Four Types of Moral Wriggle Room: Uncovering Mechanisms of Racial Discrimination*

Kai Spiekermann

December 8, 2013

Recent experiments in behavioral economics reveal that individuals frequently use so-called 'moral wriggle room' to avoid complying with costly normative demands. Different opportunities for strategic information manipulation are classified by developing a typology of 'moral wriggle rooms'. 'Moral wriggling' can often operate in an unconscious, yet systematic way. The example of racial discrimination shows how such self-serving biases promote subtle, indirect forms of discrimination.

1 Introduction

What we ought to do depends on facts. These facts can either be facts about the available actions and their context or facts about the applicable norms. Call the former 'action-facts' and the latter 'norm-facts'. If we are uncertain about these facts, we sometimes have the opportunity to acquire or ignore available information about them before acting. These opportunities can create strategic incentives. For example, you might think that you ought to comply with a norm that prohibits causing unnecessary greenhouse gas emissions. You also have an inkling that your excessive air travel might be relevant in that context. However, you tend to avoid information about the (very high) levels of emissions caused by your flying, but you like to hear about how planes become more efficient, how other emission sources are much worse, and so on. In other words, you shape your own belief system about action-facts so that you can convince yourself that your behaviour is—by and large—compliant with the norm.

In a similar fashion, you can also choose to manipulate your beliefs about norm-facts. For instance, the norm prohibiting unnecessary greenhouse gas emissions might

*I am grateful to Miranda Fricker and Michael Brady for helpful comments. I gave this paper at the Social and Political Thought Seminar, University of Sussex and would like to thank the audience for their questions.

not be uncontested. If so, you might be able to convince yourself that a more lenient norm applies, or that other people around you do not comply with the more stringent requirements anyhow, which, you might tell yourself, lowers the normative demand on you. If you succeed, you have strategically changed your beliefs about norm-facts to reduce your obligations. In both cases you have used 'moral wriggle room' (a term coined by Dana, Weber and Kuang, 2007)[1] to avoid costs of compliance.

There are two long and involved debates in epistemology that bear some resemblance to the issues addressed here, but are in fact conceptual distinct: first, the debate between doxastic voluntarists and non-voluntarists; second, the debate about the correct 'ethics of belief'. The first debate is about whether one can decide to believe something, or whether believing is an involuntary act (see, for instance, the contributions by Audi, Feldman and Ginet in Steup, 2001). Depending on one's position regarding the first question, one is then likely to give different answers to the second question, arriving at different views as to what we ought to believe or how we ought to form beliefs (cf. Chignell, 2013; Feldman, 2000; Kornblith, 1983). Contributing to these debates is not within the scope of this essay. All what matters for my discussion is the (much less contested) obligation to seek additional evidence if acquiring this evidence is easy and increases the chance of bringing about valuable outcomes. Or more precisely: the evidence is such that it enables a reasonable and well-meaning agent to perform actions with a higher expected (moral) value. In this uncontroversial sense, I maintain that there are *obligations of inquiry*; obligations to find out certain facts (with important normative implications) if there is a reasonable opportunity to do so (Hall and Johnson, 1998). These obligations of inquiry are not primarily justified because they promote true beliefs or because of a general duty to ground our beliefs on evidence (as some evidentialists would have it), they are primarily justified by prudential and moral (not epistemic) reasons—the true beliefs promote valuable outcomes (Feldman, 2000, p. 689).

Recent experiments in behavioral economics reveal that individuals frequently use moral wriggle room and often violate some fairly obvious obligations of inquiry. For example, several experiments show that 'strategic ignorance' can be reproduced in the artificial setting of a behavioural experiment. When people are put in situations in which obligations of inquiry apply, but complying with these obligations is potentially costly, a significant number of individuals choose to stay ignorant. Put in more psychological terms: individuals are subject to self-serving biases when deciding whether they obtain information that will have normative implications. They tend to avoid information that promises to increase and seek information that promises to reduce the normative demands on them. In this sense individuals are engaged in *strategic normative context shaping*.

We will see that the experimental literature tends to be focused on individual behaviour. However, insofar as failures to meet obligations of inquiry are rooted in social practices and institutions, the problem must be addressed from a social perspective, the perspective of social epistemology. Since the strategic shaping of normative context is achieved more easily in a group, the individual strategic manipulation of information

[1] Dana et al. spell 'wriggle' without the 'r'.

is often intertwined with collective biases and ignorance. For example, one's individual ability to maintain biased beliefs is much higher when all of one's peers have the same biased beliefs.

In the following, I will report the results of some experiments that create a temptation for strategic information manipulation in different ways. I will then classify the results by developing a typology of 'moral wriggle rooms', before discussing implications for social moral epistemology, using the case study of racial discrimination.

2 Some Experiments

2.1 Strategic Ignorance

Arguably the most influential paper on strategic ignorance in experimental economics is Dana, Weber and Kuang's *Exploiting Moral Wiggle Room* (2007). In their clever experimental design, subjects play a binary version of the so-called 'dictator game'. The baseline treatment is the game shown on the left-hand side of Figure 1. The 'dictator' (a more neutral term is used in the instructions for the participants) can choose between actions A and B, causing payoffs for himself and his randomly matched 'receiver', as stated in the table. A selfish dictator will choose action A, as this maximizes his own payoff. A more altruistically minded dictator will sacrifice \$1 to increase the payoff for the receiver by \$4. In line with previous results from dictator game treatments, 74% of dictators choose the 'fair' action B. In other words: considerations of equity often prevail in dictator games, even though the participants in the experiment will never learn with whom they played, and the game is played only once.

S_1 States S_2

		payoffs	
		Dictator	Receiver
Actions	A	6	1
	B	5	5

	payoffs	
	Dictator	Receiver
A	6	5
B	5	1

Figure 1: Dana, Weber, and Kuang's binary dictator games. Numbers are payoffs in US-\$.

The clever part of the experiment is the 'hidden information' treatment. In this treatment, the dictators are told that the payoff table can either be the one in state S_1 (left) or S_2 (right) with equal probability. Note that the dictator payoff does not depend on the state; the uncertainty is only about the effect of actions A and B *on the receiver*. Before the dictators choose their action, they can optionally find out which payoff table applies by clicking on a button. Revealing the information is free and is literally no more than 'one click away'. Remarkably, now that dictators have a chance to remain ignorant about the effect of their action on the receiver, 'fair' choices become much less frequent. 44% of dictators choose not to reveal which payoff table is applied.

Of those, 86% choose option A, thereby accepting a 50% chance of only giving \$1 to the receiver. By contrast, among those who reveal and find that S_1 is applicable (the right choice for S_2 is obvious since A is the better choice for both), only 25% choose the selfish action A. The results show that even a minimal barrier to full information can reduce altruistic choices dramatically. Dana et al. submit that these results suggest 'an illusory preference for fairness', that is, a willingness to show fair behaviour when the situation is clear, but to use ignorance as an excuse for selfishness when that excuse is available.

Dana et al.'s results had considerable impact among behavioural economists because they falsify many popular theories of altruistic behaviour. For example, if the dictator behaviour were driven by preferences over distributions alone (as is often assumed), we could not explain this dramatic reversal at all. Preferences-over-distributions theories have to be rejected because a dictator following such preferences would always reveal the information to satisfy his preferences. More promising are theories that appeal to self- and social image management (see Van der Weele, 2012; Bénabou and Tirole, 2011). Another approach is to think more explicitly about the prescriptions of the applicable norms that dictators feel subjected to. Spiekermann and Weiss (2013) suggest that some subjects take the applicable norms as only prescribing altruism under certainty, while uncertainty renders more selfish choices morally acceptable. If that is so, the subjects do not have 'an illusory preference for fairness'. Rather, they think that the requirements of fairness under uncertainty differ from those under certainty.

2.2 Biased Norm Perceptions

Apart from strategically choosing (not) to know about action-facts, one can also acquire self-serving norm-facts. If one anticipates a decision that might come with significant norm compliance costs, it is tempting to believe that the most lenient norm is applicable. If there are several competing norms that could be applied, or if there are more or less demanding interpretations of a norm, biased beliefs about norm-facts are to be expected.

Devising experiments to measure strategic norm-fact acquisition is challenging: one needs to design a situation of normative ambivalence, create a choice situation in which the potential norms become relevant, show that the self-serving behavioural patterns arise, and—crucially—also show that these behavioural patterns are indeed caused by self-servingly chosen norms, rather than mere selfish payoff maximization. One influential experimental design to this effect can be found in Konow (2000), another relevant article is Bicchieri and Chavez (2013). Here I only report those parts of the experiments that are most relevant for biased norm perception and omit some of the more intricate details.

In the first stage of Konow's experiment, all subjects are asked to perform what experimental economists call a 'real effort task': they have to fold letters and stuff them into envelopes. In the treatments of interest here, the subjects are provided with material for ten letters and have seven minutes to complete them, which is more than enough for all of them to finish the task. Dictator-receiver pairs are formed randomly. To create a normatively ambivalent situation, Konow introduces arbitrary levels of credit (salary) for each letter completed, ranging from 55 to 75 cents for dictators and 25 to 45 cents for

receivers, but averaging to 50 cents for every dictator-receiver pair. All subjects know the credits they and their counterpart receive. The total credits a pair has accrued can then be distributed between them by the dictator.

The subjects are told clearly that the differential credit for their work is entirely arbitrary, not related to their performance at all. This suggests that their joined credit ought to be distributed according to *output*, which is equal (both have completed ten letters). An alternative fairness norm, however, is to distribute according to *monetary contribution*, which varies because of the (arbitrary) difference in credits per envelope. Since the setup ensures that dictators always contribute more in terms of money, the latter norm[2] is attractive to them.

In a first step, the dictators are asked to distribute the credits between themselves and the receiver. Dictators keep, on average, 59% of the money to be distributed, deviating statistically significantly from an equal distribution. This is probably due to selfish preferences.[3] In a second, unannounced round, Konow puts the dictators in the position of a benevolent dictator who has to distribute money between two other subjects (which means that the dictator has no economic stake in this decision). Ingeniously, these two other subjects are exact mirror images of the dictator and his receiver from round 1: they are assigned exactly the same credits for their envelope stuffing. In other words, the two subjects in round 2 exhibit exactly the same normative status as the dictator and his receiver in round 1. Asking the dictator to distribute as a benevolent dictator in this mirror image of the previous allocation decision enables Konow to observe *what the dictator perceives as fair if she is not personally involved*. This choice will, of course, be influenced by the previous decision and the potentially selfish choices made there.

In the second round, dictators still assign, on average, more to the person with higher per-envelope credit (their mirror image), even though the credits are known to be arbitrary, and even though the dictator has no economic stakes in this distribution decision (ruling out selfishness). This shows that some dictators have managed to convince themselves that subjects with higher per-envelope credits (like themselves) are more deserving. More precisely, after excluding dictators with theory-inconsistent choices (i.e., those keeping less than 50%, or keeping less in the first than in the second round), 39% show this bias. The choices can then be compared with a control group in which subjects decide only as benevolent dictators, without a prior round 1. In the control group, giving is completely in line with an egalitarian norm, so that the choices in the second round of the treatment group can only be explained by a self-serving norm bias.[4]

In Konow's experiments, the subjects convince themselves to apply a norm that works in their favour. In Bicchieri and Chavez's (2013) experiment, the subjects convince themselves that the norm in their favour is endorsed by others, and is therefore the

[2] To be precise, Konow does not think that this behaviour can be considered norm-guided; for him the arbitrary credits offer a mere 'contextual pretense' (p. 1079) for selfishness.

[3] In fact, in standard dictator games subjects tend to be more selfish and an average giving of 41% is quite high.

[4] A skeptic could could argue that the result is due to a minimal identification effect, such that the dictator identifies with his mirror image in the second round. This alternative explanation cannot be fully dismissed.

relevant norm to comply with. Subjects play a so-called *ultimatum game*, in which the proposer can split \$10 between himself and a receiver. In contrast to the dictator games discussed above, the receiver in the ultimatum game is not totally passive: in ultimatum games, the proposer suggests a split of the money, then the receiver can either accept the split (which is then implemented as suggested) or reject the split, which leads to zero payoff for both. One can interpret the receiver's reject option as a 'veto'. If the receiver is a payoff-maximizer and the ultimatum game is played only once (ruling out any reciprocity or reputation effects), the rational choice is to accept any positive offer, which, in turn, should induce the proposer to offer the minimum amount above zero to the receiver. This is not what experimenters find in the lab: the offer rates are much higher and rejection of low offers is frequent (see Camerer, 2003, chapter 2 for details).

Returning to Bicchieri and Chavez's experiment, the available choices are to split the money as \$5-\$5, \$8-\$2, or to have a coin tossed to decide between the two aforementioned splits. It is common knowledge which actions are available and which is chosen. Unsurprisingly, most proposers and receivers agree that an equal split is fair, while \$8-\$2 is unfair. There is, however, a substantial disagreement between proposers and receivers as to whether tossing the coin is fair. 81% of proposers but only 51% of receivers consider the coin toss fair. In addition, proposers were asked to estimate how many receivers think that the coin toss is fair; the average estimate was 76%, while receivers only estimated an acceptance rate of 46% for the coin toss among themselves.

This shows that the proposers are biased towards beliefs that, if widely shared, would constitute a norm that permits choosing the coin toss, while the receivers are much less inclined to believe that such a permissive norm exists. Since the coin toss leads to a higher expected payoff for the proposer, the results show a selfishly biased perception of norm-facts.

2.3 Hiding Behind Small Cakes

The term 'hiding behind a small cake' was first introduced by the research group around Werner Güth in Güth, Huck and Ockenfels (1996) and Güth and Huck (1997) in the context of ultimatum games. 'Hiding behind a small cake' is possible when the proposer either has a small or a large amount of money available for distribution. The receiver knows the probability for the existence of the small or large 'cake' amount, but does not know which 'cake' the proposer actually has available. Given that, the proposer can trick the receiver into thinking that a split is fair by offering 50% *of the small cake amount*, even though the proposer has in fact received the large cake amount. This is indeed what some receivers appear to do (Güth, Huck and Ockenfels 1996, see also Güth and Huck, 1997 and Mitzkewitz and Nagel, 1993).[5]

The obvious problem with using the ultimatum game is that we cannot distinguish the proposer's motivation to maintain a positive image from the payoff incentive to avoid an offer rejection. In other words, we cannot determine whether the proposer gives 50% (or a little less) of the small cake for tactical reasons, or whether he chooses his offer

[5]Though most of the evidence is of an anecdotal nature—there is surprisingly little rigorous testing for 'hiding behind a small cake' in the ultimatum game.

because he likes to appear fair *regardless of the monetary payoff*. To investigate the latter motivation separately, one needs to remove the tactical incentive. This can be done by creating an opportunity to hide behind a small cake in a dictator game.

Ockenfels and Werner (2012) did precisely that. In cooperation with a large German newspaper, they had 701 members of the general public play a dictator game with two possible cake sizes, administered through the website of the daily *Die Welt*. The 'cake' was either 1000 or 3000 Euros. Participants know that one randomly selected receiver-dictator pair is paid out with real money. Ockenfels and Werner assign their subjects to one of two treatments. In NOINFO, the receivers only learn how much money the dictator assigns to them; no other information is provided. In INFO, the receivers are also told whether the dictator had the small or the large cake available (and the dictators know that the receivers will find out).[6] While dictators can hide behind a small cake in NOINFO, their cover is blown in INFO, so that one would expect a significant difference in giving between the two treatments at and just below 500 Euros (= 50% of the small cake).

This is indeed what Ockenfels and Werner find—though the effect is perhaps not as strong as one would have thought: in the INFO treatment, 10.4% of dictators choose to give 500 Euros or less when they have a large cake, compared to 14.8% in the NOINFO treatment, a statistically significant difference. The effect is more pronounced when looking at those dictators giving *exactly* 500 Euro when they have a large cake (2.6% versus 7.6%). Since 500 out of 3000 is not a particularly plausible fraction to give (unless one wants to hide behind a small cake), these results are quite revealing. They show that some subjects, given the opportunity, pretend to have a small cake when they really have a large one.

3 Four Kinds of 'Moral Wriggle Room'

The experiments described above show that self-serving information manipulation can work in different ways. It is useful to distinguish two dimensions: First, does the manipulation target information about action-facts or norm facts? Second, is the primary target of manipulation one's own belief, or someone else's belief?

Beginning with the first dimension, it is useful to describe *norm-facts* more carefully: they are facts about the norms prescribing appropriate actions in a specified context. In ambivalent situations, individuals might disagree about the applicable norms, and this disagreement can be fueled by selfish biases. In addition, in the case of conventions and social norms, the existence of social practices or social expectations partly determines the normative content, since social norms and conventions hinge on what others do or want us to do (Bicchieri, 2006; Southwood, 2011). Individuals might therefore develop a conveniently biased perception as to what others do or expect, as in Bicchieri and Chavez's experiment.

[6] To obtain as many data points as possible, all subjects are asked what they would do as dictators for two cases: if the cake is large and if the cake is small. Roles and cake sizes were assigned subsequently, and the stated strategies were then implemented for the dictators.

I take *action-facts* to be facts about the outcomes of actions and all other potentially normatively relevant facts, apart from the norm-facts. In the case of Dana, Weber and Kuang's hidden information treatment, there is an obvious action-fact: the dictator either does or does not cause a negative externality for the receiver. Avoiding this information is potentially advantageous, as one can prevent the direct confrontation with a demanding norm that prohibits the imposition of such severe losses. In the case of hiding behind a small cake, the relevant action-fact is, most immediately, about the endowment of the dictator (the small or large cake).

In both of these experiments, the relevant avoided facts pertain to the outcome, described as the payoff distribution between dictator and receiver. However, action-facts are not necessarily payoff distributions, they can also be about other normatively relevant properties. For instance, in an experimental setup in Spiekermann and Weiss (2013), dictators can decide (not) to learn about how deserving their receiver is. There, identical payoff distributions can be part of two quite different outcomes: giving, say, 10% of the endowment to an undeserving receiver may be fair, but giving 10% to a deserving receiver may be blatantly unfair. Here the relevant facts are about desert or entitlement, not payoff distributions.

The second dimension pertains to the person targeted for belief manipulation. The more obvious way is to manipulate one's own beliefs about the situation by being biased in a self-serving way. However, the 'hiding behind a small cake' literature shows that it is also possible to manipulate the beliefs of others in order to reduce expectations on oneself. By making partial or incomplete statements about the world, or by using ambiguity, subjects can create false impressions about normatively relevant facts.

Manipulating the norm-facts held by others is the least researched type of moral wriggle room. Nevertheless, I think it is more than a mere conceptual possibility. For instance, one could make self-serving statements to others about the norms that apply or make misleading suggestions about which norm is more widely accepted or which behaviour would be accepted by the relevant peer group. Such self-serving manipulations of others's beliefs might be particularly relevant when the target person is uncertain about or unfamiliar with the normative context. For instance, waiters or cab drivers might want to suggest overly generous local tipping norms to uninitiated tourists. To my best knowledge, an experiment to that effect has not yet been conducted.

Manipulate information about...	**Target own beliefs**	**Target others' beliefs**
action-facts	Strategic Ignorance	Hiding Behind a Small Cake
norm-facts	Biased Norm Perception	?

Table 1: Experiments and Four Types of Moral Wriggle Room.

Table 1 summarizes the two dimensions and the resulting four types of moral wriggle

room and positions the three types of experiments from the previous section in the appropriate cells.

4 Wriggle Room, Social Moral Epistemology and Racism

We have seen that the manipulation of information about action- and norm-facts, accessible to oneself or to others, is not only conceptually possible, but can be reproduced under the rigorous control conditions of a behavioural lab. But can we use the results in those stylized, artificial environments to learn something about self-serving tendencies in real-world knowledge acquisition? I suggest that we can. In this final section I draw links to the field of social moral epistemology. I also show that the more abstract experimental results are at least suggestive for understanding subtle forms of racism and other forms of discrimination.

Allen Buchanan (2002, p. 126) defines social moral epistemology as the 'study of the social practices and institutions that promote (or impede) the formation, preservation, and transmission of true beliefs so far as true beliefs facilitate right action or reduce the incidence of wrong action.' Moral wriggling is one important mechanism for explaining how social practices can impede the formation of true beliefs, and how the resulting biased beliefs lead to wrong actions. The self-serving biases identified above allow individuals and even whole societies to get away with acts that are objectively impermissible, sometimes without even realizing themselves.

As Miranda Fricker observes, the field of epistemology has a tendency to emphasize ideal standards of knowledge acquisition, losing sight of epistemic injustice even though '[. . .] the only way to reveal what is involved in epistemic justice (indeed, even to see that there is such a thing as epistemic justice) is by looking at the negative space that is epistemic injustice' (Fricker, 2007, p. viii). Understanding the strategies for avoiding normatively relevant knowledge will allow us to grasp more clearly how epistemic injustice can be the result of subtle biases. We will see that the problem of moral wriggling can lead to particularly serious injustice because it enables individuals to hide their failings to themselves or others, leading to entrenched, stealthy practices of injustice that need to be uncovered.

For a case study, consider the many subtle forms of racism. Sometimes we encounter transparently racist attitudes. An overt, conscious racist holds a certain set of beliefs, norms and values, and these are typically clear to himself and others. Such racists are easy to recognize and argue against. But there are also much more subtle forms of racism—racially biased attitudes and actions that are less transparent, rooted in the moral wriggle room discussed above. José Medina refers to such subtle mechanisms when he describes how 'epistemic neglect' can lead to racially biased perception:

> 'Continual epistemic neglect creates blinders that one allows to grow around one's epistemic perspective, constraining and slanting one's vantage point. As we shall see, responsible epistemic agency requires a minimum of diligence, because knowledge requires work and its acquisition will not happen without the active participation of the knower. Becoming lazy is letting oneself go

> epistemically; and it damages the objectivity of one's perspective and limits one's epistemic agency.' (Medina, 2013, p. 33-34)

Medina emphasizes that epistemic neglect is often unconscious, but that does not imply that it is harmless or exculpating:

> 'Actively ignorant subjects are those who can be blamed not just for lacking particular pieces of knowledge, but also for having epistemic attitudes and habits that contribute to create and maintain bodies of ignorance. These subjects are at fault for their complicity (often unconscious and involuntary) with epistemic injustices that support and contribute to situations of oppression.' (Medina, 2013, p. 39)

Charles Mills[7] also emphasizes the possibly unconscious but still pervasive nature of this mechanism when he observes that

> 'racialized causality can give rise to what I am calling white ignorance, straightforwardly for a racist cognizer, but indirectly for a nonracist cognizer who may form mistaken beliefs (e.g., that after the abolition of slavery in the United States, blacks generally had opportunities equal to whites) because of the social suppression of the pertinent knowledge, though without prejudice himself.'(Mills, 2007, p. 21).

Mills's 'white ignorance' is structurally similar to the strategic ignorance detected by Dana, Weber and Kuang. Their experiment demonstrates that many subjects are keen to avoid information that would clarify and potentially increase the normative demands on them, and do so even if that information is free and the potential negative implications of ignorance clear. Since information avoidance works for these subjects in such an artificial setting, it should be even easier in real-life interactions, where we make plenty of choices to avoid information. Whether we read or not read a newspaper article, watch or not watch a TV programme, talk or not talk to an acquaintance—we take all these decisions on a daily basis, and often we have an inkling whether we can expect to learn something that might be 'uncomfortable' news for us. For instance, if we were to put away Medina's 'blinders' we might be confronted with the fact that there weren't equal opportunities for African Americans after the abolition of slavery, which would, in turn, entail obligations to address such a continuing injustice.

As recognized by Medina, strategic ignorance can work in an unconscious way. The more conscious the subject becomes of the implications of ignorance, the more she may become aware of her 'epistemic laziness', and feel compelled to do something about it. However, such obligations of inquiry seem to have a comparatively weak effect on many subjects: if the effect were stronger, all subjects in Dana et al.'s experiment would have clicked the button to reveal information.

A different form of wriggle room and strategic belief manipulation suggested by the experimental literature (especially in Spiekermann and Weiss 2013) features less prominently in the literature on racism, but might also be important: the opportunity to selectively *acquire* the sort of information that justifies one's potentially problematic moral

[7]Thanks to Miranda Fricker for pointing me towards Medina and Mills.

conduct. For instance, in many field experiments researchers found robust evidence that applicants with 'foreign-sounding' names on their written application materials are less likely to be invited to job interviews despite qualifications equal to the control group (e.g., Riach and Rich, 2002, for a review). It is, of course, perfectly possible that the relevant selectors are conscious racists. Much more plausible, however, is the assumption that the racist bias is at least partially unconscious. The selectors might focus on the weaknesses (and overlook the strengths) of a perceived 'non-native' applicant more than they would otherwise, without noticing their own biased perception. They might therefore feel entirely justified in their decisions and not see any racial bias. If this is true, the problem is not rooted in an open racist attitude, but in more subtle biases that might go unnoticed.

So far the focus has been on the upper left cell of Table 1. However, other forms of moral wriggling when it comes to racism are also conceivable. Many people swimming along with the stream in a racist society reduce their cognitive dissonance by convincing themselves that norms of equal treatment apply only in certain contexts, or that they experience strong normative expectations by others that 'force' them to engage in racist practices, even though in reality the sanctions for non-compliance might be minimal or non-existent. Again, the choices here may be both to know and not to know about the relevant norms.

Finally, moral wriggling, as we have seen, might also involve manipulating the information others have. The artificial setting from 'hiding behind a small cake' is probably not directly transferable to the context of race. Nevertheless, it is not too difficult to think of mechanisms that skew the normatively relevant information others hold. For example, Medina reports that the state of Arizona decided to ban 'ethnic studies' in schools. This included removing several texts from the curriculum on the grounds that they promote 'critical race theory'. Medina interprets these interventions as an attempt to remove intellectual resources that would help to uncover racial discrimination. Banning these topics and texts prevents pupils from experiencing the necessary 'friction' that makes them question the pre-dominant 'cognitive laziness' (Medina, 2013, p. 145). For a more subtle example, consider the notion of 'color-blindness' as a supposed ideal of equal treatment. Is it a fanciful assumption that 'color-blindness' might have an epistemic-strategic function—namely, to stop others from questioning entrenched injustice? Being 'color-blind' sounds like a normatively compelling idea on first sight, but it effectively blocks any discussion of existing racial discrimination if 'color-blindness' is taken to rule out *any* reference to race in the public sphere. Such a blanket ban prevents citizens from uncovering normatively relevant facts and norms that would undermine the status quo and lead to more demanding normative-political obligations to tackle discrimination.[8]

As we have seen, 'moral wriggling' can often operate in an unconscious, yet systematic way. This chimes well with a recent turn of attention towards more subtle forms of racial biases. Elizabeth Anderson puts these biases at the centre of her analysis by emphasizing 'the roles of implicit and automatic cognition, which cause discriminatory treatment

[8]See Anderson 2010 for a discussion as to how the principle of color-blindness leads to additional discrimination outcomes under non-ideal circumstances.

even in the absence of discriminatory beliefs or a conscious intention to discriminate' (Anderson, 2010, p. 63). In a similar vein, Medina stresses the systematic and social nature of these biases:

> 'We become active participants in collective bodies of ignorance typically without knowing it and apparently without much conscious effort on our part, but this is because there is a complex set of social structures, procedures, and practices that encourage us to go on with our daily business without taking an interest in certain things' (Medina, 2013, p. 145).

As a final step in our analysis of moral wriggling as a problem of social epistemology, we can also ask which mechanisms and institutions can be used to mitigate this problem. A full analysis of counter-strategies is clearly beyond this short article (but see Anderson 2010 and Medina 2013 for book-lengh treatments of that question with regard to racism). In this paper, a short discussion of some helpful, suggestive experimental results must suffice.

The strategic avoidance or acquisition of relevant information is possible because individuals do not comply with their obligations of inquiry to be informed enough in order to choose the right actions. How far these obligations of inquiry might go is subject to debate, but in the experiments presented and in the case of 'epistemic laziness' with regard to racial discrimination it should be uncontroversial that individuals are under *some* obligation to understand the context they are acting in. The question is how one can prod individuals to meet this demand.

An experiment by Cappelen et al. (2011) suggests that structured normative reflection can be an important tool. Cappelen et al. let their subjects play a dictator game with money earned in a previous investment task. Similar to Konow's (2000) setup, the subjects have different rates of return in the investment task, but the rate is assigned at random to create normative ambiguity. The profit in the investment stage is partly due to individual choices, partly due to luck. In the treatment group, the dictators have to reflect on the fairness of several distributive principles before playing the actual dictator game themselves. Cappelen et al. suggest that relevant fairness considerations could be 'to share equally, to share in proportion to individual investment, and to share in proportion to individual production' (p. 107). Without going into the details of their analysis, their experiment shows that both the average and the mode of giving increases with the reflection treatment, providing some evidence that more selfish choices are harder to make after reflecting on fairness principles. Increasing the salience of fairness principles (in particular, making the subject's own fairness principles salient) can sometimes completely eliminate self-serving biases, as Haisley and Weber (2010) show.

An experiment that is calibrated specifically to test measures against strategic avoidance or uptake of information is, unfortunately, still missing. Nonetheless, we can take a lesson from the experiments mentioned: if we hope for altruistic behaviour based on norms that are not enforced by heavy-handed monitoring and sanctions, then the effectiveness of these norms can be improved by either removing uncertainty about the

situation, or, if that is not possible, by simply reminding individuals of their own principles. In the case of racism, this could be put into practice by making the experience and the effects of racism much more concrete in the public debate. For instance, many whites are aware that blacks tend to experience some disadvantages. What they are likely to underestimate (and perhaps ignore strategically), is the pervasiveness and strength of the discrimination. One could try to counter this with education programmes that make the experience of discrimination concrete in terms of the lived experience, but also in terms of differences in educational and job prospects, wealth, health care, life expectancy, and so on. The 'blinders' Medina talks about could be removed by making the ongoing, experienced discrimination salient and thus impossible to ignore.

Another option to counter epistemic biases is to strengthen the norms regulating our obligations of inquiry. So, rather than making people aware of the specific bads caused by their actions, one could train them to *be more aware of their obligations of inquiry*, in a general sense. To what extent this can succeed is an empirical question. At this point, little is known about possible effects of strengthening the norms of inquiry. *Ceteris paribus*, making specific normative properties of an action salient is probably easier than training people to identify these salient properties themselves—even though the latter would often be the more robust solution. Training individuals to resist self-serving biases would enhance their autonomy, but it remains unclear whether it can succeed.

5 Conclusion

Drawing on the experimental literature in behavioural economics, I have identified four different types of moral wriggle room and linked these types to some recent behavioural experiments. I have shown that the taxonomy and the experimental results reported can be applied to the problem of racial discrimination. Thinking about the effects of 'moral wriggling' in a social context reveals that the effects might be even stronger in real life and thus a graver cause for concern. The often unconscious nature of the self-serving biases investigated challenges us to think about collective counter-strategies to uncover and prevent the harm caused by 'moral wriggling'.

References

Anderson, Elizabeth. 2010. *The Imperative of Integration.* Princeton, NJ: Princeton University Press.

Bénabou, Roland and Jean Tirole. 2011. "Identity, morals, and taboos: Beliefs as assets." *The Quarterly Journal of Economics* 126(2):805–855.

Bicchieri, Christina. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms.* Cambridge et al.: Cambridge University Press.

Bicchieri, Cristina and Alex K. Chavez. 2013. "Norm Manipulation, Norm Evasion: Experimental Evidence." *Economics and Philosophy* 29(2):175–198.

Buchanan, Allen. 2002. "Social moral epistemology." *Social Philosophy and Policy* 19(2):126–152.

Camerer, Colin. 2003. *Behavioral game theory: experiments in strategic interaction.* New York, N.Y: Russell Sage Foundation.

Cappelen, Alexander W, Astri Drange Hole, Erik Ø Sørensen and Bertil Tungodden. 2011. "The importance of moral reflection and self-reported data in a dictator game with production." *Social Choice and Welfare* 36(1):105–120.

Chignell, Andrew. 2013. The Ethics of Belief. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Spring 2013 ed.
URL: *plato.stanford.edu/archives/spr2013/entries/ethics-belief*

Dana, Jason, Roberto A. Weber and Jason Xi Kuang. 2007. "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness." *Economic Theory* 33(1):67–80.

Feldman, R. 2000. "The ethics of belief." *Philosophy and Phenomenological Research* 60(3):667–695.

Fricker, Miranda. 2007. *Epistemic Injustice: Power & the Ethics of Knowing.* Oxford: Oxford University Press.

Güth, Werner and Steffen Huck. 1997. "From ultimatum bargaining to dictatorship – An experimental study of four games varying in veto power." *Metroeconomica* 48(3):262–299.

Güth, Werner, Steffen Huck and Peter Ockenfels. 1996. "Two-Level Ultimatum Bargaining with Incomplete Information: an Experimental Study." *The Economic Journal* 106(436):593–604.

Haisley, Emily C and Roberto A Weber. 2010. "Self-serving interpretations of ambiguity in other-regarding behavior." *Games and Economic Behavior* 68(2):614–625.

Hall, Richard J and Charles R Johnson. 1998. "The epistemic duty to seek more evidence." *American Philosophical Quarterly* 35(2):129–139.

Konow, J. 2000. "Fair shares: Accountability and cognitive dissonance in allocation decisions." *The American Economic Review* 90(4):1072–1091.

Kornblith, Hilary. 1983. "Justified Belief and Epistemically Responsible Action." *The Philosophical Review* 92(1):33–48.

Medina, José. 2013. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations.* Oxford and New York: Oxford University Press.

Mills, Charles. 2007. White ignorance. In *Race and epistemologies of ignorance*, ed. Shannon Sullivan and Nancy Tuana. Albany, NY: State University of New York Press pp. 11–38.

Mitzkewitz, Michael and Rosemarie Nagel. 1993. "Experimental results on ultimatum games with incomplete information." *International Journal of Game Theory* 22(2):171–198.

Ockenfels, Axel and Peter Werner. 2012. "Hiding behind a small cake in a newspaper dictator game." *Journal of Economic Behavior & Organization* 82(1):82–85.

Riach, Peter A and Judith Rich. 2002. "Field Experiments of Discrimination in the Market Place*." *The Economic Journal* 112(483):F480–F518.

Southwood, Nicholas. 2011. "The Moral/Conventional Distinction." *Mind* 120(479):761–802.

Spiekermann, Kai and Arne Weiss. 2013. "Subjective and Objective Compliance: How Moral Wriggle Room Opens.". working paper.
URL: *http://personal.lse.ac.uk/spiekerk/papers/Spiekermann_Weiss_2013-10-17_web.pdf*

Steup, Matthias, ed. 2001. *Knowledge, truth, and duty: Essays on epistemic justification, responsibility, and virtue.* New York: Oxford University Press.

Van der Weele, Joel J. 2012. "When Ignorance Is Innocence: On Information Avoidance in Moral Dilemmas." *SSRN* .
URL: *http://dx.doi.org/10.2139/ssrn.1844702*