Implicit Bias and Discrimination

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

First draft – please do not quote or circulate without the author's permission.

1 Introduction

> Elliot is head of administration at a large company, and an outspoken defender of everyone's equal rights, both in her professional and in her personal life. Still, whenever the company hires a black person, she fails to be as welcoming, helpful and curious about them as she is with newly hired white people. Elliot is unaware of these subtle, but clearly noticeable differences. When she takes the Race-IAT, Elliot is surprised to learn that she displays a moderate automatic preference for white people over people of colour.

Simply put, the Race-IAT (Implicit Association Test) is a web-based application that is supposed to test the strength of one's racial prejudices. More specifically, it presents the user mid-screen with either alternating pictures of black or white people, or with positively or negatively valenced words, such as 'smile', 'peace', 'rotten' or 'agony. The user is required to sort these items, as fast and accurately as possible, into categories presented in the upper left and right corners of the screen. These categories are disjunctive: 'African-American *or* unpleasant' and 'European-American *or* pleasant'. It turns out that, when sorting the items into these categories, a clear majority of users is faster, and makes fewer mistakes, as compared to when sorting the items into the contrasting categories 'African-American *or* pleasant' and 'European-American *or* unpleasant'. This is interpreted as a preference for white people over black people. Just as in the hypothetical case of Elliot, most of the users scoring such a (slight, moderate or strong) preference for white people over black people endorse egalitarian or anti-racist views, when asked explicitly.

How can the discrepancy between Elliot's explicit (explicitly stated) beliefs and attitudes, on the one hand, and her IAT-results, on the other, be explained? How could such an explanation be connected to her observed workplace behaviour? And how should we morally evaluate this behaviour? These are the guiding questions for this article. The article's aim is twofold. First, I seek to improve our understanding of the phenomenon implicit bias, and specifically its moral status, by examining it through the lens of a theory of discrimination and its wrongness. Second, I simultaneously seek to improve my preferred theory of discrimination by exploring the conceptual space it can provide for implicit bias discrimination.

In section 2, I will introduce the phenomenon of implicit bias more thoroughly. Section 3 will give a brief overview of the theory of discrimination I will be working with and explore two ways of distinguishing direct and indirect discrimination. In the light of this pair of distinctions, section 4 will spell out four different forms of discrimination and locate implicit bias discrimination in the resulting conceptual space. Section 5 introduces a problem with a common way of describing implicit bias discrimination: the empirical evidence seems to count against the accurateness of this description. In the light of this difficulty, section 6 sets out to

locate implicit bias on an aggregated collected level instead. However, as section 7 argues, this risks making the whole discrimination framework rather obsolete. Section 8 concludes.

2 Implicit bias

One way to interpret the discrepancy between Elliot's explicit beliefs and attitudes and her IAT-score would be to deny the credibility of her explicit statements. Behind the thin varnish of civilisation, the idea goes, deep down Elliot really *is* a racist. To support this interpretation, one could refer to empirical studies showing that social desirability issues affect attitudes measured in surveys: many people hesitate, despite knowing that their anonymity is protected, to answer in a way that deviates from the established norm on politically or morally sensitive questions.[1] So under normal circumstances, Elliot 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 Elliot to answer quickly, stripping her of the possibility to cover them up.

Another interpretation, which has the strength of granting the possibility that Elliot is truthful in reporting her convictions, is that she *also* harbours inner mental processes which in some sense contradict these convictions, and which influence her behaviour – such as the behaviour measured by the IAT. 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)".[2]

There is an illuminating analogy here with 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* we have inferred from its observed gravitational effects on its intragalactic surroundings.[3]

Analogously, basically all we know about the *Human Dark Matter* we call implicit bias is inferred from its effects on human behaviour, as measured by tests like the IAT. Just as in Elliot'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 deceit or error theory – although arguably plausible in some cases – has the unpalatable implication of pointing out Elliot and everyone else as deceiving others or

[1] E.g. (Greenwald and Banaji, 1995).
[2] Jfr. (Holroyd and Sweetman, 2016, p. 81), (Saul, 2013, p. 40).
[3] Cf. (Rubin, 1983).

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 Elliot suggests that her differential treatment of new employees, of which she is unaware, is *caused* by her implicit biases, as measured by the IAT, which contradict Elliot's explicit egalitarian convictions and attitudes. If this is so, how should we morally evaluate Elliot's behaviour? Does she commit a moral wrong in failing to be equally welcoming, helpful and curious, and if so, what exactly does this wrong consist of?

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, at first sight, appear to be morally objectionable. In the legal context, the purpose is to ascribe legal liability to the discriminator, and signal legal non-acceptability of the treatment in question. In the moral context, likewise, we use the framework to ascribe moral responsibility to the discriminator, and signal moral non-acceptability of the treatment. Elliot's differential, disadvantageous treatment of black employees must surely be classified as discrimination, and, if it's wrong (morally and/or legally), be wrong for the same reasons other wrongful discriminatory acts are.

I will in this paper focus on a *moral* framework of discrimination theory. Applying this framework to Elliot's case is not entirely straightforward, however, due to the complicated, and hitherto not fully understood nature of implicit biases – the dark matter evidenced by IAT-scores – and their relation to behaviour *beyond* the reactions measured by the test scores. In order to see the challenges that the phenomenon implicit bias poses to theories of discrimination, let's consider a recent influential such theory.

3 Discrimination

I am here specifically concerned with group discrimination, i.e. discrimination due to group membership. In the present context, I will use the following, by now rather widely occurring definition of generic group discrimination:

> "An agent, X, [group] discriminates against someone, Y, in relation to another, Z, by ϕ-ing (e.g., hiring Z rather than Y) if, and only if:
>
> (i) There is a property, P, such that Y has P or X believes that Y has P, and Z does not have P or X believes that Z does not have P,
>
> (ii) X treats Y worse than he treats or would treat Z by ϕ-ing,
>
> (iii) It is because (X believes that) Y has P and (X believes that) Z does not have P that X treats Y worse than Z by ϕ-ing" and

(iv) ”P is the property of being a member of a socially salient group (to which Z does not belong)”.[4]

Most theories of discrimination moreover distinguish between two more specific forms of discrimination: direct and indirect. Direct discrimination is often 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”.[5] Drawing on the above definition, we can bring out the idea as follows: had X interacted with Z (rather than with Y), X would have χ-ed, rather than ϕ-ed, and χ-ing toward someone constitutes better treatment of them than ϕ-ing toward someone.

Indirect discrimination, on the other hand, is often understood in terms of *disparate impact*. The idea is that “a policy or procedure is on the face of it neutral, but in fact disproportionally disadvantages members of a particular social group”.[6] To spell this out with reference to the above definition: had X interacted with Z (rather than with Y), X would still have ϕ-ed, but ϕ-ing toward someone without P (such as Z) constitutes better treatment than ϕ-ing toward someone with P (such as Y).[7]

Using this rudimentary taxonomy of forms of discrimination, one may now wonder: which form of discrimination is in play in Elliot’s case? It seems quite obvious that it cannot be indirect discrimination in the present, disparate impact sense: Elliot’s failure to be welcoming, helpful and curious towards black new employees may be very subtle, but it surely is not “on the face of it neutral”. This would have been the case, had Elliot displayed an equally non-welcoming response to all new employees, in combination with e.g. a higher hostility sensitivity in black employees. But as described above, it clearly is direct discrimination, in the disparate treatment sense: instead of χ-ing (being welcoming, helpful and curious, as she is toward white new employees), Elliot ϕ-s (acts differently toward black new employees). The comparison is between two *different* kinds of behaviour, rather than between different effects of the *same* behaviour.

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

> Kim, who was Elliot’s predecessor as head of administration, held explicit racist views. Whenever the company hired a black person, Kim failed to be as

[4] (Lippert-Rasmussen, 2014, p. 15; 26). Lippert-Rasmussen suggests an additional, rather cumbersome (and to my mind non-necessary) condition (v) (Lippert-Rasmussen, 2014, p. 28), which I here leave out, as its presence or absence will not affect my arguments. For similar definitions, cf. many of the entries in (Thomsen, 2017). I believe that a definition of group discrimination in *intra*personally comparative terms (in terms of X discriminating against the actual Y-with-P compared to a hypothetical Y-without-P) is feasible and useful (see (Berndt Rasmussen, n.d.)). But I will here go with the definition that has become somewhat of a standard in the discrimination literature (which could be easily adapted by substituting ’Z’ with ‘Y-without-P’).

[5] (Thomsen, 2017, p. 21).

[6] (Holroyd, 2017, p. 382). This interpretation 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).

[7] For this way of drawing the distinction, see also (Hellman, 2017, p. 98).

> welcoming, helpful and curious about them as she was with newly hired white people. Kim was fully aware of, and content with these differences in treatment. Had the Race-IAT been around at that time, she would not have been surprised to learn that she displayed a strong automatic preference for white people over black people, and she would rather openly have endorsed it.

Is it really plausible to classify both Elliot's and Kim's behaviour as direct discrimination? This now raises a further question: why are we even interested in classifying implicit bias discrimination according to the direct/indirect divide? Is this not merely terminological quibbling?[8] What is the point of this distinction? This question is rarely discussed in its own right in the discrimination literature, but there are some clues: direct and indirect discrimination may have different moral status, they may constitute "two different kinds of wrong",[9] or rather, since they are construed as subcategories of the same thing, "two different types of one and the same wrong".[10] According to one interpretation, direct discrimination is more severe.[11]

But if this is the reason we want to make the distinction, the suggested way of making it becomes doubtful. Why would treating Y and Z differently, and thereby treating Z better, be morally more severe than treating Y and Z the same, where this constitutes better treatment of Z due to their lack of P? In both cases, Y is left at a disadvantage, compared to Z, and P has some causal role in this. Does it really matter whether this is brought about by differential treatment of Y and Z rather than by differential effects of the same treatment?

To make sense of the idea that the first instance is, in fact, more severe, we could point out that it – unlike the other – involves a *shift* of treatment due to the presence of P. That is, it is implied that the agent, X, here has a choice between two options – ϕ-ing and χ-ing, respectively – where their choice turns on the (believed) presence or absence of P in X's counterparts Y and Z. This, in fact, 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.

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".[12] In terms of the above definition of discrimination, we can bring out the idea as follows: X intends the worse treatment of Y, compared to Z, i.e., the 'because' in clause (iii) is analysed by reference to the agent's *beliefs and desires* concerning people having or lacking P.

---

[8] Cf. (Altman, 2016, para. 3.2).
[9] (Altman, 2016, para. 3.1).
[10] (Altman, 2016, p. 4.2).
[11] 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).
[12] (Lippert-Rasmussen, 2017, p. 3); for an explanation of the omissions, see below.

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".[13] In terms of the above definition: the worse treatment of Y, compared to Z, is not intended by X; the 'because' in clause (iii) instead refers to other reasons why people having or lacking P are disadvantaged or advantaged by X's treatment.[14]

Intuitively, the distinction along this divide does have the potential to capture differences in moral status: to many people's minds, it makes a difference whether an agent *intends* to disadvantage someone, as compared to *inadvertently* disadvantaging them. I surmise that this distinction lines up with the idea that Kim should be called out as *a racist* (due to having racist intentions, beliefs or desires), while Elliot's case rather is a (possibly less pernicious) case of *racist behaviour*.

Likewise, a number of moral theories of the wrongness of discrimination are sensitive to the distinction in this sense: they make the agent's mental states directly or indirectly relevant for the wrongness of the action. Considering a number of views on the moral wrongness of direct discrimination, Andrew Altman states:

> "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."[15]

That is, Altman states that what makes direct discrimination wrong, according to a number of views, is some sort of flaw in the *practical reasoning process behind* the action in question (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).[16] 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.[17]

---

[13] (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).

[14] A more general (less malice-focussed) way of capturing this distinction is to employ the socially salient property P either as a component of X's motivating reasons for ϕ-ing, or as part of an explanatory reason for the differential impact of X's action on Y vs. Z. According to this suggestion, direct discrimination refers to the agent's *motivating reasons* for disadvantaging Y, in the sense that there is an "intention to disadvantage the members of [some salient social] group [or some] other objectionable mental state, such as indifference or bias, motivating the act" (Altman, 2016, para. 2.2). Reconnecting this to the above definition of discrimination, we can bring out the idea as follows: the 'because' in clause (iii) is analysed in terms of that "the [belief] that Y, and not Z, has P is part of X's direct, motivating reason for ϕ-ing" (Lippert-Rasmussen, 2014, p. 38). Indirect discrimination instead refers to *explanatory reasons* for disadvantaging Y. Reconnecting this to the above definition of discrimination, we can bring out the idea as follows: the 'because' in clause (iii) is analysed in terms of that "the fact that Y, and not Z, has P causally explains X's ϕ-ing and this in turn is causally explained by the fact that people with P are often treated worse than those without P in the sense given by (i)" (ibid.).

[15] (Altman, 2016, para. 4.1)

[16] (Ibid.)

[17] (Altman, 2016, para. 4.2).

(Note, though, that there are other influential theories of the wrongness of discrimination which do not seem to 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.[18] Likewise, on a harm-based – purely outcome-focussed – account of the wrongness of discrimination, the intentional/non-intentional divide carries no direct moral significance.)

We thus have two separate distinctions, each of which has been used to define direct and indirect discrimination. In the literature on discrimination, these distinctions are often conflated, especially concerning definitions of direct discrimination. 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".[19] The intention component is here combined with the disparate treatment component. 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".[20]

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'."[21] And consider Mari Mikkola, who seems to define 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."[22]

My contention is that both the disparate treatment of different groups, and the disparate impact on different groups of the same treatment, 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 elsewhere. 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".[23]

[18] (Hellman, 2008).
[19] (Lippert-Rasmussen, 2017, p. 3, my italics).
[20] (Thomsen, 2017, pp. 21, 24)
[21] (Moreau, 2017, p. 164)
[22] (Mikkola, 2017, p. 289)
[23] (Mendoza, 2017, p. 258)

To bring out this idea in terms of the above definition of discrimination: when interacting with Y, X ϕ-s, and had X interacted with Z instead, X would still have ϕ-ed – but ϕ-ing toward someone without P (such as Z) constitutes better treatment than ϕ-ing toward someone with P (such as Y). Thus, this is an instance of discrimination in the disparate impact sense: the same treatment leads to different effects for Y and Z. Notably, ϕ-ing here makes reference to some policy A: a policy that is facially neutral and thus does not make reference to P, e.g., "do not hire people lacking a high school degree". But policy A has differential impact on those with P and those without, e.g. because all or most white applicants have a high school degree, while all or most black applicants lack one (due to a segregated education system etc.). Now it's possible that policy A, in itself facially neutral, was adopted *because of this differential impact* on those having or lacking P. Then we are dealing with *intentional* disparate impact (yet same treatment) discrimination – Mendoza's option (a). Or it might be that policy A was chosen on entirely different grounds and just happens to have this disparate effect. Then we are dealing with *non-intentional* disparate impact (yet same treatment) discrimination – Mendoza's option (b).

My proposal is now that even *disparate treatment* could 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 been hidden in the theoretical shadow of this conflated category. I want to explore whether it is precisely there we should locate differential treatment from implicit bias, as described in the above case of Elliot.[24]

4 Discrimination from implicit bias – the individual level

This means that we are really dealing with four, rather than two different forms of discrimination. The following table (Table 1) includes what I take to be intuitively plausible examples for each of the four possibilities.

[24] 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.

	Disparate treatment: ϕ-ing vs. χ-ing	*Disparate impact of ϕ-ing*
Intentional	1) A university in the early 1950's US South accepts a white applicant but turns down an equally qualified black applicant with the motivation: "This is a whites-only university. Blacks are referred to apply to some separate-but-equal university for African Americans."[25]	2) An employer turns down a white applicant and an equally qualified black applicant with the words: "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.[26]
Non-intentional	3) A university accepts a white PhD-candidate but turns down an equally qualified black candidate with the motivation that even though both were equally qualified, the white candidate still somehow seemed better, where such "seeming" stems from the evaluators' implicit biases.	4) An employer turns down a white applicant and an equally qualified black applicant with the words: "We don't hire people who lack high school education", without any awareness of the criterion's ability to track politically induced, race-correlated educational deficits.[27]

Table 1

To see how these four forms of discrimination capture different senses of the direct/indirect distinction, consider them as follows. In the intentional cases, (1) and (2), there is a *direct* intention to keep out blacks from the relevant position, while in the non-intentional cases, (3) and (4), there is no such direct intention but nevertheless, *indirectly*, the same effect. In the disparate impact cases, (2) and (4), treating everyone the same *indirectly*, due to different background conditions, leads to different outcomes, while in the disparate treatment cases, (1) and (3), the different treatment of the black and white candidates is a *direct* result of their racial affiliation, in the sense of not relying on any mediating background conditions.

It may now be suggested that implicit bias discrimination still *may* fall under (1): depending on the more 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.

Indeed, various 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.[28] In a similar vein, Mandelbaum suggests a "Structured Belief" hypothesis, according to which implicit biases arise from unconscious, "propositionally structured mental representations; mental

[25] 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* 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.

[26] 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).

[27] This case resembles *Griggs v. Duke Power Company* under "absence of a discriminatory intent" (Khaitan, 2017, p. 31).

[28] (Egan, 2011).

representations that we bear the belief relation to", that create "piggybacking associations" between the frequently involved concepts.[29] Sally Haslanger proposes to understand 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".[30] 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 (more relevantly in this context) "Muslims are terrorists".[31] Any of these models could help to integrate implicit bias with belief-desire models that underpin the intentional disparate treatment form of discrimination.

Thus, if it turns out that implicit bias is best conceptualised as relevantly similar to ordinary beliefs, the intentional/non-intentional distinction, in the sense I have here developed, would risk becoming superfluous. Even then, though, we may still want to uphold a potentially morally relevant difference between the cases of Elliot and Kim, respectively. We could then resist broadening the definition of 'intentional' to allow processes beyond the agent's direct introspective accessibility, control, or endorsement. This would allow us to classify differential treatment from implicit bias as non-intentional disparate treatment discrimination. The upshot so far is that there is some, rather flexible, conceptual space to account for the phenomenon of implicit bias discrimination – as conceptually and (possibly) morally distinct from other, more orthodox forms of discrimination.

5 The causal connection problem

There is a serious worry concerning implicit bias discrimination, as I have described the phenomenon in square (3) of Table 1, and in the case of Elliot: empirical studies seem to raise doubts as to whether these really are accurate descriptions. More specifically, the causal connections between implicit bias (the mental phenomenon explaining the IAT-score) and differential treatment are anything but well-understood and uncontroversial.

In Elliot's case, we simply assumed that implicit bias causes her differential, disadvantageous treatment of black employees. Some studies may be taken to corroborate such assumptions. E.g., Allen McDonnell & Jill Leibold assessed the IAT-scores of 42 white undergraduates, as well as their behaviour toward white and black experiment leaders, respectively. They found "significant correlations" between these scores and behaviours:

> "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. In addition […], 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."[32]

[29] (Mandelbaum, 2016).
[30](Haslanger, 2013, p. 11).
[31] (Leslie, 2017).
[32] (McConnell and Leibold, 2001, p. 439, my italics).

However, this study also found "a significant correlation between the IAT and explicit reports of prejudice",[33] which casts into doubt the causal role assigned to *implicit* biases.

One large meta-study from 2009 shows that, for socially sensitive issues, and specifically interracial (black vs. white) contacts, IAT-scores are a significantly better predictor for prejudiced behaviour then explicit (self-)reports of prejudice.[34] But there are other studies contradicting these findings: a more recent meta-study from 2013 concludes that a "closer look at the IAT criterion studies in the domains of ethnic and racial discrimination revealed, however, that the IAT provides little insight into who will discriminate against whom, and provides no more insight than explicit measures of bias".[35]

Moreover, further studies indicate that IAT-scores may 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".[36] This would mean that, even if correlations between IAT-scores and biased behaviour could be established, the underlying causal mechanism could be quite different from what we expected: the dark matter we call implicit bias might be an effect, rather than a cause of discrimination.

To sum up, the strength of the correlations between implicit bias (as measured by the IAT) and differential treatment is still very much under debate. Moreover, lacking generally accepted models of the mental phenomenon, we are not in any position to convincingly account for the causal mechanisms underlying such correlations either. But all this makes talk about discrimination *from* implicit bias rather problematic, on an individual level.

In the light of these difficulties, it makes sense to approach the problem of discrimination from implicit bias on another level: the aggregate collective level. I will sketch this approach in the next section, mostly building on empirical data pertaining to Swedish society. (I trust that a similar story can be told about many other modern Western/European societies.)

6 Discrimination from implicit bias – the collective level

Sweden, like many modern Western/European states, is a liberal democracy with explicitly gender-blind and ethnically blind legislation, and with strong anti-discrimination laws, backed up by an active discrimination ombudsman, prohibiting racist and sexist discrimination. Moreover, according to recent surveys, sexist and racist attitudes are held by minorities in

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

[34] (Greenwald et al., 2009a). Cf. studies showing IAT-scores to reliably predict different types of, especially non-verbal, behaviour: e.g. voting for the black (rather than the white) political candidate (Greenwald et al., 2009b) but see (Ditonto et al., 2013) for contradictory results; calling back applicants with Arabic names for a job interview (Rooth, 2010); acting in a civil manner towards female (Cortina et al., 2013) or black (Dovidio et al., 2002) coworkers.

[35] (Oswald et al., 2013) p.188. Cf. (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.)

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

Sweden. A clear majority reports positive attitudes toward diversity.[37] And likewise, a clear majority considers gender equality to be politically important.[38]

Still, when it comes to gender relations within Swedish society, while women to a greater degree than men have a higher education, they make only 80% of men's total income; take three out of four parental leave days; are systematically underrepresented in positions of power; and the labour market is dominated by gender-segregated occupations.[39] When it comes to ethnic relations, immigrants are statistically less likely to complete primary education, compared to those born in Sweden; unemployment and (relative) poverty affect disproportional percentages of non-EU immigrants (especially from the Middle East and North Africa); moreover the housing market is dominated by highly ethnically segregated residential areas.[40] In addition, a number of field experiments show that there is extensive ethnic discrimination in various sectors, such as the labour market and the housing market.[41]

Thus, the statistics and empirical studies point at widespread segregation and inequalities along the lines of ethnicity and gender. And, as we have seen, these observations cannot be explained by reference to racist or sexist laws, policies or social values. Just as in the individual case (think about Elliot's case above), there is a discrepancy between what we believe guides our actions – explicit rules or values – and the observed behaviours or outcomes.

My proposal is that at this collective level, implicit bias can do a better job of explaining and reconciling the observed phenomena than at the individual level. The reason is that even relatively weak correlations between individual implicit bias measures and individual differential behaviour suffice to account for sizeable inequalities on the aggregated, collective level. The starting point here is of course Thomas Schelling's modelling of the systemic effects of individual preferences. His dynamic models show how e.g. even very slight individual preferences for one's neighbourhood's racial profile can quickly lead to very segregated neighbourhoods.[42]

More recently, organisation researchers have proposed further improved ways to analyse such micro- to macro-level processes. Richard Martell et al. 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 that produced it".[43]

Their basic idea is that small, barely noticeable micro-level gender biases can give rise to entirely unintended, large-scale and persistent system-level gender inequalities. This idea can be supported by computational modelling. In one very simple computational model, Martell

[37] (Ahmadi et al., 2016). Percentage points have decreased though from 74 to 64 between 2014 and 2016; cf. (Mella et al., 2014). Cf. (World Values Survey Association, 2014), (Bail 2008).
[38] (SIFO, 2014). Cf. (World Values Survey Association, 2014).
[39] (SCB, 2016).
[40] (Socialstyrelsen 2010).
[41] (Bursell, 2014), (Rooth and Agerström, 2009), (Ahmed and Hammarstedt, 2008).
[42] (Schelling, 1971)
[43] (Martell et al., 2012, p. 142)

et al. examined 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 was 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."[44]

However, in real-world scenarios, causal connections between micro- and macro-level are far more complex than the ones explored in this (or, for that matter, Schelling's own) simplistic model. This is e.g. due to much more complex rules and signalling mechanisms for promotion decisions, as well as the reciprocal effects of macro-level outcomes on micro-level bias and behaviour, which further perpetuate large-scale inequalities. To capture the dynamics of such more complex emergent phenomena, the authors call for agent-based computational modelling of a more sophisticated kind:

> "Agent-based modeling involves the creation of a virtual environment to study the behavior of individuals within group settings that can range from small (e.g., a three-person group), to moderate (e.g., a business firm), to extremely large (e.g., a nation or society). It is designed to model a system in which individuals (agents) engage in behavior (often some sort of decision-making) within an environment where a prescribed set of rules is in force. In agent-based modeling, individuals are endowed with certain probabilistic behavioral tendencies (e.g., a preference for men over women of some initial predetermined magnitude) that may change in response to the nature of the localized environment which, itself, is responsive to the collective behavior of other agents. One example is the role of downward causation in perpetuating organizational segregation at senior levels discussed earlier in this article. Thus, the interactions that unfold in response to the rules in place, the various features of the environment and the nature of the entity that ultimately emerges are most often the focus of agent-based studies. For these reasons, agent-based modeling is particularly appropriate when the mobility of individuals within a system is a central concern, the population studied is heterogeneous, questions arise regarding whether, and under what conditions, certain groups may dominate others, and it is believed that the collective action of individuals may produce outcomes that are not intended."[45]

This is fascinating research which has the potential of illuminating the systemic role of implicit biases for social inequalities and segregation as emergent phenomena.

An interesting case to model – in an extremely large, viz. societal, setting – would be the results of the following 2009 study. Brian Nosek et al. aggregated individual IAT-scores, pertaining to gender–science stereotypes, at the national levels for 34 countries. They then 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

[44] (Martell et al., 1996, p. 157)
[45] (Martell et al., 2012, p. 149)

stereotypes uniquely accounted for 19% and 24%, respectively".[46] It should be possible to devise computational models to account for these correlations, while taking into account other country-specific parameters, e.g. pertaining to the education system, downward causation factors (e.g. the occurrence of female scientists or mathematicians in public discourse), etc.

Embedding implicit biases within the framework of emergent social phenomena, however, presents a challenge to their conceptualisation within theories of discrimination. The problem is that there now seems nothing left to meaningfully label as 'discriminating'.

7 The 'discrimination' problem

First, within the large collective picture tying together aggregated micro-level bias scores and macro-level inequalities, we have lost track of any individual agents, X, whose *intentions* refer to the socially salient property P (e.g. 'being a woman'). This might not appear too damaging, since we might reasonably concede that we are here dealing with non-intentional discrimination. However, non-intentional discrimination, both in the disparate treatment and the disparate impact sense are tied to specific actions or behaviours: ϕ-ing or χ-ing, as it were. But within this large collective picture we have also lost track of any specific actions or behaviours. There is no single act that constitutes disparate treatment or causes disparate impact to be isolated and considered. On the level of taxonomy, implicit bias discrimination now belongs to neither of the four boxes of the above Table 1.

All we have is, what we might call *disparate conditions*, along socially salient lines. This is the hallmark of what is usually called *structural* discrimination. Thus, Lippert-Rasmussen suggests that "[s]tructural discrimination obtains where, and only where, the social structures are such that certain socially salient groups are disadvantaged relative to others and where at least part of the explanation why these structures are in place appeals to the fact that these groups are subjected to or have been subjected to various forms of direct [i.e. intentional disparate treatment] discrimination".[47]

Somewhat more concretely, Mikkola suggests that structural discrimination "has its causes in norms, habits, symbolic meanings, and assumptions unquestionably embedded in and underlying our institutional and social arrangements".[48] In a similar vein, Altman states: "Indirect discrimination is structural when the rules and norms of society consistently produce disproportionately disadvantageous outcomes for the members of a certain group, relative to the other groups in society, the outcomes are unjust to the members of the disadvantaged group, and the production of the outcomes is to be explained by the group membership of those individuals."[49]

These writers' references to "social structures" "norms, habits", "rules" etc., in the indeterminate plural, however, are symptomatic. They are merely hand-waiving at a diffuse array of phenomena none of which could be isolated in order to be marked as, in itself,

[46] (Nosek et al., 2009).
[47] (Lippert-Rasmussen, 2014, pp. 77–78).
[48] (Mikkola, 2017, p. 289).
[49] (Altman, 2016, para. 4.2).

discriminatory. What is discriminating here is rather the entirety or systemic whole of the disparate conditions caused by these rules, norms and habits.

This is, then, how I would situate implicit bias discrimination, understood as structural discrimination, in the taxonomical context of this article: by complementing the *disparate treatment*/*disparate impact* distinction with a *disparate conditions* addendum (Table 2).

	Disparate treatment: ϕ-ing vs. χ-ing	*Disparate impact of ϕ-ing*	*Disparate conditions*
Intentional	(1)	(2)	
Non-intentional	(3)	(4)	(5) Structural discrimination

Table 2

Now, given this way of understanding implicit bias discrimination – as structural discrimination marked out by disparate conditions along socially salient lines – my estimation is that the discrimination framework ceases to be useful. Recall what I stated to be the purpose of applying the discrimination framework in the first place (towards the end of section 2 above). The idea was to apply this term to instances of differential or disadvantageous treatment of members of socially salient groups, which, at first sight, appear to be morally objectionable. The framework should help us ascribe moral responsibility (blame) *to the discriminator*, and signal moral non-acceptability *of the treatment*. With the discriminator as well as any specific form of treatment out of the picture, the framework becomes moot.

This of course does not mean that there is nothing morally objectionable with such macro-level disparate conditions, as underpinned by (among other things) micro-level implicit biases. To the contrary. But we already have a powerful alternative framework for assessing the wrongness of such social inequalities, which I think is more straightforward and potentially more useful here: theories of justice, including theories of distributive justice.[50]

8 Conclusion

In this article, I have tried to locate discrimination from individual implicit bias within a discrimination theoretical framework. In the course of this endeavour, I have suggested a novel way of assessing existing distinctions between direct and indirect discrimination, which resulted in a taxonomy of four forms of discrimination. I have then argued that the description of implicit bias discrimination – notably the assumption about the causal connections between individual bias and behaviour – although aptly fitting into this framework, is undermined by contradictory results of empirical studies. I have then suggested a way to understand implicit bias discrimination within a collective framework, but argued that as such, it no longer has much use of the discrimination theoretical framework. Instead, we should analyse the phenomenon and its wrongness within a framework of theories of justice. This concession, of course, raises many more questions, not least how to reconcile theories of justice with theories of discrimination, given my taxonomical framework. These are, however, questions for another day.

[50] This conclusion is very much in line with (Schouten, 2017, p. 192), and specifically her claim that "Not all social moral problems must be understood in terms of discrimination. [Patterned principles of distributive justice are a case in point.]" (ibid, p. 193, 194 n. 4).

References

Ahmadi, F., Palm, I., Ahmadi, N., 2016. Mångfaldsbarometern (text). Högskolan i Gävle, Gävle.

Ahmed, A.M., Hammarstedt, M., 2008. Discrimination in the rental housing market: A field experiment on the Internet. J. Urban Econ. 64, 362–372. https://doi.org/10.1016/j.jue.2008.02.004

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., n.d. Harm and Discrimination. Ethical Theory Moral Pract.

Bursell, M., 2014. The Multiple Burdens of Foreign-Named Men—Evidence from a Field Experiment on Gendered Ethnic Hiring Discrimination in Sweden. Eur. Sociol. Rev. 30, 399–409. https://doi.org/10.1093/esr/jcu047

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

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).

Greenwald, A.G., Banaji, M.R., 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. Psychol. Rev. 102, 4–27.

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.

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., 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.
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
Mella, O., Ahmadi, F., Palm, I., 2014. Mångfaldsbarometern (text). Högskolan i Gävle, Gävle.
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., 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
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
Rooth, D.-O., Agerström, J., 2009. Implicit Prejudice and Ethnic Minorities: Arab-Muslims in Sweden. Int. J. Manpow. 30, 43–55.

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.
SCB, S.S., 2016. På tal om kvinnor och män – Lathund om jämställdhet 2016.
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.
SIFO, 2014. Allt fler svenskar är feminister [WWW Document]. URL http://opinion.se/sakfragor/allt-fler-svenskar-ar-feminister (accessed 3.8.18).
Thomsen, F.K., 2017. Direct Discrimination, in: Lippert-Rasmussen, K. (Ed.), The Routledge Handbook of the Ethics of Discrimination. Routledge, London ; New York.
World Values Survey Association, 2014. World Values Survey, Wave 6 (2010-2014): Sweden 2011 [WWW Document]. URL http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp (accessed 3.9.18).