Federico Picinali

Base-rates of negative traits: instructions for use in criminal trials

Article (Accepted version) (Refereed)

Base-rates of Negative Traits: Instructions for Use in Criminal Trials

Federico Picinali[*]

ABSTRACT: Decision-makers in institutional and non-institutional contexts are sometimes confronted with the issue of whether to use generalisations expressing the statistical incidence of a negative trait in a disadvantaged and discriminated-against social group in order to draw an inference concerning a member of that group. If a criminal court were confronted with such a question, what answer should it give? First, the paper argues that, our qualms notwithstanding, morality does not demand that these generalisations be disregarded. In doing so, the paper addresses the relationship between factual accuracy and the demands of morality in criminal trials. Second, the paper considers the implications of this conclusion for the legal question as to whether the evidence at issue should be excluded, in particular, on grounds of unfairness – pursuant to section 78(1) of the Police and Criminal Evidence Act 1984.

1. Introduction

In a possible world the Xs are a disadvantaged and discriminated-against ethnic minority. They mostly live in the suburbs of large cities, have humble jobs, and earn modest salaries; they rarely succeed in climbing the social ladder, mainly due to the unrelenting prejudices that the more advantaged citizens of different ethnic origins harbour towards them. Against this backdrop, we are not surprised to discover that the Xs engage in misconduct at a rate disproportionately higher than their fellow citizens. A rather grim picture, although a familiar one.[1]

Jack is an X man. He is facing a criminal trial with the charge of inflicting grievous bodily harm on his daughter Merena. No one witnessed the alleged event and Merena is in a coma, therefore unable to testify. However, a neighbour reports that a few hours before the event, Merena and Jack had a heated discussion: during the confrontation Merena admitted to dating a man of another ethnic group and, in response to Jack's anger at this news, she vehemently attacked the Xs' traditional patriarchal values. A series of reliable sociological studies shows that the statistical incidence of violent chastisement is substantially higher among X fathers than among non-X fathers, and that it reaches its peak when X children disparage Xs'

traditional values.[2] The prosecution introduces these base-rates[3] at trial with the aim of proving Jack's propensity to harm his daughter. More precisely, the prosecution advances the following argument: all else being equal, the fact that Jack is an X man makes it more likely that he has injured his daughter than if Jack were not an X man; therefore, Jack's membership in the class of X men is relevant information and it is probative of Jack's criminal conduct. Call the base-rates, together with the argument built upon them, the 'membership evidence'.[4] Imagine that you were the authority in charge of adjudicating the case: *should you rely on the membership evidence at all?* Importantly, the question is not whether the membership evidence could be *sufficient* to conclude in a court of law that Jack is guilty; rather, it is whether the membership evidence should be given any role whatsoever in the inference, that is, whether it should be *admitted* as evidence in court.[5] I will come back to the distinction between the question of admissibility and the question of sufficiency in due course.

Many of us have intuitions about the question in italics. In particular, many think that there is something morally problematic with using arguments such as that advanced in Jack's case. This uneasiness is experienced notwithstanding that the argument is known to be accurate. In the paper I try to unpack and test these intuitions. In section 2 I clarify the contours of the problem. In sections 3-7 I consider the problem from a moral standpoint, through elaborating and confuting arguments in favour of disregarding the evidence at issue. These arguments are predicated on ideas drawn, in particular, from literature on implicit bias, single-case inference, profiling, and generics. The upshot of this moral investigation is that there is no indefeasible moral reason against using the membership evidence in Jack's case. In order to reach this conclusion, however, some provisos will need to be introduced along the way. In section 8 I consider the implications of the moral investigation for the legal question as to whether the evidence at issue should be excluded under the relevant provisions of English law, in particular, pursuant to section 78(1) of the Police and Criminal Evidence Act 1984 (PACE) – stating that "… the court may refuse to allow evidence … if it appears to the court that … the admission of the

evidence would have such an adverse effect on the fairness of the proceedings that the court ought not to admit it." More precisely, I address the legal questions concerning whether admitting the membership evidence would render the trial unfair and, in the event of a negative answer, whether there is any other ground on which to exclude such evidence.[6] Finally, section 9 offers some concluding remarks.

2. Factual accuracy and the demands of morality

The issue posed by our hypothetical case may be ascribed to the broader topic of the relationship between factual accuracy and the demands of morality in legal adjudication. Inquiring into this relationship is relevant to the legal decision as to whether to admit items of evidence in court. It is often the case that factual accuracy is instrumental to morality: its being grounded in accurate evidence and reasoning is generally among the conditions for the moral appropriateness of a legal judgment. However, there may be instances where morality requires that the adjudicator act as if she didn't have some accurate evidence that she does have. The typical case is that of a defendant who confessed to committing a crime as a result of being tortured in a local police station or – to make the example more actual[7] – in a foreign country by foreign secret services. Many would argue that morality demands that the state does not rely on the extorted confession in order to prove the defendant's guilt, irrespective of how credible it may be. This argument may rest on the claim that using the confession would be unfair towards the defendant and would 'contaminate'[8] the decision; but it may also rest on the claim that excluding the evidence is necessary to deter misconduct on the part of public officials.[9] Arguments to the effect that morality demands the exclusion of the extorted confession are open to criticism.[10] However, they are widely shared, they feature prominently in the national and supranational case law,[11] and – at least some of them – have intuitive appeal on their side. This said, I do not intend to discuss the torture case any further. Instead, the question explored in the paper is whether in cases like Jack's morality asks that criminal adjudication dispense with accurate information, that is, the

membership evidence. Some scholars have answered the question in the affirmative.[12]

It is important to stress again that the question addressed here does not concern the sufficiency of the evidence, but its admissibility. The issue is not whether the membership evidence could be sufficient for a finding that the defendant engaged in criminal conduct; it is, instead, whether this evidence should be relied upon at all by the fact finder, whatever the weight she may give to it. The question of sufficiency of the membership evidence has been discussed in the literature.[13] This question presents some interesting theoretical aspects, but it has little practical relevance. First, it is very unlikely that someone could be brought to trial – indeed, that a case could make it to the trial! – on the basis of the membership evidence alone.[14] Nor are there reasons to expect that this will happen in the future. Second, the question of sufficiency depends on the applicable standard of proof. Even if we were to conclude that the membership evidence is not morally problematic and can, therefore, be admitted at trial, it would be very unlikely that this evidence alone would ever be strong enough to satisfy the criminal standard of proof – i.e., proof beyond a reasonable doubt.[15] Of course, we may conjure up contrived examples to disprove this conclusion.[16] However, all that these examples would do is to prove the conclusion wrong in a world that is very much unlike ours.

Before addressing the question of admissibility from the standpoint of morality, it is necessary that we highlight the main features of Jack's scenario. The considerations that follow in the next sections hinge upon one or more of these features. The first feature is the nature of the base-rates produced by the prosecution:[17] generalisations concerning the incidence of a negative trait in members of a disadvantaged[18] and discriminated-against social group.[19] A negative trait is a morally objectionable behaviour or disposition: in Jack's scenario it is the fact of engaging in violent chastisement. Importantly, the accuracy of the generalisations – and of the underlying research and reasoning – is not in doubt. The term 'accuracy' in the previous sentence refers to the generalisation's reliability, as

indicated by factors such as sample size, randomisation and experimental design. Accuracy is, therefore, unrelated to the statistical incidence expressed in the generalisation itself:[20] a universal generalisation may well be inaccurate.[21] The second feature is the epistemic task for which the base-rates are employed. The task is to show that someone's membership in a group makes her more likely to engage in particular behaviour. I have used the expression 'membership evidence' to refer to the base-rates together with the argument that is built upon them in order to fulfil this task. The third and the fourth features are typical of – but by no means exclusive to – the trial's institutional setting. The third feature is the publicity of the judgment: a considerable amount of people may have access to it through direct experience or through the media.[22] The fourth feature is the authority of the judgment: at least part of these people can be reasonably expected to defer to the decision and/or to its underlying reasoning.

As evident from the above, my primary focus is on criminal trials. Not only has the present question been the object of academic attention with reference to this specific context;[23] it is also reasonable to expect that social stratification, coupled with an increasing interest in empirical studies on the behaviour of social groups, will render the question more and more compelling in the forensic domain. Notwithstanding this main focus, the considerations that follow in sections 3-7 may apply to other instances of decision-making that present the four properties described in the previous paragraph, as well as to instances that have only some such features. In particular, I have in mind deliberations in both institutional and non-institutional contexts, such as deliberations by parliamentary commissions of inquiry, by disciplinary committees in private or public institutions, by historians and, most notably, by journalists.

3. The argument from the immorality of beliefs

Someone may argue that the membership evidence should be ignored because it is immoral to *believe on the basis of this evidence that Jack is more likely to have committed the*

crime than he would be were this evidence not available. At first glance the argument seems untenable. Since believing is something that happens to us, rather than something we choose to do, it appears that we cannot be held morally responsible for our believing that such and such is the case.[24] However, it would be wrong to claim that we have no control whatsoever over our beliefs. While we cannot help (not) believing an hypothesis given the evidence of which we are aware, it is up to us to decide whether to embark on the investigation in the first place, whether to follow one or another of the possible investigative avenues open to us, and whether to give up the investigation at some stage, or to continue looking for further evidence that may update our beliefs. All of these decisions will impact, albeit indirectly, on what we believe. Thus, we do have a form of indirect control over our beliefs and, in light of this, the claim that we may be held responsible for them is defensible.

I do not intend to enter into the debates on indirect control of beliefs and on doxastic responsibility. However, I am sceptical of their relevance to our question. The scholars who have argued for a notion of responsibility that may extend to our cognitive attitudes have done so with a particular type of attitudes in mind, to wit, false – or, at the least, misleading – representations of states of affairs.[25] Possible examples thereof are those entrenched dispositional attitudes concerning racial, ethnic, religious, etc. groups, referred to in the literature as implicit biases.[26] It may be reprehensible to hold a cognitive attitude of this sort precisely because something of epistemic relevance has gone awry in the enquiry that led to the attitude itself and because the attitude holder was somehow responsible for this failure.

However, by hypothesis no similar epistemic failure characterises the membership evidence and the belief that this evidence may prompt, described above. The generalisation produced by the sociological study is accurate and it is arrived at in compliance with statistical rules of inference. Moreover, the argument to the effect that Jack's membership in the class of X fathers makes him more likely to have committed the crime is hardly objectionable on epistemic grounds: it is a straightforward application of the logic of single-case inference.[27] Every inference

regarding individual behaviour requires looking beyond the information we possess on the individual case[28] – e.g., the fact of being an X father. In fact, this information is *silent* until we use it to ascribe the individual case to a suitable reference class – e.g., the class of X fathers. From observing the instances falling into this class we learn their behaviour. These statistical data are crucial to the single-case inference, as they allow us to correlate the evidence concerning the individual and the particular behaviour whose occurrence is at issue – e.g., whether an X father engaged in violent chastisement. In other words, they tell us whether this evidence has any probative value with respect to the fact in question.

Now, if the accuracy of the evidence and of the belief that this evidence may prompt is not in question, our case is very much unlike that of implicit biases: no room seems left to argue that it is immoral to hold the belief at issue. In fact, would it be reasonable to claim that it may be immoral for someone to hold an accurate belief and, moreover, to hold it in the absence of any epistemic failure in the reasoning that led her to such belief? I believe it wouldn't.

The discussion so far does not rule out the argument that it is immoral to use the membership evidence in order to ascribe moral responsibility to an individual. The target of this argument is not the cognitive attitude, which the evidence may produce, on whether Jack engaged in certain conduct; it is, instead, the use of such evidence to justify a moral assessment of individual conduct.[29] The arguments discussed in the next two sections are of this kind.

4. The argument from autonomy

It may be argued that using information pertaining to a group of people in order to ascribe moral responsibility for an act to an individual member of the group is immoral because it involves a lack of respect for the individual's autonomy. The literature abounds in claims that seem to advance this argument.[30] I will focus on two claims that have attracted particular scholarly attention. Taken at face value, these claims advance only an impoverished version of the argument from autonomy.

I will consider this version and confute it. Later, I will suggest a more charitable interpretation of the two claims, which allows for an improved version of the argument. This version too will be considered and, finally, rejected.

David Wasserman writes: "…what is objectionable is the reliance on others' conduct … to infer [the defendant's] commission of a wrongful act. We object to this inference because it ignores the defendant's capacity to diverge from his associates … thereby demeaning his individuality and autonomy".[31] In a similar vein Amit Pundik writes: "Inferences based on statistical evidence demean the individual's autonomy because they regard her as a predetermined mechanism whose behaviour could be learnt by observing … the behaviour of other similar mechanisms".[32] Under a literal interpretation, these passages target the inference from group to individual. They say that it is morally wrong to use information on the behaviour of a group of people in order to draw an inference concerning the behaviour of an individual member of that group. This is because *the inference* demeans the autonomy of the individual. These claims are unconvincing, to say the least.

In the previous section I pointed out that following the logic of single-case inferences is epistemically sound – and, I would add, necessary.[33] On this basis, I argued that there is no ground to deem immoral the belief in a higher probability of commission of a certain act, which the inference may prompt. The reference to autonomy does not weaken this conclusion. There is no reasonable conception of autonomy under which the mere fact of drawing a sound inference concerning individual behaviour or of holding an accurate belief as a result of such inference would demean autonomy.[34] Thus, under a literal interpretation of it, Wasserman's and Pundik's argument is not persuasive. And yet, at a closer look their words – in particular, Wasserman's reference to the moral nature of the act to be proven – would seem to suggest that the authors are not targeting the inference *per se*, but the particular use to which the inference is put, namely, ascribing moral responsibility.[35] In light of this, it seems possible to advance a more defensible interpretation of their claims. According to this interpretation, Wasserman and Pundik are not arguing that

individual autonomy is violated by the inference or by the belief *per se*; they are arguing, instead, that individual autonomy is violated by the use of the inference in order to ascribe moral responsibility to an individual for a particular behaviour. I will now try to flesh out this alternative version of the argument from autonomy.

As a starting point, I must clarify the conception of autonomy that the revised argument rests upon. Autonomy is understood as "a relationship between an agent and her motivational states".[36] It is the agent's ability to make decisions on the basis of her beliefs and desires. According to this conception, the autonomous agent exercises authority over her decision-making: she decides which motivational states to follow and which to disregard. It is because of the exercise of authority in decision-making that we can consider the agent's ensuing actions as *hers*, as an expression of her autonomy.

The motivational states of an individual are often used as evidence to prove that the individual acted in a particular way. A typical example of this is the use of motive in criminal trials. Interestingly, if we use the membership evidence in order to draw an inference about the behaviour of the individual we completely bypass the consideration of the motivational states. The membership evidence allows us to 'jump' to the conclusion that there is a certain probability that someone acted in a particular way, without inquiring into the reasons that she may have had to act thus.[37] At best, our inquiry into the motivational states may stop at the consideration that some, most or all the members of the group may share similar motivational states, whatever these may be.[38] By showing that there is a certain degree of uniformity in the behaviour of the group members, the membership evidence gives some support to this claim. As a result, we may conclude that if the individual has acted in a particular way, she may have done so because of reasons that are similar to those that motivated the members of the group who acted similarly. It goes without saying that this investigation into the individual's motivational states is superficial, to say the least.

We can understand why the use of membership evidence may appear problematic within a process that ultimately aims at ascribing moral responsibility to an individual. Ascriptions of moral responsibility cannot leave aside the consideration of the motivational states of the individual. This holds true for the central cases of criminal responsibility, e.g., responsibility for murder, rape, theft, grievous bodily harm – the crime at issue in Jack's case. Inquiring into the reasons that motivated someone to behave in a particular way is crucial to determining her responsibility. While it may not be necessary that we know the motivational states that actually played a role in the individual's decision-making, it is necessary that we exclude that the agent acted because of certain motivational states. For instance, before blaming and convicting Jack for harming Merena we should exclude the possibility that he acted in self-defence or under duress. On the other hand, it may be irrelevant to his responsibility whether he harmed Merena because he disapproved of the fact that she was seeing a non-X man or because he disapproved of Merena's dress. However, more fundamental than inquiring into the content of the agent's motivational states is inquiring into whether the agent behaved as she did because she was motivated to do so by her own states. After all, Jack may have harmed his daughter as a result of the directives given by one Dr Caligari,[39] after being put by the doctor into a hypnotic state. Until we know that the action was prompted by the agent's motivational states we cannot consider it as the expression of the agent's autonomy, that is, as *her* action – and, possibly, as an action at all. This clarifies why Wasserman writes that using the membership evidence amounts to ignoring the agent's capacity to diverge from what others do, and why Pundik resorts to the metaphor of the predetermined mechanism. Even if the membership evidence does permit us to draw inferences about the probability that the individual has acted in a particular way, this evidence does not give us the kind of information that we need in order to determine whether an action is attributable to the individual's autonomous decision.[40] It is on these grounds that the evidence should be excluded from a process that aims at ascribing responsibility: if we use it in order

to ascribe responsibility to an individual, we ignore and disrespect the individual's autonomy.

Even in its revised version, the argument from autonomy fails. The problem with this argument is that it confuses the 'narrow' aim with which the membership evidence is introduced by the prosecution – i.e., proving that Jack injured Merena – with the 'broader' aim of adjudication – i.e., determining whether Jack is to be held responsible for injuring Merena. It is true that in order to ascribe moral and criminal responsibility to Jack we must investigate into the content of his motivational states and, crucially, into the relationship between his behaviour and such states. However, this does not deny that we can find out how Jack behaved even without knowing anything about his motivational states. After all, eyewitnesses do not testify about the motivational states of the agent,[41] but we still draw inferences about the agent's behaviour based on their testimonies. When the membership evidence is used with the aim of proving a particular behaviour on the part of an individual, any concern that the individual's autonomy may be demeaned by such use seems, therefore, unwarranted. The situation is radically different when it comes to ascribing moral or criminal responsibility to the individual because of how she behaved. For the ascription of responsibility it is necessary that we ascertain – possibly by using membership evidence – that the individual behaved in a particular way. But this is not yet sufficient: we must inquire into the agent's motivational states and the relationship between such states and her behaviour, as described above. Of course, this inquiry will still be part of fact finding. The prosecution and the defence will have to produce evidence on whether the defendant's act was voluntary (i.e., the product of her motivational states), whether she acted in self-defence, under duress, in a state of insanity etc.[42] And, certainly, the membership evidence that may have been used to prove the defendant's behaviour will not be of any help here.

To conclude, proving conduct through inferences that sidestep the consideration of the motivational states and of their link with conduct does not amount to denying that inquiring into these facts is fundamental for moral and criminal responsibility.

This is because the immediate aim with which these inferences are employed – i.e., proving conduct – is but a step towards the achievement of the ultimate aim of ascribing responsibility.

The argument from autonomy fails also in its revised version. It is not obvious how the argument could be unpacked in any more convincing fashion.

5. The argument from the origin of the trait

Certain negative traits characterising a disadvantaged and discriminated-against group may be attributed to the group's condition of discrimination and deprivation. *Ex hypothesi*, the Xs are no exception to this. In fact, sociological research shows that X fathers' violent manners in the enforcement of the Xs' traditional values are a result of the mockery and contempt that these values are subjected to by members of more advantaged ethnic groups. Based on these findings it may be argued that we should not rely on Jack's membership in the class of X fathers in order to determine whether he is responsible for harming his daughter.[43] The argument would consist of two claims: (1) X fathers who engage in violent chastisement should be excused,[44] because their misconduct is ultimately determined by unjust social conditions affecting them, which they have not chosen, nor created themselves; (2), if (1), then in Jack's case the membership evidence should be disregarded because, to the extent that this evidence proves that Jack has harmed his daughter, it implies that he is excused – and, therefore, should not be held responsible – for doing so. Both claims are unconvincing for the reasons stated below.

As far as (1) is concerned, it is open to question whether the fact – which we are assuming *ex hypothesi* – that X fathers' involuntary and unjust social condition explains their misconduct is sufficient to provide them with an excuse. If anything, their excuse may be partial. Thus, to hold X fathers responsible may still be an appropriate reaction to their behaviour. To be sure, these evaluations will depend upon the nature of the social conditions and of the negative trait at issue, and would

be very hard to make with respect to the group as a whole. Importantly, even if (1) were correct, (2) would not follow. As we saw in the previous section, the membership evidence does not give us adequate information – if any – on the motivational states of the agent. In fact, the evidence is introduced to prove Jack's conduct. In order to ascertain whether Jack harmed Merena we don't need to know his motivational states. Nor do we need to consider the social or personal conditions that produced those states. Because of its failure to give us information on the motivational states of Jack, the membership evidence cannot possibly support the claim that Jack should be excused for acting as he did. Other evidence is needed in order to reach this conclusion. Thus, at a first glance (2) seems untenable. However, the proponent of the argument from the origins of the trait would point out that we do have other evidence: it is the evidence provided by the sociological research mentioned at the start of this section. By virtue of Jack's membership in the group of X fathers, the sociological evidence gives some information concerning his motivational states. On the basis of the sociological evidence, the proponent would then press the point that to the extent that we rely on the membership evidence to prove Jack's conduct, we are bound to recognise that Jack is excused. More precisely, to the extent that we use the argument that Jack's membership in the group of X fathers makes him more likely to have harmed Merena, we must accept the argument that Jack engaged in this conduct because of the involuntary and unjust social conditions that affect the Xs – call this claim (2.1). If so, we should not use the membership evidence at all, because of an alleged general principle according to which there is no good reason to invest resources into proving misconduct once we already know that the agent would be excused for it – call this claim (2.2). Not only would this be uneconomical, but it would also expose the defendant to an unnecessary painful ordeal. Jack's conduct must be proven with evidence of a different kind.

(2.1) and (2.2) are nothing but the result of breaking down and fleshing out (2). I do not intend to take issue with (2.1). I will focus, instead, on (2.2). The main

problem with this claim is that it misconceives the nature and implications of excuses. The presence of an excuse does not authorise the adjudicator to turn a blind eye to the very behaviour for which someone is excused and to the consequences thereof. Therefore, the adjudicator should not disregard evidence on the ground that this evidence, by proving misconduct, would also serve as a basis for excusing the agent.[45] Even if we knew in advance that X fathers, and Jack among them, would be excused for their misconduct, had they in fact misbehaved, this would not be a reason to abandon the investigation into whether misconduct indeed took place. Proving that misconduct occurred is important independently of the existence of an excuse. First, if there has been misconduct, the victims thereof would be no less harmed and/or wronged because of the presence of the excuse:[46] ascertaining misconduct is necessary to recognising their status as victims. Moreover, establishing the dynamics – including misconduct and the reasons thereof – that led to victimisation is a necessary step towards addressing the issue of victimisation at its root. Proving that Jack, as well as other X fathers, engage in violent chastisement, and that they do so as a result of the conditions of oppression they suffer, is the premise for taking action towards the prevention of further misconduct – where prevention, of course, involves removing those social conditions in the first place.

6. The argument from the ratchet effect

The arguments considered and rejected so far aim at proving that there is something wrong in using the membership evidence when trying Jack, irrespective of the consequences of doing so. Other arguments focus, instead, on consequences: the argument from the ratchet effect is one of them. This argument rests on the attributes of publicity and of authoritativeness characterising the judgment involved in Jack's case.

Where a public and authoritative judgment relies on generalisations that report a higher rate of misconduct on the part of a group, this encourages law-enforcement agencies to adopt these generalisations in order to profile such group, i.e., to

dedicate wider resources to monitoring those individuals shown by the generalisation to be more likely to engage in misconduct.[47] As explained by Harcourt,[48] a notable problem[49] with profiling is that it creates a gap between two proportions: the proportion of offenders of the targeted group within the overall population of offenders, and the proportion of *convicted* offenders of the targeted group within the overall population of convicted offenders. As a result of profiling, the second proportion is higher: the members of the group are over-represented within the population of convicted offenders. Moreover, if at a later stage law-enforcement agencies use the second proportion to shape new profiling practices, the gap is further widened. This is the ratchet effect. Needless to say, this effect evidences an unequal treatment of similarly positioned individuals: whereas *ceteris paribus* all offenders should be treated equally, an offender's membership in a group makes it more or less likely that she will be prosecuted and convicted. Now, given that profiling produces this deleterious effect – which, as Harcourt shows, has negative repercussions on the social conditions of the profiled group – we should avoid practices that encourage it. Using the membership evidence in a public and authoritative judgment is one such practice. This evidence – the argument concludes – should be disregarded.

The argument from the ratchet effect points us in a new interesting direction. In contrast with the previous arguments, it leads us to shift our consideration to the consequences of using the membership evidence in Jack's case and, then, to ask whether these consequences may render this usage immoral. Nevertheless, the argument from the ratchet effect is weak. As explained above, the argument claims that using membership evidence is wrong because it is conducive to profiling and because profiling produces the despised ratchet effect. The weak link lies in positing that a public and authoritative reliance on membership evidence in judgments of the sort discussed here will result in the respective base-rates being used for profiling. Such a result cannot be ruled out, of course; yet it is not a necessary consequence. Law-enforcement practices are generally subject to policymaking and supervision;

they are not – and/or should not be – at the discretion of the individual officer. This means that if policymakers and supervisors acknowledge the ratchet effect of profiling and discourage and reprimand this practice, the risk that individual officers will profile a certain group out of deference to the judgment may be reduced within acceptable limits. In conclusion, while the argument from the ratchet effect may provide strong evidence against profiling, it does not extend its purview to the use of membership evidence in cases similar to Jack's, insofar as the appropriate measures are in place.[50]

7. The argument from the biased audience

A more promising argument focussing on consequences is the argument from the biased audience. As mentioned in section 2, the attributes of publicity and of authoritativeness signify, respectively, that a considerable amount of people may have access to the judgment, and that some of them may defer to it and to the underlying reasoning. I shall refer to this part as 'the audience'. Considering the social conditions of the Xs as depicted in the introduction, it is reasonable to expect that a segment of the audience – in particular, the non-X audience – is biased against the Xs. The biased audience generally attributes to some or all the Xs negative traits that none of them possesses or it exaggerates the statistical incidence within the X population of negative traits that some Xs do possess. A public and authoritative judgment reporting a negative trait of the Xs creates the conditions for the bias to operate.

As implausible as it seems, imagine that the base-rate in Jack's case expressed a probability of 1, therefore being universal: *all* X fathers resort to violent chastisement. In other words, given that someone is an X father he will *certainly* engage in this practice. If the adjudicator decides to use this generalisation as evidence of Jack's misconduct, there is no risk that the biased audience will misread the generalisation by exaggerating the statistical incidence of the negative trait attributed to X fathers. In fact, by hypothesis, all X fathers engage in violent chastisement: no room is left for

a distortion of reality through overestimating the incidence of the trait. Now consider the more realistic case in which the base-rate expresses a probability that is lower than 1. Depending on how much lower than 1 the probability is, there will be more or less opportunity for the biased audience to form a distorted representation of reality by attributing a stronger probative value to the generalisation. For instance, generalisations to the effect that "55 per cent" of X fathers resort to violent chastisement may be distorted, and may be understood as if they expressed a considerably higher probability. This is even more likely to occur if the generalisation is conveyed using quantifiers rather than numbers: the former are vaguer and leave greater room for (mis)interpretation. Importantly, the claim is not that if a member of the biased audience were asked to state the number or quantifier expressed in the generalisation used in the judgment she would be likely to mis-state it by erring on the side of excess. This is possible, of course, and may be evidence of a particularly strong bias. Rather, the claim is that the biased audience would probably form an inaccurate understanding as to what that very number or quantifier signifies. The mechanism is similar to that of the 'bias of imaginability' described by Tversky & Kahneman.[51] Even in the lack of the experience of a particular event, the ease with which this event is imagined may determine the belief in the probability that the event indeed occurs. For an individual biased against the Xs, it will be relatively easy to imagine that they present a negative trait. Once a judgment has explicitly associated the Xs with a negative trait, the biased individual may believe that the trait is more probable within the population of the Xs than the judgment actually suggests.[52] In fact, this mechanism may also be ascribed to the category of confirmation biases, being an instance in which the individual affected by the bias selectively reinterprets evidence, so as to bring it into accord with, and to confirm, the bias itself.

There is an additional problem. Not only may the biased audience magnify the statistical incidence of the trait; especially when the generalisation used in the judgment already expresses a high incidence, there is a risk that the biased audience

may slide into the language and logic of 'generics'. Generics are generalisations of a particular kind: they omit quantifiers, let alone numerical probabilities. As a result, they are liable to be understood as claims about the essential – or natural – traits of a category,[53] that is, traits that *define* the members of that category. The generic 'the Xs have trait *f*' may convey that having *f* is rooted in what it is to be an X. Particularly when it comes to negative traits, it is easy to appreciate the distorting effects of this discourse.

Why should we worry about this at all? The answer is simple: we don't want the Xs to be further burdened by the spreading of false negative stereotypes. If our conduct increases the probability of this happening, morality provides reasons for behaving differently. Now, is there a way to use the membership evidence and, at the same time, to prevent the biased audience from misunderstanding the statistical incidence of the negative trait and the nature of such trait? There may be a way to do so: providing lay people with a clear explanation of the significance of the base-rate, also mentioning in peremptory terms the interpretations and uses thereof that are not epistemically warranted.[54] The expectation is that conveying this information would make it harder for the bias to operate, if not neutralise its effect altogether. By this I do not mean to say that the bias would be eradicated: this aim cannot possibly be achieved by the 'explanatory tactic' suggested here, nor is it necessary to achieve it in order to offset the argument from the biased audience. The debiasing effect of this tactic, instead, would be limited to preventing a pejorative misinterpretation of the base-rate in the particular case.[55] Notably, a bias may be explicit or implicit, depending on whether or not a person consciously relies on it when forming beliefs and choosing courses of action.[56] Whatever the form of the bias, the tactic provides compelling information that – absent a *mala fide* endorsement of inaccuracies – would seem to leave no room for misunderstanding the particular base-rate to the detriment of the disadvantaged and discriminated-against social group. Bias-induced false beliefs on the significance and the implications of the base-rate may, thus, be averted. To the extent that this can be achieved, the argument from the

biased audience loses its bite. True, critics of the adequacy of the explanatory tactic as a response to the argument from the biased audience may point out that the tactic should be subjected to appropriate testing. However, the same need for testing applies to the argument from the biased audience in the first place![57]

8. An unfair trial for Jack?

As previously pointed out the moral arguments against using the membership evidence in Jack's case may be grouped into two classes. On the one hand, there are arguments that insist on the immorality of using this evidence irrespective of the consequences of doing so. On the other, there are arguments to the effect that the immorality derives from the consequences of such use. The former are unsuccessful. The arguments from consequences are more compelling. However, these arguments may be offset through implementing measures to prevent the use of the evidence from leading to undesirable ramifications. Is it possible to formulate other reasonable and more successful arguments against using the membership evidence? I have not (yet) managed to do so, but other scholars are warmly invited to try.

Provided that there are cases where morality demands that we disregard what is true and relevant to decision-making, it is doubtful that Jack's is one of them. What are the implications of this conclusion for the question of whether in a criminal trial the membership evidence at issue should be excluded pursuant to section 78(1) PACE 1984? As you will recall, this section states that "… the court may refuse to allow evidence … if it appears to the court that … the admission of the evidence would have such an adverse effect on the fairness of the proceedings that the court ought not to admit it." Is it the case that admitting the membership evidence would render the trial unfair? In this section I shall answer this question in the negative. However, I shall also show that this is not the end of the matter: there is still an argument, unrelated to the issue of trial fairness, that suggests excluding the evidence. This argument will be assessed and, ultimately, rejected.

The notion of trial fairness that is relevant to the application of s78(1) is a 'defendant-centred' notion.[58] With this expression I mean to say that whether the trial is rendered unfair by the decision to admit the evidence depends on whether that decision is found to have wronged the defendant and to have affected negatively her situation at trial. This is not to claim that, while assessing the fairness of the trial, the European Court of Human Rights and the English courts do not consider factors other than the wrong suffered by the defendant. Whether rightfully or not, these courts do consider other factors – for instance, the seriousness of the crime charged and the related factor consisting in the public interest in achieving a conviction.[59] My point is merely that in order for a claim of unfairness to succeed, it must be shown that the defendant was wronged by the decision to admit the evidence and that her case was made worse off by such decision.

We saw that of the arguments explored above only the two arguments from consequences seem to carry some weight – for the time being I leave aside the consideration of their respective countermeasures. I shall, therefore, concentrate exclusively on these two arguments.[60] Notably, neither of them is defendant-centred: they aim to show that admitting the evidence would (or may) amount to wronging the entire social group to which the defendant belongs. According to these arguments, what is (or may be) wrongly affected by using the evidence is not the defendant's situation at trial, but the group's social condition.[61] If this is correct, it follows that, irrespective of the weight that is given to such arguments, a court could not exclude the membership evidence on grounds of unfairness following either of them. It would seem, therefore, that section 78(1) PACE 1984 and the related notion of trial fairness do not provide the appropriate conceptual framework through which to funnel the only promising arguments against admission. This conclusion, however, would be too hasty. On closer examination, we find that at least one of the arguments from consequences does raise important concerns about using the membership evidence and may possibly raise the issue of trial fairness as well. This is due to the particular features of the criminal trial, which cast doubt on whether the

countermeasures that would offset the argument are sufficient and, indeed, practicable.

The argument from the ratchet effect is straightforwardly handled. It is, after all, an argument against profiling rather than against relying on the membership evidence in trial fact finding. As was shown, to the extent that measures can be implemented in order to prevent profiling practices on the part of law-enforcement authorities, the argument loses all its strength. We have no reason to assume that these measures cannot be implemented.

Offsetting the argument from the biased audience in the trial context is, however, more complicated. In the previous section I argued that an effective countermeasure against the operation of biases harboured by the audience would be to explain the significance of the generalisation at issue, clearly mentioning the interpretations and uses thereof that are not epistemically warranted. The problem is that criminal trials present certain features that may render this explanatory tactic insufficient and impracticable.[62] First, when fact finding is the task of a jury of lay people, it is reasonable to expect that fact finders themselves may harbour biases against the disadvantaged and discriminated-against social group to which the defendant belongs. In light of this it is possible to formulate a defendant-centred version of the argument from the biased audience – which we may call, in the absence of a more inventive name, 'the argument from the biased jury' – susceptible therefore to raising the question of trial fairness. If jurors may be biased, admitting evidence likely to trigger such bias may have the effect of distorting fact finding to the detriment of the defendant: jurors may interpret the evidence as being stronger than it really is. Therefore, such evidence should be excluded pursuant to section 78(1) PACE 1984. However, the criminal trial already has devices rendering the exclusion of the membership evidence superfluous as a safeguard against the unfair outcome envisaged in this argument, and thus, all things considered, a negative measure. It is, in fact, possible to implement the explanatory tactic – targeted, this time, on jurors – through using jury directions and judicial comment, and to enforce this tactic

through appellate review. By hindering the operation of the bias as argued in the previous section, these devices seem to provide sufficient ground on which to rebut the defendant-centred argument from the biased jury.[63]

Second, notoriously, juries do not give reasons for their verdict; also, the jury room is a 'black box', the inviolability of which is protected through criminal sanctions. This seems to cast doubts on the adequacy of the explanatory tactic as a countermeasure to the argument from the biased audience. The audience has access to information concerning the evidence admitted at trial and the verdict. Instead, the reasoning underlying the verdict, which would be the ideal source of information on the epistemic significance of the membership evidence and its epistemically impermissible uses, is beyond reach. As a result, it would seem that the audience can receive no guidance concerning these important matters. Nothing precludes, though, that jury directions could perform as explanatory a function with respect to the audience as they do with respect to the jury. To the extent that the public is exposed to the trial and the evidence presented therein, it is also exposed to jury directions, as these are publicly given in a (crucial) trial phase. True, jury directions are by definition instituted and devised for juries. However, if the communication between the court and the jury can produce a valuable epiphenomenon in terms of sending a message to members of the public, it is reasonable to take this into account and acknowledge its worth when addressing the problem of public misunderstanding. Even in the case of audiences of jury trials it seems, therefore, that the explanatory countermeasure can be implemented: the reasons for exclusion are weakened, if not defeated.

9. Conclusion

The upshot of this paper is that there is no indefeasible moral reason to disregard base-rate of a negative trait in a disadvantaged and discriminated-against social group, when drawing an inference concerning a member of that group in a criminal trial. In light of the results of our moral investigation, I have made a *prima facie* case

for the claim that section 78(1) PACE 1984 does not provide the right framework for deciding the question of admissibility: it is incorrect to address the problem of admissibility as being a problem of trial fairness. However, the argument from the biased audience seems to present peculiar issues in the context of jury trials, given that in such context the explanatory tactic – which would offset the argument – seems insufficient and impracticable. This provides grounds for excluding the evidence, one of which raises the issue of trial fairness. I have shown that these reasons for exclusion are not insuperable: jury directions can be conceived of as the device through which to implement the explanatory tactic towards both jurors and the public.

The foregoing discussion – and the argument from the biased audience, in particular – suggests that the question of admissibility may be triggered and determined by considerations unrelated to the impact that certain evidence may have on the defendant's situation at trial. It seems reasonable to argue that the dynamics of the trial and its epilogue should also be informed by a certain degree of consideration for external factors, such as the public perception of criminal justice.[64] The above suggestion, however, raises a question that is not limited to the public perception of criminal justice. It is the question whether and to what extent any decision-making taking place during the criminal process – e.g., the decision on the admission of evidence, the decision whether to convict, the decision on sentence – should be carried out with an eye to preventing or resolving problems that pertain outside the criminal process and concern people other than those directly involved in it. This is a broad and difficult question that I leave for further research.

Federico Picinali
Law Department, London School of Economics and Political Science
f.picinali@lse.ac.uk

[*] I am indebted to Giambattista Picinali, Mike Redmayne, Nicola Lacey, Grégoire Webber, and Sarah Paterson for the invaluable exchanges I had with them during the research for this paper. Special

thanks go to Jules Holroyd: without our 'philosophical' dog walks and cups of coffee this paper would be little more than a collection of words.

[1] See Michael Tonry, 'Race, Ethnicity, Crime, and Immigration' in S. M. Bucerius, M. Tonry (eds.) *The Oxford Handbook of Ethnicity, Crime, and Immigration* (Oxford and New York: OUP, 2014): 1-17.

[2] For the sake of simplicity I am avoiding the use of numbers. In reality, however, generalisations of this sort are likely to include numerical probabilities.

[3] I shall often use this expression throughout the paper. 'Base-rate' refers to statistical information: numerical or non-numerical probabilistic assertions pertaining to a class of instances rather than to an individual instance. The expression is thus used here as a synonym for generalization. See Federico Picinali, 'Structuring Inferential Reasoning in Criminal Fact Finding: An Analogical Theory', *Law, Probability & Risk* 11 (2012): pp. 199-201.

[4] Throughout the paper I will often use this expression to identify a type of evidence, rather than just the membership evidence in Jack's case.

[5] This highlights the main difference between the problem raised by Jack's case and that raised by the oft-debated hypotheticals of the 'gatecrashers', the 'blue bus' and the 'prisoners in the yard' – where the main issue is the sufficiency of the statistical evidence. For a critical discussion of these hypotheticals see, in particular: Laurence J. Cohen, *The Probable and the Provable* (Oxford: Clarendon Press, 1977): pp. 74-76, 216 ff.; Ronald J. Allen, Michael S. Pardo, 'The Problematic Value of Mathematical Models of Evidence', *Journal of Legal Studies* 36 (2007): 107-140; and Mike Redmayne, 'Exploring the Proof Paradoxes', *Legal Theory* 14 (2008): 281-309. Further important characteristics of the problem addressed here are discussed in the next section.

[6] It would not be far-fetched to qualify the base-rate evidence in Jack's case as character evidence, under section 98 of the Criminal Justice Act 2003 (CJA). In fact, it is evidence "of a disposition towards" misconduct. According to section 101 of the Act, evidence of this sort is admissible only if certain conditions are satisfied; in particular, it may be admitted if "it is relevant to an important matter in issue between the defendant and the prosecution" (section 101(1)(d) CJA). Among the important matters at issue is "the question whether the defendant has a propensity to commit offences of the kind with which he is charged" (section 103(1)(a) CJA). Now, membership evidence of the sort discussed here may be relevant to such question and, apparently, may be admitted at trial. However, according to section 101(3) CJA – which replicates *verbatim* the final part of section 78(1) PACE – "[t]he court must not admit evidence … if … it appears to the court that the admission of the evidence would have such an adverse effect on the fairness of the proceedings that the court ought not to admit it." Moreover, were this provision absent – and all other things being equal – section 78(1) PACE would anyway apply to character evidence. Whether or not the membership evidence is qualified as character evidence, the question of admissibility would always require that the court assess whether admitting such evidence would render the trial unfair.

[7] See A v Home Secretary [2005] UKHL 71.

[8] The metaphor of contamination is used in A v Home Secretary, *supra* note 7.

[9] These and other arguments for exclusion are discussed in Andrew Ashworth, Mike Redmayne, *The Criminal Process* (Oxford and New York: OUP, 2010): pp. 342-362. See also Ian Dennis, 'Instrumental Protection, Human Right or Functional Necessity? Reassessing the Privilege Against Self-Incrimination', *Cambridge Law Journal* 54 (2) (1995): pp. 352-353.

[10] Cf. Larry Laudan, *Truth, Error and Criminal Law: an Essay in Legal Epistemology* (Cambridge: Cambridge University Press, 2006): pp. 226-230.

[11] Consider, in particular, the case law of the European Court of Human Rights. For arguments to the effect that evidence obtained through torture should never be used against the victim see Jalloh v Germany (2007) 44 EHRR 32 and Gäfgen v Germany (2011) 52 EHRR 1.

[12] I am referring, in particular, to those scholars who advanced versions of the argument from autonomy discussed in section 4.

[13] See note 5 above.

[14] Consider, for instance, that in Jack's case the testimony of the neighbour plays a crucial evidential role.

[15] On this point see Picinali, op. cit., pp. 221-222.

[16] Consider the 'prisoners in the yard' hypothetical, discussed in the literature referenced in note 5.

[17] The paper does not address the question as to whether membership evidence should be admitted were it part of the defence case – i.e., if it were exculpatory evidence. Most of the arguments discussed here rest on the assumption that the evidence is produced as part of the prosecution case. There may be grounds to maintain that some of these arguments would equally apply if the evidence were part of the defence case. Indeed, there may be further arguments against admission in the latter case. These questions, however, are left for future research.

[18] With this attribute I refer to economic disadvantages. Notably, a disadvantaged group is not necessarily discriminated-against, and *vice versa*. My reason to focus on groups that are both disadvantaged and discriminated-against is that the combination of these attributes triggers the strongest intuitions against using membership evidence of negative traits. Thus, if I succeed in showing that, notwithstanding that both attributes are at play, there is nothing morally objectionable in using the evidence, this conclusion should apply *a fortiori* to the case of groups that are either disadvantaged or discriminated-against.

[19] I use the term 'social group' to indicate any group of people – ethnic, national, religious, racial, etc. – that is identifiable within a certain society through certain characteristic traits.

[20] See Picinali, op. cit., pp. 213-215.

[21] Aside from accuracy and statistical incidence, another variable that helps in characterising generalisations is the 'degree of specificity' of the group or class to which they refer. E.g., generalisations could be about X men, but also about X men who are fathers, or about X fathers who have kids that disparage Xs' traditional values. The degree of specificity is an important factor to take into account when selecting the reference class from which to draw the most informative generalisation. In fact, although in Jack's case all of the three generalisations above would be informative, the third generalisation would be the most informative, as it encompasses more of Jack's traits than the other two do. In saying this I am assuming that the traits characterising the class that each generalisation refers to are all relevant to our inquiry - where relevance consists, roughly, in the fact that the presence of each trait changes the statistical incidence of the fact at issue. The degree of specificity need not concern us here: it bears on the question of sufficiency rather than on that of admissibility. The problem discussed in the paper regards generalisations irrespective of their degree of specificity. What matters, of course, is that the generalisation is informative, that is, relevant to the inquiry: it must involve traits that are relevant in the sense clarified above. If the generalisation were not informative, it would be excluded on grounds of irrelevance and, therefore, there would be no further question of admissibility left to answer. For an explanation of how the concepts of relevance, of degree of specificity, of reliability (i.e., accuracy), and of evidential strength (i.e., statistical incidence) may fit together into a theory of proof see Picinali, op. cit.

[22] However, see section 8 below for aspects concerning the publicity of deliberations in jury trials.

[23] In addition to the works referenced in section 4, see Peter Tillers, 'If Wishes Were Horses: Discursive Comments on Attempts to Prevent Individuals from Being Unfairly Burdened by their Reference Classes', *Law, Probability & Risk* 4 (2005): pp. 44-46 and Mike Redmayne, *Character in the Criminal Trial* (Oxford: OUP, forthcoming 2015): Ch. 4.

[24] See Michael P. Lynch, *True to Life: Why Truth Matters* (Cambridge, Mass: MIT Press, 2004): pp. 49, 137.

[25] See Lawrence Blum, 'Stereotypes and Stereotyping: A Moral Analysis', *Philosophical Papers* 33(3) (2004): 251-289.

[26] See Tamar Szabó Gendler, 'On the Epistemic Costs of Implicit Bias', *Philosophical Studies* 156 (2011): 33-63 and Jules Holroyd, 'Implicit Bias, Awareness, and Other Imperfect Cognitions', *Consciousness and Cognition* (2014): doi:10.1016/j.concog.2014.08.024.

[27] See Hans Reichenbach, *The Theory of Probability* (Berkeley: University of California Press, 1949): pp. 372-378 and Wesley C. Salmon, 'Statistical Explanation' in W. C. Salmon (ed.) *Statistical Explanation and Statistical Relevance* (Pittsburgh: University of Pittsburgh Press, 1971): 29-87.

[28] On this point see Tillers, op. cit.

[29] In other words, the target of the argument is not the belief that Jack (is more likely to have) harmed Merena, but the belief that Jack is *guilty* for harming Merena. The argument in the next section suggests that the membership evidence is not the kind of evidence that is needed to support the latter belief.

[30] Some of these claims have appeared within the long-standing debate on the admissibility of so-called 'naked statistics' in civil and criminal trials. As far as the question addressed here is concerned, they represents the most important aspect of this multi-faceted scholarly dispute. More information and references on this debate can be found in Redmayne (2008), op. cit., and Picinali, op. cit., pp. 221-222.

[31] David T. Wasserman, 'The Morality of Statistical Proof and the Risk of Mistaken Liability', *Cardozo Law Review* 13 (1991): pp. 942-943. Cf. Hock Lai Ho, *A Philosophy of Evidence Law: Justice in the Search for Truth* (Oxford: Oxford University Press, 2008): pp. 301-302.

[32] Amit Pundik, 'Statistical Evidence and Individual Litigants: a Reconsideration of Wasserman's Argument from Autonomy', *International Journal of Evidence & Proof* 12 (2008): p. 306. To be sure, Pundik writes this sentence while describing the argument from Wasserman. However, at 315 he paraphrases the sentence when expressing his own view.

[33] Even when witnesses are available, we must resort to the logic of single-case inference to draw conclusions on their credibility. In this case, however, it is the witness, rather than the defendant, who is ascribed to a reference class in order to draw inferences on her behavior. According to the above claims, however, this should not make a difference: the individual who is wronged is a different one, but she is wronged nonetheless. But see Pundik, op. cit., pp. 314-315.

[34] Cf. Redmayne (forthcoming 2015), op. cit., Ch. 4.

[35] This suggestion is also supported by the consideration of several passages of Wasserman's and Pundik's articles, where the authors refer to the problem of the membership evidence in connection to the problem of 'imposing liability'.

[36] Nomy Arpaly, *Unprincipled Virtue: An Inquiry into Moral Agency* (Oxford and New York: OUP, 2002): p. 118.

[37] Consider that also in order to draw an inference from motivational states to the behaviour of the individual we need to use membership evidence: we need to look at generalisations concerning the behaviour of people with those motivational states. However, for the reasons given below, this case would not be problematic for the proponents of the argument from autonomy.

[38] It is true that there may be base-rates concerning the beliefs or desires of members of a certain social group. If so, the corresponding membership evidence would give us some information on the motivational states of an individual member of the group. Importantly, this evidence would fall outside the scope of this paper. As clarified in section 2, the paper deals with membership evidence of negative traits, defined as morally objectionable behaviour or dispositions.

[39] The reference is to the famous film 'The cabinet of Dr Caligari', where a man is kept in a state of hypnosis by the captivating and mysterious Dr Caligari, and murders people as a result of the doctor's orders. Or so it seems…

[40] Cf. Antony Duff, 'Dangerousness and Citizenship' in A. Ashworth, M. Wasik (eds.), *Fundamentals of Sentencing Theory: Essays in Honour of Andrew von Hirsch* (Oxford and New York: OUP, 1998): pp. 154-156.

[41] Sure, an eyewitness may state that on a certain occasion the defendant appeared angered, sad, etc. and this information is indicative of the defendant's motivational states. However, eyewitnesses often limit themselves to stating that the defendant was in a particular place at a particular time and/or that the defendant behaved in a particular way.

[42] Whether the evidence must be produced by the prosecution or the defence depends on the allocation of the burden of proof for each claim.

[43] An argument similar in some important respects to that considered in this section is discussed in Frederick Schauer, *Profiles, Probabilities, and Stereotypes* (Cambridge, Mass: Harvard University Press, 2006: pp. 138-141. Cf. Kasper Lippert-Rasmussen, 'Racial Profiling Versus Community', *Journal of Applied Philosophy* 23 (2006): pp. 195-197.

[44] Here I am referring to a possible moral conception of excuses, not to any existing legal excuse.

[45] Imagine an assault case where the crucial issue is identity. The victim was punched from behind in a nightclub and the defendant was in that nightclub at the time of the assault. We know that the defendant was forced to ingest an anger-inducing drug moments before the victim was attacked. In this scenario, the fact that the defendant was under the influence of such a drug may support both the claim that she was the perpetrator and the claim that she should be excused for committing the crime. Is this a reason to disregard this evidence?

[46] On this point cf. George P. Fletcher, *Rethinking Criminal Law* (Oxford and New York: OUP, 2000): p. 759 and John Gardner, 'Justifications and Reasons' in A. P. Simester, A. T. H. Smith (eds.), *Harm and Culpability* (Oxford and New York: OUP, 1996): pp. 118-122.

[47] Consider that Jack's case is not a case of profiling. *Ex hypothesi*, he is already on trial on the basis of evidence other than the membership evidence (e.g., the testimony of the neighbour). Indeed, the question addressed in the paper is whether the membership evidence should be used to prove Jack's guilt, not whether it should be used to profile Jack or any other X father. As the following shows, it is important to appreciate the difference between the two questions.

[48] See Bernard E. Harcourt, *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age* (Chicago: Chicago University Press, 2007): pp. 145-171.

[49] I am advancing no claim that this is the only problematic aspect of profiling. In any case – as the following shows – an argument against profiling need not necessarily be an argument against using membership evidence in criminal trials.

[50] Someone may argue that a frequent use of membership evidence would increase the probability that X fathers (or other social groups for which membership evidence is available) are convicted and, therefore, would produce the ratchet effect. This argument, however, rests on three assumptions that would need to be tested. The first assumption is that the membership evidence would be used very frequently, if considered admissible at trial. The second assumption is that X fathers would be convicted more often if the membership evidence were admitted. This depends on the strength of the membership evidence vis-à-vis that of the other evidence available. It may be that in many cases the other evidence is sufficiently strong and, therefore, the membership evidence does not play any crucial role in securing conviction. The third assumption is that membership evidence would not be used – or would not be used as often and successfully – in trials of members of other social groups.

[51] See Amos Tversky, Daniel Kahneman, 'Judgment under Uncertainty: Heuristics and Biases', *Science* 185 (1974): pp. 1127-1128.

[52] In addition to its apparent plausibility, the claim that the biased audience is likely to misinterpret the base-rate through magnifying the statistical incidence of the negative trait is clearly consistent with the empirical studies that have been conducted on the operation of implicit biases (see John T. Jost, Laurie A. Rudman, Irene V. Blair, Dana R. Carney, Nilanjana Dasgupta, Jack Glaser, Curtis D. Hardin, 'The Existence of Implicit Bias is Beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of Ten Studies that no Manager Should Ignore', *Research in Organizational Behaviour* 29 (2009): 39-69). Nonetheless, it would certainly be interesting and useful to test this particular claim through an apposite experiment.

[53] See Sally Haslanger, 'Ideology, Generics, and Common Ground' in C. Witt (ed.) *Feminist Metaphysics: Explorations in the Ontology of Sex, Gender and the Self* (Dordrecht: Springer, 2011): pp. 182, 189-190.

[54] E.g., an explanation to the effect that 'a generalisation stating that 70% of X fathers engage in violent chastisement is not sufficient evidence to conclude that any given X father has engaged in such practice; furthermore, it does not allow for the conclusion that there is something about the nature of X fathers which brings them violently to chastise their children.' This, of course, is just a sketch: the more precise and clear the explanation, the better.

[55] Understanding the limited scope of the explanatory tactic is crucial to appreciate its viability. For the success of this tactic it is not necessary that biased individuals acknowledge their bias, which is certainly an important condition for the eradication of biases. As suggested by the psychological research, an individual's acknowledgment of her biases is rare, because biases may not be available to introspection (see Brian A. Nosek, Anthony G. Greenwald, Mahzarin R. Banaji, 'The Implicit Association Test at Age 7: A Methodological and Conceptual Review' in J. A. Bargh (ed.), *Automatic Processes in Social Thinking and Behavior* (New York: Psychology Press, 2007): pp. 284-285) and because we tend to over-rely on our 'introspective contents' (see Emily Pronin, Matthew B. Kugler, 'Valuing Thoughts, Ignoring Behavior: The Introspection Illusion as a Source of the Bias Blind Spot', *Journal of Experimental Social Psychology* 43 (2007): p. 566). As a result of these two factors, our biases may remain hidden behind a 'blind spot' that causes us to be readier to impute them to others, than to ourselves. This makes it all the more difficult to eradicate biases, but it may not be equally problematic when the aim is merely that of neutralising the operation of the bias in a particular instance.

[56] Cf. Holroyd (2014), op. cit., arguing that implicit biases do not exclude some kind of awareness of the distorting cognitive processes characterising biased reasoning.

[57] Lilienfeld et al. write that the literature on debiasing techniques against confirmation bias is characterised "by three glaring facts: the paucity of research on the topic, the lack of theoretical coherence among differing debiasing techniques, and the decidedly mixed research evidence concerning their efficacy" (see Scott O. Lilienfeld, Rachel Ammirati, Kristin Landfield, 'Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare?', *Perspectives on Psychological Science* 4 (2009): p. 393). They add that "there is a pressing need for additional research on concerted efforts to combat confirmation bias and related biases" (Lilienfeld et al., op. cit., p. 395), because "a plausible case can be made that debiasing people against errors in thinking could be among psychology's most enduring legacies to the promotion of human welfare" (Lilienfeld et al., op. cit., p. 391). A test of the accuracy of the argument from the biased audience and of the efficacy of the explanatory tactic devised to defeat it may be part of this welfare-promoting project.

[58] Courts have recognized that trial fairness is a 'two-way street', i.e., that the trial should be fair for the prosecution as well. However, as far as s78(1) PACE is concerned, the issue is to determine whether the admission of an item of evidence would render the trial unfair *for the defendant*. In fact, s78(1) applies only to evidence "on which the prosecution proposes to rely".

[59] Cf. Laura Hoyano, 'What is Balanced on the Scales of Justice? In Search of the Essence of the Right to a Fair Trial', *Criminal Law Review* 1 (2014): 4-29.

[60] As was shown, both the argument from the ratchet effect and the argument from the biased audience rely on the publicity (and authority) of the adjudicative process. As far as trials are concerned, publicity is not just a reality – see the considerable amount of high-profile cases that are meticulously reported in the news and sometimes broadcast, attracting vivid public attention – but also a legal safeguard – see article 6 (1) of the European Convention on Human Rights.

[61] One may argue – based on the arguments from consequences – that the defendant is wronged by virtue of being a member of the social group (and that her case at trial is made worse off by the introduction of new incriminating evidence). It is highly doubtful that this would be considered sufficient by any court to establish unfairness pursuant to s78(1) PACE.

[62] Here I work under the assumption that it would be inadequate simply to trust that the party producing the base-rate evidence (in particular, the prosecution) satisfies the above explanatory conditions, especially in the absence of some form of control or incentive in place. It is safer to entrust

the adjudicator (in particular, the court) with implementing the explanatory tactic. The question is whether the adjudicator has the means to do so. As I will argue, she does.

[63] Again, whether the argument from the biased audience and the explanatory tactic work should be subject to testing. And again, if jurors are prone to accepting inaccuracies *mala fide*, neither jury directions nor the exclusion of the membership evidence will preserve trial fairness.

[64] See: Ashworth & Redmayne, op. cit., p. 346; Ian Dennis, *The Law of Evidence* (London: Sweet & Maxwell, 2010): pp. 49-56; and Kenworthey Bilz, 'Dirty Hands or Deterrence? An Experimental Examination of the Exclusionary Rule', *Journal of Empirical Legal Studies* 9(1) (2012): 149-171.