reasonable explanation was that these galaxies' masses must be much greater, and differently distributed, than their observable parts had suggested. They concluded that there must be more – non-observable – matter, whose mass and distribution made sense of the equations in accordance with the Newtonian theory of gravity. Basically, all we know about this *Astrophysical Dark Matter* is inferred from its observed gravitational effects on its intragalactic surroundings.[4]

Analogously, basically all we know about the *Human Dark Matter* called implicit bias is inferred from its effects on human behaviour, as measured by tests like the IAT. Just as in Alex's case, IAT-scores are often inconsistent with the test subjects' explicit convictions, and thus cannot be explained by them. An alternative explanation in terms of the above-mentioned error (or deceit) theory – although arguably plausible in some cases – has the unpalatable implication of pointing out Alex and everyone else as deceiving others or themselves, regardless of their insistent claims to the contrary. Psychologists have instead suggested that the test scores reflect mental processes that are automatic, not directly accessible to introspection, and beyond our direct control, and that can be in conflict with other, more easily accessible mental processes: implicit biases or implicit associations. By taking the IAT, we thus indirectly learn about our own unobservable mental dark matter.

The case of Alex and Billie suggests that their differential treatment, of which they are unaware, is *caused* by implicit bias. If this is so, how should we morally evaluate their behaviour? Do they act wrongly in failing to be equally nice, or to rank employees equitably? And if so, what exactly makes their actions morally wrong?

One natural way to address these questions is to turn to theories of discrimination. We usually apply this term to such instances of differential or disadvantageous treatment of members of socially salient groups, which appear as objectionable. In legal contexts, the purpose is to signal legal non-acceptability of the treatment, and ascribe legal liability to the agent. In a moral context, likewise, we use the term within accounts of such treatment's moral wrongness and of the agent's moral responsibility. Alex's and Billie's differential, disadvantageous treatment of black and female employees, respectively, surely looks like discrimination. And if it is wrong (morally and/or legally), it's presumably wrong for the same reasons other wrongful discriminatory acts are.

I will in this paper focus on a *moral* framework of discrimination. Applying this framework to Alex's case is not entirely straightforward, however, due to the complicated and contested nature of implicit biases – the dark matter evidenced by IAT-scores – and their relation to

[4] Cf. (Rubin, 1983).

behaviour beyond the reaction patterns measured by the test. In the remainder of this paper, I will explore the challenges that implicit bias poses to discrimination theory.

3 Discrimination

I am here specifically concerned with group discrimination, i.e. discrimination due to group membership, according to the following definition:

> (GD) An agent, X, group discriminates against someone, Y, by ϕ-ing if, and only if:
> (i) There is a property, P, such that Y has P or X believes that Y has P,
> (ii) By ϕ-ing, X treats Y worse than X would have treated Y, had Y not had P or had X not believed Y to have P,
> (iii) It is because (X believes that) Y has P that X treats Y worse by ϕ-ing, and
> (iv) P is the property of being a member of a socially salient group [i.e., a group perceived membership of which is important to the structure of social interactions across a wide range of social contexts].[5]

In the literature on discrimination, it is common to distinguish between two specific forms of discrimination: direct and indirect. Direct discrimination is typically taken to refer to *disparate treatment*, in the sense that "an agent treats a person or group of persons in a way in which she does not treat other persons".[6] Drawing on the above definition, we can bring out the idea as a precisification of (GDii):

> (GDii') Had Y not had P, X would have χ*-ed*, rather than ϕ-ed, and ϕ-ing toward someone constitutes worse treatment of them than χ-ing.

Indirect discrimination, on the other hand, is often understood in terms of *disparate impact*. The idea is then that "a policy or procedure is on the face of it neutral, but in fact disproportionally disadvantages members of a particular social group".[7] This amounts to an alternative precisification of (GDii):

> (GDii'') Had Y not had P, X would still have ϕ-ed, but ϕ-ing toward someone with P constitutes worse treatment than ϕ-ing toward someone without P.[8]

With this basic taxonomy of forms of discrimination, which one is in play in Alex's and Billie's case? It can hardly be indirect discrimination in the disparate impact sense: Alex ranks female employees *lower*; Billie is *less* nice towards black co-workers. Their differential treatment may be subtle, but it surely is not "on the face of it neutral". To compare: this might have been the case, had e.g. Billie displayed an *equally* non-welcoming response to all co-workers, in combination with a higher hostility sensitivity in black employees. But as described above, the

[5] (Berndt Rasmussen, 2018, para. 7). For similar definitions, see (Lippert-Rasmussen, 2014) and many of the entries in (Lippert-Rasmussen, 2017).
[6] (Thomsen, 2017, p. 21).
[7] (Holroyd, 2017, p. 382). This interpretation also closely aligns with the widespread legal understanding of indirect discrimination involving "an apparently neutral practice or policy which puts members of a protected group (say, women) at a disproportionate disadvantage compared with members of a cognate group (say, men)" (Khaitan, 2017, p. 31).
[8] For this way of drawing the distinction, see also (Hellman, 2017, p. 98).

comparison is between two different kinds of behaviour, rather than between different effects of the same behaviour. This clearly is direct discrimination, in the disparate treatment sense.

This may seem straightforward, but it has a noteworthy implication: discrimination from implicit bias is now classified under the same heading as discrimination from explicit bias. That is, there is no taxonomical difference between Billie's case and the following:

> *Charlie*, one of Billie's co-workers, holds explicit racist views. Because of these views, he is less nice – less welcoming, helpful and curious – toward newly hired black people compared to newly hired white people. Charlie is fully aware of, and endorses, his views and the resulting differences in treatment.

But is it really plausible to classify both Billie's and Charlie's behaviour as the same form of discrimination? This prompts a further question: why are we interested in classifying discrimination according to the direct/indirect divide in the first place?[9] The normative relevance of the distinction is rarely discussed in its own right in the discrimination literature. But there are some clues: the different forms of discrimination may have different moral status, they may constitute "two different kinds of wrong",[10] or "two different types of one and the same wrong".[11] One suggestion is that direct discrimination is more severe.[12]

Intuitively, there is a moral difference between Billie's and Charlie's way of treating their co-workers: Charlie's behaviour seems more objectionable than Billie's. We might therefore expect these cases to be classified differently by a plausible theory of discrimination: Charlie's case as morally more severe direct, Billie's case as less severe indirect discrimination.

Yet now the moral relevance of the distinction, as spelled out above, becomes doubtful. Why would differential treatment, which amounts to treating Y worse due to P, be morally more severe than equal treatment, which amounts to treating Y worse due to P? In both cases, Y is treated worse, compared to what had been the case, had Y lacked P.

To make sense of the idea that disparate treatment discrimination is more severe, we could point out that it – unlike disparate impact discrimination – involves a *shift* in treatment, due to the presence of P. That is, it is implied that the agent, X, has a choice between two options (ϕ-ing and χ-ing) where their choice turns on the (believed) presence or absence of P in X's counterpart Y. This, however, suggests a different way to distinguish between direct and indirect discrimination: viz., in terms of the agent's intentions. And this alternative distinction, too, appears frequently in the literature on discrimination.

[9] Cf. (Altman, 2016, para. 3.2).
[10] (Altman, 2016, para. 3.1).
[11] (Altman, 2016, p. 4.2).
[12] Cf. Khaitan for the difference in severity of direct and indirect discrimination in the American and British legal contexts: "courts treat indirect discrimination as *almost always* justifiable, i.e. they are open to the possibility that indirect discrimination in a given case might be justified [as a] necessary and proportionate means to pursue a sufficiently important objective", and they apply to that a "less exacting" standard than the one required to justify "direct discrimination, *if that is permitted at all*" (Khaitan, 2017, p. 34, my italics).

Lippert-Rasmussen characterises direct discrimination as "[…] treatment where the discriminator […] intended to exclude people on the basis of membership of a particular socially salient group, whose members he thought inferior in certain ways or to whom he was hostile".[13] In terms of the above definition of discrimination, this amounts to a precisification of (GDiii):

> (GDiii') it is because X has P-related intentions – e.g. *dislikes* people with P and *believes* that Y has P – that X treats Y worse.

Indirect discrimination, on the other hand, "does not involve any intentions to exclude, but does in fact exclude because of how rules, practices, institutions etc. have been designed in a context where they serve the needs and match the capacities of particular groups".[14] This amounts to an alternative precisification of (GDiii):

> (GDiii'') It is *not* because X has P-related intentions that X treats Y worse, but rather because of some other reason.[15]

According to this way of making the distinction, implicit bias discrimination would rather be of the indirect variety. Intuitively, this distinction does have some potential to capture differences in moral status. To many people's minds, at least, it makes a difference whether an agent *intends* to treat someone worse, as compared to *inadvertently* doing so. A theory of the moral wrongness of discrimination which draws on this distinction may then explain the intuition that that Charlie's behaviour is more objectionable than Billie's.

There are theories which are sensitive to the distinction in this sense: they make the agent's mental states relevant for the wrongness of the action. Andrew Altman goes so far as to state:

> There is general agreement that the wrong [of direct discrimination] concerns the kind of reason or motive that guides the action of the agent of discrimination: the agent is acting on a reason or motive that is in some way illegitimate or morally tainted.[16]

Thus, according to Altman, what makes direct discrimination wrong is some moral flaw in the *practical reasoning process behind* the action in question.[17] Indirect discrimination, Altman states, is wrong due to something else: a feature of the *outcome* of the action, or a flaw in the *social processes behind* whatever brings about the outcome.[18]

---

[13] (Lippert-Rasmussen, 2017, p. 3); for an explanation of the omissions, see section 4 below.
[14] (Lippert-Rasmussen, 2017, p. 3). For this way of drawing the distinction, see also Altman §3.1: "Direct discrimination is essentially a matter of the reasons or motives that guide the act or policy of a particular agent, while indirect discrimination is not about such reasons or motives." Cf. even (Moreau, 2017, pp. 166–167).
[15] For this distinction, in terms of motivating vs. explanatory reasons, cf. (Lippert-Rasmussen, 2014, p. 38), (Altman, 2016, para. 2.2).
[16] (Altman, 2016, para. 4.1).
[17] E.g., the motive that guides the action refers to some immutable trait of, or to an inaccurate stereotype about its victim, or the motive is irrational or arbitrary, or it fails to take the victim's merits into proper account (ibid.).
[18] (Altman, 2016, para. 4.2).

Still, there are other influential theories of the wrongness of discrimination which do not easily align with the intentional/non-intentional distinction. On an objective social meaning account, discrimination (of any variety) is wrong because it demeans its victims – irrespective of the agent's motives.[19] Likewise, on a harm-based – purely outcome-focussed – account of the wrongness of discrimination, the intentional/non-intentional divide carries no immediate moral significance.[20]

The divide, however, may be made morally relevant in a further sense, when considering the agent's moral responsibility for acting wrongly.[21] Consider theories which spell out moral responsibility in terms of attributability, in the sense that the agent is the source of her action.[22] According to two such theories, we are morally responsible only for actions which are motivated by (higher-order) mental states we endorse,[23] or which are brought about by a reasons-responsive mechanism (which, had there been sufficient reason to do otherwise, would have resulted in an alternative action).[24] The present discrimination framework can be made sensitive to such preconditions for moral responsibility, by making the intentional/non-intentional distinction thicker, adding either condition (a) or (b):

> (GDiii'*) it is because X has P-related intentions – e.g. *dislikes* people with P and *believes* that Y has P – that X treats Y worse, and either
>
> (a) X endorses these P-related intentions, or
>
> (b) these P-related intentions are not reasons-responsive.
>
> (GDiii''*) It is *not* because X has P-related intentions that are
>
> (a) endorsed, or
>
> (b) reasons-responsive,
>
> that X treats Y worse, but rather because of some other reason.[25]

Then, since Charlie's failure to be equally nice to black co-workers result from (racist) beliefs and desires he endorses, or which are appropriately reasons-responsive, his discriminatory actions are attributable to him. This does not hold for Billie's actions, which result from mental states she is unaware of and whose content she explicitly rejects and, some claim, form part of a mechanism that is fundamentally not reasons-responsive.[26] Another way of phrasing this is that Charlie, but not Billie, can be said to be acting fully as an agent, in treating their black and white co-workers differently.[27]

[19] (Hellman, 2008), (Hellman, 2017).
[20] (Lippert-Rasmussen, 2014), Arneson (2013), Arneson (2017).
[21] The groundbreaking article on implicit bias and moral responsibility is (Holroyd, 2012).
[22] Cf. Zheng, Robin (2016) *Implicit Bias and Philosophy* Vol II.
[23] Cf. (Frankfurt, 1971). (Brownstein, 2016) argues from a Frankfurt-inspired agent care-based account of attributability for agents' moral responsibility for implicit bias discrimination.
[24] Cf. (Fischer, 2018). (Levy, 2017) argues that agents are not morally responsible for implicit bias discrimination, in part due to attributability reasons.
[25] For this distinction, in terms of motivating vs. explanatory reasons, cf. (Lippert-Rasmussen, 2014, p. 38), (Altman, 2016, para. 2.2).
[26] See (Levy, 2017).
[27] Cf. Zheng (2016).

4 An improved taxonomy

In the literature on discrimination, the two above distinctions have been used to define direct and indirect discrimination, albeit often in conflated ways. Consider Lippert-Rasmussen's above characterisation of direct discrimination, once the omitted parts are filled in: direct discrimination is "*differential treatment* where the discriminator treated people – say, job applicants – differently, because he *intended* to exclude people on the basis of membership of a particular socially salient group, whose members he thought inferior in certain ways or to whom he was hostile".[28] And consider Frej Klem Thomsen, who introduces the distinction between indirect and direct discrimination in terms of "[d]istinguishing equal from *differential* treatment", but states subsequently that "only *intentional* discrimination is direct".[29] Further, consider Sophia Moreau: "Sometimes [disadvantaging certain individuals] occurs *intentionally* or explicitly, and we call it 'direct discrimination' or '*disparate treatment*'; sometimes it is a side-effect of a policy adopted for quite different and perhaps even beneficial reasons, and we call it 'indirect discrimination' or 'disparate impact'."[30] And consider Mari Mikkola, who defines direct discrimination in intentional terms, and indirect discrimination in disparate impact terms: "discrimination may be direct (some people are explicitly and *intentionally* singled out for disadvantaging treatment due to socially salient features), or it may be indirect […]. In the latter case, some rule *disproportionally affects* a group of people due to a socially salient feature they possess, although the rule is at face value neutral."[31]

My proposal is that both the disparate treatment of different groups and the disparate impact of the same treatment on different groups can be either intentionally or non-intentionally brought about by the agent. The idea that *disparate impact* can be brought about either intentionally or non-intentionally has been proposed before. José Jorge Mendoza writes that a "policy indirectly discriminates [in the disparate impact sense] when the policy is facially neutral, […] where either (a) this facially neutral criterion is covertly used to target members of a protected class or unfairly benefit members of historically advantaged groups or (b) this facially neutral criterion has a disparate impact that leads to a similarly discriminatory outcome – even when that outcome was not the intention of the policymakers or enforcers".[32]

My proposal goes further in that even *disparate treatment* can be either intentional or non-intentional. The intentional version is the one often presented as direct discrimination proper, in the above conflated sense. Its non-intentional version has hitherto been hidden by the theoretical shadow of this conflated category. My suggestion is that it is precisely here we should locate differential treatment from implicit bias, as described in the above case of Alex and Billie.[33] I.e., the phrase 'because of some other reason' in condition (GDiii'') is precisified by reference to implicit bias:

[28] (Lippert-Rasmussen, 2017, p. 3, my italics).
[29] (Thomsen, 2017, pp. 21, 24, my italics).
[30] (Moreau, 2017, p. 164, my italics).
[31] (Mikkola, 2017, p. 289, my italics).
[32] (Mendoza, 2017, p. 258)
[33] Note that this is in line with some theorising in legal theory: while e.g. Katya Hosking and Roseanne Russell characterise indirect discrimination in disparate impact terms (Hosking and Russell, 2016, p. 262), they state that direct discrimination is "less favourable [i.e. disparate] treatment because of a protected characteristic" which "does not have to be, and usually is not, intentional or deliberate" (Hosking and Russell, 2016, pp. 257, 258). They thus allow for both intentional and non-intentional forms of disparate treatment discrimination.

(GDiii''°) It is *not* because X has P-related intentions that X treats Y worse, but rather because X has a P-related implicit bias.

We are thus dealing with four, rather than two forms of discrimination. *Table 1* systematises and states paradigmatic examples for each of them. (For real-life evidence, see the footnotes).

	Disparate treatment: ϕ-ing vs. χ-ing – condition (GDii')	*Disparate impact of ϕ-ing – condition (GDii'')*
Intentional – condition (GDiii')	1) A university in the early 1950's US South accepts a white applicant but turns down an *equally* qualified black applicant, stating: "This is a whites-only university. Blacks are referred to apply to some 'separate-but-equal' university for African Americans."[34]	2) An employer turns down a *qualified* black applicant, stating: "We don't hire people who lack high school education", while intentionally using this criterion because of its ability to track politically induced, race-correlated educational deficits.[35]
Non-intentional – condition (GDiii'')	3) A university accepts a white PhD-candidate but turns down an *equally* qualified black PhD-candidate, stating that the latter was less qualified, where the unequitable ranking stems from the evaluators' implicit biases.[36]	4) An employer turns down a *qualified* black applicant, stating: "We don't hire people who lack high school education", without awareness of the criterion's ability to track politically induced, race-correlated educational deficits.[37]

Table 1

This taxonomy thus provides conceptual space for implicit bias discrimination, under (3) *non-intentional disparate treatment discrimination*. The classification also helps solve the following puzzle. Jules Holroyd notes that "we should expect discrimination due to implicit bias to be captured by an analysis of direct discrimination [since] it is a matter of an individual treating another individual disadvantageously" – i.e. of disparate treatment. Yet, it does not "involve [depreciating] beliefs or judgments that are taken to justify differential treatment" – i.e., it is not intentional treatment.[38] The puzzle of whether or not to apply the label 'direct discrimination' simply stems from the conflation of the intentional and disparate treatment component. In sum, the phenomenon of implicit bias serves to highlight the inept taxonomy of existing discrimination theory, and to devise a better one.

However, one could object that implicit bias discrimination might as well fall under (1) – depending on the precise meaning of 'intentional' and on the precise conceptualisation of implicit bias. If implicit biases, as has been proposed, are belief-like states, and if 'intentional' is defined broadly, as not necessarily presupposing mental processes that are directly introspectively accessible to the agent or under her direct control, the intentional/non-intentional distinction might well be of little use here. This constitutes a first challenge to my

[34] This case resembles *Sweatt v. Painter*; cf. (Lavergne, 2010). Note that there may but needn't be disparate impact under disparate treatment: if (contrary to historical fact) the educational facilities had been separate *and* relevantly equal, blacks would not have been worse off than whites, but such non-disadvantageous yet differential treatment would still constitute discrimination and may still be marked as morally wrong as such.
[35] This case resembles *Griggs v. Duke Power Company*; cf. (Khaitan, 2017, p. 31), but with the addition that the criterion "is covertly used to target members of a protected class" (Mendoza, 2017, p. 258). Cf. even Altman's "Jim Crow era" example, (Altman, 2016, para. 2.1).
[36] This case resembles instances of unequal rankings of identical CV's under different (racially or gender coded) names, cf. (Rooth, 2010).
[37] This case resembles *Griggs v. Duke Power Company* under "absence of a discriminatory intent" (Khaitan, 2017, p. 31). There is, of course, the separate but related problem of discrimination at the educational level.
[38] (Holroyd, 2017, p. 387).

proposal of capturing implicit bias within a discrimination framework. I call it the metaphysical challenge: to deal with it, we need to get clear on the metaphysics of implicit bias.

4.1 The metaphysical challenge

Tamar Gendler's ground-breaking philosophical account of the nature of implicit bias proposes it as a *sui generis* kind of mental state, called "alief". According to her, an alief is a mental state with an intentional content of representational, affective and behavioural components (such as "Black man! – Scary! – Avoid!"), clustered in such a way that activation of one component activates the others.[39] Neil Levy proposes another *sui generis* account of implicit biases as "patchy endorsements": while they do have a belief-like propositional structure, they are too fragmented and poorly integrated with the agent's other beliefs and desires to properly be considered beliefs.[40] Such sui generis mental states do not fit squarely with the belief-desire model of intentional action, but could figure as an explanatory reason for an agent's discriminating behaviour (e.g., not calling a qualified black applicant back for an interview), thus rendering implicit bias discrimination as (3) non-intentional disparate treatment discrimination.

Other theorists have proposed accounts of implicit bias within an extended belief-desire framework of cognition and motivation. Inspired by theories of the fragmentation or "compartmentalization" of the mind, Andy Egan proposes the adoption of an account of compartmentalized beliefs (and other propositional attitudes), in order to account for phenomena of implicit bias without positing novel, queer entities.[41] In a similar vein, Eric Mandelbaum suggests a "Structured Belief" hypothesis, according to which implicit biases arise from unconscious, "propositionally structured mental representations; mental representations that we bear the belief relation to", that create "piggybacking associations" between the frequently involved concepts.[42] Sally Haslanger analyses implicit biases in terms of cognitive schemas: "clusters of culturally shared concepts, beliefs, and other attitudes that enable us to interpret and organize information and co-ordinate action, thought, and affect".[43] Or we could see them, with Sarah-Jane Leslie, as "striking property generics", a primitive form of generalisation with negatively valenced predicates such as "the big white shark attacks surfers", or "Muslims are terrorists".[44] If implicit bias is best conceptualised as relevantly similar to ordinary beliefs, implicit bias discrimination might rather constitute (1) intentional disparate treatment discrimination.

Even if we accept one of the latter accounts, we may want to uphold a potentially morally relevant difference between Billie and Charlie. My proposed taxonomy is open to a potential solution: we can, again, adjust the intentional/non-intentional distinction and e.g. reformulate it in terms of the agent's introspective accessibility endorsement – vs. the lack thereof. This would allow us to classify Billie's differential treatment from implicit bias under square (3), while Charlie's would fall on the other side of the divide under square (1). The way the framework deals with the metaphysical challenge is thus to offer some rather flexible

[39] (Gendler, 2008)
[40] (Levy, 2015)
[41] (Egan, 2011).
[42] (Mandelbaum, 2016).
[43](Haslanger, 2013, p. 11).
[44] (Leslie, 2017).

conceptual space, to account for the phenomenon of implicit bias discrimination in accordance with our intuitions. Depending on which account of implicit bias we eventually accept, the framework has resources to classify implicit bias discrimination as conceptually and (potentially) morally distinct from other, more orthodox forms of discrimination.

4.2 The moral insignificance challenge

This is a second challenge to my attempt to capture implicit bias within a discrimination framework. Consider again Billie's case: we assumed that her implicit bias causes her differential, disadvantageous treatment of black co-workers. There is some empirical evidence corroborating such a causal connection when it comes to the micro-behaviours ("being nice") considered in Billie's case. E.g., one study assessed the IAT-scores of white undergraduates, as well as their behaviour toward white and black experiment leaders, respectively, and found "significant correlations":

> Specifically, as participants' IAT scores reflected relatively more positive attitudes toward Whites than Blacks, social interactions were more positive toward the White experimenter than the Black experimenter as assessed both by trained [external observers] and by the experimenters themselves. […] larger IAT effect scores predicted greater *speaking time*, more *smiling*, more *extemporaneous social comments*, fewer *speech errors*, and fewer *speech hesitations* in interactions with the White (vs Black) experimenter.[45]

The challenge is now that each micro-behaviour in itself seems morally insignificant. Recall what I stated to be the purpose of applying the discrimination framework in the first place (section 2 above). The idea was to apply the term to instances of differential or disadvantageous treatment of members of socially salient groups, which seem morally objectionable. The framework should help us ascribe moral responsibility (blame) to the discriminator, and signal moral non-acceptability of the treatment. Yet, labelling Billie's failing to be nice in these microscopic ways as morally wrongful discrimination, or holding her morally responsible for them, seems like using a sledgehammer to crack a nut, as it were.

On the other hand, we know that insignificant differences on the individual level may aggregate to sizeable inequalities across many individuals or longer time spans. As Thomas Schelling's modelling of the systemic effects of individual preferences has shown, even very slight individual preferences for e.g. one's neighbourhood's racial profile can quickly lead to very segregated neighbourhoods.[46] More recently, organisation researchers have proposed further improved ways to analyse such micro- to macro-level processes. Richard Martell and colleagues propose an analysis of hierarchic upper-level gender segregation within organisations within a theoretical framework of *emergent phenomena*, where "emergence is concerned with the consequence of interactions among individuals within a system, the product of which defies prediction by a simple aggregation of individual-level behavior, as there is no simple relationship between the nature of what emerges and the individual actions

[45] (McConnell and Leibold, 2001, p. 439, my italics). Cf. (Cortina et al., 2013) who show similar results for acting in a civil manner towards female co-workers; cf. (Dovidio et al., 2002) for "friendliness" towards black co-workers.
[46] (Schelling, 1971).

that produced it".[47] They propose, and computationally model, that a barely noticeable micro-level gender bias can give rise to entirely unintended, large-scale and persistent system-level gender inequalities. In one model, they examine the effects of a very slight male bias (5% or even just 1% – meaning that male performance scores were boosted 5% or 1%, compared to female scores) for employees in a hierarchical promotional structure. Unsurprisingly, the result is that "a very high percentage of upper-level positions were filled by men, whereas women tended to cluster at the lower levels of the organization."[48] Thus, the moral insignificance challenge can be dealt with by pointing out that our intuition of the moral insignificance of Billie's micro-behaviour simply is due to our failure to see its causal role within a larger picture.

4.3 The causal connection challenge

In dealing with implicit bias within a discrimination framework, I have assumed that this mental phenomenon causes differential treatment. For micro-behaviour, such as Billie's, there is some corroborating evidence, as just stated (section 4.2). However, the above cited study, which found that "larger IAT effect scores predicted greater *speaking time*, more *smiling*", etc., also found "a significant correlation between the IAT and explicit reports of prejudice",[49] casting into doubt the causal role of specifically *implicit* biases for discriminating behaviour. For macro-behaviourcases of the Alex variety, the evidence is even more damning: a 2013 meta-study concludes that a "closer look at the IAT criterion studies in the domains of ethnic and racial discrimination revealed […] that the IAT provides little insight into who will discriminate against whom, and provides no more insight than explicit measures of bias".[50] On the other hand, a 2009 meta-study found that, for socially sensitive issues (specifically interracial contacts), IAT-scores are a significantly better predictor for prejudiced behaviour then explicit (self-)reports of prejudice.[51] Yet still other studies indicate that IAT-scores reflect not individual bias but rather cultural prejudices and social inequalities, finding "preliminary support for the environmental association model of the IAT [according to which] the IAT taps the associations a person has been exposed to in his or her environment, not that individual's level of endorsement regarding the attitude object".[52] This means that, even if correlations between IAT-scores and biased behaviour can be established, the underlying causal mechanism may be quite different from what we expected: the dark matter we call implicit bias may be an effect, rather than a cause of discrimination.

---

[47] (Martell et al., 2012, p. 142).

[48] (Martell et al., 1996, p. 157). Cf. (Greenwald et al., 2015) who consider the cumulative impact of biases affecting large groups of people, or the same person repeatedly. Cf. Brennan (2016) on accumulalted micro-inequities.

[49] (McConnell and Leibold, 2001, p. 440).

[50] (Oswald et al., 2013) p.188. For similar results, see (Oswald et al., 2015). Cf. even (Forscher et al., n.d.) who show that IAT-scores can be changed by a number of interventions, but that those changes in turn do not effect measurable changes in behaviour. This casts doubt on the idea that there is a direct causal relation between the implicit biases allegedly measured by the IAT and discriminating behaviour. (Note though that their study, although cited by different media outlets, has not been published in any peer-reviewed journal yet.)

[51] (Greenwald et al., 2009a). Cf. studies showing IAT-scores to reliably predict different types of, especially non-verbal, behaviour: e.g., calling back applicants with Arabic names for a job interview (Rooth, 2010); voting for the black (rather than the white) political candidate (Greenwald et al., 2009b); but see (Ditonto et al., 2013) for contradictory results.

[52] (Karpinski and Hilton, 2001, p. 786). Cf. (Arkes and Tetlock, 2004).

Thus, the empirical evidence, even if far from conclusive, makes talk about discrimination *from* implicit bias rather problematic. There might not be a clear-cut phenomenon of implicit bias discrimination, which falls under the above proposed, precisified condition:

> (GDiii''°) It is *not* because X has P-related intentions that X treats Y worse, but rather because X has a P-related implicit bias.

Of course, the more general condition still covers the cases here discussed:

> (GDiii'') It is *not* because X has P-related intentions that X treats Y worse, but rather because of *some other* reason.

This means, however, that if the causal connection challenge bears out empirically, we lose sight of any specific phenomenon of *implicit bias* discrimination, as captured by square (3).

4.4 The challenge from irreducibly collective bias

There is an alternative strategy for capturing implicit bias discrimination within my framework. This strategy takes into account a relatively new approach to understanding implicit bias, viz., as an irreducibly collective phenomenon. More specifically, the "bias of crowds" model[53] analyses implicit bias as "a social phenomenon that passes through individual minds, rather than residing in them" – much like a "wave" passes through the crowd of spectators in a soccer stadium, that cannot be assessed by measuring individual spectators' "personal propensity to stand or sit".[54]

The model builds on earlier studies such as the following. A 2009 study aggregated individual (Gender–Science) IAT-scores at the national levels for 34 countries and showed that these aggregated scores predict national sex differences in science and math achievements for eight-graders. The researchers also analysed the predictive role of aggregated explicit stereotypes and concluded that "explicit stereotypes uniquely accounted for 2% of variance in the science sex gap and 1% of the math sex gap, whereas implicit stereotypes uniquely accounted for 19% and 24%, respectively".[55] A 2015 study showed that, for different regions of the US, the average implicit (Race) bias among White residents correlated with racial disparities in use of lethal police force, when controlling for other relevant factors.[56]

What is puzzling is that, on these collective levels, average biases are very stable, and strongly correlated to these social outcomes – while on the individual level, IAT-scores fluctuate considerably and are only weakly correlated to discriminating behaviour. This occasions Keith Payne and colleagues to "propose that implicit bias reflects the accessibility of concepts [the likelihood that a thought, evaluation, stereotype, trait, or other piece of information will be retrieved for use] linked to a social category", when such category is activated, rather than individual mental states.[57] The authors note that the individual accessibility of such concepts

[53] For their ground-breaking article, see (Payne et al., 2017). For critical discussions, see the rest of this special issue of *Psychological Inquiry*.
[54] (Payne et al., 2017, p. 236).
[55] (Nosek et al., 2009).
[56] (Hehman et al., 2018). For further studies, see (Payne et al., 2017).
[57] (ibid., p. 235)

varies with the specific situation of the individual, e.g. when taking the IAT. Still, when "aggregated across a sample of subjects, the average bias score will reflect the knowledge [of stereotypes and prejudices] with the most widely shared accessibility", while idiosyncratic variations will be averaged away because they are randomly distributed across persons".[58] The "bias of crowds" model proposes that individual implicit bias is "a psychological marker of systemic prejudice in the environment", i.e. of the *systemic* bias that is constituted by inequalities along socially salient lines.[59]

On this model, individual implicit bias – as assumed in Alex' and Billie's cases – is not the real problem. Systemic bias is. But does this not mean that the discrimination framework, with its focus on individual agents and actions, ceases to be useful? It seems that that there is nothing left to meaningfully label as 'discriminating', according to my definition. There is no discriminating individual agent, X, who acts for P-related reasons, be they intentional mental states like racist or sexist beliefs and desires or non-intentional, race- or gender-related mental states beyond the agent's introspective accessibility or endorsement. Indeed, there is no specific individual action or behaviour, ϕ-ing or χ-ing, constitutive of disparate treatment. On the level of taxonomy, implicit bias has no place in any of the four boxes of the above Table 1.

However, systemic inequalities along socially salient lines are, of course, the hallmark of what is usually called *structural* discrimination. And there is a way to capture this systemic bias within my framework: variables X and Y do not need to stand for individuals. They can denote collectives: a society X, a group Y of members of this society, united by their property P. Likewise ϕ-ing or χ-ing need not refer to individual actions, but can denote e.g. alternative societal setups, ways of organising society. This, then, is how my framework can accommodate the irreducibly collective phenomenon of systemic bias discrimination:

> (GD') A society, X, group discriminates against a group, Y, by social ϕ-ing if, and only if:
>
> > (i) There is a property, P, such that all and only members of Y have P (or are believed to have P),
> > (ii') Had the members of Y not had P (or not been believed to have P), X would have had societal setup χ, constituting better treatment of members of Y, rather than societal setup ϕ,
> > (iii'') It is not because X has P-related intentions that X has societal setup ϕ rather than χ, but rather because of some other reason, and
> > (iv) P is the property of being a member of a socially salient group [i.e., a group perceived membership of which is important to the structure of social interactions across a wide range of social contexts].

Analysing the moral implications of this classification in detail will require an article of its own. What is clear at this stage is that, if we take the challenge from irreducibly collective bias seriously, the assessment of the moral wrongness must take place at the collective level,

[58] (Payne et al., 2017, p. 237).
[59] As Payne and colleagues write, "these [inequalities] constitute the racism and sexism itself"(Payne et al., 2017, p. 238). As an aside, this approach should also move our attention from bias-reduction by individual intervention (such as anonymizing applications) towards systemic solutions (such as affirmative action).

concerning the societal setup. And moral responsibility will be irreducibly collective, assignable to society X, rather than to the individuals constituting it.

5 Conclusion

In this article, I have tried to connect the phenomenon of implicit bias to a discrimination theoretical framework. In the course of this endeavour, I suggested a novel way of assessing existing distinctions between direct and indirect discrimination, resulting in a taxonomy of four forms of discrimination, which makes room for – and thereby makes sense of – implicit bias discrimination. I then dealt with four challenges to my proposal of capturing implicit bias within my discrimination framework: the metaphysical challenge, the moral insignificance challenge, the causal connection challenge, and the challenge from irreducibly collective bias. There is work left to do, mainly regarding the more specific moral implications of the distinctions constituting my framework, and regarding the switch from individual to collective agents which I proposed as a way of dealing with the challenge from irreducibly collective bias. There is, moreover, the urgent question of what, if anything, we can and should do about implicit bias discrimination in its different forms, including systemic bias. These are, however, questions for another day.

References

Altman, A., 2016. Discrimination, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.
Arkes, H.R., Tetlock, P.E., 2004. Attributions of Implicit Prejudice, or "Would Jesse Jackson 'Fail' the Implicit Association Test?" Psychol. Inq. 15, 257–278.
Berndt Rasmussen, K., 2018. Harm and Discrimination. Ethical Theory Moral Pract. https://doi.org/10.1007/s10677-018-9908-4
Brownstein, M., 2016. Attributionism and Moral Responsibility for Implicit Bias. Rev. Philos. Psychol. 7, 765–786. https://doi.org/10.1007/s13164-015-0287-7
Cortina, L.M., Kabat-Farr, D., Leskinen, E.A., Huerta, M., Magley, V.J., 2013. Selective Incivility as Modern Discrimination in Organizations: Evidence and Impact. J. Manag. 39, 1579–1605. https://doi.org/10.1177/0149206311418835
Ditonto, T.M., Lau, R.R., Sears, D.O., 2013. AMPing Racial Attitudes: Comparing the Power of Explicit and Implicit Racism Measures in 2008. Polit. Psychol. 34, 487–510. https://doi.org/10.1111/pops.12013
Dovidio, J.F., Kawakami, K., Gaertner, S.L., 2002. Implicit and explicit prejudice and interracial interaction. J. Pers. Soc. Psychol. 82, 62–68.
Egan, A., 2011. Comments on Gendler's, "the epistemic costs of implicit bias." Philos. Stud. 156, 65. https://doi.org/10.1007/s11098-011-9803-5
Fischer, J.M., 2018. The Freedom Required for Moral Responsibility, in: Virtue, Happiness, Knowledge: Themes from the Work of Gail Fine and Terence Irwin. Oxford University Press.
Forscher, P., Lai, C.K., Axt, J.R., Ebersole, C.R., Herman, M., Devine, P.G., Nosek, B.A., n.d. A Meta-Analysis of Change in Implicit Bias [WWW Document]. ResearchGate. URL https://www.researchgate.net/publication/308926636_A_Meta-Analysis_of_Change_in_Implicit_Bias (accessed 2.15.18).
Frankfurt, H.G., 1971. Freedom of the Will and the Concept of a Person. J. Philos. 68, 5–20. https://doi.org/10.2307/2024717
Gendler, T.S., 2008. Alief and Belief. J. Philos. 105, 634–663.
Greenwald, A.G., Banaji, M.R., 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. Psychol. Rev. 102, 4–27.
Greenwald, A.G., Banaji, M.R., Nosek, B.A., Link to external site, this link will open in a

new window, 2015. Statistically small effects of the Implicit Association Test can have societally large effects. J. Pers. Soc. Psychol. 108, 553–561. http://dx.doi.org.ezp.sub.su.se/10.1037/pspa0000016
Greenwald, A.G., Poehlman, T.A., Uhlmann, E.L., Banaji, M.R., 2009a. Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. J. Pers. Soc. Psychol. 97, 17–41. https://doi.org/10.1037/a0015575
Greenwald, A.G., Smith, C.T., Sriram, N., Bar-Anan, Y., Nosek, B.A., 2009b. Implicit Race Attitudes Predicted Vote in the 2008 U.S. Presidential Election. Anal. Soc. Issues Public Policy 9, 241–253. https://doi.org/10.1111/j.1530-2415.2009.01195.x
Haslanger, S., 2013. Social Meaning and Philosophical Method.
Hehman, E., Flake, J.K., Calanchini, J., 2018. Disproportionate Use of Lethal Force in Policing Is Associated With Regional Racial Biases of Residents. Soc. Psychol. Personal. Sci. 9, 393–401. https://doi.org/10.1177/1948550617711229
Hellman, D., 2017. Discrimination and Social Meaning, in: Lippert-Rasmussen, K. (Ed.), The Routledge Handbook of the Ethics of Discrimination. Routledge, London ; New York.
Hellman, D., 2008. When Is Discrimination Wrong? Harvard University Press, Cambridge, MA.
Holroyd, J., 2017. The Social Psychology of Discrmination, in: The Routledge Handbook of the Ethics of Discrimination. Routledge, London ; New York.
Holroyd, J., 2012. Responsibility for Implicit Bias. J. Soc. Philos. 43, 274–306. https://doi.org/10.1111/j.1467-9833.2012.01565.x
Holroyd, J., Sweetman, J., 2016. The Heterogeneity of Implicit Bias, in: Implicit Bias and Philosophy, Vol. I: Metaphysics and Epistemology. Oxford University Press.
Hosking, K., Russell, R., 2016. Discrimination Law, Equality Law, and Implicit Bias, in: Brownstein, M., Saul, J. (Eds.), Implicit Bias and Philosophy. Oxford University Press, Oxford.
Karpinski, A., Hilton, J.L., 2001. Attitudes and the Implicit Association Test. J. Pers. Soc. Psychol. 81, 774–788. https://doi.org/10.1037//0022-3514.81.5.774
Khaitan, T., 2017. Indirect Discrimination, in: Lippert-Rasmussen, K. (Ed.), The Routledge Handbook of the Ethics of Discrimination. Routledge, London ; New York.
Lavergne, G.M., 2010. Before Brown: Heman Marion Sweatt, Thurgood Marshall, and the Long Road to Justice. University of Texas Press.
Leslie, S.-J., 2017. The Original Sin of Cognition: Fear, Prejudice, and Generalization. J. Philos. 114, 393–421.
Levy, N., 2017. Implicit Bias and Moral Responsibility: Probing the Data. Philos. Phenomenol. Res. 94, 3–26. https://doi.org/10.1111/phpr.12352
Levy, N., 2015. Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. Noûs 49, 800–823. https://doi.org/10.1111/nous.12074
Lippert-Rasmussen, K., 2017. The Philosophy of Discrimination: An Introduction, in: Lippert-Rasmussen, K. (Ed.), The Routledge Handbook of the Ethics of Discrimination. Routledge, London ; New York.
Lippert-Rasmussen, K., 2014. Born Free and Equal?: A Philosophical Inquiry into the Nature of Discrimination. Oxford University Press, Oxford; New York.
Mandelbaum, E., 2016. Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. Noûs 50, 629–658. https://doi.org/10.1111/nous.12089
Martell, R.F., Emrich, C.G., Robison-Cox, J., 2012. From bias to exclusion: A multilevel emergent theory of gender segregation in organizations. Res. Organ. Behav. 32, 137–162. https://doi.org/10.1016/j.riob.2012.10.001
Martell, R.F., Lane, D.M., Emrich, C.G., 1996. Male-female differences: A computer simulation. Am. Psychol. 51, 157–158.

McConnell, A.R., Leibold, J.M., 2001. Relations among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Attitudes. J. Exp. Soc. Psychol. 37, 435–442. https://doi.org/10.1006/jesp.2000.1470
Mendoza, J.J., 2017. Discrimination and Immigration, in: Lippert-Rasmussen, K. (Ed.), The Routledge Handbook of the Ethics of Discrimination. Routledge, London ; New York.
Mikkola, M., 2017. Discrimination and Trans Identities, in: Lippert-Rasmussen, K. (Ed.), The Routledge Handbook of the Ethics of Discrimination. Routledge, London ; New York.
Moreau, S., 2017. Discrimination and Freedom, in: Lippert-Rasmussen, K. (Ed.), The Routledge Handbook of the Ethics of Discrimination. Routledge, London ; New York.
Nosek, B.A., Smyth, F.L., Sriram, N., Lindner, N.M., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., Kesebir, S., Maliszewski, N., Neto, F., Olli, E., Park, J., Schnabel, K., Shiomura, K., Tulbure, B.T., Wiers, R.W., Somogyi, M., Akrami, N., Ekehammar, B., Vianello, M., Banaji, M.R., Greenwald, A.G., 2009. National differences in gender–science stereotypes predict national sex differences in science and math achievement. Proc. Natl. Acad. Sci. 106, 10593–10597. https://doi.org/10.1073/pnas.0809921106
Oswald, F.L., Mitchell, G., Blanton, H., Jaccard, J., Tetlock, P.E., 2015. Using the IAT to predict ethnic and racial discrimination: small effect sizes of unknown societal significance. J. Pers. Soc. Psychol. 108, 562–571. https://doi.org/10.1037/pspa0000023
Oswald, F.L., Mitchell, G., Blanton, H., Jaccard, J., Tetlock, P.E., 2013. Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. J. Pers. Soc. Psychol. 105, 171–192. http://dx.doi.org.ezp.sub.su.se/10.1037/a0032734
Payne, B.K., Vuletich, H.A., Lundberg, K.B., 2017. The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice. Psychol. Inq. 28, 233–248. https://doi.org/10.1080/1047840X.2017.1335568
Rooth, D.-O., 2010. Automatic associations and discrimination in hiring: Real world evidence. Labour Econ. 17, 523–534. https://doi.org/10.1016/j.labeco.2009.04.005
Rubin, V.C., 1983. Dark Matter in Spiral Galaxies. Sci. Am. 248, 96–109.
Saul, J., 2013. Implicit Bias, Stereotype Threat, and Women in Philosophy - Oxford Scholarship, in: Women in Philosophy: What Needs to Change? Oxford University Press, Oxford.
Schelling, T.C., 1971. Dynamic models of segregation. J. Math. Sociol. 1, 143–186. https://doi.org/10.1080/0022250X.1971.9989794
Schouten, G., 2017. Discrimination and Gender, in: Lippert-Rasmussen, K. (Ed.), The Routledge Handbook of the Ethics of Discrimination. Routledge, London ; New York.
Thomsen, F.K., 2017. Direct Discrimination, in: Lippert-Rasmussen, K. (Ed.), The Routledge Handbook of the Ethics of Discrimination. Routledge, London ; New York.