Katharina Berndt Rasmussen & Nicolas Olsson Yaouzis
katharina.berndt.rasmussen@iffs.se; niklas.olsson-yaouzis@philosophy.su.se

"If you're all egalitarians, how come you're so racist?" Social norms, implicit bias and discrimination

§1 Introduction

Swedish society today is marked by ethnic discrimination in the job market (Bursell 2014, Rooth 2010, Rooth & Agerström 2009), as well as segregation in the housing market (Börjesson 2018, Ahmed & Hammarstedt 2008). At the same time, surveys show that a majority of Swedes endorses egalitarian values (WVS, Ahmadi & Palm 2018), values that are mirrored in jurisdiction (e.g., antidiscrimination law) as well as politics (e.g., the programs of all governing political parties during the last decades). Research in social psychology has suggested an unexpected answer to the question of why racial inequalities and discrimination persist, in spite of prevalent egalitarian values: they might be due to implicit racial biases – roughly, unconscious and automatic race-related stereotypes and associations. However, the mechanisms connecting implicit bias to discrimination and social inequalities, and to social values and norms in general, are hitherto poorly analysed and understood. Moreover, the unconscious, automatic, and unendorsed nature of the phenomenon seems to undermine the justification for holding people responsible for their biases and resulting behaviour. Yet, to therefore "simply let people off the hook" (Zheng 2016: 62) may make it harder to address and reduce racial inequalities and discrimination.

The aim of our project is to improve our understanding of implicit biases' effects on social norm followers, and to analyse the implications for moral responsibility. More specifically, we aim to (A) develop a dynamic game-theoretic model of social norms, which can accommodate the agents' implicit biases and model their effects on the social outcome. This will provide a new way of understanding the micro-level mechanisms pertaining to the macro-level problems of racial inequalities and discrimination. Moreover, we thereby propose (B) a new basis for ascribing moral responsibility to these agents. Ultimately, we expect this framework to provide us with a new set of, and justification for, tools to address and reduce these injustices.

§2 State-of-the-art

The project seeks to bring the game-theoretical literature on social norms into conversation with the philosophical literature on implicit bias and responsibility, while being sensitive to the empirical findings on implicit bias.

There is a large body of empirical implicit bias literature, originating from Greenwald & Banaji 1995 and Greenwald et al. 1998. Most relevant for our project are recent meta studies concerning replicability of implicit measures and their predictive value for discriminating behaviour (e.g. Greenwald et al. 2009, Oswald et al. 2013, Gawronski et al. 2017, Forscher et al. ms), as well as the ongoing discussion of the "bias of crowds" paradigm, originating from the special journal issue surrounding Payne et al. (2017).

The philosophical literature on implicit bias draws on these empirical findings within social psychology, with pioneering explorations into the metaphysics of implicit bias by Gendler (2008a, 2008b), Egan (2008), Schwitzgebel (2010, 2013), Madva (2012), Levy (2015),

Mandelbaum (2016). Most relevant for our purposes is the emerging debate on moral responsibility for implicit bias, with notable contributions by Holroyd (2012), Saul (2013), Levy (2016), Madva (2018), and contributions in Brownstein & Saul (2016); as well as the debate on the epistemology of implicit bias and epistemic/ethical dilemmas (Gendler 2011, Egan 2011, Beeghly 2014). A further emerging debate concerns the question of whether individual interventions can contribute to decreasing social injustice or whether structural change is required (Madva 2016, Haslanger 2015, 2017, Zheng 2018, Davidsson & Kelly 2018).

There is an extensive philosophical literature concerning the nature and function of conventions and social norms. Some notable historical sources include Hobbes (1651), Hume (1738), Rousseau (1984). More recent contributions include Lewis (1969), Schelling (2006), Axelrod (2009), Skyrms (2003), and Alexander & Skyrms (1999). For our purposes, Bicchieri's (2005, 2017) recent work on the nature of social norms is the most relevant.

§3 Purpose

The project aims to answer two urgent questions: (I) Why do racism and discrimination persist, despite widespread egalitarian social norms? We seek to answer this question by integrating implicit bias into a game theoretic model of social norms. (II) What can and should we do about it? We develop the model's implications for issues of moral responsibility and interventions.

In the course of developing the model and its implications, we moreover aim to fill in some of the gaps in current debates: (i) Although there is a recent attempt to connect the discussion of implicit bias to the literature on the psychology of social norms (Davidsson & Kelly 2018), we still lack an analysis of the phenomenon within a game-theoretic framework of social norms, which can provide micro-to-macro mechanisms. This project attempt to bridge this gap by modelling implicit bias as a tremble in a dynamic model of social norm following. This will provide a new way of understanding the mechanisms pertaining to racism and discrimination.

(ii) Most of the debate concerning moral responsibility and implicit bias is based on the assumption that individual implicit bias contributes to causing, or at least reliably predicts, discriminating behaviour. This assumption, however, is challenged by empirical studies indicating that the predictive value of implicit bias is low (or at least not higher than the predictive value of explicit prejudice), and that individual measures such as the IAT face serious replicability issues. (see, e.g., Oswald et al 2013, Gawronski et al. 2017, Forscher et al ms) On the other hand, there is a growing evidence that on a collective level, average test scores are remarkably stable over time, as well as significantly correlated to social inequalities (Payne et al. 2017). However, a challenge for the current emerging "bias of crowds" paradigm is to identify the mechanisms connecting aggregated test scores to social outcomes.

(iii)The philosophical debate on bias and responsibility has hitherto focused on individuals as such, or individuals qua group members, and has not yet been informed by the recent collectivist turn of the empirical "bias of crowds" paradigm. Our game theoretical model will provide on this new understanding of implicit bias and thereby build a new basis for assigning moral responsibility to individuals qua social norm followers.

(iv) Laying bare the effects of implicit racial bias on groups adhering to egalitarian social norms will allow for a new analysis of what can be done to rectify these effects. We will discuss the normative status of these effects and the possible interventions, and analyse them in light of our new understanding of social norm followers' moral responsibility.

§4 Project

The purpose of the project is to answer two broad questions. First, in what way, if any, are we responsible for our implicit bias? We believe that in order to answer this question we have to analyse how implicit bias works. This motivates the second question: what are the micro-mechanisms that connect implicit bias to social outcomes? Our model proposes a much-needed mechanism connecting individual implicit bias with social outcomes. This provides a new foundation for moral responsibility for implicit bias.

A natural place to look for mechanisms connecting individual phenomena with unintended social outcomes is game theory. Any interaction between two or more agents where the outcome is determined by what the players jointly do can be described as a game, according to game theory. More formally, a game consists of a set of players, a set of strategies for each player, and a set of payoffs for each player and each possible combination of individual strategies (Dixit et al. 2015). A first, methodological part of the project focuses on the objection that game theoretic accounts of social norms fail to capture the phenomenology of norms (Gilbert 2008, Anderson 2000). We concede that, if the goal is to understand how people think of social norms, then it is important to provide a theory that captures its phenomenology. However, our main interest is not to understand how ordinary people think and talk about social norms. Our ultimate goal is to examine what, if anything, can be done to get rid of racism and discrimination. For this purpose, a rational reconstruction of what social norms are may suffice (Bicchieri 2005; Hedström & Bearman 2009). We argue that, in so far as the reconstruction allows us to make good predictions about what will happen when we make certain social interventions, it does not really matter whether the reconstruction provides understanding of how people think of their social norms. Within this part of the project, we will also compare our instrumentalist approach with more phenomenologically accurate approaches, such as Davidson and Kelly's (2018) discussion on implicit bias and social norms.

Our project builds on our very recent work on sexual harassment and social norms (Berndt Rasmussen & Olsson Yaouzis ms). In that work we expand Bicchieri's game theoretic account of social norms and use it to analyse sexual harassment. We believe that the model provides an intuitive and fruitful way of thinking about social norms and how behaviour can be manipulated (see §4.1). Thus, the second part of the project will work to expand the model to include the effects of implicit bias as asymmetric so-called "trembles" (see §4.2). That is, we suggest that implicit bias can be modelled as a propensity to make mistakes and deviate from a norm that one in is in fact willing to follow. In addition to including implicit bias and examining its effects, the particularities of racism and discrimination seem to have some surprising effects. Unlike sexual harassment, much racism and discrimination are nigh unnoticeable. Because the willingness to follow a social norm is conditioned on the belief that enough others follow the norm, the difficulty of noticing norm violations may contribute to upholding the norm even in the face of widespread violations. This creates a problem for policy makers (akin to, but separate from, so-called epistemic/ethical dilemmas, see §2). On

the one hand, they need to inform individuals about the effects of their implicit biases. On the other hand, they need to avoid spreading the impression that there are many norm violations, since this may lead to a breakdown of the norm.

A third part of the project is open-ended and explores further extensions of the model. For example, does the case where agents interact on a grid rather than at random affect the longevity of egalitarian norms? (cf. Alexander and Skyrms 1999) If the agents are able to send costless signals to each other, will this help uphold norms? (cf. Wärneryd 1991) What happens if different types of norm-followers have the ability to identify other norm-followers and only interact with them? (cf. Skyrms and Pemantle 2000).

A fourth part will relate the model and its implications to the discussion of moral responsibility and implicit bias (see 4.3). Are people praiseworthy for sticking to egalitarian norms, or blameworthy for trembling? Does the special nature of implicit bias exculpate them from moral responsibility?

§ 4.1 The model

The model connecting social norms and implicit bias, on the one hand, with racism and discrimination, on the other, is easiest to grasp by considering a simple game. Suppose that a large firm consists of n sections and that each section's middle manager has to hire a new worker. Furthermore, suppose Ada, a middle manager, has the choice between hiring an ethnic Swede or a non-Swede who has a crucial skill-set. She believes that the firm needs at least one employee with that skill-set. However, she also believes that it would be a hassle to employ a non-Swede in her section, and that the hassle would be even worse if her section was the only section in the firm with a non-Swede. This may be because Ada believes that culturally homogenous groups get along better than culturally heterogenous groups.

Therefore, Ada prefers the outcome where she is the only one to hire a Swede to the outcome where everyone hires a non-Swede; she prefers the outcome where everyone hires a Swede to the outcome where she is the only one to hire a non-Swede; and, finally, Ada prefers the outcome where everyone hires a non-Swede to the outcome where nobody hires a non-Swede. Ada's preference ordering is shared by the other middle managers of the firm, and if their preference orderings are transitive and complete then the payoffs can be represented by numbers where a higher number represents a greater payoff.

		Others	
		Hire Swede	Hire Non-Swede
Ada	Hire Swede	2, 2	4, 1
	Hire Non-Swede	1, 4	3, 3

The employer's dilemma (first number represents Ada's payoff, and second number the payoff others get)

The strategy set where everyone hires a Swede is the game's unique Nash equilibrium. That is, it is the only outcome where all players play their best response to the other players' strategies. Furthermore, hiring a Swede dominates hiring a non-Swede, so whatever the other middle managers do, Ada will be better off by hiring a Swede. If Ada and the other middle managers are utility maximisers, they will hire Swedes although they prefer the outcome

where non-Swedes are hired to the outcome where no non-Swedes are hired. A game with this structure is called a prisoner's dilemma (e.g., Rapoport 1974) and it, and the related tragedy of the commons (Hardin 1968), are often used to describe situations where individual rationality leads to a catastrophic but unintended collective outcome.

Explaining how people manage to co-operate (e.g, hire a non-Swede) in prisoner's dilemma situations has been one of the great challenges for the social sciences. Historically, Hobbes (1651) and Hume (1738) appealed to the shadow of the future to explain why it was rational to co-operate: rational agents co-operate because they are afraid that they will be excluded from future co-operative endeavours if they fail to co-operate. More recently, economists (Schelling 2006), biologists (Maynard Smith and Price 1973), political scientists (Axelrod 2009), and philosophers (Gauthier 1986) have suggested that focal points, conventions, or norms can explain the existence of pro-social behaviour. A problem with these accounts is that they identify social norms with a behavioural disposition and have little to say about norm-followers' beliefs and preferences. Recently, however, Bicchieri (2005, 2017) has provided a more detailed account of social norms using the building blocks of game theory, i.e., beliefs, preferences, and behaviour.

Consider the rule, "when hiring, do not discriminate against non-Swedes". Following, Bicchieri (2005) we suggest that a population can be separated into four types with respect to their attitude regarding this rule.

1. Mores consider non-discrimination a moral norm in the sense that they are unconditional followers. They would refrain from discriminating against non-Swedes even if others discriminated.
2. Normies consider non-discrimination a legitimate social norm and are conditional rule followers. They prefer to follow the rule if and only if
    a. Empirical expectations: they believe that enough others follow the rule, and
    b. Normative expectations:[1] they believe that enough others want them to follow the rule.
3. Jerkies consider non-discrimination an illegitimate social norm but are also conditional followers. They prefer to follow the rule if and only if
    a. Empirical expectations: they believe that enough others follow the rule, and
    b. Normative expectations with sanctions: they believe that enough others want them to follow the rule and are willing to sanction their behaviour.
4. Non-Confies consider discrimination to be a moral norm in the sense that they unconditionally discriminate against non-Swedes.

Suppose Ada and the other middle managers belong to the Normie-category and therefore accept as a legitimate norm the rule, "when hiring, do not discriminate against non-Swedes". Thus, their willingness to follow the rule is conditioned on their belief about whether others

[1] Normative expectations may be a bit confusing since it refers to the social norm followers' beliefs about what other expect them to do. We acknowledge this but stick to Bicchieri's labels to make comparisons between our models easier.

will comply and their belief about whether others expects them to comply. If the conditions are satisfied, the middle managers' preference orderings will change.

		Others	
		Hire Swede	Hire Non-Swede
Ada	Hire Swede	3, 3	2, 1
	Hire Non-Swede	1, 2	4, 4

The employer's dilemma from the view of the social norm-follower

When the conditions are satisfied, the cost of violating a legitimate norm exceeds the perceived benefit of employing a culturally homogenous group. Therefore, Ada now prefers the outcome where she too hires a non-Swede to the outcome where she is the only one to hire a Swede. However, if contrary to Ada's expectations, the others were to discriminate, then Ada would also prefer to hire a non-Swede. After all, she does not want to end up a sucker.

The new game has two Nash equilibria. The first is the strategy profile where everyone hires a Swede, and the second is the strategy profile where everyone hires a non-Swede. Normally, it is impossible to determine what equilibrium will be played by merely analysing the information provided by the game (e.g., Schelling 2006; Bicchieri 1993). However, since social norm followers are assumed to have beliefs about what the other players will do it is easy to determine what rational players will do. Another way of saying this is that if social norms are followed, they do not only transform the prisoner's dilemmas into coordination games, they also create focal points (Schelling 2006) that cause the players' expectations to converge on the Pareto optimal equilibrium. So, if Ada and the others are social norm followers, we can expect them to hire non-Swedes.

The players' actual norm-following thus depends on their beliefs about how many others follow the norm, and their beliefs about how many others expect them to follow the norm. So, by identifying the rules individuals use for updating their beliefs based on past observations it is possible to provide a dynamic model that describes how a population changes over time with respect to behaviour, beliefs and preferences.

Suppose that the game is played a number of rounds. Furthermore, suppose each round consists of two stages. At the first stage, the players decide what to do based on their beliefs. At the second stage, they update their beliefs, based on their observations of what the other players did at the first stage. Then, the next round begins, and the players use their new beliefs to decide what to do.

Furthermore, suppose that each individual uses a simple rule for updating their beliefs. They increase their expectation about the proportion of others who will follow the norm ("when hiring, do not discriminate against non-Swedes") next round, if their observation of this round exceeded their expectations. They lower their expectation for the next round if their observation fell short of their expectations in this round. Finally, they retain their expectation if their observation matches their expectations. It is also assumed that, if increased, the new expectation will not exceed the observed proportion, and if lowered, the new expectation will not fall below the observed proportion.

If the players' beliefs about the others' normative expectations do not change and no one makes a mistake, then, once they have accepted the norm as legitimate, they will stick to it indefinitely: their observation during the initial round will match their expectation that enough others follow the norm; therefore, nobody changes their expectation about the proportion who will follow the norm in the second round, and so on.

Now, suppose that players sometimes "tremble" (Selten 1975) and choose the wrong strategy by mistake. That is, at some round, *t*, a small proportion of the players who intend to not discriminate against non-Swedes happen to do so nonetheless, e.g., they make an honest mistake, or fall back into old habits. Upon observing this, the players will, at the end of the round, update their expectations about what the others will do in the next round. If the proportion of tremblers is large enough, players may cease to believe (i) that enough others conform and hence they will be unwilling to conform next round. However, even if the proportion of those who tremble is small enough, such that the player still believes (i) that enough others conform, a player may cease to believe (ii) that enough others expect her to conform, and hence be unwilling to conform next round. This might be the case if, e.g., the trembles are not met with expressions of blame or reproach, but rather with indifference by those who observe them, such that no reaffirmation of the expectation to conform is communicated to those observing the trembles. In either case, the player would stop conforming to the norm. If the observation of her subsequent defection causes further players to stop conforming, the social norm would soon break down.

§ 4.2 Integrating implicit bias

Now, suppose that what explains occasional "trembles" is the phenomenon of implicit bias. Thus, all group members intend (explicitly) to not discriminate against non-Swedes in their hiring decisions, but a proportion of this "crowd" happens to do so nonetheless – and this proportion is reliably predicted by the group's average result on implicit attitude measures (such as the IAT).

If each individual tremble is nigh unnoticeable, we might at first sight be tempted to infer the following. If no one notices these trembles from implicit bias, they will not be met with expressions of blame or reproach, such that no reaffirmation of the expectation to conform to the nondiscrimination norm is communicated to those who trembled. Thus, according to the above line of reasoning, the social norm might ultimately break down. However, we should notice that if these trembles are unnoticeable, updating beliefs at the end of each round will never result in any changes: each player will continue to believe (i) that all others conform to the norm. Moreover, since there will never be any occasion for the players to reaffirm their expectation of others to conform, e.g. by blaming them for failing to live up to this expectation, each player's belief (ii) that enough others expect her to conform will not change either.

An interesting implication of this is that it points to reasons to keep quiet about deviations from the social norm. As an illustration consider again the firm where Ada is one of n middle managers. Suppose that all middle managers at Ada's firm demand that at least a $0<d<n$ middle managers comply in order for them to be willing to follow the non-discrimination norm. Furthermore, suppose that as a result of implicit bias $0<e<n$ middle managers

unintentionally fail to follow the norm. The social optimum is of course the scenario where no middle managers tremble, i.e., e=0. The first best outcome, so to speak, is the outcome where everyone is willing to follow the norm and nobody suffers from implicit bias.

So, if e>0, there are reasons to attempt to reduce e. Since trembling is due to implicit bias, the middle managers may not even be aware that they discriminate. It may, therefore, be tempting to inform the middle managers of the effects of their implicit bias. The idea is that once they have become aware of their bias, they will, because they consider the norm legitimate, try to reduce its effects.

However, because of conditional nature of social norms an awareness raising campaign may in fact by counterproductive. Suppose that the number of trembles is so large, e>n-d, that Ada and her colleagues would be unwilling to follow the norm if they became aware of e. In this scenario the only thing that keeps the non-discrimination norm from completely breaking down is the fact that the middle managers suffer from "false consciousness" and falsely believe that most of them are in fact conforming to the norm. Due to the conditional nature of social norms and the difficulty of identifying effects of implicit bias, the second-best outcome may be an outcome where n-e comply but everyone falsely believes that (almost) all n comply.

This analysis may provide us with a new tool to understand the tension glossed in our introduction (§1), of a society adhering to egalitarian norms and values, yet plagued by racial inequalities and discrimination.

§ 4.3 Implications for moral responsibility

If egalitarian norms can be – sincerely and explicitly – upheld, though they are constantly undermined by nigh unnoticeable trembles, what are the lessons for the moral responsibility of the purported norm followers? Are they praiseworthy for sticking to the norm, even though they unknowingly fail to live up to it? Are they blameworthy for their propensity to tremble (their implicit biases), or for each instance in which they actually tremble? Or does their lack of control, awareness or endorsement of these biases exculpate them?

The philosophical debate on implicit bias and moral responsibility has provided us with a number of important distinctions (Holroyd 2012, Holroyd et al. 2017). There are the backward-looking senses of (1) being blameworthy for one's bias (Saul 2012) and of (2) holding someone responsible, through practices and acts of expressing blame (Holroyd et al. 2017) or other kinds of responses (Zheng 2016). And there is the forward-looking sense of (3) taking responsibility for addressing one's biases (Holroyd 2012: 278). In addition, moral responsibility (in any of these senses) can regard (i) having implicit biases, (ii) manifesting them in behaviour, and (iii) responding to the knowledge that one (most likely) is biased (Holroyd 2012: 277f.).

A lion part of the debate on implicit bias and moral responsibility (see §2) has focused on the necessary conditions (e.g., awareness, control, or endorsement of the mental states motivating the action in question) for moral responsibility, in these various senses. Considering these conditions in light of the empirical evidence about implicit bias, some writers have denied that agents are blameworthy for acting from implicit bias (Saul 2012) or that they should be blamed (Zheng 2016). Many others, however, have proposed to revise

these conditions in light of this evidence, such that e.g. blameworthiness for acting from bias is retained (Holroyd 2012, Washington & Kelly 2016, Glasgow 2016, Faucher 2016). There are also some attempts to move the debate towards a forwardlooking sense of taking responsibility for manifesting implicit bias, by examining the consequences of the practice of blaming people for biased behaviour (e.g., Holroyd 2015). Sie & Voorst Vader-Bours (2016) also move in this direction. Their arguments, however, come with a collective twist: since implicit biases are "based on culturally and historically influenced [stereotypes and prejudices]", and changing the latter requires a collective effort, any individual responsibility for manifesting implicit biases is of an indirect kind (ibid: 103). Individuals, qua members of society, collectively uphold and contribute to these underlying stereotypes and prejudices and are thus indirectly responsible for addressing implicit bias by collective change.

Our model moves further in this direction by building on a collective framing of implicit bias within a context of social norms, in the spirit of the recent "bias of crowds" paradigm (Payne et al. 2017), which contends that bias measures say more about social context than any specific individual within it. This allows for a new and fruitful discussion of moral responsibility for implicit bias that addresses individuals in a – in this context hitherto neglected – capacity: as social norm followers. As a result, we will contribute to ideas concerning the forward-looking sense of taking responsibility, for (iii*) responding to the knowledge that one's group is biased. Thus, this sense of moral responsibility is new both in the sense of taking seriously the "bias of crowds" paradigm and in the sense of applying not to individuals simpliciter, or to individuals qua members of society but rather to individuals qua agents within a context of social norms.

We will consider the contradicting implications from our model: in one sense, taking responsibility for bias might imply refraining from acquiring knowledge that one's group is biased, in order to prevent the norm from breaking down. Might it even imply actively preventing dissemination of such knowledge within the group in general? Should we stop talking so much about race, racism and racial inequalities? In another sense, taking responsibility for bias might imply actively acquiring (and helping disseminate) such knowledge, in order to facilitate to implementing safeguards which can prevent trembling in the first place (e.g., anonymisation of job applications). We will connect this conundrum to the debate on epistemic/ethical dilemmas, though from a social norms framework. Moreover, we will connect our discussion to the debate between individualists (Madva 2016) and structuralists (Haslanger 2015, 2017, Zheng 2018) regarding what is needed for reducing racial inequalities and discrimination: individual improvement or structural change; one of the most recent attempts of reconciling camps through social norms – we add a related but different road to reconciliation within game theoretic framework of social norms.

Finally, providing a new way to understand the mechanisms connecting implicit bias to social outcomes, and the role they play within a social norm context, paves the way for considering new tools to combat racism and discrimination.

References

Ahmadi, F. & I. Palm (2018) Mångfaldsbarometern, Högskolan i Gävle.
Ahmed, A. M. & M. Hammarstedt. (2008) "Discrimination in the Rental Housing Market: A Field Experiment on the Internet". Journal of Urban Economics 64(2).

Alexander, J. & B. Skyrms (1999). “Bargaining with Neighbors: Is Justice Contagious?” Journal of Philosophy 96 (11).
Anderson, E. (2000). “Beyond Homo Economicus: New Developments in Theories of Social Norms” Philosophy and Public Affairs 29 (2).
Axelrod, R. (2009). The Evolution of Cooperation. Basic Books.
Beeghly, E. (2014) Seeing Difference: The Epistemology and Ethics of Stereotyping, PhD diss., University of California, Berkeley, California.
Berndt Rasmussen, K. (2018a) ”Harm and Discrimination”, Ethical Theory and Moral Practice
Berndt Rasmussen, K. (2018b) ”Människans mörka materia: så styrs vi av implicit bias”, Arena Essä.
Berndt Rasmussen, K. (ms) ”Implicit Bias and Discrimination”.
Berndt Rasmussen, K. & Å. Burman (forthcoming) “Människans mörka materia: Om implicit bias och moraliskt ansvar”, Filosofisk Tidskrift.
Berndt Rasmussen, K. & N. Olsson Yaouzis (ms) ”#metoo, social norms and sanctions”.
Bicchieri, C. (1993). Rationality and Coordination. Cambridge University Press.
Bicchieri, C. (2005). The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge University Press.
Bicchieri, C. (2017). Norms in the Wild: How to Diagnose, Measure, and Change Social Norms. Oxford University Press USA.
Brownstein, M. & J. Saul (2016) (eds.), Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics. Oxford University Press UK.
Bursell, M. (2014) “The Multiple Burdens of Foreign-Named Men—Evidence from a Field Experiment on Gendered Ethnic Hiring Discrimination in Sweden”. European Sociological Review 30(3).
Börjeson, L. (2018) ”Diversity and Segregation in Sweden”, Hyresgästföreningen.
Davidson, L. & D. Kelly (2018) “Minding the Gap: Bias, Soft Structures, and the Double Life of Social Norms”, Journal of Applied Philosophy, DOI: 10.1111/japp.12351.
Dixit, A., Skeath, S., Reiley, D. (2014) Games of Strategy. W.W. Norton Company.
Egan, A. (2008), “Seeing and believing: perception, belief formation and the divided mind”, Philosophical Studies, 140(1).
Egan, A. (2011) “Comments on Gendler’s ‘The epistemic costs of implicit bias’”, Philosophical Studies, 156.
Faucher, L. (2016) “Revisionism and Moral Responsibility”, in Brownstein & Saul (2016).
Forscher, P. S., et al. (ms) “A meta-analysis of change in implicit bias” (under review).
Gauthier, D. (1986) Morals by Agreement. Oxford University Press.
Gawronski, B., et al. (2017) “Temporal Stability of Implicit and Explicit Measures: A Longitudinal Analysis”, Personality and Social Psychology Bulletin 43(3).
Gendler, T. (2008a) “Alief and belief”, The Journal of Philosophy, 105(10).
Gendler, T. (2008b) “Alief in action (and reaction)”, Mind and Language, 23(5).
Gendler, T. (2011) “On the epistemic costs of implicit bias”, Philosophical Studies, 156.
Gilbert, M. (2008) “Social convention revisited” Topoi (1-2).
Glasgow, J. (2016) “Alienation and Responsibility”, in Brownstein & Saul (2016).
Greenwald, A. & M. Banaji (1995) “Implicit social cognition: attitudes, self-esteem, and stereotypes”, Psychological review, 102(1).
Greenwald, A., D. McGhee, & J. Schwartz (1998) “Measuring individual differences in implicit cognition: The implicit association test”, Journal of Personality and Social Psychology, 74.

Greenwald, A. G., et al. (2009) “Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity”. Journal of Personality and Social Psychology 97(1).
Hardin, G. (1968) “The Tragedy of the Commons”. Science. 162.
Haslanger (2017) “Injustice within Systems of Coordination and Cognition: Comment on Madva for Brains Blog” http://philosophyofbrains.com/2017/03/06/symposium-on-alex-madvas-a-plea-foranti- anti-individualism.aspx.
Haslanger, S. (2015) “Distinguished Lecture: Social Structure, Narrative and Explanation” Canadian Journal of Philosophy 45 (1).
Hedström, P. & P. Bearman (2009) “What is analytical sociology all about? An introductory essay”. In Hedström & Bearman (eds.) The Oxford Handbook of Analytical Sociology. Oxford University Press.
Hobbes, Thomas (1651). Leviathan. Harmondsworth, Penguin.
Holroyd, J., R. Scaife & T. Stafford (2017) “Responsibility for Implicit Bias”, Philosophy Compass 12(3).
Holroyd, J. (2015) “Holding each other accountable for implicitly biased behavior”, seminar presentation at IF, https://www.iffs.se/kalendarium/jules-holroyd-holding-each-otheraccountable-for-implicitly-biased-behaviour/.
Holroyd, J. (2012) “Responsibility for Implicit Bias”, Journal of Social Philosophy, 43(3).
Hume, David (1738) A Treatise of Human Nature. Oxford University Press.
Levy, N. (2016) “Implicit Bias and Moral Responsibility: Probing the Data”. Philosophy and Phenomenological Research 93 (3).
Levy, N., (2015) “Neither fish nor fowl: Implicit attitudes as patchy endorsements”, Noûs 49(4).
Lewis, David (1969) Convention: A Philosophical Study. Wiley-Blackwell.
Madva, A. (2016) “A Plea for Anti-Anti-Individualism: How Oversimple Psychology Misleads Social Policy” Ergo: An Open Access Journal of Philosophy, 3(27).
Madva, A. (2012) The hidden mechanisms of prejudice: Implicit bias and interpersonal fluency, PhD dissertation, Columbia University.
Madva, A. (2016) “Why Implicit Attitudes Are (Probably) not Beliefs", Synthese, 193.
Madva, A. (2018) "Implicit Bias, Moods, and Moral Responsibility", Pacific Philosophical Quarterly, DOI: 10.1111/papq.12212.
Mandelbaum, E. (2016) “Attitude, Association, and Inference: On the Propositional Structure of Implicit Bias”, Noûs 50(3).
Olsson Yaouzis, N. (2010) “Revolutionaries, despots, and rationality”. Rationality and Society, 22 (3).
Olsson Yaouzis, N. (2012a) “An evolutionary dynamic of revolutions” Public Choice 151(3-4).
Olsson Yaouzis, N. (2012b) Ideology, Rationality, and Revolution: An Essay on the Persistence of Oppression. PhD thesis, Stockholm University.
Olsson Yaouzis, N. (2018) “‘That is just what they want you to believe’: A modest defence of Marxist paranoia”. European Journal of Philosophy 26(2).
Olsson Yaouzis, N. (2019) “Morality and oppression”. In Richard Garner (ed.) The End of Morality. Routledge.
Oswald F.L., et al. (2013) “Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies”, Journal of Personality and Social Psychology 105(2).
Payne, B.K., H.A. Vuletich & K.B. Lundberg (2017) “The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice”, Psychological Inquiry: An International Journal for the Advancement of Psychological Theory 28:4.

Rapoport, A. (1974) “Prisoner’s Dilemma – Recollections and Observations”. In: Rapoport A. (ed.) Game Theory as a Theory of a Conflict Resolution, vol 2. Springer, Dordrecht.
Rooth, D.-O. (2010) “Automatic Associations and Discrimination in Hiring: Real World Evidence”. Labour Economics 17(3).
Rooth, D. & J. Agerström (2009) "Implicit Prejudice and Ethnic Minorities: Arab-Muslims in Sweden", International journal of manpower 30.
Rousseau, J.-J.(1984) A Discourse on Inequality. Penguin Books.
Saul, J. (2012) “Skepticism and Implicit Bias”, Disputatio, Lecture, 5(37).
Schelling, T. (2006) Micromotives and Macrobehaviour. W.W. Norton &Company.
Schwitzgebel, E. (2010) “Acting Contrary to Our Professed Beliefs, or The Gulf Between Occurrent Judgment and Dispositional Belief”, Pacific Philosophical Quarterly, 91.
Schwitzgebel, E. (2013) “A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box”, in Nottelmann (ed.) New Essays on Belief, Palgrave Macmillan.
Selten, R. (1975) “Reexamination of the perfectness concept for equilibrium points in extensive games”. Int J Game Theory 4.
Sie, M. & N. Vorst Vader-Bours (2016) “Personal Responsibility vis-à-vis Prejudice Resulting from Implicit Bias”, in Brownstein & Saul (2016).
Skyrms, B., & R. Pemantle (2000) “A dynamic model of social network formation”, Proceedings to the national academy of sciences of the USA. 97: 9340-9346.
Skyrms, Brian (2003) The Stag Hunt and the Evolution of Social Structure. Cambridge University Press.
Smith, J. & G. R. Price (1973) “The logic of animal conflict”, Nature, 246.
Warneryd, K. (1991) “Evolutionary stability in unanimity games with cheap talk”. Economics Letters 36(4).
Washington, N. & D. Kelly (2016) “Who’s Responsible for This? Moral Responsibility, Externalism, and Knowledge about Implicit Bias”. In Saul & Brownstein (2016).
WVS (World Values Survey), Waves 1–6, http://www.worldvaluessurvey.org/WVSOnline.jsp.
Zheng, R. (2016) “Attributability, Accountability and Implicit Attitudes”, in Brownstein & Saul (2016).
Zheng, R. (2018) "Bias, Structure, and Injustice: A Reply to Haslanger" Feminist Philosophy Quarterly 4(1).