Journal of Practical Ethics

Volume 4, Number 1. June 2016

CONTENTS

The Journal of Practical Ethics is available online, free of charge, at:
http://jpe.ox.ac.uk

Editorial Policy
The *Journal of Practical Ethics* is an invitation only journal. Papers are anonymously appraised prior to publication by expert reviewers who are not part of the editorial staff. It is entirely open access online, and print copies may be ordered at cost price via a print-on-demand service. Authors and reviewers are offered an honorarium for accepted articles. The journal aims to bring the best in academic moral and political philosophy, applied to practical matters, to a broader student or interested public audience. It seeks to promote informed, rational debate, and is not tied to any one particular viewpoint. the journal will present a range of views and conclusions within the analytic philosophy tradition. It is funded through the generous support of the *Uehiro Foundation in Ethics and Education*.

ISSN: 2051-655X

Dilemmas of Political Correctness

DAN MOLLER

University of Maryland

ABSTRACT

Debates about political correctness often proceed as if proponents see nothing to fear in erecting norms that inhibit expression on the one side, and opponents see nothing but misguided efforts to silence political enemies on the other.[1] Both views are mistaken. Political correctness, as I argue, is an important attempt to advance the legitimate interests of certain groups in the public sphere. However, this type of norm comes with costs that mustn't be neglected–sometimes in the form of conflict with other values we hold dear, but often by creating an internal schism that threatens us with collective irrationality. Political correctness thus sets up dilemmas I wish to set out (but not, alas, resolve). The cliché is that political correctness tramples on rights to free-speech, as if the potential loss were merely expressive; the real issue is that in filtering public discourse, political correctness may defeat our own substantive aims.

WHAT IS POLITICAL CORRECTNESS?

Political correctness, as I will understand it, is the attempt to establish norms of speech (or sometimes behavior) that are thought to (a) protect vulnerable, marginalized or historically victimized groups, and which (b) function by shaping public discourse, often by inhibiting speech or other forms of social signaling, and that (c) are supposed to avoid insult and outrage, a lowered sense of self-esteem, or otherwise offending the sensibilities of such groups or their allies. The concept, we should note, is one used by its enemies; dubbing something politically incorrect implies there is something worrisome or objectionable at work, though not necessarily that the po-

1. Earlier philosophical debates illustrate this. See, e.g., Friedman and Narveson 1995

litically correct option is wrong all things considered. But to avoid verbal disputes, let us simply take on board the language of "political correctness" and concentrate on the substantive merits of the doubts that are implicit in the pejorative tone.

According to this characterization, merely advocating for substantive policy changes is not itself a reflection of political correctness, except in a vague, by-association sense of the term. Criticizing someone for referring to an administrative assistant as a "secretary" is a manifestation of political correctness, but advocating for higher wages for assistants is not; insisting on trigger-warnings on syllabi or deleting offending material is again a form of political correctness, but arguing for rape-prevention security measures is not. Certain fringe environmental positions might themselves be loosely dubbed "PC" views, but I suspect this is only because the people who adopt such positions often advocate politically correct norms alongside. Symmetrically, it isn't politically incorrect to make a donation to fight gay marriage–though of course many would respond to doing so with outrage–but it is politically incorrect to write an op-ed making a careful, dispassionate argument against gay marriage. Political correctness thus isn't about private choices deviating from some norm; the notion doesn't refer to a distinctive personal morality, but to a system for moulding public discourse.

The norms involved are mainly, if not universally, negative and inhibitive, and many cases that initially seem positive are more complex, as when campus advocates urge multicultural curricula that move away from the Western canon, or argue for including more women or members of various other categories on syllabi, which doesn't initially seem inhibiting. But the underlying goal, even in these cases, is to avoid the sense that certain groups are marginalized or devalued because members of their group aren't represented in the canon or syllabus. What is being resisted in this kind of case is a certain *implication* that would otherwise inform public discourse, an implication that proponents of political correctness wish to eliminate. We might worry that this isn't what is happening when people simply point out that the Bronze Head from Ife, say, has intrinsic aesthetic merit that warrants study on par with comparable European art, but then this doesn't look much like an appeal to political correctness. The account, it's important to emphasize, isn't supposed to capture just any revisionary or vaguely "left" policy, but rather those with the flavor of responding to the sensibilities of marginalized groups by blocking an offending element. Making a case for the Ife head as envisioned above is an aesthetic argument motivated by independent, positive enthusiasm for the features of the work; appeal-

ing to the negative effects on the self-esteem of certain students when asked to study Phidias, Michelangelo and Picasso is an appeal to political correctness. The same goes for historians' arguments for the revision of inaccurately jingoistic textbooks versus those concerned with avoiding any implication that certain groups are inferior or that their grievances aren't worth addressing. (Of course, the distinction can be a difficult one to draw).

On this definition, moreover, it is significant that what makes a statement politically incorrect is not whether it in fact serves to promote the interests of certain people overall, but whether it threatens their public standing, as typically manifested in a sense of insult and outrage, or a lowered sense of self-esteem and inclusion. Notice, for instance, that no one is willing to retract judgments about what look like politically incorrect statements if they later turn out to promote the interests of the groups in question. If the president of the university says "Members of group G are underrepresented in field F because of unflattering trait T" this may well be judged politically incorrect, and that judgment won't change if it turns out that this was just what G needed to hear. The objective likelihood of advancing the cause of G is beside the point when it comes to political correctness. The air of political incorrectness is brought about by the insult itself, and thus the usual way of overcoming substantive criticism–by showing that the local harm of the insult was outweighed–is ineffective. Calling someone by some group-epithet does not become less politically incorrect if that turns out to be motivating and helpful to the individual, as insulting a friend at a tennis match ("Come on, you jerk!") can evade criticism if it turns out to be helpful overall. That is because the target of political correctness is the insult itself (along with the corresponding threat to the public standing of the group), not the overall interests of the people involved.

This might be resisted on grounds that there is evidence of "stereotype threat," i.e., that issuing politically incorrect statements like the university president's often negatively impact the actual performance of members of marginalized groups, simply by raising the salience of their group-membership and the perception that members are less good at a given task. (Subtly reminding test-takers that they are members of a marginalized group can cause them to perform worse than control groups that take the same test without a priming-cue.)[2] My characterization may then seem falsely to suggest that the concern is for something trivial-sounding ("not hurting people's feel-

2. For a summary and discussion, see Gendler 2011, 48-51. See also research on the possible impact of emphasizing native talent or brilliance, Leslie et al. 2015.

ings") whereas the ultimate concern is to prevent the real and documented damage that the relevant speech and behavior does. However, my claim is not that politically incorrect speech cannot have objectively damaging effects on others, or that such effects might not motivate politically correct norms. The idea is that what *makes* something politically incorrect is a certain kind of offense in virtue of undermining public standing, not that offending people in such a way need be trivial or that blocking such offenses might not have a deeper underlying motivation–just the opposite, as we will see.

As a final elaboration, my gloss emphasizes marginalized groups as the intended beneficiaries of politically correct norms. It is this aspect that leads me to differ with the economist Glenn Loury's otherwise searching analysis. Loury treats political correctness as a far more general phenomenon than I have, suggesting that as "an implicit social convention of restraint on public expression, operating within a given community" it applies to any such restraint, left or right, including, say, fascist censorship. (1994 p.430) The key for Loury is that political correctness, left or right, culminates in a kind of *self*-censorship through a feedback loop in which, first, there are sanctions for those who violate the communal norms, and then those who are willing nevertheless to risk such sanctions come to seem especially refractory.

Suspicious speech signals deviance because once the practice of punishing those who express certain ideas is well established, the only ones who risk ostracism by speaking recklessly are those who place so little value on sharing our community that they must be presumed not to share our dearest common values.(Loury 1994 p.436)

Loury emphasizes such examples as the German politician Phillipp Jenninger, who fell into disgrace after a speech that engaged rhetorically with the perspective of Nazi-era Germans, even though it was unambiguously clear that both the speech and Jenninger's prior life and work were devoid of Nazi-sympathies or anti-semitism. It is worth noting that after Loury's article was published, a Jewish leader gave the same speech in a synagogue in order to demonstrate what he rightly predicted would be the non-response.(*Die Welt* 1995) The worry, clearly, wasn't the substance of what Jenninger had to say but the *signal*, Loury would underscore, that is sent by a German politician (but not a Jew in a synagogue) being willing to take up, if only for rhetorical purposes, the perspective of Nazi-sympathizers *after* it has been established that taking up the Nazi-era point of view is taboo.

Loury is surely right about the impressive degree of self-censorship political correctness can achieve or demand. Standard examples examples include the white

Washington DC mayoral appointee who abjectly apologized to his colleagues who were outraged by his use of the word "niggardly" leading the mayor to accept his resignation. (*The Washington Post* 1999) But although self-censorship is a kind of ultimate victory for those wishing to eliminate some form of expression, actual censorship of various sorts is on the table as well. The same term has been the focus of college speech codes, as when a student objected to it in an academic setting even after its meaning and use by Chaucer were clarified ("It's not up to the rest of the class to decide whether my feelings are valid"). (*Reason* 1999) And of course it's natural to start with formal censorship in order to induce self-censorship. More importantly, I believe there is something distinctively *left* about political correctness, something connected to the concern for victims' groups. This may sound semantic–Loury and others could just announce their conception of political correctness is a bit broader than mine. But there are important differences between how and why various norms shaping public discourse originate and are enforced that are worth recognizing.

Take right-wing attempts to delegitimize opposition to war by suggesting dissenters are insulting "the brave men and women who fight on our behalf," or attempts to shape discourse concerning torture by insisting on Orwellian euphemisms like "enhanced interrogation." Such maneuvers are important to analyze in their own right, since they may work to inhibit speech in disastrous ways, often with outcomes far worse than anything to emerge from petty squabbles over how to refer to an office assistant. But that doesn't make the cases any less distinct. What motivates these right-leaning efforts, baleful though they are, is usually different from and nearly opposite to what motivates norms against questioning affirmative action or syllabi with too many dead white men. In a typical case, what motivates an effort to suppress dissent about war or torture is concern for security, not sympathy or feeling sorry for marginalized or oppressed groups. And the target of the norms is typically what is seen as a display of weakness rather than insults or offenses against the sensibilities of those marginalized, while the response tends toward contempt for the weak rather than outrage at the insult; accusations of disloyalty or spinelessness are more likely than those of being insensitive or cruel. There is a common danger that these attempts at molding discourse will backfire in ways we'll explore below, but these sub-differences, summarized in the table below, still result in an important overall difference in the character of what takes place. (Of course, these are just one set of possible differences; the right-list would differ for those motivated by a concern for individual liberty rather than security.)

	Left	*Right*
Motive	Sympathy	Security
Target	Offense	Weakness
Response to violations	Outrage	Contempt
Accusation	Cruelty	Disloyalty
Danger	Backfire etc.	Backfire etc.

Political correctness, then, is far from unique in trying to influence public discourse and in trying to compel people to speak or think in certain categories or terms. In discussing the problems associated with political correctness we are not singling out left-leaning concerns for special scrutiny. All kinds of social institutions, both left and right, shape which arguments get made, including libel and national security laws, and informal conventions that govern clubs or associations, each with their own profile of burdens and benefits. But political correctness *is* distinctive, and a distinctively left phenomenon, I want to insist. Those attempting to shape discourse on the right are rarely moved by feeling sorry for some group and rarely make corresponding objections focused on avoiding offense. And when they go wrong and undercut their own aims, as when their attempts to shape debate about a war turn out to undermine national security in the long run, they do so by exhibiting a characteristic series of mistakes that are distinct from those most common on the left.

LEGITIMATE ENDS

Political correctness is dismissed by its opponents as if it were either a bizarre and trivial insistence on redefining words, or else an insidious attempt to advance an ideology by silencing the competition. "Yes, but..." is not the typical response of those with reservations. Loury, for example, speaks of the "superficial moralism" of political correctness. This certainly applies to cases like "niggardly" that we can dismiss as childish. But it is less easy to dismiss the taboo on the N-word itself (herewith observed), harder still to dismiss certain taboos regarding racial science, and impossible, I think, to dismiss the underlying worries animating such strictures.

Consider as a historical example the response to the controversial book *The Bell Curve*, which claimed that there is such a thing as general intelligence, that IQ-differences are partly heritable, that they have a significant effect on social outcomes,

and that there are racial differences in average IQ as measured by a standardized test. That reactions to books like *The Bell Curve* really are manifestations of political correctness, not just anodyne scientific disagreement (though there was plenty of that as well), seems hard to deny. The clearest way of making such a diagnosis is to observe that the concern is overwhelmingly rooted in anxiety about offending or insulting a historically marginalized group. The key test is how similar research fares that either doesn't insult but rather tends to praise the group in question, or insults but not a group that we collectively feel sensitive toward. What is striking is that there is no widespread outrage about research into the cognitive *advantages* that Jews or certain Asian groups are sometimes said to enjoy (in scoring higher on average on certain kinds of tests), or the flipside to such research which is that various European groups are inferior in some respect. Here, again, there is plenty of scientific disagreement about the validity of the categorizations involved and the specific experiments or tests underlying the claims, but there isn't the collective shock of a taboo being breached and the accompanying outrage, what amount to public trials, excommunications, and so on.[3] Whatever the scientific flaws critics detected in *The Bell Curve,* political correctness is the only plausible explanation for the asymmetric treatment of parallel work that just isn't insulting or else insults groups no one feels much sympathy for.

So much critics of political correctness get right in a case like this. But they neglect the perfectly good *reasons* for cultivating and enforcing various politically correct norms. In this instance, the root concern is clearly that there exists a horrific record of violence and injustice directed toward African-Americans, as well as a record of promoting such violence by superficially respectable means (including racial pseudo-science), and enlightened moral thinking has thus converged on a *default norm against advancing ideas associated with the oppression or marginalization of African-Americans*. This is why leading responses to *The Bell Curve* focused on associating it with earlier instances of debunked racial science.(Gould 1996) Political correctness thus represents the evolution of public standards with the praiseworthy tendency to protect and promote the interests of historically oppressed groups. These standards work by introducing a *high barrier of entry* to those wishing to enter

3. In 1994 *The New York Times* published an entire series of articles denouncing *The Bell Curve*, a representative opinion piece being "The 'Bell Curve' Agenda" (Oct. 24, 1994). By contrast, the Times headline on research suggesting the heritability of high Jewish IQ scores was "Researchers say Intelligence and Disease May Be Linked in Ashkenazic Gene" (June 3, 2005). This piece, too, raised doubts about the thesis, but the difference in tenor is obvious from the headlines and the texts themselves. *New York Magazine* published an article entitled "Are Jews Smarter?", which also expressed substantive skepticism, but it is hard to imagine a symmetric article with an inverted title about some gentile group.

public discourse in a way that that threatens to undermine moral progress. By maintaining the norms, we acknowledge that such threats exist and that it is important to us collectively to signal to new entrants into public discourse that they must observe the norms carved out to protect the status of groups potentially under threat. And what is true in this case is true of many other examples of political correctness, such as censoring stereotyped depiction of Asians, the German anxiety over displays of sympathy for National Socialism, calls for including more women and other groups on syllabi, or suggestions that the poor are to blame for their plight.

In this respect, I am entirely in agreement with Richard Rorty that political correctness has made "the casual infliction of humiliation…much less socially acceptable than it was," and even that "encouraging students to be what mocking neoconservatives call 'politically correct' has made our country a far better place."(Rorty 1998 pp.81-2) There is no denying that norms to avoid insulting or otherwise attacking the status of women or gay people have brought huge benefits, and critics of political correctness who ignore them are simply mistaken. There are, to be sure, limits on the pursuit of these worthy ends, and inevitably disagreement about where to locate those borders. Barriers to entering the arena of public discourse can be higher or lower–at one end of the spectrum are minor conventions and taboos, the sense of collective shock when someone "dares" to utter certain things. At the other end are explicit laws, say prohibiting hate-speech, which may themselves be narrowly or very broadly defined. Some Canadian jurisdictions, for example, have made it a human rights violation to make any "vexatious comment" known to be "unwelcome by the individual or class" on grounds that include "political belief," which one might reasonably fear as absurdly overbroad.(Northwest Territories Human Rights Act, sec. 14(2), as of 2015.) One may acknowledge the legitimate ends of political correctness without endorsing any and all barriers to public discourse. Correspondingly, we may resist sticks-and-stones maxims suggesting anything-goes in public discourse while symmetrically resisting attempts to shape public discourse by certain agents (such as the state) or to certain degrees (refusing to hire anyone who says anything vaguely "un-PC"). These may, after all, produce costs or pose dangers in their own right.[4] And of course any particular instance of political correctness may be wrongheaded or petty, just as individual applications of patriotic norms. We must not, as Rorty ultimately does, lose sight of the potential drawbacks to political correctness so as to

4. See Waldron 2012 for an argument that the state should in fact pursue heavy-handed tactics like hate-speech laws in pursuit of the sort of legitimate ends I have been acknowledging.

arrive at a reasonable estimate of what, all-in, we gain and suffer, in upholding these norms.

These, then, are legitimate ends for political correctness. Political correctness in itself needn't be mistaken in its fundamental aspirations. Proponents of PC-norms aren't confused to think that racial pseudo-science has had enormous, damaging effects in the past; they aren't mistaken to regard any revival of racial science as potentially disastrous and in any case accompanied by huge costs. It is not true that opposition to any such revival is (or need be) rooted in mere "superficial moralism," and there are good reasons for maintaining collective default-norms that signal certain kinds of discussion out of bounds in the normal course of things.

There are two wrinkles in this story that bear mentioning, however. One is that politically correct norms have a distinctive *content* that makes emphasis on language inevitable. The whole point of such norms, as I have described them, is to generate a set of default-presumptions that those participating in public *discourse* are expected to observe in order to ward off threats to a certain kind of moral progress, and so naturally terminology and word-choice features prominently in the marshaling of such norms. This can then give rise to the absurd cases already noted that often revolve around what really are morally superficial *applications* of reasonable norms. Norms pertaining to language-use are perhaps especially liable to misuse in ways that will strike some as preposterous since they inevitably implicate what can always be ridiculed as "mere" labels. Relatedly, we noted that political correctness concerns offense and sensibilities, not the objective interests of those involved. It might seem surprising that the norm to evolve was one that focused on offense and not simply on promoting whatever was in the people's objective interest. But this is again similar to other norms, like love of country. In both cases there is a core goal of promoting the interests of some entity, but part of this is taken to involve discouraging insults and other threats to the publicly recognized status of the people or thing in question. Failing to acknowledge the values in question by a lack of reverence or deviance from certain standards are thus punished, even when what is at stake seems superficially to be only symbolic. Political correctness is one face of a deeper concern for the oppressed comparable to the dimension of patriotism associated with denouncing insults to country.

DILEMMAS

There is nothing wrong with promoting a presumption that historically oppressed or marginalized groups should not be insulted or subjected to discourse threatening to undermine their status, and it is puzzling that critics of political correctness seem frequently unwilling to acknowledge its legitimate ends. That leaves the door open to a second kind of criticism, the misguided application of the relevant norms, but whatever the damage to individuals losing their jobs or being publicly anathematized, it cannot be said that mere misapplication of values raises interesting philosophical problems. It is rather a third kind of problem with political correctness that should anchor our attention, the problem of *conflicts* among values, whether between those associated with political correctness and other things we care about, or even internal conflicts within the former. We can enumerate several different kinds of dilemma-engendering conflicts.

Orwellian discourse: One kind of conflict occurs when politically correct norms lead to the kind of abuse of language that Orwell criticizes in "Politics of the English Language." We noted earlier that the petty misapplication of language-norms isn't worth making a fuss over, but as Orwell points out, the vague and imprecise use of terms like "fascism" can come to serve as a "defense of the indefensible."(1946 p.162) Contemporary versions of this on the right are easily recognized, as when a kill list becomes a "disposition matrix," but political correctness seems to involve a similar tendency. Loury draws the connection to phrases like "disadvantaged minorities" (nowadays the term would be "diversity") which he says is "used in educational philanthropy circles when the speaker really means 'non-Whites, excluding Asians.'... Such linguistic imprecision impairs analysis. But that is often its purpose," among other reasons, he suggests, because announcing that a scholarship was to be offered to "non-whites excluding Asians" would, by its very accuracy, render the proposal impossible.(Loury 1996 p.447) Another policy-shaping example is the increasing tendency to reject official government terms like "illegal alien" in favor of "undocumented immigrant" or even "undocumented citizen," with the implication that refusing to do so implies reactionary or hateful views. These campaigns aren't just the one-off ideas of random individuals; the phrase "undocumented citizen" is encouraged by administrators at a major state university in the United States, and others urge that

the statement "America is a melting pot" constitutes a form of "microaggression."[5] Regardless of what the right immigration policy is, and notwithstanding the legitimate interest in avoiding various forms of marginalization, this kind of discourse once again "impairs analysis." "Undocumented immigrant" is meant to make it harder to focus on the fact that there are laws and procedures governing entry to the country that were flouted by the persons in question, while the Orwellian "undocumented citizen" seeks to present a political aspiration as a *fait accompli*. To the extent that we recognize both the legitimate ends of political correctness and the undesirable effects that Orwell drew our attention to, we should see these as cases that present a dilemma.

Causal structures: More substantively, fears of politically incorrect stereotyping threaten to subvert our understanding of the world even without Orwellian word-games, as when there is resistance in the public sphere to the suggestion that a stereotypical trait is causally implicated in some negative outcome. The stereotype of deference to authority in many East Asian (and other) societies and its role in causing accidents is an example.[6] When Korean airline flight Asiana 214 crashed in San Francisco, the suggestion was made that such deference made a difference, as the pilot was relatively junior and was being supervised by an instructor, possibly leading the pilot to be reluctant to assert the need for a go-around. This suggestion was in turn widely derided for succumbing to cultural stereotypes. *Atlantic Monthly* author James Fallows, for example, introduced the claim under the heading "Confucius in the cockpit," alongside a comical depiction of the sage, and noted that he was, "highly skeptical of this whole line of thinking...If an (apparently) mishandled approach shows something about Korea–or East Asia, or Confucius, or rote-learning systems–then what do we make of the many thousands of Asian-piloted flights that land smoothly and safely throughout Asia every single day?" (Fallows 2013) Needless to say, the safe landings Fallows refers to aren't reasons to discount or mock the suggestion that a trait exhibited with a higher prevalence in some cultures

5. "Undocumented citizen" occurs for example in an official publicity campaign of the University of Maryland, which places "illegal alien" alongside expressions like "retarded" and "no homo." http://thestamp.umd.edu/multicultural_involvement_community_advocacy/programs/inclusive_language/phrases, accessed 2/3/15. The "microaggression" point is from "Tool: Recognizing Microaggressions and the Messages they Send," part of the materials for a leadership seminar sponsored and extensively promoted by UCLA. http://www.ucop.edu/academic-personnel-programs/programs-and-initiatives/faculty-diversity-initiatives/faculty-leadership-seminars.html, accessed 7/20/2015. For more details, see Volokh 2015.

6. For a review of how accurate stereotypes are, see Lee et al. 1995. The stereotype that stereotypes are generally wrong is itself dubious, as the authors point out. For an accessible historical survey focusing on Asian aviation safety, see Gladwell 2008, ch. 7; I focus on a more recent example.

than others might contribute to the explanation of an accident in this case. What appears to be at work in this writing as well as in other public denunciations of the hypothesis is anxiety about reinforcing stereotypes. The important thing for our purposes isn't whether authority-deference actually *did* play a role, only that politically correct norms threaten rational analysis of the cause of a plane crash, assuming that public ridicule counts as a cost those analyzing such crashes must reckon with. As it happens, "Interviews with pilots indicate that Korean culture may have played a role in the crash...Captain Lee told investigators that any of the three pilots on the plane could have decided to break off the approach, but he said it was 'very hard' for him to do so because he was a 'low-level' person being supervised by an instructor pilot."(*The New York Times* 2013) The NTSB report states that "the PF's [pilot flying's] deference to authority likely played some role in the fact that he did not initiate a go-around."(NTSB/AAR-14/01, 92)

Against this, the fact that there are sources to cite discussing the role of cultural differences in accidents may seem to undercut the idea that there is some politically correct taboo surrounding the subject. But politically correct norms are graded–some topics are widely off-limits in public discourse, but others merely get subjected to "heightened scrutiny." These introduce a filtering effect. The thought isn't that it is impossible to discuss publicly the arguments involved, but knowing one will be subjected to moralized criticism introduces an initial barrier serving as a partial filter on public speech. Similarly, no one thinks it was impossible to criticize the Iraq war, but patriots seeding suspicion of dissent in effect raised the barrier to entrants to the public discussion. And it's worth observing in passing that the barrier introduced by stereotype-aversion extends beyond political correctness strictly speaking, to avoiding stereotypes more broadly, again with worrying effects. Neanderthal research, at least any that informs public discourse is inevitably along the lines of, "surprising new study upends stereotype that Neanderthals were dim-witted." This in itself might seem to reflect a random piece of scientific progress. The trouble is that it is difficult to imagine an article title, let alone a newspaper headline, along the lines of "Neanderthals: as dumb as we thought." Such research would thus need to overcome both bias in favor of novelty and the quasi-politically correct bias against saying anything nasty about the underdog (whom the extinct presumably exemplify).

Backfire: Other conflicts are internal to the concerns the community has for those marginalized, particularly conflicts arising between the public-facing desire not to insult or offend, on the one hand, and the substantive concern actually to advance

people's interests on the other. We see this in the case of patriotism when jingoistic zeal interferes with frank and open attempts to improve the life of the country. Dissent in war is the obvious case, but there are many others, as when critics on the right refuse to accept "revisionist" histories that attempt to wrestle with an ugly past so as to improve national culture, or when national pride leads to a denial that core values are being undermined by various policies. These conflicts represent a set of norms backfiring against those who apply them so that the core-values the norms emerged from are actually undermined as a net result. Politically correct backfiring includes pressure for trigger warnings in courses and attendant pressure to leave off "sensitive" materials from the syllabus that may upset students who have been traumatized. This has the predictable consequence that instructors are less likely to teach material relevant to topics like sexual violence, with the unintended consequence of decreasing knowledge of relevant law that might be used to protect women: "asking students to challenge each other in discussions of rape law has become so difficult that teachers are starting to give up on the subject" leading instructors to omit "rape law in their courses, arguing that it's not worth the risk of complaints of discomfort by students."(Suk 2014)

The so-called mismatch hypothesis concerning American-style affirmative action furnishes another example of the costs of political correctness. (For an overview, see Sander and Taylor 2012.) Once again, my point does not turn on whether the empirical claim is true; as with any social science work, there is bound to be some controversy and my goal isn't to establish the validity of a particular scientific claim. But according to a significant body of research, affirmative action does immense damage to the "beneficiaries" of the program by tending to shift students from academic environments in which they might well flourish toward harsher, more demanding environments for which they may not be as well prepared, and in which they consequently do worse. The main problem appears to occur not at the higher echelon of elite institutions, but in somewhat less selective schools, who as a result of an under-supply of suitable students are left with fewer and fewer candidates, as those there are get scooped up by the more elite schools. Diversity-pressure at higher echelons, in other words, is said to have disastrous consequences at lower echelons that fill their ranks with candidates who would benefit more at less selective institutions. These benefits are said to include better grades, greater learning, better bar-exam results, greater likelihood of going into the sciences, and better careers. For example, following the natural experiment introduced by California's proposition

209 barring affirmative action, the number of African-American and Hispanic freshmen who went on to graduate in four years rose 55% and the number who went on to get a STEM degree in four years rose 51%. (Sander and Taylor 2012 p.154)

Obviously, it is reasonable to wonder whether the benefits of being admitted to an elite institution outweigh the benefits of ending up at a less elite but better "fit" institution, or how common such dilemmas really are, and so on. But a body of academic research had accumulated that this might be so by the early 2010s. According to this work, there were large net benefits to attending a school for which students were well prepared academically instead of a fancier school in which they were more likely to struggle. This research had a reasonable hypothesis as its target, was performed by multiple, well-respected faculty at prominent institutions and was published in serious, peer-reviewed journals. Nevertheless, there was (and is) a profound resistance to taking any of this research seriously, despite the fact that it purported to show that a set of policies was backfiring so as to cause the community to fail to achieve its own aims.[7] At Duke, research showed that non-Asian minorities tended to self-select out of the hard sciences because of poor performance as a consequence of mismatch, but instead of prompting corrective action, school officials reacted with lukewarm affirmations of academic freedom, and the comment that "We understand how the conclusions of the research paper can be interpreted in ways that reinforce negative stereotypes." (Lange et al. 2012) Any sense that there was powerful evidence that our policies might be irrational (in the formal sense that they caused us to act contrary to our own self-given aims) was (and still is) almost entirely absent. And at the second order, the authors received the scathing denunciations characteristic of political correctness, including at campus protests, signaling that research into what would promote the substantive interests of historically oppressed groups would not be tolerated if the results conflicted with norms against insult and offense.

I am not claiming that universities and public intellectuals were wrong, all things considered, to ignore or deny this research. We have seen that there are legitimate reasons to adopt a default norm of straight-arming ideas tending to insult the status of marginalized groups. The point is rather that doing so comes with *costs*, setting up a *dilemma*. My central contention isn't that we ought to do away with the supposed superficial moralism of political correctness, but rather that we ought to focus instead

7. For a striking illustration, see the high-profile public debate on the subject "Affirmative Action on Campus Does more Harm than Good" (Intelligence Squared US), widely available online.

on the dilemmas political correctness introduces, and face up to the costs incurred in being gored on either horn.

NORM-DEPENDENT RESPONSES AND REVERSE-HYPOCRISY

It is difficult to tally up or even to compare the costs of having or not having politically correct norms, but it is clear that both can be high. To dwell on the costs of enforcing them, in a high-cost scenario they can lead to what Timur Kuran calls widespread "preference falsification" in which what people believe in private becomes increasingly detached from what is spoken in public, which in the case of East-bloc communism made it impossible to discuss pervasive dysfunctions urgently requiring reform. Worse, Kuran identifies an "intergenerational process through which the unthinkable becomes the unthought," making such dysfunctions unidentifiable even if they could be discussed. (Kuran 1995 ch.13 and p.186) Alternatively, Loury points out that preference falsification can lead to polarization, by a process analogous to Gresham's law, whereby the bad money (extreme opinion consonant with politically correct ideology or else violently opposed to it) drives out the good (moderately heterodox opinion), and so comes to dominate what circulates in public. (Loury 1996 pp.435-6)

This is the high-cost scenario for political correctness. We can illustrate it in the arena of distributive justice, which offers ample scope for the relevant norms. Attributing poor social outcomes to factors *external* to the person (to society, the state, etc.) sounds "nice," since we don't feel like we're blaming the underdog for their already unpleasant position; attributing them to factors *internal* to the person (e.g., to poor choices) sounds "mean" and is likely to trigger charges of "blaming the victim." This makes the latter less politically correct. And on the international stage, claiming that poor countries are in part worse off due to endogenous factors like institutions or culture similarly has a un-PC quality to it that blaming multinational corporations or the rich countries does not. Since there is a long and ugly history of rich countries invading poor countries and an even longer history of richer citizens taking advantage of poorer citizens in politics, law and business, it isn't unreasonable to accept a norm discouraging theories threatening to undermine the status of the poor. But against this, if it turns out that people are capable of significantly influencing social outcomes in the course of educational, fertility or work decisions, and that absent these, statist policies are likely to be ineffective, it will be disastrous for such facts

not to inform public discourse, or for them to face ridicule. (Similarly in the international variant.) The high-cost scenario in these cases, then, is one in which there is a widespread belief that, say, social pathologies play an important role in explaining bad social outcomes, but there is reluctance to discuss that belief; or in which the thought doesn't seem even a live option to many (it's "unthought"); or that damaging polarization sets in because those with moderate views face penalties for voicing their opinions in the public arena ("Most people can exercise significant influence on whether they end up poor and should be criticized for not doing so positively, but of course violence-prone slums or abandoned rural areas are another story").

This high-cost scenario seems to me more than overblown fear-mongering, though it would be difficult to establish the extent to which it is or may be realized. Instead, let me make two specific suggestions about how to think about the costs on either side of the ledger. On the one side, there is a curious problem that arises when one is concerned to promulgate norms so as to avoid insult or offense, but those very norms play a role in shaping the nature of the insult or offense. In Germany Americans are sometimes referred to as "Amis," short for Amerikaner (not the French term). Suppose a headmaster notices that this expression is used of a small minority of expatriot students who are sometimes bullied. In sympathy, the headmaster forbids what he sees as a possibly condescending use of "Ami," insisting instead on the full "Amerikaner," and goes on to lecture students on how to treat their foreign guests. Is this the proper *remedy* for oppressed ex-patriots? It may be, and the old-fashioned approach of telling the ex-pats to buck up and pay no mind to the rotten kids may turn out to be ineffective or cruel. But the danger is that the headmaster's sympathetic norm itself sensitizes the ex-pats to what they formerly paid little mind to, but now interpret as major offenses that they ought to dwell on, talk about, feel traumatized by, and so on. A parent might reasonably judge the headmaster's approach a mistake, once the subtle point about the feedback loop implicit in such norms is recognized. In promulgating norms designed to benefit marginalized groups we both help and hurt them.

There is ample empirical evidence for this idea in relation to serious trauma. Victims of sexual abuse and combat veterans fare worse the more they see their traumatic event as a central, defining moment; norms tending to *downplay* the importance of the event would thus be expected to help. (Robinaugh and McNally 2011) A *New England Journal of Medicine* piece on victimhood and resilience points out that immediate counseling after trauma, which tends to highlight that the victims *are* victims

who should be *expect* to feel traumatized, often seems to make things worse, increasing the likelihood of mental disorders.[8] Given all this, we must be cautious in thinking about how to assess the costs of having or not having some norm that superficially promotes some victimized groups' interests; assessing the overall effects is far from straightforward.

On the other side of the ledger, an important metric to pay attention to (and which social scientists could attempt to measure) is the prevalence of reverse hypocrisy. This is the practice of applying high standards in one's private life, especially toward one's children or other loved ones, while publicly promulgating low standards for others, either explicitly or by withholding public criticism. Reverse hypocrisy is evidence of something like Kuran's preference falsification. The savvy communist party member publicly signals agreement with low standards for productivity, urging that the state should provide for everyone's needs without anyone doing "extra" work not officially assigned to him; but privately he urges his kids to work long hours in the informal sector and to accumulate savings. Examples closer to home include private insistence on personal responsibility in domains like fertility decisions or work ethos, while ignoring or even mocking these as public norms. In this sense, it's reverse hypocrisy to make it clear that you expect your children to make sensible decisions about family and to work hard in school whatever the excuses they are tempted to make, while criticizing or lampooning old-fashioned sounding public norms to the same effect.

It might be objected that there are substantive reasons to uphold different standards in the public arena than in the personal. A liberal tolerance for differences or even just politeness might seem to dictate as much, and of course if one judges that others are less fortunate in their capacity for making the relevant discriminations, or are less well positioned to act on them, it may seem inappropriate to uphold what would be unreasonable standards. Arguments from liberal toleration or etiquette are less persuasive, though, when the stakes include the wellbeing of someone else's family. And the view that *we* or those close to us can adhere to high standards that will promote our wellbeing but *they* cannot, has a worrying ring of condescension. Short of extreme circumstances, many successful families simply will not tolerate children doing poorly in school (let alone not finishing), making poor fertility decisions, or failing to work. But many of the same people are reluctant to assert these as public norms or to issue criticisms based on them. Since asserting such norms would

8. Wessely 2005. For philosophical reflection on our propensity to underestimate resilience in the face of trauma, see Moller 2007.

often involve criticizing marginalized groups–those on the receiving end of such criticism would almost by definition be worse off–this sort of reluctance looks like a good measure of political correctness, and its prevalence would be a useful barometer of what I call the high-cost scenario.

IS POLITICAL CORRECTNESS A MYTH?

Given the many examples cited up to this point, it may seem surprising that some have doubted whether political correctness exists at all, at least to any significant extent. But writers have in fact expressed two kinds of doubt along these lines.[9] One is the denial that there are socially significant instances of public speech being shaped in the ways I have outlined, so that it is denied that public speech about race or gender say is subject to anti-marginalizing norms to any significant extent. The other form that denial takes is insisting that, while there really are norms informing this discourse, these aren't motivated by the kinds of concerns I have singled out. If it turns out that "political correctness" is just the neutral struggle for truth and justice that the recalcitrant wish to rebrand and demonize, then perhaps the phenomenon is, once again, a kind of myth. (There is a tendency to combine these two forms of denial, but notice that they are inconsistent and so we should stick with one or the other.)

Presumably we can demonstrate that X exerts a non-trivial influence on public discourse by showing that the discussion of major social institutions is in part shaped by X. It should be sufficient to dispel doubts about the existence of political correctness that we demonstrate non-trivial instances of public discourse being subjected to its influence. And it seems to me that we have seen ample evidence that this condition is satisfied. Immigration laws are important social institutions. So are university admission policies and government investigative agencies. In each instance, as we have seen, there have been non-trivial exertions of influence by powerful entities such as university administrators and members of the press in order to shape discussion of the relevant issues. That just is political correctness, provided it is motivated in the way I have defined the phrase. There may be reasonable disagreement about how much political correctness there is (compare: "Exactly how much jingoism or

9. They include an anonymous reviewer, Feldstein 1997, Wilson 1995, and Fish 1994, in varying degrees. The latter three are responding to culture-war polemics from the right (e.g., Feldstein 1997, 116-120, Wilson 1995, 10-15, Fish 1994, 53-79), not careful analysis like Loury's, and so their doubts should perhaps be taken in that light.

sexism is there?"), or whether it is a major concern in the grand scheme of things ("Compared to all the other problems in the world, how bad is jingoism or sexism?"). But we shouldn't move from views about how important, exactly, political correctness is to denying its existence outright.

Alternatively, there is the suggestion that what is deemed political correctness is just a pejoratively described, politically neutral attempt to fight for truth and justice. But on reflection, this too succumbs to the evidence assembled up this point. The problem is that the entities seeking to influence public discourse seem specifically motivated to defend historically oppressed or marginalized groups and not other kinds of groups. Thus, if the objections to works like the *The Bell Curve* were rooted in neutral concerns about shoddy science, we would expect to see symmetrical concern for claims about inferiority and superiority, and among those that are supposedly inferior, similar concern for historically oppressed and historically dominant groups. But as noted earlier, that is not the case. Findings of high IQ scores among Ashkenazi Jews do not produce the same degree of social anguish and institutional ostracism, and no one is worried by the implication that gentiles are inferior. This doesn't show the scientific objections aren't correct–I am not drawing the invalid inference that because outrage was triggered by political correctness, therefore charges of shoddy science are wrong or can be dismissed. Let us just assume all of the scientific objections were correct. The point is that there's a distinctive concern for the status of oppressed and marginalized groups at work here, not that such motivations cannot serve to uncover the truth. A similar symmetry-test applied to the other cases discussed produces similar results. University officials urging us to refer to illegal aliens as "undocumented citizens" make similar suggestions concerning other marginalized groups, we noted, not concerning dominant groups who might be deemed mislabeled. Those concerned about the role of stereotypes in causal explanations aren't symmetrically concerned to stamp out stereotypes about dominant groups. Nor is any of this surprising. It would, if anything, be strange if a well-meaning public failed to have some norms about public discourse concerning historically marginalized members. As long as the public does, we should expect these kinds of asymmetric norms which, as I have argued, have legitimate ends but also pose difficult dilemmas.

Pressing on such asymmetries may seem misguided if there are real differences between the cases. Discourse that suggests that historically marginalized people somehow deserve to be marginalized is obviously harmful in a way that insulting dominant groups is not. It is no wonder that we respond differently to these differ-

ent cases, we may think, which involve harms that can hardly be compared. This, however, is to make my point for me. The differences involved are precisely those that make for political correctness. To say that "it's different" when what is at stake is the public standing of a group that has been persistently wronged in the past is just to say that it's different when it's politically incorrect. This is what I have tried to argue all along.

REFERENCES

Fallows, James. 2013. Confucius in the Cockpit, and Other Items to Read, and Ignore, on Asiana 214. The Atlantic. [Available at: http://www.theatlantic.com/technology/archive/2013/07/confucius-in-the-cockpit-and-other-items-to-read-and-ignore-on-asiana-214/277703/, accessed 1/14/15].

Feldstein, Richard. 1996. *Political Correctness: A Response from the Cultural Left* (Minneapolis: University of Minnesota Press).

Fish, Stanley. 1994. *There's No Such Thing as Free Speech* (New York: Oxford University Press).

Friedman, Marilyn and Jan Narveson. 1995. *Political Correctness: For and Against* (London, Rowman and Littlefield).

Gendler, Tamar. 2011. "On the Epistemic Costs of Implicit Bias," *Philosophical Studies* 156, 33-63.

Gladwell, Maclcolm. 2008. *Outliers* (New York: Little Brown).

Gould, Stephen Jay. 1996. *The Mismeasure of Man* (New York: WW Norton).

Kuran, Timur. 1995. *Private Truths, Public Lies* (Cambridge: Harvard University Press).

Lange, Peter et al. 2012. "A message from Administrators Regarding New Study." *The Duke Chronicle*, January 18. [Available at: http://www.dukechronicle.com/article/2012/01/message-administrators-regarding-new-study, accessed 28 June 2016].

Lee, Yueh-Ting, et al. 1995. *Stereotype Accuracy: Toward Appreciating Group Differences* (Washington DC: American Psychological Association).

Leslie, Sarah Jane et al. "Expectations of Brilliance Underlie Gender Distributions Across Academic Disciplines" *Science* 347 no. 6219 (2015), 262-265.

Loury, Glenn. 1994. "Self-Censorship in Public Discourse," *Rationality and Society* 6, 428-461.

National Academies. 2008. *Treatment of Posttraumatic Stress Disorder: An Assessment of the Evidence*

The New York Times. 2013. "Pilots in Crash Were Confused About Control Systems, Experts Say," December 13. [Available at: http://www.nytimes.com/2013/12/12/us/asiana-airlines-crash-san-francisco-airport.html, accessed 1/22/15].

Moller, Dan. 2007. "Love and Death," *The Journal of Philosophy* 104, 301-316.

Orwell, George. 1946. "Politics and the English Language." In his *Collection of Essays* (New York: Harcourt Brace).

Reason. 1999. "Cracking the Speech Code." July. [Available at: http://reason.com/archives/1999/07/01/cracking-the-speech-code, accessed online 1/14/15].

Robinaugh, Donald and Richard McNally. 2011. "Trauma Centrality and PTSD Symptom Severity in Adult Survivors of Childhood Sexual Abuse." *Journal of Traumatic Stress* 24, 483–486.

Rorty, Richard. 1998. *Achieving our Country* (Cambridge: Harvard University Press).

Sander, Richard and Stuart Taylor. 2012. *Mismatch* (New York: Basic Books).

Suk, Jeannie. 2014. "The Trouble with Teaching Rape Law." *The New Yorker* Dec. 15. [Available at: http://www.newyorker.com/news/news-desk/trouble-teaching-rape-law, accessed 28 June 2016].

Vallone, Robert, Lee Ross, and Mark Lepper. 1985. "The Hostile Media Phenomenon." *Journal of Personality and Social Psychology* 49, 577-585.

Volokh, Eugene. 2015. "UC Teaching Faculty Members not to Criticize Race-based Affirmative Action, Call America 'Melting Pot,' and More." *Washington Post* 6/16. [Available at: https://www.washington-post.com/news/volokh-conspiracy/wp/2015/06/16/uc-teaching-faculty-members-not-to-criticize-race-based-affirmative-action-call-america-melting-pot-and-more/, accessed online 28 June 2016].

Die Welt. 1995. "Keiner hat etwas gemerkt." January 12. [Available at: http://www.welt.de/print-welt/article664397/Keiner-hat-etwas-gemerkt.html, accessed online 1/14/15].

The Washington Post. 1999. "Williams Aide Resigns in Language Dispute." January 27. [Available at: http://www.washingtonpost.com/wp-srv/local/daily/jan99/district27.htm, accessed online 1/14/15].

Waldron, Jeremy. 2012. T*he Harm in Hate Speech* (Cambridge: Harvard University Press)

Wessely, Simon. 2005. "Victimhood and Resilience." *New England Journal of Medicine* 353;6, 548-550.

Wilson, John. 1995. *The Myth of Political Correctness* (Durham: Duke University Press).

Offsetting Class Privilege

HOLLY LAWFORD-SMITH

University of Sheffield

ABSTRACT

The UK is an unequal society. Societies like these raise significant ethical questions for those who live in them. One is how they should respond to such inequality, and in particular, to its effects on those who are worst-off. In this article, I'll approach this question by focusing on the obligations of a particular group of those who are best-off. I'll defend the idea of morally objectionable class-based advantage, which I'll call 'class privilege', argue that class privilege can be non-culpable, and put forward an account of the obligations those with class privilege have. My main claim will be that those with class privilege have obligations to 'offset' their privilege, in something like the same way high emitters have obligations to offset their greenhouse gas emissions.

INTRODUCTION

The UK is an unequal society. Take income, for example. At one end of this spectrum of inequality, there are people who cannot secure employment at all, who are either destitute, or who rely on benefits of £72.40 per week.[1] There are people on apprenticeships earning a weekly wage of £158.40, and there are people who work routine jobs for a minimum weekly wage of £345.60.[2] At the other end, there are

1. This figure is the Jobseekers Allowance (JSA) for people aged 25 years and over, and is the same for both contribution-based and income-based JSAs (which a person is entitled to depends on whether she has made sufficient past contributions to National Insurance) (totaljobs.com).

2. These numbers are based on the (hourly) National Living Wage and National Minimum Wage rates that apply from 1st April 2016, for the category of people aged 25 years and over, and calculated to a weekly wage on the assumption of the UK's legal maximum of a 48-hour working week (GOV.UK 2016).

around 2.93 million people earning more than £1,117 per week, a figure which represents the bottom of the top 5% of earners in the UK (Jenkins 2015, p. 4). Or take occupation. Some people have jobs that come with low levels of social prestige and recognition, while others have jobs that come with high levels of social status and prestige. A poll in the United States revealed Banker, Actor, and Real Estate Agent to have the least prestige, and Firefighter, Scientist, and Teacher to have the most prestige, of the occupations surveyed (HarrisInteractive, 2007).

Or take education. 26% of the UK's jobs require a degree, but most of the UK's population do not go to university—the percentage who do is between 27.2% (based on data from the Office for National Statistics, 2013) and 40.2% (based on data from the Annual Population Survey) (Ball, 2013). At the high-school level, only 7% of the UK population goes to private schools, but graduates of private schools make up 75% of the UK's judges, 70% of the UK's finance directors, 53% of the UK's journalists, 45% of the UK's top civil servants, and 32% of the UK's Members of Parliament (Monbiot 2010). Finally, the children of higher professionals are three times more likely than the children of people in routine work to get five good GCSE grades (ibid). Or take social capital. Some people have extended networks of friends, colleagues, and contacts in influential social positions. These people can be called upon for favours, or to assist in difficult times such as transitions in employment, or to alleviate financial pressure. Others have smaller networks, consisting of people in non-influential social positions.

All of this is hardly surprising from a descriptive perspective, given the country's long feudalist history. But societies like these raise huge numbers of ethical questions for those who live in them. One such question is how we should respond to such inequality, and in particular, to its effects on certain members of the society. Many will be troubled by the situation of those at the bottom end of this spectrum. The broadest version of the issue I'm interested in here is whether there is anything that people in such societies owe to each other, as a result of these inequalities. Whether they do—and what it is they owe *if* they do—depends on a number of things.

Chief among them is whether some people are culpable in the fact of this inequality and its effects. Culpability is usually assigned on the basis of a person's intentionally (or at least forseeably) doing harm. So there would be culpability for class privilege if, for example, some of the people at the top have intentionally made it the case that some of the people at the bottom are at the bottom. If there is culpability, either for the inequality itself or for the fact that certain people end up at the bottom,

then much of the story about what some owe to others can be told in terms of the obligations of the culpable to make reparation for, or pay compensation to, those who they have harmed. (Or those who have been harmed as a foreseeable result of what they have done).

If there is no culpability, an answer to the question of what people in such a society owe each other may yet be given in other terms. Some might owe others *assistance*, on the simple grounds that some have the resources to provide assistance, and others need assistance. Or we might all owe particular things to anyone with whom we share a particular kind of *association*, such as the political association residents of the UK share with one another.

In this paper, I'm interested in pursuing a very different way of telling the story, namely in terms of *benefiting*. I want to ask specifically about the obligations of those at the top. This is to take seriously the intuition defended in Daniel Butt's paper 'On Benefiting From Injustice', that beneficiaries of injustice[3] have obligations that are stronger or more extensive than those that everyone has in virtue of either shared association, or capacity to provide assistance (Butt 2007; see also Barry 2003).

That will require doing three things: (i) defending the idea of morally objectionable class-based advantage, which I'll call *class privilege*[4] (Section II below), (ii) arguing that class privilege can be non-culpable (this keeps the story about who has what obligations in the domain of beneficiaries rather than shifting it to the domain of redress for culpable harm) (Section III), and (iii) putting forward an account of the obligations those with class privilege have (Section IV). After that, I'll address an important objection to do with people being complicit in their own disadvantage (Section V). I'll argue that class privilege is best understood as a failure of social mobility; that there are many class-privileged people who are not culpable in the fact of class privilege; and that nevertheless the class-privileged ought to 'offset' their privilege by taking on cost to undermine the current failures of social mobility. This is in just the same way that high emitters of greenhouses gases ought to offset their emissions.

One caveat before that. It is possible to give a range of different answers to the question of what the class privileged owe, because it is possible to approach the ques-

3. Here I'm extending the idea of benefiting from (discrete, identifiable acts of) injustice (i.e. perpetrated by one individual against another and sending benefits to a third), to cover benefiting from structural injustice, social inequality, and other states of affairs that are morally problematic and yet may fall short of being unjust.

4. Different features than class—for example race, and gender—will be more or less relevant to social inequality in the context of different countries. Class is one very important feature in the UK, which is why I'm focusing on it here.

tion in a more and less utopian way. The *best* moral answer to class privilege might be the dismantling of the very fact of class-based differences between people. A *good* answer might be to ensure that class position is decided fairly, for example, by lottery, or by choice, or by effort alone. I do not claim to be defending the best moral answer in this paper. I claim to be defending a good answer, one which is sensitive to changes that might be politically feasible as well as ethical.

CLASS PRIVILEGE

1. Class, class advantage, & class privilege

When philosophers talk about concepts—like 'class'—they typically try to capture as much of the ordinary understanding of those concepts as possible, although sometimes what they want to do with those concepts will lead them to propose revisions. A good place for us to start, then, is with the way 'class' is ordinarily understood. The question we're starting with is 'what is class'?

Traditionally, a British person was understood as belonging to one of three social classes: Upper Class, Middle Class, or Working Class. The upper classes were the aristocracy, the middle class were landowners, and the working class were those engaged in manual work. A recent BBC survey with over 160,000 respondents collected information about economic, cultural, and social capital, and concluded that there are now seven social classes in the UK: Elite, Established Middle Class, Technical Middle Class, New Affluent Workers, Emergent Service Workers, Traditional Working Class, and Precariat (Savage & Devine 2011).[5] The UK Office for National Statistics uses a division based solely on occupations, and they present an eight-class, five-class, and three-class grouping, commenting that only the three-class grouping should be taken to be hierarchical. UK Geographics presents a six-class occupational grouping, made on the basis of Occupational Code, Employment status, Qualification, Tenure, and Full-Time Equivalent, see (UKGeographics 2014)).

5. The total number of respondents was 161,458. These were mostly from England (86%) with small proportions from Scotland (8%), Wales (3%) and Northern Ireland (1%). 56% of respondents were men and 43% were women, the average age was 35, and 90% of participants described themselves as white. (These figures are not fully representative; the 2001 census put the proportion of white people in the British population at 81.9% (Office of National Statistics 2011), and the proportion of women as being slightly higher than men: 32.2 million women compared to 31 million men (ibid)).

These are four different ways of understanding class, all of which include *occupation*, some of which include much more. They distinguish between 3 to 8 class groupings respectively. They create both *relations* and *hierarchies* between groupings, because any given group stands in a particular relation to another, and the groups can be ranked in order of which has more and which has less of a particular good. For example, the Elite on the BBC understanding are at the top of the hierarchy when it comes to the possession of economic, cultural, and social capital, and they are better-off in relation to each of the six remaining class groups. Next, what is 'class advantage'? Advantage is a simple matter of being better-off. Only the class group at the bottom of the hierarchy—for example the Precariat on the BBC understanding—lacks class advantage. The rest are better-off than at least one other group. Groups in the middle of the hierarchy will be advantaged relative to some and disadvantaged relative to others. What we've got so far is a story about class that permits an understanding of class advantage. What we're missing is a moral dimension. Does it matter if some classes of people are advantaged? Let me first explain why that's missing, and then go on to extend these initial suggestions in a way that makes class-based advantage morally objectionable.

If we care about equality *per se*, then these facts about social hierarchies and relations of advantage and disadvantage will be enough to start talking about what these people owe to each other. But there is a strong tendency in contemporary liberal political philosophy to think that *some* inequality can be permissible. Some defend this as being necessary to incentivize greater productivity, creativity, or entrepreneurship in society, which in turn can 'trickle down' to make the worst-off better off; some see it as an appropriate response to social contributions that require different levels of skill, training, effort, stress, or responsibility. So long as inequality is permissible, then there's nothing wrong with the mere fact that there are social classes. So we need more than just the story about what groups there are, and which people are in which group. We need something that suggests *unfairness* or *injustice* in the fact that certain people are in certain groups.[6]

Of course, we can't go and look into the backstory for every person in the UK, to check how each has ended up in the group they're in, and whether this history involved any unfairness or injustice. But we can check for unfairness or injustice in a more general way. For example, we can look at data on the social distribution of par-

6. For other accounts of privilege which similarly look for morally objectionable aspects of advantage, but instead focus on race or gender, see (McIntosh 1989; Bailey 1998; Frye 1983).

ticular things—all and any of the things discussed earlier, like occupation, income, social capital, cultural capital, or indeed property holdings, honorifics, education—and check whether these things cluster in an improbable way according to particular traits or features. We should expect them to cluster according to each other: there will tend to be a correlation between education and income, for example, or between income and property holdings. What we shouldn't expect is them to cluster according to some feature of a person that should be *irrelevant from a moral point of view*.

This kind of approach is often taken when it comes to features like race and gender. For example, we might make a graph showing the distribution of income between people in the UK, and then we might check this distribution for clusterings by gender, which is to say, whether there are more men then women in the top income categories, and more women than men in the bottom income categories. If we observe this clustering, there *might*, of course, be a perfectly reasonable explanation. For example, it might turn out that more women than men are working part-time, and the higher-income jobs require a full-time commitment; or that more women than men have chosen occupations that come with lower levels of remuneration; or that the highest-income jobs are those that were historically the most exclusionary of women, and this has resulted in there being more women in junior positions (the men who are now in the most senior positions were junior at a time where there were few if any women in the companies).

As implausible as these explanations might be in the case of gender, the more general point is that there *can* be such explanations. While distributions of particular things might look at first glance to be clustered in a problematic way, this will not always turn out to be morally objectionable. The problem that remains is to say what the traits or features are that we look for when we want to check whether a social distribution of something like social capital reveals improbable clusterings. If the distribution reveals clusterings by gender, we might say there's gender privilege; if it reveals clusterings by race, we might say there's race privilege. What would a clustering by class look like? In other words, what is *class privilege*?

There are two very different ways to answer this question. The first involves taking a cue from research into other forms of social discrimination. Familiar forms include discrimination on the basis of race, and discrimination on the basis of gender. There are particular social markers and social signals of race and gender, and these can trigger stereotypes and generalizations about race and gender groups. A person who has negative beliefs about women may encounter a particular individual, read

off her appearance that she is a woman, and then apply those negative beliefs to her. For example, a man might believe that single women over 30 are desperate to marry and have children, meet a woman who signals sexual interest in him, and decide that what she wants from him is a tenure-track to marriage and family. They can also result in negative treatment, for example testimonial injustice where what a person says is less likely to be believed (Fricker 1999). Are there negative beliefs—stereotypes and generalizations—about some or all of the cluster of features we've identified as determining or relating to social class?

It's clear that there are. Owen Jones catalogues a number of these in his book *Chavs*, with its revealing subtitle 'The Demonization of the Working Class' (Jones [2011] 2012). The most pervasive of these is perhaps that instead of extending sympathy to those in Working Class groups whose industries were destroyed under Thatcher, leading to high levels of unemployment and desperation, many in the Middle Class groups believe that unemployment or dependence on benefits is a *preference*. But there are many different features that might act as markers or signals of class. Consider employment, where initial selection of candidates works through CVs. Both *names* and *addresses* may signal one's class group, as they have been found to do for racial groups.

In their (2004) experiments on racial discrimination in the United States, Marianne Bertrand and Sendhil Mullainathan found that job applicants listing addresses in whiter, more educated, or higher-income neighbourhoods had a higher probability of being called to interview (Bertrand & Mullainathan 2004, p. 1003). They also found that those applicants with typically White names ('Allison', 'Brad') were 50% more likely to be called to interview than those applicants with typically African American names ('Aisha', 'Darnell') (ibid, p. 998 & p. 1012). In a field experiment of UK employers' social class discrimination, Michelle Jackson (2009) found that applicants with a name, school type, and interests associated with the social elite were more likely to receive a reply from employers, and that the single feature that made the most difference was name (Jackson 2009, p. 680 & p. 681).

Or consider face-to-face interactions. The following markers may all act as class markers: conventions of appearance (e.g. clothing, grooming), regional dialect, vocabulary, etiquette, and ability to converse on particular topics. If a person has negative beliefs about 'the poor', or 'the working class', and meets a person who has one or more markers of being in these groups, then she may apply her negative beliefs to this individual. This suggests that in just the same way as there can be gender- or race-based discrimination, on the basis of harmful stereotypes and generalizations about

gender or about race, there can be class-based discrimination, on the basis of harmful stereotypes or generalizations about class groups.

Jean-Claude Croizet and Theresa Claire conducted research designed to test whether the idea of 'stereotype threat'—namely that a person can be caused to underperform merely by being made aware of stereotypes that predict members of her social group to underperform, demonstrated in the case of both race and gender (Steele 1997)—can be extended to the case of class (Croizet & Claire 1998). They showed that it can. (Notice that this can go in both directions; to again use the BBC understanding, those in the Precariat might have negative beliefs about the Elite, and apply these to particular individuals on the basis of markers of appearance or social interaction that signal membership in the Elite. Stereotypes are bad for everyone, but the *effects* of class stereotypes are much worse for those at the bottom than those at the top).

The second way to answer the question looks instead at *determinants* of class position, rather than *markers* of class position. The way Croizet & Claire measured class is interesting for us. They equated class with socioeconomic status, and grouped students into either high or low socioeconomic status groups. But they did the groupings by accessing the students' administrative records, and looking at the *occupation of the parent who is the main provider for the student's family*. Students assigned to the low socioeconomic status condition had parents who were manual labourers, unemployed, and in administrative jobs, while students assigned to the high socioeconomic condition had parents who were managers, professionals, researchers, and college professors (ibid, p. 590).

This gives us a feature we might use: the occupation of a person's parents. If greater numbers of the people with high social capital or high occupational prestige have a parent with a high socioeconomic status job, and this correlation cannot be explained away, then we might well have *morally problematic class-based advantage*, namely, *class privilege*. This correlation is also demonstrated in a study tracking the relationship between fathers' incomes at the time their sons are born, and sons' incomes at age 30. Fathers' incomes are *highly predictive* of sons' incomes in the UK (see discussion in Pickett & Wilkinson 2010, p. 160 & p. 289). One study found correlations of between 0.4 - 0.6 for fathers' and sons' incomes, and between 0.45 - 0.7 for fathers' and daughters' incomes, where 1.0 is complete determination of one by the other (see discussion in Aldridge 2004, pp. 20-27; Paxton & Dixon 2004).

Thus we have two different ways of understanding class privilege, both which

capture intuitive features of our ordinary understanding of class, and both which involve unfairness or injustice; the first by way of explicit or implicit discrimination on the basis of class stereotypes and generalizations, the second by way of political and institutional obstacles to social mobility. The unfairness or injustice is what takes us from class and class advantage to class *privilege*. A person is privileged when she has markers which lead others to treat her favourably, or equivalently, lacks markers that lead others to treat her unfavourably. She is privileged when she has a parent in a better-off class group, which predicts that she herself will end up in a better-off class group.

II. Social *Group* Privilege, and Group *Member* Privilege

Are you privileged? You can refer to one of the sources mentioned earlier in this section, and depending which you choose, use your occupation, income, or other information to determine your social classification.[7] How do you know whether that group has privilege, and whether you have privilege as a member of that group? The second part of this question is more difficult than the first. You know the group has privilege when it's one of the better-off groups, and when rates of relative social mobility are low, because that means a major part of the explanation for why you're in that group is that one or both of your parents had a certain occupation, or income, or social status, etc.

(There's a further complication here: which groups count as 'better-off' and which count as 'worse-off'? There are many ways to divide the two, for example, taking the middle group (on the BBC understanding the New Affluent Workers) as the dividing line, so that the three above them are the better-off and the three below are the worse-off; or putting the line at a particular point we take to represent 'a life worth living'. We could say that every group higher in the hierarchy than the worst-off (on the BBC understanding, every group except the Precariat) counts as better-off. Conversely, we could say that only the best-off class group counts as better-off, because it is better-off than all the others. Any such decision would be more or less arbitrary. A *principled way* to divide the two would be to use a 'hypothetical baseline', which is to say, a non-actual distribution of goods against which we can compare the current distribution of goods—whichever goods we're interested in. A good baseline would be one where a lack of social mobility does not preserve class advantage and

7. E.g. http://www.bbc.com/news/magazine-22000973 accessed 23rd May 2016.

disadvantage. Take occupation, for example. We might compare the current distribution of occupations and its clusterings by class group to a hypothetical distribution of occupations distributed fairly, for example by chance, across class groups. Each individual could then compare her actual position to the likelihood of her hypothetical position. The closer these two are to each other, the less she would be 'better-off' in the actual distribution, and the further apart these two are, the more she would be 'better-off' in the actual distribution).

You also know the group has privilege when the goods whose distribution we're interested in are 'zero-sum', which is to say, there are a fixed amount of them, so that one person's having more means another's having less.[8] Social prestige is *not* zero-sum: we can imagine a world in which all different kinds of jobs are accorded respect and recognition. Income is not zero-sum either. But occupation might be, because there is a more or less fixed range of things that need to be done, and number of people needed to do them. In specific instances, it certainly is: a company wants to hire three new people, and nine people apply. Any one person's being hired means there is one less position available to the others. If the competition for the positions is fair, there's no problem. If three are unfairly or unjustly disadvantaged by features they have and the remaining six lack, then the remaining six have privilege, regardless of which of their number is hired.

It's this last thought that's the most difficult to make sense of, and takes us back to the second part of the question raised above. Are all members of a group privileged? The best way to unpack it is to consider four different individuals. Two have fathers belonging to the Traditional Working Class and two have fathers belonging to the Established Middle Class (again making use of the BBC understanding). Let's say these individuals are all daughters, and so have as much as a 0.7 chance of ending up in the same class group as their fathers. But that is not the same thing as their class group being *determined* by their fathers: they could be one of the 0.3 who shift between class groups.

Now imagine that two of the daughters, one from each group, *in fact* end up in the same class groups as their fathers. The daughter of the Established Middle Class parent experiences a wide range of opportunities which would not have been extended to her if she were not in that group, and she makes use of those opportunities. The daughter of the Traditional Working Class parent experiences a much narrower

8. See also the interesting discussion of benefiting 'at the expense' of another, in (Anwander 2005).

range of opportunities, a range which would have been more expansive if she were not in that group, and she makes use of those opportunities. And finally imagine that the remaining two of the daughters, one from each group, *in fact* manage to shift groups.

This happens entirely by luck: although the daughter of the Established Middle Class father has all the markers that would usually lead to favourable treatment, she happens to pursue an interest in which those markers do not translate into actual advantage. She experiences less opportunity for that reason, and as a result is 'downwardly mobile', which is to say, ends up in a class group lower in the hierarchy than her father. Symmetrically, although the daughter of the Traditional Middle Class father has all the markers that would usually lead to unfavourable treatment, she happens to pursue an interest in which those markers do not translate into actual disadvantage. She experiences more opportunity for that reason, and as a result is 'upwardly mobile', which is to say, ends up in a class group higher in the hierarchy than her father. Note that no one in this story *squanders* any opportunity, or does anything else that would make her responsible for where she ends up.

The daughter of the Traditional Working Class father who remained in the same group, and the daughter of the Established Middle Class father who shifted down between groups, might end up in a comparable situation. Likewise, the daughter of the Established Middle Class father who remained in the same group and the daughter of the Traditional Working Class father who shifted up between groups might end up in similar situations. If we were to consider each pair of daughters who have ended up with similar holdings of the relevant goods—occupation, income, social prestige, cultural capital, social capital—we might find it undesirable to describe one as having class privilege and the other as having class-based disadvantage.

Having a father in the Established Middle Class might have given the downwardly mobile daugher *better chances*, but those chances didn't materialize into actual holdings. Given that she's ended up in a comparable situation to a person we describe as disadvantaged, shouldn't we rather think she's disadvantaged too? And similarly, having a father in the Traditional Working Class might have given the upwardly mobile daughter *worse chances*, but those chances didn't materialize into actual holdings. Given that she's ended up in a comparable situation to a person we describe as advantaged, shouldn't we rather think she's advantaged too?

Against this thought, we can see that those in better-off class groups enjoy greater *security* over their positions. Even though the daughters of the Established

Middle Class *can* end up in class groups lower in the hierarchy, they are much less likely to; even though the daughters of the Traditional Working Class can end up in class groups higher in the hierarchy, they are much less likely to. Considering the differences between these four daughters allows us to see two things.

First, both daughters of Established Middle Class fathers have class privilege. The first daughter has *more* privilege than the second, because her likelihood of ending up well-off translated into actually being well-off, while the second daughter's likelihood did not. This will be important later in the paper, because I'll suggest that the class privileged ought to 'offset' their privilege. The first daughter will have more privilege to offset than the second, so the first will have to take on more cost in order to satisfy her obligations than the second (see discussion in §IV).

Second, neither of the daughters of the Traditional Working Class have class privilege. The first has more class-based disadvantage than the second, because her likelihood of ending up with roughly the same income as her father translated into her actually ending up with roughly the same income as her father, while the second daughter's likelihood did not materialize, and she in fact shifted into a better-off class group. This has the nice implication that even though she might end up in a class group which we'd think of as better-off, she doesn't have obligations to offset, because she's not class privileged. Privilege is determined not by mere membership, but by the backstory about membership and security in access to holdings.

The lack of social mobility that keeps the disadvantaged in place across multiple generations simultaneously keeps the advantaged in place across multiple generations. In other words: obstacles to equality of opportunity work out well for some. The idea behind there being unique obligations for beneficiaries of injustice is that those who benefit from unjust, unfair, or otherwise morally objectionable actions, events, states of affairs, histories, etc. have—or have *stronger*—obligations than others who are not beneficiaries (and who are not culpable, which I'll say more about in the next section). In the rest of the paper, I'll focus on the content and strength of these obligations.

III. Is Class Privilege Culpable or Non-Culpable?

When we culpably cause others to be badly-off, we will generally have very strong obligations to redress their situation. Some think our obligations can be so strong that they require us to take on cost *in excess* of the good it would do the badly-

off person to have redress made. We don't need to enter into that discussion here; but we can accept for the sake of argument that if the class privileged are culpable in their having class privilege, or in others' lacking class privilege, then their obligations will be substantially stronger than if they are merely beneficiaries of obstacles to equality of opportunity. In this section, I assess whether the class privileged can be understood as culpable, first for *having* privilege, and second for the ways in which they are *disposed* toward that privilege.

1. Culpability for having class privilege

I talked above about employers discriminating against job applicants. If this discrimination is explicit, then they're obviously culpable. If it's implicit, then they may not be (see discussion in Holroyd 2012). Certainly there may be specific individuals and groups who are culpable, either in what they have, or in what others lack. These individuals should certainly be held accountable. But is there anything we can say about class groups as a whole, or about most members of particular class groups?

It's easy enough to see what people might want from a concept of privilege that would require it to involve culpability. They might be interested in privilege understood as the receipt of stolen goods, or at least the possession of goods that are the legacy of colonial theft, violence, and injustice. While this is a plausible way to think of many goods in a country like the UK, it's not clear that it will allocate privilege along class lines (presumably *all* UK residents are privileged in this sense, rather than only those in the better-off class groups). They might be interested in privilege understood as profiting from political injustice against co-nationals, where e.g. failing to support mining communities in a transition to new employment industries is a political injustice, and those who profit are those whose interests are supported instead. The difficulty here is that government spending goes into a broad range of areas, so it's again unclear that those who profit from this injustice are those in the better-off class groups—even if it's clear that those people profit *more* than others.

They might be interested in privilege understood as complicity in a system designed deliberately to protect the advantage of some at the expense of others, or privilege understood as the sustaining, perpetuating, enabling, or upholding of that system. Supporting private schools by sending one's children to them might be a good example of this kind of complicity. They might be thinking of privilege as a

'club good', exclusion from which is a harm to non-members. Owen Jones has described the British Conservative Party in this way, as a 'coalition of privileged interests' (Jones 2012). Michael Monahan writes in *The South African Journal of Philosophy* that privilege requires active participation on the part of the privileged (Monahan 2014).

George Yancy, writing recently in the *New York Times*, argues that the privileged can be culpable simply in virtue of group membership: men in virtue of being members of the group of all men, white people in virtue of being members of the group of all white people. For him, to be white in a race-divided society is to be racist; to be a man in a gender-divided society is to be sexist. He gives a range of disparate justifications for this claim.

For gender, they include: that despite men's best intentions they perpetuate sexism; that men are complicit in industries that objectify women; that men see women through the male gaze despite intentions not to objectify women; that men share collective erotic feelings and fantasies which themselves are complicit in the degradation of women; that even if men fight against their sexism there will be moments of failure, and they will oppress women, so they cannot be fully innocent (Yancy 2015).

For race, his justifications include: that white people perpetuate racism; white people 'harbour' racism; white people benefit from racism; white people are part of, and reap comfort from, a system that gives them advantages while giving black people disadvantages; that white people are tied to forms of domination; that white people are wilfully ignorant of their ties to forms of domination; that white people have 'signed a contract' that guarantees them, but not others, social safety (Yancy 2015). Yancy says explicitly that not doing these things intentionally is not enough to free people from responsibility for them.

That's about as much as I can offer in favour of the culpability of *having* privilege. Against such culpability, I can offer two arguments. The first is that we can cause harm, yet not in a way that we are morally responsible for. Think of actions that fall below some threshold of moral accountability, such as individual instances of rudeness; or actions that are not known to be (or even merely widely recognized to be) harms, as individual greenhouse-gas emitting actions before 1990[9] were not; or actions taken with care and without malice that nonetheless by luck turn out to do damage to another person. In these kinds of cases we can be a cause (or part of

9. This is a generous date; some think it is much too late. See discussion in (Bell 2011).

the cause) of harm without meeting the stronger conditions required to be *morally responsible* for what we cause, like intention,[10] knowledge, foreseeability, ability to do otherwise, and so on (these vary between accounts).

The former kinds of case are particularly interesting because these 'below the threshold' actions can add up to social harms that are particularly damaging for those upon whom they fall, and yet ethicists struggle to account for any moral responsibility to remedy the harms (see e.g. Glover & Scott-Taggart 1975). Many see climate change in this way, because individuals' greenhouse gas emissions don't appear to be intentional causes of harm, but their cumulative effects involve great harm for a great number of people. Causing harm with one's actions taken alone is not the same as causing harm through one's actions taken together with many other people's actions (any action taken alone may not amount to a harm while the actions taken together may do), and causing a harm—whether alone or together with others—is not generally thought to be sufficient for culpability, if the other conditions are not met.

Perhaps what is in the background is the thought that the privileged could get together and take action to make it the case that they weren't privileged anymore, and the fact that they don't do so is an omission for which they are culpable. I have argued elsewhere against the culpability of these kinds of groups, on the grounds that they lack the control necessary to describe what they cause or don't cause as intentional actions or omissions (Lawford-Smith, 2015). Of course, Monahan and Yancy could be tacitly suggesting a revision to the requirements for moral culpability, following something like a 'strict liability' model as exists in tort law. But they'd need to make an argument for this, and as far as I have been able to find, they haven't.

The second argument against the culpability of having privilege is that we can be entirely uninvolved in the causation of harm, and yet be a beneficiary of it. There are plenty of cases of this kind of 'innocent' benefiting given in the literature on benefiting from injustice (see e.g. Butt 2007; Anwander 2005; and the papers collected in Page & Pasternak 2014). Without the relevant kind of causal involvement, there's no question of culpability. We might still be interested in these kinds of advantages, because they may yet establish obligations. Most of the discussion about benefiting from injustice has been about articulating the obligations of innocent beneficiaries, whether in cases of historical wrongdoing, or in the contemporary case of climate

10. At least, this is a condition for moral responsibility in what behavioural economists call 'WEIRD' societies: Western, educated, industrial, rich, and democratic. Significantly less importance is placed on intention in non-WEIRD societies. See discussion in (Barrett et al. 2016).

change (with the exception of Pasternak 2014 who takes up the issue of beneficiaries who might in various ways fail to be fully innocent, which I'll come back to in the next sub-section).

We already have well worked-out moral theories that tell us about the normative implications of those who cause harm, contribute to the causing of harm, are complicit in harm, and so on. For those in privileged class groups who count as harming in one of these ways, we can simply apply what we already know about those kinds of cases. For example, Christian Barry and Gerhard Øverland have done a lot of work on the responsibilities that follow from a person's contributing to harm (Barry & Øverland 2015); Chiara Lepora and Robert Goodin have provided a very thorough discussion of the ways a person can be complicit in harm, and what might follow from that in terms of holding the complicit responsible (Lepora & Goodin 2013). We don't, however, have a well-worked out moral theory that tells us about those who merely benefit from harm, especially in the more distinctive ways typically involved when we think about privilege.

The disagreement with Monahan and Yancy, and any others who think that having privilege is culpable, is over the proportion of privileged people who can plausibly be classified as culpably involved in the preservation of their own advantage, compared with the proportion who cannot be. My suspicion—as naïve and charitable as it may be!—is that when it comes to class, there are a great many people who cannot be classified as culpable, even if there are some who can. I'll proceed by focusing on the obligations of the larger non-culpable group.

II. Culpability for Dispositions Toward One's Class Privilege

Avia Pasternak is one of the few people who talks about ways of benefiting that are not fully innocent (Pasternak 2014). She makes a set of distinctions about the ways people can benefit that are useful in thinking about class privilege. She distinguishes (1) being unaware, and not reasonably able to be aware, that you're benefiting from wrongdoing; (2) receiving benefits passively rather than actively seeking them; (3) not desiring the benefit; and (4) not being able to avoid receiving the benefit without unreasonable cost (see discussion in her paper for references to authors who discuss the moral upshot of each of these). On this view, what it would mean to be truly innocent in one's class privilege would be to lack knowledge, desire, activity, and freedom in being privileged. Being implicated in one or more of these ways can, Pasternak

argues, change the content and strength of the obligations one has in virtue of one's benefiting.

Although it's an empirical question, it does seem plausible that there are a majority of people in the better-off UK class groups who would meet (2) and (3)—not actively seeking the advantages they have, and not desiring them. Presumably many people in the UK desire a more equal and more socially mobile society. (1) and (4) are more difficult. How many people meet (1)—being aware or not reasonably able to be aware that they're benefiting from wrongdoing—depends on how much of the UK's class-based inequality can be attributed to wrongdoing by specific actors, compared to how much is a matter of wrongs emerging from long-established policies, systems, institutions, norms, and so on, and the extent to which the former, if true, is common knowledge. The more that it is the former, and the more that is common knowledge, the more it is open that those who are aware of this will not be able to escape the charge of knowledge of benefiting from wrongdoing.

How many people meet (4)—not being able to avoid receiving the benefit without unreasonable cost—depends on facts about what the class-privileged person's secure opportunities are. Obviously no child can choose to walk away from her class-privileged parents in order to neutralize her starting position. But class-privileged parents often choose to game the educational system, and could obviously choose to send their children to state-funded schools instead of private schools. In that sense, many class-privileged children benefit from *others*' wrongdoing on their behalf, and some have argued that this can also be a ground of very strong obligations (see discussion in Goodin & Barry 2014). The upshot is that some class-privileged people will have stronger obligations than others depending on how many of these conditions she meets.

MORAL OBLIGATIONS FROM CLASS PRIVILEGE

Now that we have a decent handle on non-culpable class privilege, we can start to think about the obligations that those with class privilege might have. In a recent thread in Legal Theory, some have approached the issue of class privilege by proposing more stringent legal responsibilities for the 'gatekeepers of social advantage', including landlords, employers, and university admissions boards (see discussion in Khaitan 2015). This is a sensible *legal* approach, but as an articulation of the *moral* responsibilities coming from class privilege it won't be precise enough. Many who

happen to be gatekeepers will not themselves have privilege, and because even for those gatekeepers who do, measures designed to ensure fair equality of opportunity in access to advantage do not do anything to affect the privilege of the gatekeepers themselves. (At best, such measures will make access to advantage fairer in the future, and so affect who the future gatekeepers of advantage are). We're asking what obligations arise for whoever has class privilege, not for whoever controls the future distribution of privilege—although the former might end up being partially directed towards the latter.

Discussions about the obligations of beneficiaries have tended to focus on individuals' actual benefits (see e.g. Anwander 2005; Butt 2007; Goodin 2013; Goodin & Barry 2014; Haydar & Øverland 2014; Heyward 2014; Pasternak 2014). In developing the concept of class privilege, I have argued that people can have class privilege because they are more likely to receive benefits, even if they actually do not benefit. Those born to parents in better-off class groups are more likely to end up with more economic, social, and cultural goods than those born to parents in worse-off class groups.

Discussions about the obligations of beneficiaries have also tended to focus on discrete identifiable acts of injustice, from which specific kinds of benefits can be traced to specific individuals (although cf. Barry & Wiens 2014; Heyward 2014). In developing the concept of class privilege, I extended the scope of what people can benefit *from*, to cover structural injustice and social inequality.[11] And I extended the scope of what benefits can *consist in*, to include e.g. social and cultural capital. So unlike when benefits are held in the form of *money* or *material goods* to some discrete degree, the class-privileged person often cannot simply 'give up' her privilege and be done with the matter.

In fact, characterizing the obligations of the class privileged in the way others have characterized beneficiaries' obligations—for example to *disgorge* benefits (Goodin & Barry 2014, pp. 371-372), or to relinquish benefits to the subjective extent that you value them (Butt 2007, pp. 140-143)—would seem to misfire. Giving up benefits might mean cutting family ties, or throwing away educational opportunities, or walking away from challenging and rewarding jobs, or simply trading places with another class privileged person, or otherwise making oneself comparatively worse-

11. Existing accounts of obligations to address structural injustice are given on the basis of social interdependence, and do not assign unique obligations to those who do well out of the injustice. See discussion in (Young 2003).

off for no obvious net gain in advantage to someone else (merely trading places in an *ad hoc* way with someone worse-off doesn't obviously serve the cause of justice, because it might not be the best or the fairest way to compensate a person who has been unfairly disadvantaged, and it might not do anything to change the distribution of goods into the future).

It makes more sense to think in terms of the privileged having obligations to *offset* their privilege. I borrow the metaphor of 'offsetting' from discussions about climate change, where it is accepted that it would be very difficult for individuals in many contemporary domestic institutional settings to fully eliminate their greenhouse gas (GHG) emissions. 'Offsetting' captures the idea that emitting to some degree, although surely not to just any degree, is non-culpable (see discussion in §III), and yet can and should nonetheless be neutralized. Applied to class privilege, it suggests that it would be very difficult, and in some cases undesirable, for individuals to avoid *having* privilege, and yet that such individuals can and should neutralize the privilege that they have.

The explanation for why emitters should offset is that climate change threatens serious harm to persons, animals, and the environment; the explanation for why the class privileged should offset is that a lack of social mobility is a serious harm to those with parents in worse-off class groups. Worlds characterized by class privilege are bad, even if the people who have class privilege are not (necessarily) themselves bad. The idea of offsetting also makes the *object* of the obligations clear. When we offset our GHG emissions, we neutralize the harms they might otherwise do by removing or preventing GHG emissions elsewhere (e.g. by planting new trees, or preventing deforestation), and thereby make a small contribution to the mitigation of a major harm.

Similarly, then, when a person offsets her class-privilege, she must be attempting to neutralize the harm done by a *system* that distorts the distribution of goods in a society according to class. She can do this to maximum effect by channeling her offsetting to undermine the *source* of her class privilege and others' corresponding class-based disadvantage. In summary:

> What the class-privileged owe: *Members of class-privileged groups must offset their privilege by taking on costs in order to undermine the sources of class privilege and class-based disadvantage.*

The class-privileged can offset their privilege by taking on costs up to a point that is commensurate with their group-based advantage, either as time, effort, money, or other material resources (see also discussion in Barry & Lawford-Smith, ms.) What kinds of things count as ways to take on the relevant costs, and thereby offset privilege? The following are potential contributions that go to the source of class privilege (although this list is not exhaustive):

Challenge classist comments made in social situations

Show social respect and recognition to members of worse-off class groups

Engage in leisure activities where you are likely to interact with people from a range of different class groups

Take steps to collectivize into groups organized against class injustice (Young 2003)

Publicly boycott companies and corporations known to be involved in classist hiring or employment practices

Stand in solidarity with members of class groups experiencing discrimination or oppression (e.g. the working class) (Kolers 2014)

Undertake research into class-based social differences and whether they have alternative explanations, and share findings

Write to MPs, sign petitions, raise awareness about morally problematic class-based social differences

Encourage workplaces (your own and others') to use anonymized CVs when hiring to mitigate class bias

Encourage workplaces (your own and others') to add 'class' to existing diversity policies for hiring

Donate money, goods, or labour hours to charities and organizations working against class injustice

Vote for political parties whose platforms include action against class injustice

If you are a parent, send your children to state-funded schools (see also discussion in Swift & Brighouse 2016)

Notice while these would all look fairly uncontroversial as normative implications of race or gender privilege, they are slightly more surprising when it comes to class. In the first case, there's not yet much of a public consensus on what kinds of comments count as classist (as Owen Jones nicely demonstrates in *Chavs* (2012)). So we might need to add an item to the list, to first figure out exactly what kinds of comments count as disparaging, discriminatory, prejudiced etc. against people based on class. Social norms are part of what maintain class privilege, and norms can be partly eroded with minor social sanctions that threaten esteem, such as verbal challenge. Recognition can make a difference to those who have been unfairly disadvantaged, and greater integration between class groups can provide opportunities for giving recognition, and more generally for challenging and breaking down stereotypes.

Whether we need anonymized CVs (as it has been argued that we do to combat implicit race and gender bias), will depend on whether we can read class off names, addresses, educational institutions, etc.—more empirical work needs to be done on this. If class can be read off appearance, dialect, or other features made visible in social interaction, then there will be further issues of implicit bias to be faced up to (there may also be *explicit* bias, but as explained, this takes us back to culpability). The main struggle will be to change long-standing institutions and policies, those related to education perhaps chief among them.

Conscientious readers might wonder how *important* obligations relating to class privilege are, compared to other kinds of obligations we might have. This is a huge issue so I can't say much about it here. But the most important point is to make is that there's continuing disagreement about the extent to which it's permissible to show *partiality* in moral matters to those within one's own national borders. To the extent that it is, class injustice is one of the most prominent sources of injustice to persons in the UK, so the obligations I've outlined above will be very important. To the extent that it isn't, the conclusion will be very different. After all, the UK is a rich country, and there are many people in the world who are much worse-off than the worst-off here. For those who deny that it's permissible to show partiality to those within one's own national borders, rather than remedying class privilege in one's own rich country, it might be more important to take action against climate change, or against global poverty (for more on this question of making moral tradeoffs see Lawford-Smith forthcoming).

MAIN CHALLENGE

In this final section of the paper I want to address a challenge to this account of class privilege and the moral obligations that it comes with. What has been driving this whole story is the fact that some people are likely to end up in a worse-off position in a distribution of a given good, simply because of their social class group. I've argued that this can give particular kinds of obligations to those likely to end up in better-off positions simply because of their social class groups. But what if there's an explanation of people ending up in a worse-off position, that *isn't* simply the lack of social mobility in the UK? What if members of worse-off class groups are *complicit in their own disadvantage*, for example because they have internalized preferences against taking up certain kinds of opportunities?

(This challenge is not unique to class, it shows up in particular for gender as well. We might notice that there are fewer women than men in full-time employment,[12] and be concerned that this suggests morally objectionable gender inequality in the workplace. A critic might counter that large numbers of women *prefer* not to work, so they can care full-time for their young children).

I said earlier that the correlation between fathers' incomes and daughters' incomes is between 0.45 and 0.7 (where 1.0 would mean fathers' incomes fully determine daughters' incomes). Another way to think about this is that for every 100 daughters of fathers in worse-off class groups, between 45 and 70 of these daughters will end up with roughly the same income as their fathers, and between 55 and 30 of these daughters will end up with incomes significantly different to their fathers'. But notice that we're reading this data on the assumption that the daughters' preferences lead them to take up the opportunities they are presented with, so that the explanation of as many as 70% of the daughters of fathers in worse-off class groups ending up with the same income as their fathers is the UK's lack of social mobility. If the daughters' preferences lead them to *reject* some or many of the opportunities they're presented with, then their disadvantage will have an alternative (or additional) explanation.

There is at least anecdotal evidence in the UK to suggest that at least some members of worse-off class groups have been complicit in their own disadvantage,

12. In the UK, roughly 90% of men aged 28-44 are in full-time work, compared to roughly 70% of work (these numbers change slightly for different age groups). (Office for National Statistics 2013, p. 5).

for example by internalizing preferences against taking up opportunities that would provide more social mobility. Examples include being disposed against higher education, or disposed in favour of certain kinds of manual or routine occupations which generally come with less social standing and less remuneration. *To the extent that this is true*, a member of a worse-off class group could be making a free choice to adopt the norms or values of her class group, in which case she is not *only* being disadvantaged by a lack of social mobility, but is also determining her own disadvantage.

I say 'could' rather than 'would' because it's not clear whether we should see this as a free choice. It matters whether the choice is made reflectively, with knowledge of what is at stake. Preferences can fail to count as genuine when they are the result of coercion or social conditioning. Group identification dynamics can be like this: others ascribe negative characteristics to a group, and members of the group adopt and affirm these characteristics in order to reclaim social esteem; or the group itself adopts certain values, perhaps in opposition to other groups, and conditions them into new members (in particular children). But in those cases the preferences do not *explain away* the disadvantage; they're part of the disadvantage.

What is tricky about this challenge is that it puts us between a rock and a hard place. We could agree that those in worse-off class groups sometimes have preferences that lead them to reject opportunities, but say that these preferences are coerced or conditioned, so that we are not forced to agree that they're complicit in their own disadvantage. Only genuine preferences, those endorsed reflectively and with knowledge of the consequences, could make them complicit. But there's something uncomfortable about looking at someone's preferences and telling her that they're not her *real preferences*—that she would prefer different things if she hadn't been conditioned by her class group to want those things. Doesn't it add insult to injury to tell those who deny the value of education and prefer to make an earlier start in the labour market that they're simply mistaken about what's good for them?

On the other hand, if these preferences are genuine, then it's hard to see how the disadvantage counts as *morally objectionable* at all. The challenge from earlier was to move from class and class advantage to class privilege, which we did by locating unfairness or injustice in the backstory of who got to be in which class groups. If those who end up in worse-off class groups are there because of the choices they made, rather than the opportunities that were not made available to them but were made available to others, then unfairness or injustice drops out of the picture.

Must we choose between adding insult to injury, and denying that there is class

privilege? In fact we can squeeze out of this difficult position entirely. It matters whether the disadvantage of the person in the worse-off class group is 'overdetermined', which is to say, caused by two different things either one of which would have been sufficient. If the *sole* cause of the disadvantage is her own choice—if her preferences are genuine, she prefers a job that comes with less social prestige and less remuneration, and *if* she had preferred differently then she could have ended up in a different job—then we're forced to deny that there is class privilege. The disadvantage is not morally problematic; so there's no injustice or unfairness in the backstory about the distribution that compromises the advantages; so there's no class privilege and corresponding obligations.

But if her choice is only *one of the* causes, then we neither have to deny that her preferences are genuine, nor give up on the idea of class privilege. Whether her preferences are genuine or not, the fact remains that were she to have chosen differently, she would *still* have been disadvantaged. Her disadvantage is overdetermined because the external obstacles to social mobility remain in place. (One thing this does imply, though, is that undermining obstacles to social mobility might not be *sufficient* to equal opportunity in class-group determination. If people have preferences that lead them to reject particular kinds of opportunities, we might see similar patterns to those we see now, even in a society with full equality of opportunity.)

In summary, the lack of social mobility in the UK causes inequalities between people that are not solely a matter of individuals' choices. Those who do well out of these inequalities are privileged. Even when they are not culpable for having privilege, or for the dispositions they have toward their privilege, such people have obligations to take steps to address this inequality. One effective and politically achievable way for them to do so is through offsetting their privilege in one or more of the ways suggested above. Offsetting gives the privileged a concrete way to address a serious moral problem in their own society. While I have focused this discussion on class inequality in the UK, none of the ethical issues are restricted to the UK context. So this discussion should be useful to anyone worried about the ethical implications of class-based societies.

Acknowledgements: This work is funded by the European Commission. I'm grateful to audiences in Philosophy Departments at the University of Stirling and the University of Melbourne, and the MSPT Graduate Workshop at the Australian National University, for discussion on this paper; to Komarine Romdenh-Romluc, Ten-Herng Lai, RJ Leland, Sarah

Hannan, and Nancy Yang, for comments on the written version; and to Dominic Wilkinson, two anonymous reviewers, and four anonymous editors of the Journal of Practical Ethics for comments and suggestions.

REFERENCES

Aldridge, S. 'Life Chances and Social Mobility: An Overview of the Evidence', Cabinet Office, Prime Minister's Strategy Unit (2004). Online at http://www.swslim.org.uk/documents/themes/lt10_life-chances_socialmobility.pdf accessed 23rd May 2016.

Anwander, Norbert. "Contributing and Benefiting: Two Grounds for Duties to the Victims of Injustice", *Ethics and International Affairs* 19 (2005), 39-45.

Bailey, Alison. 'Privilege: Expanding on Marilyn Frye's 'Oppression'', *Journal of Social Philosophy* 29/3 (1998), pp. 104-119.

Ball, Charlie. 'Most people in the UK do not go to university - and maybe never will', *The Guardian*, Tuesday 4th June, 2013. Online at http://www.theguardian.com/higher-education-network/blog/2013/jun/04/higher-education-participation-data-analysis accessed 23rd May 2016.

Barrett, H. Clark., Bolyanatzc, Alexander., Crittendend, Alyssa N., Fesslera, Daniel M. T., Fitzpatricke, Simon. Gurvenf, Michael., Henrich, Joseph., Kanovskyj, Martin., Kushnickk, Geoff., Pisorf, Anne., Scelzaa, Brooke A., Stich, Stephen., von Ruedenn, Chris., Zhaoh, Wanying., Laurence, and Stephen. 'Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment', *Proceedings of the National Academy of Sciences* (2016) [online first].

Barry, Christian. 'Global Justice: Aims, Arrangements, and Responsibilities' in Toni Erskine (Ed.) *Can Institutions Have Responsibilities?* (Basingstoke: Palgrave, 2003).

Barry, Christian. & Kirby, Robert. 'Scepticism about Beneficiary Pays: A Critique', *Journal of Applied Philosophy* 32/4 (2015) [early view].

Barry, Christian. & Lawford-Smith, Holly. 'On Satisfying Duties to Assist', manuscript at 25th May 2016.

Barry, Christian. & Øverland, Gerhard. *Responding to Global Poverty* (Cambridge: Cambridge University Press, 2015).

Barry, Christian. & Wiens, David. 'Benefiting from Wrongdoing and Sustaining Wrongful Harm', *Journal of Moral Philosophy* (2014), pp. 1-23.

Bell, Derek. 'Global Climate Justice, Historic Emissions, and Excusable Ignorance', *The Monist* 94/3 (2011), pp. 391-411.

Bertrand, Marianne. & Mullainathan, Sendhil. 'Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labour Market Discrimination', *The American Economic Review* 94/4 (2004), pp. 991-1013.

Brighouse, Harry. & Swift, Adam. *Family Values: The Ethics of Parent-Child Relationships.* (New Haven: Princeton University Press, 2016).

Butt, Daniel. "On Benefiting From Injustice", *Canadian Journal of Philosophy 37* (2007), 129-152.

Croizet, Jean-Claude., & Claire, Theresa. 'Extending the Concept of Stereotype Threat to Social Class: The Intellectual Underperformance of Students from Low Socioeconomic Backgrounds', *Personality and Social Psychology Bulletin* 24/6 (1998), pp. 588-594.

Fricker, Miranda. 'Epistemic Oppression and Epistemic Privilege', *Canadian Journal of Philosophy* 29/1 (1999), pp. 191-210.

Frye, Marilyn. 'Oppression', in *The Politics of Reality*. (Freedom, California: Crossing Press, 1983), pp. 1-16.

Glover, Jonathan and M. Scott-Taggart. 1975. 'It Makes No Difference Whether or Not I Do It'. *Aristotelian Society Supplementary Volume* xlix, pp: 171–209.

Goodin, Robert. & Barry, Christian. 'Benefiting from the Wrongdoing of Others', *Journal of Applied Philosophy* 31/4 (2014), pp. 363-376.

GOV.UK. 'National Minimum Wage and National Living Wage rates' (2016). Online at www.gov.uk/national-minimum-wage-rates accessed 21st May 2016.

HarrisInteractive. 'Firefighters, Scientists and Teachers Top List as "Most Prestigious Occupations" According to New Harris Poll', August 1st 2007. Online at http://media.theharrispoll.com/documents/Harris-Interactive-Poll-Research-Pres-Occupations-2007-08.pdf accessed 21st May 2016.

Heyward, Clare. 'Benefiting from Climate Geoengineering and Corresponding Remedial Duties', *Journal of Applied Philosophy* 31/4 (2014), pp. 405-419.

Holroyd, Jules. 'Responsibility for Implicit Bias', *Journal of Social Philosophy* 43/3 (2012), pp. 274-306.

Jackson, Michelle. 'Disadvantaged by discrimination? The role of employers in social stratification', *The British Journal of Sociology*, 60/4 (2009), pp. 669-692.

Jenkins, Stephen P. 'The income distribution in the UK: A picture of advantage and disadvantage', *Centre for Analysis of Social Exclusion* 186 (2015), pp. i-31.

Jones, Owen. *Chavs: The Demonization of the Working Class* (London: Verso, [2011] 2012).

Khaitan, Tarunabh. *A Theory of Discrimination Law* (Oxford: Oxford University Press, 2015).

Lawford-Smith, Holly. 'Climate Matters Pro Tanto, Does It Matter All-Things-Considered?' Midwest Studies in Philosophy, forthcoming.

———. 'What 'we'?' *Journal of Social Ontology* 1/2 (2015), pp. 225-250.

Lepora, Chiara. & Goodin, Robert. *Complicity and Compromise* (Oxford: Oxford University Press, 2013).

McIntosh, Peggy. 'White Privilege: Unpacking the Invisible Knapsack', *Peace and Freedom*, July/August (1989).

Monahan, Michael. 'The Concept of Privilege: A Critical Appraisal', *The South African Journal of Philosophy* 33/1 (2014), pp. 73-83.

Monbiot, George. 'Plan after Plan Fails to Make Oxbridge Access Fair. There is Another Way', *The Guardian*, 24th May 2010. Online at http://www.theguardian.com/commentisfree/2010/may/24/oxbridge-access-fair-top-universities accessed 1st November 2015.

Office for National Statistics, 2011 Census. *Key Statistics for Local Authorities in England and Wales*, 11th December 2012. Online at http://www.ons.gov.uk/ons/rel/census/2011-census/key-statistics-for-local-authorities-in-england-and-wales/rft-table-ks201ew.xls accessed 1st November 2015.

Office for National Statistics. Full Report: Women in the Labour Market, 25th September 2013. Online at http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/dcp171776_328352.pdf accessed 25th May 2016.

Page, Edward. & Pasternak, Avia. (Eds.). 'Special Issue: Benefiting from Injustice'. *Journal of Applied Philosophy* 31/4 (2014).

Pasternak, Avia. 'Voluntary Benefits from Wrongdoing', *Journal of Applied Philosophy* 31/4 (2014), pp. 377-391.

Paxton, W., & Dixon, M. *State of the Nation: An Audit of Injustice in the UK* (London: IPPR, 2004).

Pickett, Kate. & Wilkinson, Richard. *The Spirit Level: Why Equality is Better for Everyone.* (London: Penguin, 2010).

Savage, Mike. & Devine, Fiona. 'The Great British Class Survey - Results', *BBC Science,* 3rd April 2013. Online at http://www.bbc.co.uk/science/0/21970879 accessed 1st November 2015.

Steele, C.M. 'A threat in the air: How stereotypes shape the intellectual identities and performance of women and African Americans', *American Psychologist* 52 (1997), pp. 613-629.

Totaljobs.com. 'How to claim jobseekers allowance' (2016). Online at http://www.totaljobs.com/careers-advice/money-and-legal/how-to-claim-jobseekers-allowance accessed 21st May 2016.

UKGeographics. 'Blog: Social Grade A, B, C1, C2, D, E', *UK Geographics* 23rd February 2014. Online at http://www.ukgeographics.co.uk/blog/social-grade-a-b-c1-c2-d-e accessed 1st November 2015.

Yancy, George. 'Dear White America', *New York Times* December 24th, 2015. Online at http://opinionator.blogs.nytimes.com/2015/12/24/dear-white-america/ accessed 22nd January 2016.

Young, Iris Marion. 'Political Responsibility and Structural Injustice', *The Lindley Lecture 2003,* University of Kansas, May 5th 2003

Unjust Wars Worth Fighting For

VICTOR TADROS

School of Law, University of Warwick

ABSTRACT

I argue that people are sometimes justified in participating in unjust wars. I consider a range of reasons why war might be unjust, including the cause which it is fought for, whether it is proportionate, and whether it wrongly uses resources that could help others in dire need. These considerations sometimes make fighting in the war unjust, but sometimes not. In developing these claims, I focus especially on the 2003 Iraq war.

The laws of war, like the criminal law, have many ambitions. Here are three of the most important: to guide conduct, to guide courts in determining which conduct ought to be publicly condemned, and to guide officials in determining whom to punish. These ambitions sometimes come apart. We sometimes wish to condemn wrongdoers, but we do not wish to prevent them from acting wrongly, or to prevent others from doing the same thing. This seems paradoxical. If something is wrong, why not prevent and deter it? Surely it is better that the world contains less wrongdoing. Not always. Sometimes wrongdoing makes the world a better place: not better than it *could* be, were we able to encourage the wrongdoer to do what is best instead of acting wrongly, but better than it *would* be were the wrongdoing not to occur.

Consider:

> *Racists Three Options. The lives of fifty white people and fifty non-whites are in peril. Derek, a racist, is the only person capable of saving anyone. He has three options—*
>
> *Rescue no one;*

Rescue only the whites, causing Derek to lose his foot.

Rescue everyone, causing Derek to lose his foot.

Derek, let us suppose, is not required to rescue everyone. He would not act wrongly if he did nothing, for rescuing anyone comes at a very high cost to him. Nevertheless, it would be wrong for him to rescue only the whites. If Derek saves the whites he must save the non-whites as well. This is because if he chooses to save the whites, he can also save the non-whites at no extra cost to himself. If Derek is willing to bear the cost of losing a foot for the sake of rescuing the whites, he must also rescue the non-whites. This is so simply in virtue of the fact that if he could rescue 50 people at no cost to himself, he must do so. If he has chosen to rescue the whites at the cost of a foot, he can rescue 50 people at no additional cost to himself. That is sufficient to require him to do so.[1]

Now suppose that we can't get Derek to rescue everyone. Derek is inclined to rescue only the whites, and if he rescues the whites there will be no way of getting him to rescue the non-whites as well. What should we do? Well, whilst it is wrong for Derek to rescue only the whites, and it is not wrong for Derek to rescue no one, we would prefer it that he rescues only the whites to rescuing no one. In that case, we should encourage Derek to do the only thing that is wrong and rescue the 50 whites.

So here is an interesting fact about this case: rescuing only the whites and not the non-whites is the only wrongful act amongst Derek's three options. But that is also the act that we ought to encourage him to perform. Sometimes, we have good reason to encourage people to act wrongly. We can understand this as follows. It is because Derek has the option of rescuing the non-whites at no additional cost that rescuing the whites alone is wrong. But whilst Derek has that option, we lack it. Our options are either to cause Derek to rescue the Whites or to rescue no one. Between those options, rescuing the Whites is permissible and preferable. The fact that Derek has an option that we lack does not render our act of ensuring that the Whites are saved wrong.

In the light of this, consider how the law ought to respond to circumstances like these, assuming that they sometimes occur. We face a dilemma. On the one hand, we owe it to the non-whites and their families to condemn Derek for his failure to rescue the non-whites. But on the other hand, we don't want to deter those in

1. For a similar analysis of a closely related case, see (Parfit 1982, 131)

Derek's position from rescuing the whites. Passing a law warranting condemning and punishing Derek for his failure risks deterring people in Derek's position from rescuing anyone. After all, given that it would not have been wrong for Derek to rescue no one, we can hardly condemn and punish Derek for rescuing no one. If Derek will be condemned and punished for rescuing the whites, and he is not at all inclined to rescue the non-whites, Derek may well prefer to rescue no one. But that is the last thing that we want. We would rather see people in Derek's circumstances lose their feet if that helps to secure the saving of many lives. We would prefer it even more, of course, that those in Derek's position would rescue everyone. But in a world of racists that may not be a result that we can secure. We might conclude, then, that whilst Derek is liable to be condemned and punished for failing to rescue the non-whites, we ought not to actually condemn and punish Derek, nor to indicate that we will do so by prohibiting rescuing only the whites in law.

Fortunately, dilemmas of the kind just discussed are rare in the real world of domestic criminal law. Not so in war. It is not at all uncommon that people act wrongly in declaring war, and in participating in wars, and yet that we ought not to condemn and punish them for their actions. We might even have good reason to assist those who wrongfully declare, organise, and participate in wars to help them achieve their aims. This provides one reason why the laws of war ought to reflect the morality of war only crudely: we do not always wish to deter people from acting wrongly, if permitting or encouraging their wrongful actions serves our ambitions to protect people against being harmed.

I will explore these ideas by focusing on the Iraq war that commenced in 2003. Let us call those who invaded Iraq the Invaders. Let us call those fighting against the invasion the Resisters. Many people believe it was unjust for the governments of the Invaders to declare a war against Iraq and to orchestrate the invasion. The fact, if it is a fact, that this was wrong does not establish that it was also wrong for individual Invader soldiers to go to war. I will explore a range of reasons offered for the view that declaring and orchestrating the war was wrong, and consider both a) whether this makes it wrong for Invader soldiers to go to war; and b) whether we ought either to have assisted or deterred individual Invader soldiers from going to war.

The war in Iraq might be thought wrong because it was conducted for the wrong reason—to get oil rather than to prevent serious rights violations. I explore this consideration in Section I, concluding that even if the war was wrong for this reason it does not supply a reason against participating in the war. In Section II, I

consider whether the war was wrong because it was conducted with the intention of imposing democracy on a foreign country. I suggest that this is not itself a reason against going to war, and could not render it wrong. However, the war may have been wrong if it was conducted in order to establish unjust political institutions in Iraq that would better serve illegitimate western interests. Again, though, even if this rendered the war wrong, it would not supply a reason against participating in the war if the institutions imposed were less unjust than those they replaced. Section III is concerned with the question whether the war was wrong in virtue of the number of casualties it caused. This, I suggest, is the most powerful argument against the permissibility of the war. However, it also does not support the verdict that it was wrong for combatants to participate in it, as long as their own contributions are not disproportionate. Section IV considers an objection to the conclusions established in earlier Sections—that participating in unjust wars is typically wrong because doing so encourages future unjust wars. It is shown that this objection is not necessarily decisive, both because participating in unjust wars might not have these consequences, and because it is not always permissible to allow some people to die in order to deter the future deaths of others. Section V is concerned with a less familiar reason why going to war in Iraq might have been wrong—because the resources used for war could better be spent on non-military humanitarian aid. It is suggested that this may well have undermined the humanitarian case for war, but that this does not necessarily supply a reason for combatants not to engage in war. It may also not supply a reason to deter humanitarian wars, or participation in them.

Overall, there is no simple account of the permissibility of participating in unjust wars, or the wisdom of preventing such participation. If we are required to assess each individual act performed during a war on its own merits, and I believe that we ought to do this, whether a person acts wrongly in participating in a war depends on the particular contributions that he will make to the war. This can sometimes justify participating in unjust wars, including wars that lack a just cause, wars conducted for the wrong reason and disproportionate wars. And that is because the reasons that apply to those who decided to go to war, given the options available to them, need not be reasons that apply to potential participants in wars. Hence, sometimes it is permissible for individuals to join unjust wars, and sometimes we ought not to deter them from doing so even if doing so is wrong.

THE WAR FOR OIL

The main reason offered by Invader governments for going to war in Iraq was preventive. Political leaders of the Invader governments claimed that going to war was necessary to provide security against the potential threat that Iraq posed to the international community, either through military action or by assisting terrorists. Few believe that Iraq posed an *imminent* threat to other nations. But that fact may not count decisively against war, at least in theory if not in law.[2] What counts against this rationale for going to war is that war may well have been unnecessary to pursue the Invaders' defensive aims, even given a charitable judgement of the Invader's assessment of the likelihood that Iraq possessed weapons of mass destruction. At best, defensive aims could provide only part of the justification for going to war along with humanitarian goals.

Regardless of whether there were defensive *grounds* for going to war, though, some believe that the Invaders acted unjustly in going to war because they had the wrong intentions. Whilst Invader governments sought to claim that they had purely preventive aims, or pointed to a combination of preventive and humanitarian goals, many suspected that at least part of the reason why war was declared against Iraq was to secure cheap oil. Suppose that the Invaders went to war to secure oil rather than for defensive or humanitarian reasons. Would that render it wrong for the Invaders to declare war on Iraq?[3]

1) Intentions and Permissibility

It might seem obvious that if the Invaders went to war for oil, and not to achieve preventive or humanitarian goals, going to war was wrong. Not everyone, though, will draw this conclusion. For not everyone accepts that an action can be wrong in virtue of the bad intentions of the person performing that action. Some claim that the permissibility of warfare depends not on the intentions of participants, but rather on what they will in fact achieve. If the Invaders would prevent Iraq from attacking other countries and prevent the Iraqi government from killing their own civil-

2. For a good discussion, see (Buchanan 2007)

3. It is not essential to my argument here that intentions can be attributed to collectives, though I think it plausible that they can be.

ians, these facts can render the war permissible, even if the Invaders did not intend to secure these ends, and were only motivated to secure cheap oil.

One prominent scholar who endorses such a view—that all that matters is what one will in fact achieve, and not what one's intentions are—is Frances Kamm. For Kamm, if the Invaders went to war for cheap oil, we have good reason to criticize their characters and motivations. We also have reason to think that the *meaning* of the war was unappealing. But these facts do not make going to war wrong.

Kamm also considers the following question: does it make a difference to the permissibility of going to war that those going to war do so only on condition that they will secure preventive and humanitarian goals? For example, suppose that the invaders go to war in order to secure cheap oil. But they are willing to pursue this ambition only because securing it will also avert threats that Iraq would otherwise pose, and would prevent the killing of civilians by the Iraqi government. Does their greedy motive, in that case, render their action wrong? (see Kamm 2011, 119 - 124)[4] She concludes that acting for the wrong reason does not render the declaration of war wrong. It is not wrong for the Invaders to go to war to secure oil as long as the preventive and humanitarian causes that they might have fought for were sufficiently important to justify the war. Furthermore, she argues that declaring war may have been permissible even if the warmongers would have gone to war regardless of the achievement of preventive and humanitarian goals.[5]

I doubt that Kamm is right to think that permissibility is independent of intentions. The fact that the war could have been justified on preventive humanitarian grounds, if indeed it could, is insufficient to render their conduct permissible. The fact that these ends would be secured might render the war *justifiable*—it has the potential to be justified. But the acts of the Invaders were in fact justified only if they acted for the right reasons. If their acts of war were not justified, they were wrongful.[6]

There is a great deal to say about the difficult and important question whether and when a person's intentions can make a difference to whether her act was wrong-

4. Jeff McMahan is unsure whether the right intention is required to render a war just. (2005, 5)

5. She draws on the more general discussion in (Kamm 2007, ch. 5)

6. To explain the distinction between what is justified and what is justifiable more clearly: a pro tanto wrongful act is justifiable if a fact supplies a sufficient normative reason for a person to perform that act. It is justified if this fact was a motivating reason for the person who acted.

ful.[7] Here I will only respond to the two arguments that Kamm offers for her view. She considers: (Kamm 2011, 119-130)[8]

> *Weden Oil Case: Suppose it is permissible for some country called Weden to begin a war against Germany to stop its invasion of Norway and also its genocide of Norwegians. However, neither Weden nor any other country is interested in starting a war for these purposes, but Weden knows that if it does stop the aggression and genocide, Norway will favour Weden in the sale of its oil resources. Getting such resources is not an aim that could justify Weden in starting a war, but that it will get resources is also not a reason against starting a war that stops aggression and genocide. Suppose Weden intervenes and stops aggression and genocide, but it does so only in order to get access to the resources that Norway will grant it, and it would not have intervened had it not had this aim.*

Kamm believes that as long as Weden in fact stops the aggression and genocide, it permissibly goes to war. This is true even if stopping aggression and genocide did not motivate Weden in going to war. Furthermore, she thinks that it would not be wrong for Weden to go to war even if they would have been willing to go to war simply in order to get the resources in circumstances where they would not have prevented aggression and genocide (Kamm 2011, 128).

One potentially important difference between this case and the Iraq case is that Weden's intention—to be favoured by Norway in the sale of oil resources - is not a bad intention in itself. In contrast, it is plausible that if the Invaders went to war for oil, they did so to unjustly secure cheap oil for themselves when others were entitled to control the resources. Perhaps it is important whether one's intentions are inherently bad. (See, further McMahan 2009, 345).

But even acting on intentions that are not inherently bad may be insufficient to render Weden's actions permissible. Kamm offers two arguments to support her view. We can call the first *Scanlon's Argument*.[9] The argument draws on the idea that the primary role that the judgement that an act is wrong plays is in our practical reasoning. When we decide that an act is wrong, we also decide that we have decisive

7. I offer a more complete defence of the idea that intentions are relevant to permissibility in Tadros 2011, ch 6, 7, and 'Responses') and (Tadros 2013, 282-91) .

8. See also her discussion of Baby Killer Nation(Kamm 2011, 79-85).

9. As Kamm follows the argument in (Scanlon 2008). This kind of argument was also suggested in (Thomson 1999)

moral reasons not to perform that act. Following Scanlon, Kamm argues that in most cases, when we are deciding whether we ought or ought not to perform some act, we do not consider the intentions with which we will act. We consider other properties of the act, such as whether the act will cause another person to be harmed, or prevent harm.

This argument, though, does not seem decisive. One reason is this. When we determine how to act in cases where there are powerful considerations that might militate in favour of and against acting, we do not normally begin with a blank sheet of paper, and address all of the considerations together. Rather, we have a particular thing in mind that we want to achieve, and we determine whether it would be permissible to achieve that thing given the reasons against doing it.

For example, on seeing the Iraqi government killing its civilians, we might wish to do something to prevent this occurring in the future. However, preventing this will be very costly both to our own combatants, and to Iraqi soldiers and civilians. The question that we face is whether we have sufficient reason to do what we want to do, given the costs. When we reason like this, we have already determined the intentions with which we will act, for we are already focused on the goal of preventing the loss of life. We do not need to reason about the intentions we will act with. We start with an aim, and then consider whether pursuing this aim is to be done all things considered. If we determine that it is, and we decide, we form an intention to pursue the aim.

In the *Weden Oil Case*, though, Weden deliberates in a quite different way. Its question is not whether it would be justified to secure its aim of preventing genocide and aggression by going to war, but rather whether it would be justified in securing its aim to secure cheap oil by going to war. But if *that* is the question, we should ask how it could possibly justify its aim to secure cheap oil given that killing people is necessary to secure the aim.

It is true, of course, that good things—the prevention of aggression and genocide—will in fact occur if it goes to war. Because of these facts Weden's acts were *justifiable*—they were capable of justifying the decision to go to war. But Weden cannot appeal to these facts to justify its decision to go to war, given that these facts played no role in Weden's decision. Justifying a decision involves showing that the considerations that featured in that decision rendered it permissible. And if that is right, Weden performs unjustified acts of killing. Their acts of killing are unjustified because the facts that could have played a role in rendering the decision to kill justi-

fied did not play such a role. This draws on the plausible idea that a fact that plays no role in guiding a decision cannot justify that decision. It is a fact that could have justified it, but as it played no causal role in it, it did not do so.[10]

In other words, the way in which *Scanlon's Argument* assesses moral deliberation is very artificial. It does as though we begin with the question whether going to war is justified. But in deliberating about war we do not begin with the question whether to go to war. We begin with some other aims, such as the aim to prevent aggression or the aim to secure cheap oil. We then work out whether we would be permitted to secure these aims given the cost. Whether we are justified in securing these aims depends on how good the aims are, and how heavy the costs are in pursuing them. But this way of reasoning naturally brings intentions into the picture, for we are trying to determine whether we are justified in executing the intentions with which we begin.

Hence, if it was true that the Invaders went to war in Iraq only in order to secure oil, their acts of war were necessarily wrongful, for one cannot justify going to war to secure oil. This is true regardless of whether some other facts about the war, such as the fact that it prevented the Iraqi government killing civilians, would have been sufficient to justify going to war.

A second argument that Kamm offers can be called *The Rights Argument*. Kamm plausibly suggests that the normal reason why it is wrong to harm and terrorize people is that the rights of those people will be violated. However, Kamm suggests, people normally have rights that certain things be done to them. They do not normally have rights that what is done to them is done for the right reason.

Again, this argument misleads. First, notice that at least some of the people who will be killed in war have rights not to be killed. Even if a person has a right not to be killed, though, it does not follow that it is wrong all things considered to kill the person. Sometimes a person's rights are insufficiently important to make it wrong for another person to infringe those rights all things considered. If this is true, infringing the person's rights can be justified. But it will be justified only if the person infringing the rights is motivated by a consideration important enough to provide that justification.

Hence, it is misleading to ask whether a person has a right that others act on

10. This argument leaves open the possibility that it is sufficient for Weden to act on condition that the relevant facts obtain to render the decision permissible. If it acts on such a condition, the relevant facts do play a role in shaping the decision to go to war. I have some doubts about this being sufficient, but I leave the problem aside here.

certain reasons. The right question to ask is whether a person is justified in infringing the rights of another. The reasons for which she acts will determine whether she is. If she is not, she has wronged the person whose rights she has infringed. This is not because the person has a right that others act on certain reasons. It is rather because those infringing rights need to be able to justify their actions, and whether they can do so depends on the facts that featured in their deliberations about whether to do so.

ii) Participating in Wrongly Motivated Wars

Defensive or humanitarian reasons, that could have justified going to war, do not justify the war if these facts did not guide Invader decisions. Going to war may have been justifiable, in that case, but it was not justified. The question to be addressed now concerns the implications of this idea for the acts of soldiers who participate in the war.

Let us assume something that McMahan argues for persuasively: the fact that a soldier is commanded by her government to go to war cannot in itself normally make going to war justifiable (McMahan 2011, chs. 1 &2). If it is wrong to act in a certain way, acting in that way remains wrong even if one has been commanded to act in that way. After all, if it is wrong for me to act in a certain way, it is normally wrong for others to command me to act in that way. It would then be surprising if a command that it was wrong for another person to issue to me could transform what would otherwise be a wrongful act into a permissible act.

Nevertheless, if it was justifiable to go to war, the bad motivations of those declaring war may not make it wrong for soldiers to participate in the war. The reason is that the facts that wrongly motivated the Invader politicians in declaring war need not motivate individual Invader combatants who participate in the war. My government commands me to go to war for oil. If I decide to participate in the war, though, securing oil may not motivate me. If the war was justifiable on other grounds, say humanitarian grounds, I may be able to rely on those grounds in my own decision whether to fight.

For this reason, it is sometimes permissible to participate in a war that is fought for the wrong reasons, and hence to fight in a war that lacks a just cause, at least in one sense of that term. The idea of 'just cause', I should say, is somewhat contested. Whilst there was no just cause for the war in the sense that those declaring war did not act for the right reason, there may be a just cause in the sense that there was a

cause that *could have* rendered the declaration of war just. There are people who are liable to be killed as a result of their actions.[11] The 'objective' condition of a just war may have obtained, but not the 'subjective' correlate of it.[12]

To illustrate the idea that it is sometimes permissible to help a person to do something that they do for bad reasons, consider:

> *Wrong Reason: Harry launches a lethal attack on James. Debbie hates Harry. She attacks Harry, preventing Harry's lethal attack being completed. However, she attacks Harry only because she wants to kill him, and not to defend James, whose death she is indifferent to.*

Suppose her wrongful intention renders Debbie's action wrong. However, it would have been permissible for Debbie to attack Harry in the same way to defend James. Now suppose that I can save James only by assisting Debbie. Doing so is permissible. If I assist Debbie, Harry may complain that I will have assisted Debbie in violating his rights. But even if what Harry says is true, his complaint lacks force. Harry is liable to be killed to protect James. It is permissible for me to help Debbie in virtue of this fact even if Debbie is acting for other reasons. I will be motivated by my desire to protect James, and protecting James is what I will achieve. If the motivation of the Invader governments was the *only thing* that rendered the Iraq war wrong, participating in the war may well not have been wrong.

It is, of course, unlikely that a war conducted for oil will be identical in all respects to a war conducted for defensive or humanitarian reasons. The motivations of Invader political leaders will affect the course of the war—for example, they may prioritize securing the oil fields over sparing civilian lives. The post war actions of the Invaders will also be different than they would have been had they been properly motivated. This may render some conduct, even some well-motivated conduct, in support of the war wrong.

These facts may still not make it wrong for individual soldiers to join the war though. Suppose that the war overall prevented the Iraqi government murdering, torturing and raping Iraqi civilians. I contribute to the prevention of these acts, and

11. McMahan (2011, 27) emphasises this feature of the requirement of just cause

12. Stephen Neff (s2005, 50-51) suggests that these features of just war were traditionally distinguished. Just cause (justa causa) was the name given to the objective element. This is odd terminology, though, as the fact that some objective criterion was fulfilled could neither cause a war nor be a cause fought for.

that is my aim. In doing so I facilitate other Invader soldiers in securing Iraqi oil fields as a side-effect. I may not have acted wrongly all things considered. I have a reason not to facilitate other Invader soldiers to act wrongly. But facilitating these wrongful acts might be justified all things considered, given the contribution that I will make to defensive and humanitarian goals.[13]

Now consider the position of soldiers who act in order to secure the oil, but whose actions contribute, as a side-effect, to the humanitarian goal. If I am right that motivations are relevant to permissibility, these soldiers act wrongly. And yet we may not wish to deter them from joining the war. This is so for similar reasons to those outlined in the introduction to the permissibility of encouraging Derek to act wrongly in *Racist's Three Options*. These soldiers may act wrongly in joining the war to secure oil, but we would prefer them to join the war in order that they can assist us in pursuing our humanitarian goals. We may prefer this even if they are not acting in order to secure humanitarian goals. There may be no way of achieving our humanitarian goals without also facilitating the wrongful participation of soldiers in the war.

Hence, even though what these soldiers do is worthy of condemnation, given the reasons for their actions, we have strong reasons not to deter them from acting. A law that prohibited them from acting, were it abided by, may hamper the pursuit of humanitarian goals. It is typically more important that such goals are achieved than that wrongdoers are condemned. Although they are liable to condemnation and punishment for what they have done, we have some good reason not to condemn and punish them.

IMPOSING DEMOCRACY

One aim of the Iraq war was 'regime change'. Many people think that this could not supply a reason to go to war. Some think this on the grounds that it is wrong to impose Western-style democratic institutions on another country. The imposition of democracy, on this view, is another form of imperialism. This view is sometimes defended on the (at best) silly basis that the Iraqi people have a cultural antipathy to democracy, and that democratic institutions may justly be imposed on a people only if they have the appropriate cultural tendencies. There is no reason to believe that Iraqis have a cultural antipathy to democracy, and little reason to think that the

13. For incisive discussion of this issue, see (Bazargan 2011, 513)

justification of setting up democratic institutions depends on people being culturally orientated to those institutions.

Some cultural tendency towards democracy may be required in order to ensure that democratic institutions function. Furthermore, imposition of democratic institutions may itself generate instability because those institutions are seen as serving the interests of outside agents and not the population itself.[14] But if a people cannot create democratic institutions for themselves, the reasons against other people establishing them are, I think, quite weak.

In the context of Iraq, the view that we should be sceptical about the imposition of democracy is even more difficult to take seriously. The political institutions that the Iraqi people suffered under prior to the war can only be understood as imposed on them. Pre-war Iraqi political institutions do not represent the legitimate ambitions of the Iraqi people to shape their lives collectively. They certainly did not represent the ambitions of the Kurds and the Shiites, or of many Sunnis. Whilst imposed democracy may be to a degree unstable, and perhaps even illegitimate, because it is seen as serving the interests of outside forces, it is no less stable or legitimate than the tyranny that Iraqis had lived under for many years.

Furthermore, it is not obvious how *any* set of non-democratic institutions can be seen as representing the ambitions of the population in any reasonably large jurisdiction. We can imagine circumstances where a whole population endorses a non-democratic form of government. But in any reasonably large country there will be disagreement as to the proper shape of political institutions. Any set of political institutions in a country will be imposed on some by others. Suppose that it is important that political institutions represent the will of the people. Only democratic institutions will secure this value, because only democratic institutions can legitimately claim to represent the will of the people. It is difficult to understand what 'the will of the people' might be independently of the shaping of that will through political institutions, and it is difficult to see how that will can appropriately be shaped by non-democratic institutions in the real world. When non-democratic institutions govern a country, the vast majority of the population have no role in shaping the political arrangements that they live under.

Even if non-democratic institutions did reflect the will of a majority of Iraqis, we would have no reason to endorse institutions that foster abuse of basic rights. The invaders could justify the imposition of democratic institutions simply on the basis

14. Michael Walzer (2004, 68-69) gives this as a reason against intervention

that these institutions would do better than pre-war Iraqi institutions in protecting and promoting basic rights. As the Iraqi people have an enforceable duty to protect and promote these basic rights insofar as they are able to do so, it need not be shown that they wish to protect and promote them. Of course, there is a question whether democratic institutions serve these ends well, but it is difficult to imagine that even defective democratic institutions will have a worse record on this score than pre-war Iraqi institutions in the long term, at least given appropriate resources to provide peace and security.

Perhaps it might be argued, as Michael Walzer does, that the value of a set of political institutions depends on whether the population that is governed by them has secured them themselves, and hence that imposed democracy lacks value. (Walzer 2008, ch .6) But whilst the idea that the value of a set of political institutions depends to some degree on how those political institutions have emerged is reasonably attractive, it is very difficult to believe that this always provides a decisive reason against imposing democracy on other countries. A set of institutions that are imposed on a people may be less valuable than the same set of institutions would be had they been fought for and won by the people themselves. If there is some prospect of a population developing decent institutions for itself, this provides a reason against intervention. Whether this provides a decisive reason against intervention in fact, though, depends on the prospects of the population developing decent institutions of their own. It is, of course, difficult to know whether, and how quickly, such institutions might have been established in Iraq without military intervention.

Let us explore a more sophisticated version of the view that going to war to impose democracy on the Iraqi people was wrong. Although the political institutions that are imposed by the Invaders on the Iraqi people are likely to be preferable to the institutions imposed on them by the Resisters, it might be argued, they are still unjust. The Invaders imposed unjust institutions on the Iraqi people in order to secure their influence in the region, and to help them to have greater access to Iraqi oil.

Let us suppose, as is plausible, that this is true. Even if it would have been permissible to oust Saddam Hussein to create just democratic institutions in Iraq, it might be argued, it was wrong to oust Saddam to impose unjust institutions on the Iraqi people. That is true even if those institutions are preferable to those that existed in Iraq prior to the invasion.

It is true that we have reason to condemn the Invaders for imposing unjust

institutions on the Iraqis for their own interests, if they did so. But this fact would not provide a powerful reason against fighting on the Invader side in the war. When considering the question whom to support in a war, we unfortunately typically lack the option of supporting a group with even reasonably just ambitions. In most conflicts, no side fights for anything close to justice. When determining whether to fight and whom to fight for, the right question to ask is not normally which side is acting justly, but rather which side is likely to be the best of a bad bunch.

None of this is to say that imposing democracy could itself counts as a just cause for war. Perhaps if the imposition of democracy would prevent many deaths that would otherwise be caused, fighting for democracy might be justified. But democracy might not in itself be important enough to be worth killing people for. The imposition of democracy on Iraq may not count as a just cause because no one is liable to be killed simply for running a country in a non-democratic way.

Even if it is true that imposing democracy can never be a sufficient just cause for going to war, though, it may nevertheless provide a legitimate reason to go to war, and this may contribute to an overall assessment whether to go to war. For example, suppose that two states are perpetrating human rights abuses that would be sufficient to render it permissible to go to war against either. This could be done at very minimal cost to the invading state. It is feasible that democratic institutions might be set up in one of these states post war but not the other. That factor can determine which state ought to be attacked. Hence, even if democracy cannot provide a just cause for war, it can provide a legitimate motivation of those engaging in war.

Overall, I doubt that the attempt to instil democracy in Iraq was a powerful reason either for the Invaders to fight, or not to do so. The idea that we ought to spread democracy through the world by military intervention is problematic, not because there is something suspect about democracy, but because war is typically a disproportionate and ineffective way to pursue the aim of establishing democratic institutions. But at the same time, the fact that the Invaders claimed to be fighting for democracy does not provide a powerful reason against participating on the Invader side.

THE CASUALTIES OF WAR

A better reason to think that the Iraq war was unjust is that it caused far too much death and destruction, both to combatants and especially to non-combatants,

to justify the aims the Invaders were pursuing, or even aims that they could have pursued.

It is worth noting at the outset that a large number of the civilian casualties that occurred during the war were caused by the fact that the Resisters, when they were attacked, went to war with the Invaders. When the Resisters defended the territory they controlled, the number of civilians killed may well have become too large to justify the war—both because of deaths caused by Invaders and Resisters, and because of deaths caused by insurgents. Were Saddam Hussein immediately to have surrendered before numerous civilian casualties were caused, as he ought to have done, the war would have been closer to being justified. For, were that to have occurred, the Invaders could have improved the political institutions of Iraq without causing as much loss of life.

Even if the Invaders acted unjustly in going to war, the Resisters also acted unjustly in defending themselves. If the number of civilian casualties caused was the main reason why the war was unjust, the injustice of Invader acts of war was in virtue of the injustice of the Resister acts of war that Invader acts predictably gave rise to.

1) Responsibility for Casualties

Assuming that I am right that it was wrong for the Resisters to defend themselves against the Invaders, it might be tempting to conclude that the main responsibility for the deaths of non-combatants in the war lay with the Resisters. It might be thought to follow that these deaths cannot make it wrong for the Invaders to go to war. This is hard to believe. In deciding whether to go to war, significant weight must be given to lives that will be lost, even if those lives will be lost as a result of other people's wrongdoing.

A more plausible view treats deaths caused by intervening agents as less significant in the decision whether to go to war than deaths caused directly as a side-effect by combatants on the just side. Even this more plausible view is, I think, false. Sometimes, if I act in a certain way, and that leads other people to act wrongly, my responsibility for the harm caused by the wrongdoing is diminished. For example, suppose that I marry a person of another race. It might be predictable that racists will riot as a result. Even if this is predictable, I am not heavily responsible for the harm caused by the racists. Responsibility for the harm they cause lies with them and not with me.

Even though this is true in the example just offered, I doubt that it is more generally true that we can evade responsibility for deaths that others wrongfully cause as a result of our actions, even in part. If I act in a way that creates a new opportunity for others to act wrongly, and they act wrongly as a result, I bear very significant responsibility for the harm caused by their wrongdoing. For example, if I leave your front door unlocked, and it is predictable that thieves will steal your property as a result, I must take a great deal of responsibility for the theft. I cannot claim, in that case, that as the thieves are primarily responsible for the theft, I am significantly less responsible.

To see this even more clearly, consider:

> *Which Route: A nuclear reactor is about to explode. If it does, several thousand lives will be lost, including everyone in this example. I can get to the reactor and prevent the explosion by taking either the low road or the high road. If I take the low road, I will dislodge a boulder, which will crush X. If I take the high road, I will dislodge a different boulder, which a villain will then use to murder Y.*

I ought to take either the low road or the high road. Which ought I to take? On the view that deaths that result from my actions due to intervening wrongdoers are much less significant in my decisions than deaths I directly cause as a side-effect, I ought to take the high road. X's death, on this view, is more significant than Y's in virtue of the fact that I will directly cause X's death.

Now compare two other views. Some might think it makes no difference which road I take, and I should flip a coin to decide. Others might think that I ought to take the high road, in virtue of the fact that Y will be murdered if I take the high road, whereas X will not be murdered if I take the low road. I think that the least intuitive view of the three is the first view. I am unsure which of the two other views are to be preferred.

For a similar reason, I suspect that the Invaders must bear a significant amount of responsibility both for the deaths they cause, and also for the deaths that result from at least some of the wrongful actions that occurred as a result of the war. For example, deaths caused by insurgents arose because these insurgents were provided with a new opportunity to kill as a result of the war. Overall, civilian casualties that arose during war count powerfully against the permissibility of going to war, even if

they were caused by the wrongful acts of others, who bear full responsibility for their wrongdoing.

ii) Supporting Decisive Victories

Suppose that the humanitarian and defensive benefits of going to war could not justify the number of deaths that the Invaders caused. I think that this was the main reason to object to the Iraq war. If I am right, how should we assess the contributions that individual combatants made to the war? It might seem that if it was disproportionate for the Invaders to go to war, it was also typically disproportionate for individual Invader combatants to contribute to the war. This does not follow.

After the war began, the effects of the acts of any individual combatant on the war were typically modest. Suppose that I was deciding whether to participate in the war. If I did so I might at most have hastened victory. I could not have prevented the war from occurring, nor could I have prevented an Invader victory, by refusing to participate.

It is quite likely, then, that the main difference that acts of any Invader combatant, taken individually, made to the war was to hasten its end. Each combatant may have helped to ensure that victory was more decisive. Each combatant made no difference to whether or not victory was achieved. Given this, it seems that there are powerful reasons for each Invader combatant to participate in the war once it has commenced. The reason is that we ought to prefer a swift and decisive Invader victory to a long drawn out war. The number of casualties will only increase if the war is long and drawn out.

It may be argued, in response, that I have been presuming that it was certain that the Invaders would win the war. Perhaps at no point in the war was this certain. Even if this is true, it may not provide a decisive reason against participating in the war. First, putting aside considerations of deterrence and related considerations, we should prefer a drawn out Invader victory to a drawn out Invader defeat. Even if too many lives were going to be lost in the war to make it justified, we would prefer that if these lives are lost, they are lost in the course of securing the defensive and humanitarian aims that might have contributed to a justification of the decision to go to war.

Secondly, if it is permissible to act to support an Invader victory where that victory is certain, it is probably also permissible to support it where that victory is highly likely. The issue of risk is complicated, and is beyond the scope of this paper. I

doubt that the possibility that the Invaders might have lost was sufficient to render it wrong for Invader soldiers to participate in the Iraq war.

Against this it might be argued that although each combatant's actions cause the war to end more swiftly and decisively, and that is preferable, we should nevertheless doubt that each combatant is permitted to go to war. A reason for us to doubt this is that each combatant on the Invader side will kill particular individuals whom they are responsible for killing. Consider a particular combatant, who kills some non-combatants as a side-effect of his contribution to the war. The non-combatants that he killed may have survived the war had he not participated in it.

A full analysis of this complicated issue is beyond the scope of this paper. I doubt that it provides a decisive reason against Invader combatants going to war. One reason is that, even though the Invader combatant is responsible for particular deaths that he causes, each Invader combatant may have improved the *ex ante* prospects of survival of all non-combatants, including those who they actually killed. The Invader combatant can then rightly claim that they have no reason to believe that their actions will make this worse for those who might be killed during the war. Furthermore, the particular deaths that any combatant causes might be a proportionate side-effect of the actions that the combatant performs in the war, given the contribution that the combatant makes to ending the war swiftly. If a particular combatant helps to end the war more swiftly, lives will be saved overall. This may be sufficient to justify the deaths of any non-combatants that come about as a side-effect of his contribution.

PREVENTING FUTURE UNJUST WARS

In the previous section I offered some reasons for us to prefer a decisive, to a drawn out, Invader victory. Let us explore some objections to this view.

First, the war-making capabilities of the Invaders will be reduced if the war is drawn out. This will diminish their ability to engage in other unjust wars. Secondly, a drawn out victory might deter the Invaders from engaging in further unjust wars. As the war is drawn out there are more casualties and that will tend to discourage the Invaders from engaging in future wars, not least because it will be more difficult for political leaders to motivate citizens to accept going to war. It follows, some may argue, that there is some reason to fight against a country that engages in an unjust war even if we know that this country will win the war, and even though a greater number of civilian casualties will be caused as a result.

It is not at all unlikely that unjust aggressors will be more likely to refrain from aggressing in the future if they discover, through the harsh experience of war, that it is difficult to achieve swift and decisive victories. For example, had the Invaders scored a more decisive victory in Iraq, perhaps war with Iran would have been more likely. It would have been easier to persuade the American public that war with Iran would be just. And the US would have had greater resources to prosecute such a war. Yet war with Iran would almost certainly have been unjust.

Of course, the effects of a drawn out war may prevent the Invaders from engaging in just wars as well as unjust wars. There probably have been conflicts which Invader countries ought to have intervened in—the conflict in Rwanda being perhaps the most obvious example. Their experience of the drawn out post-war struggle against insurgents in Iraq will likely deter interventions in such conflicts in the future. I will consider whether humanitarian wars are typically just in a moment. But even if they are not we may have very good reasons to hope that they occur and to advocate for them if they do.

Furthermore, even if it is a good thing that the Invaders do not engage in future wars, because the wars they will engage in will tend to be unjust, it does not follow that it is permissible to refrain from joining the war for this reason. Much depends on why the Invaders will fail to engage in future wars.

There is little objection to a potential combatant failing to intervene because the military capacity of the Invaders will be reduced if the war is drawn out. But if the reason why it will be more difficult for the Invaders to wage war in the future is that people are horrified at the greater number of Iraqi civilian casualties, or Invader casualties, things are different. It seems wrong to allow an increase in the casualties to occur in the Iraq conflict because allowing these casualties to occur will horrify the general public, making them less likely to support future conflicts. To refrain from participating in a war for this reason would amount to acting on an intention that some suffer as a means to save others from the horrors of potential future wars. This would have been wrongfully to use the deaths of those in the Iraq conflict as a means to protect others.

To reinforce this conclusion, compare Warren Quinn's *Guinea Pig* case, in which I refrain from treating a person with a serious illness in order that I can learn more about the illness. Suppose that if I do this I can prevent other people from contracting the illness, and hence reduce the number of people who suffer from the illness overall (See Quinn 1993, 177 and Foot 2002, 92). Refraining from treating a person

for this reason seems wrong. For a similar reason we cannot justify making or allowing one war to become more horrific in order to prevent horrific wars in the future.[15] Were we to do so, we would wrongly exploit the suffering of some people in order to prevent others from suffering the horrors of war.

HUMANITARIAN WARS[16]

Recall the possibility that the war in Iraq was justified for humanitarian reasons. Suppose that the number of civilians and soldiers killed during the war was insufficiently large to render the war disproportionate given the number of lives saved. Suppose that this is so for the reason that the Resisters would have killed many people had they retained power in Iraq. I do not claim that this is true. But it is not completely implausible that it is true, especially given not only the tendency of Saddam Hussein's regime to kill Iraqi civilians, but also its tendency to engage in wars that were unjust and unnecessary, such as the lengthy war with Iran conducted between 1980 and 1988.

It may seem that if the number of lives saved was sufficiently greater than the number of people killed, the Iraq war was just. This does not follow. It is one thing to claim that the number of people killed is proportionate to the number of lives saved by killing them. It is another thing to claim that acting in this way is permitted when we compare the option of going to war with other things that we might have done with the resources that we expend on war.

I. Saving Victims of Injustice?

Evaluating humanitarian wars is a complicated affair. There are many considerations that help to determine whether such wars are justified. One is the cost that soldiers and other citizens of the intervening country will bear for the sake of preventing serious injustices in other countries. Another is whether those suffering from the humanitarian crisis welcome the intervention. Still another concerns whether it is permissible to kill innocent soldiers and civilians of the country to be intervened in

15. That is not to say that the means principle is without exception. For implications for going to war for punitive reasons, see, further, (Tadros 2014)

16. Some of the arguments in this section are similar to those of Jeff McMahan (2013-14)

to prevent the wrongdoing occurring. None of these considerations decisively rules out humanitarian war in every case.[17]

Nevertheless, humanitarian wars may well often be unjust for the following reason. A state must ensure that it fulfils its humanitarian duties generally. Humanitarian wars must be evaluated in that context. When we determine whether it is proportionate to go to war we should consider the claims that others have on the resources that are used in fighting the war—what we might call 'the war chest'.[18] There are many others in dire need of these resources, and the money ought to be spent on them.

Consider the fact that the government of any country must spend a certain amount of money on humanitarian aid of various different kinds. When we identify that amount of money, we take into consideration special responsibilities to avert threats that the country is responsible for creating, for example because it exploited citizens of other countries, used their natural resources, imposed unfair burdens on them through trade agreements and international organisations, or created environmental threats through polluting the atmosphere. This will leave a certain amount of money to spend on humanitarian aid that is not directed to addressing problems that the state under consideration is responsible for creating.

There is a limit to how much this country may spend. Spending more would violate the rights of its citizens to these resources. I make no claims about how large this amount is, though I am sure that it outstrips the foreign aid budgets of most Western states. How should the state spend these resources?

Let's start with the obvious suggestion - that it should spend its resources to save the greatest number of people that it can from the worst kinds of suffering. This suggestion is wrong. For in an effort to save the greatest number, the state might also kill some people. There is a more stringent prohibition on harming people than on failing to rescue people. Broadly speaking, then, the state should give priority to methods of saving people from harm in ways that do not harm others.

This provides a powerful reason to think that it is typically wrong to engage in humanitarian wars. The state should rather spend its resources preventing and curing illness and disease, for example by providing mosquito nets, clean water or access to essential medicines. Doing these things saves a great number of people from

17. For an excellent discussion of these factors and others, see (Fabre 2012, ch. 5) . Fabre does not focus on the issue that I explore below.

18. For further analysis to show that this is the right way of understanding comparative proportionality questions, see (Tadros 2011, ch.15)

death and serious illness, killing almost no one. Benjamin Valentino estimates that provision of medical aid is around 3000 times as cost effective as military intervention with respect to lives saved per dollar (see Valentino 2011, 60). As we will see, even if this radically overestimates the difference, there are decisive arguments against humanitarian intervention.

The most familiar argument against the view that we ought to spend all of the resources that are devoted to humanitarian causes on humanitarian aid rather than humanitarian intervention is as follows. We have a special obligation to prevent systematic wrongdoing by a state against some group of people. It might be argued that it is much more important to prevent wrongful killing by tyrants than it is to cure disease. Many people feel the force of this complaint when presented with tyrants who systematically attack their own people, as we have seen recently in both Libya and Syria.

There are two responses to this argument. First, it is not very plausible that there are much stronger reasons to prevent harmful wrongdoing than there are to prevent the equivalent non-wrongful harms. Here is an example that helps to demonstrate this. Suppose that a number of children have contracted HIV. An evil person has intentionally infected some of these children. Others have contracted the disease by misfortune. I have a stock of retroviral drugs. It seems abhorrent to push those who have been infected wrongfully to the front of the queue for treatment (see McMahan 2010 and Tadros 2011, chs. 5 & 6) . It is even more abhorrent to treat fewer people who have been infected as a result of injustice when we could treat more people who have been infected without the relevant injustice.

What this example suggests is that people's rights to an equal chance of being saved from harm are quite robust. Seemingly important differences in how the harm will come about are insufficiently important to justify departing from a practice which gives each person a fair chance of survival.

Secondly, even if we have more powerful reasons to prevent deaths that arise as a result of injustice than to prevent deaths that arise naturally, almost all humanitarian disasters arise from injustice. Almost all of the people who contract serious illnesses and diseases, or who cannot feed themselves, are victims of injustice. Their states, or others states, have perpetrated injustices against them, for example by failing to create just institutions, by violating obligations to provide humanitarian aid, by exploiting the natural resources in their countries, or by propping up dictators.

Perhaps it might be argued that humanitarian wars have long-term advantages

over other kinds of humanitarian aid. Through humanitarian wars the belligerents may be able to establish just institutions in currently unjust countries, improving the lives of citizens a great deal for the long term. If this is possible, it provides some reason to conduct humanitarian wars. It is worth noting, though, that the record of establishing just institutions through war is not especially good. It will often be just as likely that just institutions will emerge without anyone engaging in a humanitarian war. If a barbaric regime can be toppled and replaced with a democratic regime at relatively little cost of life in circumstances where this is unlikely otherwise to occur, humanitarian war may be permissible. But these circumstances are rare.

Alternatively, it might be argued that humanitarian wars have a powerful deterrent effect. Dictators will be less likely to perpetrate humanitarian abuses if they realise that doing so will bring with it the risk of invasion. This suggestion would have more force if resources were available to ensure that humanitarian intervention would follow predictably, swiftly and decisively in response to humanitarian abuses. We are, at present, a long way from this situation. Global policing of humanitarian abuses is much less effective than domestic policing of crime, and this is unlikely to change in the short term.

Perhaps it might be argued that even if it is better not to engage in humanitarian wars, a state does not act wrongly if it engages in such wars. Doing so may not be best, but it is not wrong. One reason why this might be thought true is that states have a right to shape their own identities. They can choose to shape these identities by picking the disasters that they should respond to.

I agree that, within some limits, states are permitted to shape their own identities by spending their resources on some projects rather than others. However, it is very difficult to justify killing many people, and saving the lives of far fewer people than could be saved, on this basis. The rights that citizens have to live in a state that can shape its own identity by making choices about which people to save seems insufficiently powerful to justify killing non-liable people, which is what inevitably happens in humanitarian wars.

II. Should Humanitarian Wars be Prohibited?

In the light of this discussion, do we have strong reasons to prevent humanitarian wars from occurring? I doubt it. Although these wars are often unjust, given other

things that the belligerents could achieve, we may sometimes have good reason to permit or promote them.

Notice that on the account that I have given, humanitarian wars are typically unjust on what we might call a 'comparative' rather than on an 'internal' basis. It is wrong to perpetrate them because others have claims on the resources used to pursue them, and not necessarily because the lives that are lost cannot be justified by securing humanitarian goals.

Given this, whether we have reason to prevent countries from engaging in humanitarian wars depends on whether the belligerents will pursue better options if they are prevented from engaging in humanitarian wars. International organisations ought to prevent states from engaging in humanitarian wars if doing so will lead them to spend the same resources on other kinds of humanitarian aid. For in that way, these international organisations could ensure that many more lives are saved. It would be wrong to prohibit internally proportionate humanitarian wars if doing so would not lead to the same resources being spent on humanitarian aid.

It is often easier to motivate people and states to support and engage in humanitarian wars to prevent genocide, or other seriously wrongful acts, than it is to motivate people to provide resources for other kinds of humanitarian aid (see McMahan 2010). Perhaps we ought not to deter humanitarian wars, in that case, even though it is wrong to engage in them. Deterring humanitarian wars would not motivate states to do what they ought to do - to spend the resources that they would have spent on humanitarian wars on other humanitarian projects. Given that, we might reasonably prefer that they engage in some kinds of humanitarian intervention. Hence, perhaps international law ought to permit 'internally just' humanitarian wars even if they are 'comparatively unjust'. It also ought not to deter or condemn those who participate in such wars.

CONCLUSIONS

A criminal law that prohibits most serious wrongdoing has relatively few costs if it is abided by. This is because in the domestic context a person's wrongful acts normally make things worse than they would have been had they not acted. The kinds of dilemma that I have been exploring in the context of war are rarely significant in a domestic context.

In war it is often very difficult to determine who is acting wrongly. Even if we

could do this, it is often difficult to determine whether we have reason to prevent them from so acting. Much more stringent, authoritative and well-enforced laws to prevent unjust wars and wrongful actions during war would have great advantages. They would prevent a great deal of unjust killing. But they would also have severe costs. They would hamper the ambition of those with humanitarian aims to achieve their goals. I do not argue that, all things considered, it is wrong to think that we should typically deter unjust action in war. But we should also be aware of the costs in doing so. Sometimes unjust wars are wars worth fighting for, or at least with.

Acknowledgements: I am grateful to the audience at the Ethics, Law and Armed Conflict Conference 2011, where a version of this paper was first presented, and to Helen Frowe and Jeff McMahan for helpful comments.

REFERENCES

Buchanan, A 'Justifying Preventive War' in H Shue and D Rodin *Preemption: Military Action and Moral Justification* (Oxford: OUP, 2007).

Bazargan, S 'The Permissibility of Aiding and Abetting Unjust Wars' (2011) 8 *Journal of Moral Philosophy* 513.

Fabre, C *Cosmopolitan War* (Oxford: OUP, 2012).

Foot, P 'Morality, Action, and Outcome' in *Moral Dilemmas* (Oxford: OUP, 2002).

Kamm, F *Ethics for Enemies: Terror, Torture, and War* (Oxford: OUP, 2011).

——— *Intricate Ethics: Rights, Responsibilities, and Permissible Harm* (Oxford: OUP, 2007).

McMahan, J 'Just Cause for War' (2005) 19 *Ethics and International Affairs* 1.

——— 'Intention, Permissibility, Terrorism, and War' (2009) 23 *Philosophical Perspectives* 345.

———'Humanitarian Intervention, Consent, and Proportionality' in N Ann Davis, R Keshen, and J

———*Ethics and Humanity: Themes from the Philosophy of Jonathan Glover* (Oxford: OUP, 2010).

——— *Killing in War* (Oxford: OUP, 2011).

——— 'Proportionate Defense' (2013-14) 23 *Journal of Transnational Law and Policy* 1.

Neff, S *War and the Law of Nations* (Cambridge: CUP, 2005).

Parfit, D 'Future Generations: Further Problems' (1982) 11 *Philosophy and Public Affairs* 113.

Quinn, W 'Actions, Intentions, and Consequences: the Doctrine of Double Effect' in *Morality and Action* (Cambridge: CUP, 1993).

Scanlon, TM *Moral Dimensions: Meaning, Permissibility, Blame* (Cambridge, Mass.: Harvard UP, 2008).

Tadros, V *The Ends of Harm: The Moral Foundations of Criminal Law* (Oxford: OUP, 2011)

——— 'Responses' (2013) 32 *Law and Philosophy* 241.

——— 'Punitive War' in H Frowe and G Lang *How We Fight* (Oxford: OUP, 2014).

Thomson, JJ 'Physician-Assisted Suicide: Two Moral Arguments' (1999) 109 *Ethics* 497.

Valentino, B 'The True Costs of Humanitarian Intervention' (2011) 90 *Foreign Affairs* 60.

Walzer, M *Just and Unjust Wars: A Moral Argument with Historical Illustrations* 4th Edition (New York: Basic Books, 2006).

——— 'The Politics of Rescue' in Arguing about War (New Haven: Yale University Press, 2004).

Oxford Uehiro Prize in Practical Ethics Winning Essays

In this special two- part series for the *Journal of Practical Ethics,* we present the winning essays from the 2014-15 *Oxford Uehiro Prize in Practical Ethics*.

For the prize, graduate and undergraduate students enrolled at the University of Oxford were invited to submit a short essay on any topic in Practical Ethics, with two winners from each category giving a presentation of their essay to an open audience as the deciding round for first and second places in the competition.

In this issue we present the runners-up from each category, Dillon Bowen (Graduate), and Miles Unterreiner (Undergraduate). The essays have been revised in the light of reviewer comments.

The prize is an annual event, and we hope to continue this series in future issues.

The Economics of Morality

OXFORD UEHIRO PRIZE IN PRACTICAL ETHICS 2014-15

DILLON BOWEN

Tufts University

ABSTRACT

Altruism is embedded in our biology and in our culture. We offer our bus seats to the disabled and elderly, give directions to disoriented tourists, and donate a portion of our income charity. Yet for all the good it does, there are deep problems with altruism as it is practiced today. Nearly all of us, when asked, will say that we care about practicing altruism in a way that effectively improves the lives of others. Almost none of us, when asked, can honestly say that we have made a serious effort to ensure that we are practicing altruism in a way that effectively improves the lives of others. Disparities like these are indicative of flaws in our cognitive architecture - biases which ensure that the traditional practice of altruism is incongruous with our own values. This disconnect between our values and our actions causes our altruistic efforts to help fewer people to a lesser extent than they otherwise could. I argue that traditional altruism is in need of reformation and defend a social and philosophical movement aimed at achieving this reformation known as effective altruism. The reason effective altruism is such a promising alternative to traditional altruism is its application of economic thinking to the realm of altruism and morality. An economist's mentality is, I suggest, a necessary instrument for bridging the gap between our values and our actions, allowing us to practice altruism in a way that more effectively improves the lives of others.

INTRODUCTION

People perform acts of altruism every day. When I describe an act as *altruistic*, I mean that the person performing the act (the *donor*) makes a personal sacrifice—perhaps in terms of time or money—for the sake of improving the well-being of another conscious creature (the *recipient*). In this context, we will find it helpful to narrow the definition of *altruism* to describe only those altruistic actions in which the recipient is not a member of the donor's family, friends, or community. For the purposes of this paper, an action can be altruistic only if the donor has little expectation that she will have a personal or economic relationship with the recipient. Altruism can be anything from holding the door for a stranger to donating a substantial amount of money to charity. Almost everyone, I wager, behaves altruistically from time to time —some of us on a daily basis.

The problem with altruism, as it is currently practiced, is that it is ineffective at improving the lives of conscious creatures. In what sense is the ineffectiveness of altruism 'problematic'? Instead of appealing to moral obligations or duties, I will argue that the ineffectiveness of altruism is problematic in the sense that most of those who practice altruism would, on reflection, prefer to do so more effectively. When we behave as *ineffective altruists*, we are therefore failing to behave in accordance with our own preferences. The alternative is an ethical framework known as *effective altruism*, which is most concisely described as "aiming to do the most good that one can". (Singer & MacAskill 2015, p.viii)

This paper is divided into four sections. The first gives a more rigorous definition and explanation of effective altruism. Following this, I explore the implications of effective altruism for population ethics, and show it to be a milder and more intuitive philosophy than its close cousin, classical utilitarianism. The third section explains how cognitive biases cause us to behave as ineffective altruists, and suggests that our preferences would be better served by practicing altruism more effectively. Finally, I draw an analogy between how we think about altruism and how we think about economics. As I hope to show, thinking of altruism economically will aid us in overcoming the cognitive biases that make altruism so ineffective.

EFFECTIVE ALTRUISM

Singer & MacAskill 2015 describes effective altruism as "aiming to do the most good that one can". (ibid) While this definition succeeds in its concision and popular appeal, it leaves something to be desired in terms of specificity. We might wonder, for example, what is meant by *doing good*, and if there are any bounds on the amount of time and money effective altruists should devote to doing good. I offer my own definition here in hopes that it will help clarify some of these questions.

Effective altruism is the belief that we should endeavour to spend whatever resources we plan to devote to valuable creatures who are unlikely to have a substantial impact on our lives in such a way as to maximize their aggregate well-being, provided we do not sacrifice anything else of importance in doing so.

Suppose I plan to donate $100 to charity, and that for some reason I have to choose between two charities—A and B. Both A and B provide deworming treatments for people in Kenya. For the same $100, A can deworm two people, but B can deworm only one. Assuming A and B have similar externalities, I ought to donate to the charity which provides deworming treatments for two people rather than one. All else being equal, effective altruism holds that we should improve the lives of as many people as we possibly can. Call this the *helping more people (HMP)* imperative.

Now imagine I am faced with a different choice of charities—C and D. Both C and D feed families in Uganda. For the same $100, C can feed a family for two months, but D can feed a family for only one month. Assuming C and D have similar externalities, I ought to donate to the charity which feeds a family for a longer period of time. All else being equal, effective altruism holds that we should improve people's lives to the greatest extent we can. Call this the *helping people more (HPM)* imperative.

Presented this way, effective altruism seems like a straightforward and appealing ethical philosophy. These are, of course, the easy cases. To think about more difficult cases, it will help to examine each piece of my definition in turn.

We should endeavour to spend whatever resources we plan to devote...

My view of effective altruism is weaker than what Singer wants to propose. Singer has argued, in previous works, that we should devote as much of our time and money to others as possible, stopping only when the marginal utility of keeping money for ourselves outweighs the marginal utility of donating money to others. (see,

e.g. Singer 1972) Though I strongly believe we ought to devote more of our time and money to helping others than we currently do, all I want to claim here is that whatever resources we would have spent helping others in any case should be spent in such a way as to maximize the aggregate well-being of *valuable creatures*.

...to valuable creatures...

Who is included in the set of creatures whose aggregate well-being we are trying to maximize? In other words, who should be the recipients of our altruism? I designate a set which I call *valuable creatures*. Who exactly is included in this set may depend on the donor's preferences. For example, a classical utilitarian would consider all beings capable of experiencing happiness and suffering—both those that currently exist and all those that could potentially exist in the future—as morally important. An anti-natalist, by contrast, values only creatures who currently exist and whose birth cannot be prevented. We may, as I do, wish to include nonhuman animals and artificial intelligences in this set, or we may not. I leave this category purposefully vague.

In order to avoid repeating the awkward verbiage *valuable creatures*, I will often refer to the recipients of our altruism simply as *people*. Please understand that this term is not meant to exclude nonhuman creatures.

...who are unlikely to have a substantial impact on our lives...

To reiterate a point I made in the introduction, the altruistic actions under consideration here are those in which the donor does not expect to have a personal or economic relationship with the recipient.

...in such a way as to maximize their aggregate well-being...

When evaluating the impact of an altruistic action, effective altruists care about 1) how many people it helps (HMP imperative) and 2) how much it helps them (HPM imperative). But what happens when these measures come into conflict? For example, imagine I have to choose between charities E and F, both of which fight malaria by providing long lasting insecticidal bed-nets to villages in Malawi. Charity E will use my $100 donation to provide bed-nets for two villages for one year. Charity F will use my $100 donation to provide bed-nets for one village for two years. E helps more

people, but F helps people more. Assuming the externalities of both these charities are the same, which should an effective altruist donate to?

To address questions like these, I collapse the measures of the HMP and the HPM imperatives into a single scale—aggregate well-being. If there are no further considerations that would weigh in favour of charities E or F, an effective altruist should be indifferent between them.

...provided we do not sacrifice anything else of importance in doing so.

However, we might think that there *are* further considerations which would allow us to choose between E and F. One could argue that E is a fairer charity, because it increases living standards of as many at-risk communities as it can. F is behaving unfairly, the argument goes, in providing a single village with two years of security given that children in surrounding villages are dying from malaria every day. If a particular donor, compelled by this line of reasoning, chose E over F, would that disqualify her as an effective altruist?

No. This is the purpose of the final clause of my definition. Most people profess to hold values which are not reducible to measures of well-being. If there are other important considerations weighing against aggregate well-being, it may be rational for a donor to prefer one altruistic action over another, despite them being equally effective. It may even be rational for a donor to prefer a less effective altruistic action over one which is more effective if these considerations are sufficiently compelling.

I hasten to clarify that this clause is meant to make room for ineffective altruism only when it is based on what a donor would rationally endorse as an *important* consideration. For instance, men tend to donate more generously to a charity when solicited by an attractive female. (Raihani & Smith 2015) Presumably the gender and aesthetic appeal of a charity solicitor does not qualify as an important consideration for most people, and therefore donating to an ineffective charity on this basis would be out of keeping with effective altruism.

Now that we have gone into some detail about what effective altruism is, we can discuss its implications for difficult cases—specifically its implications for two of population ethics' most obstinate problems—the repugnant conclusion and the non-identity problem. In doing this, I intend to show the plausibility of effective altruism in even the thorniest of philosophical issues, and distinguish it from its counterintuitive cousin, classical utilitarianism. Most of the theoretical objections I have

encountered to effective altruism centre around its ostensibly objectionable stance on population ethics, so it is important to set the record straight on this matter before moving on to pragmatic considerations.

EFFECTIVE ALTRUISM AND POPULATION ETHICS

The repugnant conclusion

One argument I frequently encounter against effective altruism runs like this:

If I accept effective altruism, I must accept the repugnant conclusion

I reject the repugnant conclusion

Therefore, I reject effective altruism

Just what is the repugnant conclusion, and why might we believe that effective altruism entails it? The repugnant conclusion (Parfit 1984) was first raised as an objection to classical utilitarianism, which holds that the one and only good is to maximize *aggregate* well-being. The objection attempts to invalidate classical utilitarianism on the grounds that it concerns itself solely with aggregate well-being and ignores *average* well-being. To see why we might desire an ethical philosophy that concerns itself with average well-being, imagine three worlds—A, B, and C. World A is home to only a few people (say, 10 people, or n=10), all of whom are extremely happy (whose level of well-being is 10, or u=10). By contrast, world B is home to very many people (n=100) whose lives are barely worth living (u=1). The inhabitants of world C also have lives that are barely worth living (u=1), but there are more of them than in world B (n=101). According to classical utilitarianism, we should be indifferent between worlds A and B (total utility=100), and prefer C to both of them (total utility=101). The repugnant conclusion is that, for any given world, a classical utilitarian will always prefer a world full of people whose lives are just barely worth living, so long as there are enough of them to offset the decrease in average happiness. Surely, the argument goes, we must reject the repugnant conclusion, and therefore classical utilitarianism.

Effective altruism is similar to classical utilitarianism in that it advocates max-

imizing the aggregate well-being of valuable creatures. In fact, classical utilitarianism is a form of effective altruism. The concern is that, by focusing only on aggregate well-being to the exclusion of average well-being, effective altruism makes the same mistake classical utilitarianism does. However, there is an important difference between effective altruism and utilitarianism which makes effective altruism compatible with a rejection of the repugnant conclusion.

Recall that effective altruism holds that we should maximize the well-being of 'valuable creatures', while being purposefully vague about which creatures are included in this set. A classical utilitarian has a precise view of which creatures are morally important—all of them, including all creatures alive today and which may potentially exist in the future. Even if a classical utilitarian would prefer to prevent someone from being born—say, a child who would have a debilitating illness with a high mortality rate in the first years of life—she would still consider this child a valuable creature. If it were possible, the classical utilitarian would rather see this child born and live a happy, healthy life.

But effective altruists are not committed to adopting such a broad set of valuable creatures. Take, for example, average utilitarianism, which holds that the one and only good is to maximize the average well-being of existing creatures. To the average utilitarian, the set of valuable creatures consists of those who will have a positive impact on average utility and those whose existence cannot be terminated or prevented without diminishing average utility by an even greater amount. Imagine we live in a world with a few (n=10) very happy people (u=10). Further imagine that one couple is considering having a child whose life, for whatever reason, will barely be worth living (u=1). An average utilitarian would prefer that this couple refrained from having a child.

By contrast, a classical utilitarian would prefer the couple *did* have their child. After all, it will increase aggregate utility, if only by a small increment. The difference of opinion between average and classical utilitarianism results from how they view the set of valuable creatures. Since the child's life diminishes average well-being, the average utilitarian considers it morally important if and only if its existence cannot be prevented without an even greater decrease to average utility. But the classical utilitarian views the potential child as morally important whether or not it is actually born. Whereas the average utilitarian would view the couple's choice not to have the child as excluding it from the set of valuable creatures, the classical utilitarian would view this choice as diminishing the child's utility from small to zero.

Is average utilitarianism a version of effective altruism? Yes it is. For any finite set of valuable creatures, average utility is maximized when aggregate utility is maximized. Does average utilitarianism avoid the repugnant conclusion? Again, the answer is yes. An average utilitarian would prefer world B (n=10, u=10) to worlds A (n=100, u=1) and C (n=101, u=1), and in all cases prefers a world with fewer, happier people to a world with more people whose lives are barely worth living.

Bringing the discussion back to the larger picture, I should add that I am not an advocate of average utilitarianism, which yields many counter-intuitive conclusions of its own. I bring up this ethical view because it is an example of how we can be effective altruists and still reject the repugnant conclusion. Furthermore, we can see that the way to do this is by limiting the set of valuable creatures. Effective altruism entails the repugnant conclusion if and only if we consider all people currently alive and all people with the potential to be born morally important. But such a position is not logically entailed by effective altruism.

The non-identity problem

Another objection to effective altruism which similarly relies on population ethics considerations, relies on the non-identity problem[1]:

> *If I accept effective altruism, I must accept that I can be morally blameworthy for actions which are not bad for anyone*
>
> *I reject the idea that I can be morally blameworthy for actions which are not bad for anyone*
>
> *Therefore, I reject effective altruism*

The non-identity problem involves a conflict of intuitions. At first, it seems that an action can only be bad if it is bad *for* someone. An action that neither harms nor is in any way bad for someone seems as if it cannot be wrong. But now consider a 14-year old girl who is thinking of having a child. If she decides to go through with the pregnancy, her child would live a worthwhile life. However, given her age and socioeconomic status, she will not be able to provide as good a life for her baby as she

1. The non-identity problem was first discussed in Parfit 1984.

would be able to if she waited until, say, age 26 to start a family. The intuition here is that getting pregnant at her age would be wrong.

But supposing the girl's own well-being is not affected, for whom would this action be wrong? The tempting answer is to say that it is wrong for her child. Yet the child she would have at age 14 would live a worthwhile life, and the child she would have at age 26 would be a fundamentally different person, having a different genetic structure and growing up in a different environment. So postponing pregnancy would not so much make life better for her child as it would change the identity of her child. In other words, the decision to wait to have a baby would not make life better for the child she would have had, but rather would create a different child who would lead a better life. Having a child at age 14, then, is not bad *for* anyone.

The same line of reasoning can apply to all future people. Many of the ways to 'improve' the lives of future people do not improve the lives of the future people who would have existed anyway, but rather create a different set of future people who would lead better lives. There may be very few ways to improve or diminish the quality of life of future people without changing their identities. Combine the fact that future people have undetermined identities with the moral principle that actions can only be good or bad if they are good or bad *for someone*, and we might conclude that the moral obligations we have to future people are highly limited.

What does this have to do with effective altruism? The idea is that most effective altruists include future people in their set of valuable creatures, and believe that our actions can be good or bad in relation to future people. But such a view contradicts the moral principle that actions can only be bad if they are bad *for someone*.

I believe the best way to respond to this objection is by referencing a point I made in the introduction to this piece. Ineffective altruism, I said, is problematic in the sense that it violates our preferences. When people behave as an ineffective altruists, I do not necessarily think they are violating a moral duty so much as behaving in a way I disapprove of, and a way they themselves would probably disapprove of in light of their own values. We could censure a 14-year old girl who decides to have a child on similar grounds. It may not be the case that she is violating a moral duty, but it is the case that we would prefer she made a different decision and, on reflection, she probably would as well.

This standard applies to considerations of future people in general. Imagine you can press either a red or blue button. The red button will determine that, a century from now, the world will be filled with extremely happy people. The blue button will

determine that, in the same amount of time, the world will be filled with the same number of people whose lives are only moderately happy. Further suppose the identities of the people in both these worlds are fundamentally different. If someone chose to push the blue button, it would be entirely reasonable to conclude that she has done something bad. And what makes this action bad is not necessarily that it is bad *for someone*, but that it creates a suboptimal world as judged by our values.

There are two senses of *bad* at play here. One sense implies a violation a moral duty and thereby moral blameworthiness. The other implies a violation of our preferences, and thereby social disapprobation. I would argue for an interpretation of effective altruism in which a disregard for future people is bad in the latter sense but not necessarily the former. Effective altruism does not imply moral blameworthiness for actions which are not bad for anyone, but rather strongly suggests that, in light of our own values, we should perform actions which maximize the aggregate well-being of future as well as existing people.

Effective altruism does not logically entail counterintuitive conclusions about population ethics. We do not need to accept the repugnant conclusion or believe that we are morally blameworthy for actions which are not bad for anyone in order to be effective altruists. It is interesting to note that the philosopher who first discussed the repugnant conclusion and the non-identity problem, Derek Parfit, is one of effective altruism's most vocal proponents today. Effective altruism is a much less radical proposition than utilitarianism and, as I hope I have shown, an extremely sensible moral philosophy. However, we might wonder, if effective altruism is so intuitively and logically appealing, why is altruism today so ineffective at improving the well-being of valuable creatures?

ALTRUISM AS PRACTICED TODAY

Most People are Ineffective Altruists

Altruism can take many forms, but for this section I will focus on charitable giving. Many people act as if under the impression that all charities are equally good. But if 'equally good' is taken to mean 'equally effective at improving people's lives', the claim becomes immensely implausible. The notion that all charities are equally good at helping people is about as likely to be true as the notion that all companies

are equally good at producing quality commodities. Why would it be the case that all charities currently in existence just happen to be equally effective at alleviating suffering?

Suppose we reject the belief that all charities are equally good. There is still the epistemic problem of determining which charities are better than others, and particularly, which charities are the best of them all. Those wishing to object here might claim that there are, at present, no means by which to determine how effective charities are. The claim that we have no way of knowing which charities are better than others is only slightly more plausible than the claim that no charity is, in fact, better than another. To maintain such a belief, we would have to conclude that Homeopaths Without Borders (yes, this is a real charity) is, for all we know, just as effective at improving well-being as any other charity in existence.

Here is a concrete example to illustrate the difference between effective altruism and ineffective altruism. Suppose we plan to donate $40,000 to prevent or alleviate the symptoms of blindness. Providing a single blind person with a guide dog will cost the entire $40,000 (Ord 2013 p.1) By contrast, the cost of surgery to cure trachoma-induced blindness is less than $20.(Ibid) With $40,000 one could either provide a single blind person with a guide dog, or cure 2,000 people in the developing world of trachoma-induced blindness. Conservatively estimating that the quality of life improvement of providing someone with a guide dog is equal to that of curing someone of trachoma-induced blindness, the choice is clear.

Proponents of the 'uncertainty argument' outlined above would have to believe these estimates so inaccurate as to have misassessed the situation by three orders of magnitude. Hopefully this possibility is sufficiently unlikely to compel us to accept two conclusions. First, charities differ in the degree to which they improve the lives of conscious creatures. And second, that the information needed to accurately assess cost-effectiveness is at least partially available.

If people were genuinely motivated to give to charity based on an intrinsic desire to improve the well-being of others, we might assume they would spend at least a bit of time and effort attempting to find this information. But this is not the pattern of behavior we observe. 83% of Americans donate to charity. (Gallup Editors 2013) Of them, 10% say they do not care at all about non-profit performance. (Hope Consulting 2011) The rest *say* they care about non-profit performance, but only 3% have done any research to find the highest performing charities. (Ibid)

However, you might wish to object, maximizing your impact does not neces-

sarily require researching high-impact charities. For instance, you might think that, instead of spending an hour googling effective charities, you could spend another hour at work to earn more money to donate. This is an interesting possibility, but highly implausible. Given the amount of time people spend working and the amount of money people donate to charity, such a move would only be rational if donors expected an hour's worth of research to yield less than a 0.05% increase in the effectiveness of their giving[2]. It is also worth noting that donors who do not research never cite anything like this as their reason for not conducting research—the closest equivalent being that 4% of them say they are too lazy.(Hope Consulting 2011)

Despite professing to care about effectiveness when asked, most people practice altruism ineffectively. This means that those who claim to care about effectiveness either hold beliefs about charity which are fantastically detached from reality or are being insincere. My vote is for the latter. The cost of providing one guide dog for one blind person is the equivalent of curing 2,000 people of trachoma-induced blindness. Every dollar we donate to someone in poverty in the developed world could have been donated to someone 20 times as destitute in the developing world[3]. The money required to grant a single wish for a terminally ill child could have saved five children from dying in the first place[4]. Yet we continue to donate massive sums of money to ineffective charities, and our donations will achieve only a small fraction of their potential to reduce suffering.

2. In 2014, US donors gave $358 billion to charity, or about 2% of annual GDP (Giving USA 2014). Adjusting for the fact that only 83% of Americans, donate, this makes 2.5% per donor on average. My calculations assume that individuals give at this rate throughout their lives. The average person works for about 80,000 hours—40 hour work week with 2 weeks annual vacation over 40 years. This would mean the average donor gives the equivalent of 2,000 hours salary. For one hour of research conducted before any donation has been given to yield a negative impact, it would have to have less than a 1/2000 or %0.05 increase in effectiveness.

3. More precisely, the poorest 19% of Americans live on less than $27.40 a day (US Census Bureau 2013). The poorest 17% of the world's population live on less than $1.50 a day, meaning they are 18 times as destitute (World Bank 2015). Dollar amounts adjusted for purchasing power. Calculations assume income is flat or normally distributed.

4. Between August 2012 and August 2013, the Make A Wish Foundation of America spent over $246 million (Make a Wish Foundation 'Combined Financial Statements'). In 2014, the foundation granted 14,200 wishes. Assuming expenses for 2014 were approximately equal to 2013, this amounts to $17 thousand per wish (Make A Wish Foundation 'Wish Impact & Facts'). By contrast, donations to the Against Malaria Foundation can save a child's life for $3,340 (GiveWell 2014). This means that the cost of granting a wish is equal to the cost of saving 5 lives.

Why People Donate

Hopefully this evidence is enough to convince us that the overwhelming majority of people are ineffective altruists who behave as if they are mostly indifferent to the effectiveness of their charitable donations. But if people do not donate to charity to minimize suffering, why do they donate to charity? Research in moral psychology has identified two predominant factors—the warm glow of giving and signalling effects. However, while both of these factors influence people to give to charity, they have only a limited ability to influence which charities people give to. As we will see, the cognitive mechanisms responsible for charity choice respond to cues which many of us would consider arbitrary and unimportant.

The warm glow of giving is the subjective feeling of satisfaction we experience when we make a personal sacrifice to help someone else. (see, e.g. Andreoni 1989 and Crumpler & Grossman 2008) We can experience this feeling whether or not we can expect to receive material rewards from our action, suggesting that humans have evolved or acquired an intrinsic motivation to make personal sacrifices for the sake of helping others. This feeling can even be induced when we know ahead of time that our sacrifice will do nothing to further the well-being of the intended recipients. Simply giving is enough to make us feel good about ourselves.

Another reason we give is to show off our moral rectitude. (see, e.g. Lacetera & Macis 2010; Dean & McConnell 2012; and Rand & Nowak 2013) It is important to us that our family, friends, and community members believe we are good people. Giving to charity is one way to demonstrate our altruistic character. This is called a *signalling effect*—when one of the benefits of an action is the signal it communicates to others. In this case, the action is donating to charity, and the signal it sends is that we are kind and caring individuals. As the turn of phrase goes, *be good to seem good*.

These are the two main factors that motivate people to donate to charity. Of course, this psychological evidence does not eliminate the role of helping others as a motivational factor. It is not a coincidence that we experience a warm glow when making a sacrifice *for the sake of helping others,* even when this sacrifice is entirely symbolic, or that the best way to signal we are good people is by doing something *for the sake of helping others*. The evidence simply suggests that helping others is more of an instrumental goal, and holds limited force as an intrinsic motivation.

Charity Choice

For most people, reducing suffering and improving well-being provides little intrinsic motivation to give to charity. But what motivates us to give to certain charities and not others? One third of donors report researching charities before they donate to them, but only 3% report researching cost-effectiveness. (Hope Consulting 2011) Of the donors who do research, only 17% of them aim to find information to compare charities and determine which of them to donate to.(Ibid) And of the donors who do comparison research, just over half of them research cost-effectiveness as a decisive factor. (Ibid) This means that two thirds of individual donors do no research at all, and that 90% of those who do fail to consider cost-effectiveness. So what information *do* we use to decide between charities?

Charity choice for unresearched donations are determined largely by cognitive biases. For example, when we see posters on the metro advertising for a charity you can donate to with via text message, what factors determine whether or not we will do so? Moral psychology has provided us with an extensive list of biases, but I will mention only a few of the most important here:

Physical proximity bias (Musen 2010 as described in Greene 2013)—How far away from me are the recipients of my donation?

Identifiable victim effect (e.g. Loewenstein et. al. 2006)—Do I know any personal information, especially the name and face, of the recipients of my donation?

In-group bias (e.g. Henri & Turner 1979)—Are the recipients of my donation members of my country, or another group I belong to?

These biases may also serve as a heuristic for which charities donors decide to research. For example, imagine a commuter sees one of these advertisements, but never donates to a charity without going on its website. I would conjecture that the commuter is more likely to look up a charity which helps people nearby, shows a picture of an identifiable victim, and works in her own country. When conducting research, a different set of biases come into play, including:

Evaluability bias (Caviola et. al. 2014)—Does the charity score well on easily evaluated measures, particularly low overhead?

Basic- and subordinate- level bias[5]—Does the charity work on a problem that was similar on a basic or subordinate level to a problem that affected me or a loved one?

Even though donors overwhelmingly claim to care about cost-effectiveness when prompted, helping others effectively plays a minimal role in motivating them to donate or determining which charities they donate to. This evidence should lead us to wonder whether ineffective altruism is irrational at all. Perhaps helping others has very little to do with altruism. And perhaps all of these supposed 'biases' we have been discussing are perfectly rational features of our decision-making processes.

I believe such a conclusion would be a mistake. If we had access to better information and took time to reflect on how we choose between charities, I expect most people would realize that what they actually care about is improving the lives of as many people as they possibly can by as much as they possibly can. By contrast, I would wager that most of the factors that currently determine charity choice would seem at best minimally important. The aforementioned psychological mechanisms really are *biases* in the sense that they cause us to behave in ways that we ourselves would disapprove of upon reflection.

We have already seen this revealed preference structure in tests on the evaluability bias.(Caviola et. al. 2014) When asked how much a subject wishes to donate to a charity presented in isolation, subjects' donations correlate more strongly with overhead ratio than cost-effectiveness. However, when subjects are provided with more information and are allowed to compare charities side by side, their donations correlate more strongly with cost-effectiveness than overhead ratio. The conclusion we should draw from this study is that, although people behave as if they care more about overhead than effectiveness, they do so only because of a lack of information. In fact, people care more about helping others effectively, but this preference is only revealed under conditions of better information and reflection. Though the relevant studies have yet to be conducted, I predict there will be similar findings for all of the biases I have just mentioned. To see why, ask yourself about each one in turn:

5. I hasten to add that I know of no experimental evidence for this bias, so my mention of it here should be taken as speculation based on personal observations about people's motivation for charity choice. I would encourage researchers to explore this bias experimentally.
For example, imagine a woman's child has died of leukemia. There are several levels of abstraction at which she could think about this tragedy, each of which may result in different patterns of charitable giving. She may think, 'I have lost my child to leukemia; therefore I will donate to charities which fight leukemia' (subordinate level), or 'I have lost my child to cancer, therefore I will donate to charities which fight cancer' (basic level), or 'I have lost my child; therefore I will donate to charities which fight the most prevalent causes of child mortality' (superordinate level). However, people tend to focus on the subordinate and basic levels while failing to abstract to the superordinate level.

Physical proximity bias: Does someone's suffering become less important to you as a function of geographic displacement? Would you be willing to pay $100 to save a child's life when she is a mile away from you? What about two, twenty, or one hundred miles? How far away does this child have to be before you would consider it acceptable to let her die for $100?

Identifiable victim effect: Does someone's suffering become less important to you as a function of not knowing her name? What if you determined to donate $100 to save a child whose name you were told but forgot before making the donation? Would this be an acceptable reason to let her die?

In-group bias: Does someone's suffering become less important to you because you happen to have been born in different countries? Would you be willing to donate $100 to save the life of a child from your own country? What if the child moved to a different country? Would this be an acceptable reason to let her die?

Evaluability bias (specifically overhead aversion): Is it worth letting people suffer and die to ensure that the employees and CEOs of a charity get paid less? How much less would a charity's employees and CEOs have to get paid in order for it to be worth letting a child die?

Basic- and subordinate-level bias: Does someone's suffering become less important because they suffer from something that no one you care about has experienced? For example, if a loved one of yours were to die from cancer, would this make children who die from malaria less important than children who die from cancer? Would you be willing to donate $100 to save a child from dying of cancer? How dissimilar does a cause of mortality have to be from cancer in order for you to consider it acceptable to let it kill a child for $100?

When confronted with these sorts of questions, I imagine most people would realize how arbitrary and unimportant factors like physical proximity are to them. By contrast, I predict that cost-effectiveness strikes people as an important factor even when subjected to similar scrutiny.

Effectiveness: Is the suffering of one person less important than the suffering of five people? Would you be willing to pay $100 to save one child's life? If so, does this imply you would be willing to pay more to save the lives of five children? Given the choice between donating to a charity which would use your money to save the life of one child and a charity which would use your money to save the lives of five children, would you choose to save one and let five die, or save five and let one die?

We can subject our biases to the same sort of scrutiny for any type of suffering.

Here I have chosen to focus on child mortality as a prototype cause of misery. But we could equally well ask these questions about, say, rape. For the physical proximity bias we might ask, *How far away does a woman have to be before you would consider it acceptable to allow her to be raped for $100?* My intuition is that it does not matter how far away this woman is—suffering is equally important no matter where it occurs. What *does* matter to me is that I do whatever I can to most effectively mitigate suffering and foster well-being. If you share this intuition, you ought to be an effective altruist as well.

In sum, here is the explanation for why most people are not effective altruists, but should be:

We have psychological incentives to donate to charity, even if only a small part of these incentives is a desire to improve well-being as effectively as possible. While these incentives determine that we should donate to charity, they do not fully specify which charities we should donate to.

Given proper information and rational reflection, we recognize that we would prefer to choose the most effective charities.

However, the psychological mechanisms we currently use to determine our choice of charities rely on factors which, to many of us, seem arbitrary upon reflection.

Therefore, instead of relying on psychological biases, our preferences are better served by choosing charities based mostly if not entirely on effectiveness.

OVERCOMING ALTRUISTIC BIASES

At present, most people give to charity because it gives them a warm glow and a positive reputation, and choose which charities to give to based on cognitive biases. I expect similar psychological mechanisms determine other altruistic decisions, which for some people include volunteering and formulating opinions on how government should litigate for the public good. As a result, there is much more suffering in the world than there would be if only we would act on our altruistic impulses in ways that effectively improved people's lives. Fortunately, there are ways to overcome these biases.

To illustrate my proposal, it will helpful to draw an analogy between how we think about economics and how we think about altruism. Like altruistic decision-making, economic decision-making suffers from a host of cognitive biases. But unlike altruistic decision-making, we have developed methods for recognizing and

overcoming biases in economic decision-making. In what follows, I explicate this analogy further and suggest that the methods we employ to think about economics can be used to think about altruism as well.

Overcoming Economic Biases

Consider the life-cycle hypothesis in economics, which holds that individuals prefer smooth consumption throughout the course of their lifetime. (For an early example, see Modigliani 1966) Standard economic theory predicts that, all else being equal, we prefer to consume more rather than less in any given period of time. However, there are diminishing marginal returns on consumption. In any given year, we prefer to consume $75,000 worth of goods to $50,000 worth of goods and $50,000 worth of goods to $25,000 worth of goods, but we more strongly prefer $50,000 to $25,000 than $75,000 to $50,000. Supposing we have a fixed amount of wealth which we can consume at any rate we choose, maximizing utility over the course of our lives requires that we consume at a constant rate.

The extent to which we practice consumption smoothing in real life is constrained by, among other things, psychological biases. We spend impulsively, take on more debt than we can afford, and consistently underestimate how much we need to save for our long-term financial goals. One of the biases that precipitate this behaviour is known as *hyperbolic* temporal discounting.(e.g Madden et. al. 2003 and Green et. al. 1994) Our reflective preferences dictate that we should smooth consumption, but we have an intuitive drive to consume more now and leave less for later. The conflict between immediate and delayed gratification is mediated by two largely independent cognitive processes. (e.g. McClure et. al. 2004 and Metcalfe & Mischel 1999) One—the faster, emotionally charged process—generates a strong, visceral desire to spend now. The other—the slower, emotionally cooler process—implores us to engage in long-term financial planning. Things like saving for retirement require our slower, reasoning processes to direct or perhaps supersede our faster, intuitive processes.

What this means in practice is that we should explicitly recognize our preference for smooth consumption, determine the best way of satisfying this preference using the best epistemic norms available to us, and act according to the conclusions we reach. Many people, for example, hire a financial consultant to help them plan for retirement and attempt to implement her advice by saving and investing accordingly. Not everyone thinks about retirement or relies on epistemically reliable information

such as expert advice when doing so. But we all recognize that these are the sort of steps we ought to take if we care about being financially solvent in our later years.

Economic thinking and altruistic thinking have much in common. We are capable of recognizing certain preferences in economic and altruistic decision-making, such as having a smooth consumption curve and donating to effective charities. In both domains, our preferences are hindered by cognitive biases, such as hyperbolic temporal discounting and the physical proximity bias. The conflict between our rationally endorsed preferences and our biases is mediated by similar cognitive processes with similar neural underpinnings. (Greene et. al. 2004; Greene et. al. 2001) It is therefore reasonable to expect that the same mode of thought which allows us to overcome our economic biases can allow us to overcome our altruistic biases as well.

What this involves is a procedure whereby we:

Explicitly recognize our preferences,

Use epistemically reliable methods to decide how best to satisfy these preferences, and

Act on our decisions

In the example of consumption smoothing, we realize that we need to save for retirement, rely on information provided by financial experts, and save and invest accordingly. We can follow a similar process when it comes to altruism. To begin with, we need to recognize our preference for altruistic actions which most effectively improve well-being. The next step is gather information on how best to satisfy this preference. Just as most of us rely on financial experts for advice, the most reliable way to do so—apart from conducting our own extensive research—is to rely on experts such as those at the Centre for Effective Altruism. Finally, we need to implement this advice, perhaps by switching our donations to more effective charities or considering high-impact career options.

Is Effective Altruism Killing the Love?

Before concluding, there is at least one more concern that needs to be addressed. Studies have shown that the employment of reasoning processes in pro-social de-

cision-making tasks correlates negatively with generosity. It is empirically possible, then, that employing reasoning processes in altruistic decision-making will decrease altruism to such an extent that it will more than offset its increase in effectiveness. Paradoxically, it may be more effective to make altruistic decisions based on the very cognitive biases that make our altruism ineffective.

This is an interesting possibility, but empirically implausible. Though no studies have tested this directly, related research shows that employing reasoning processes under certain conditions can decrease altruism by 15-50%[6]. But considering some charities are thousands of times more effective than others—for example, with donations to guide dog charities versus trachoma charities—it would be surprising to learn that rational thinking increases the effectiveness of our giving by less than a factor of two. On empirical grounds, the expected increase in effectiveness eclipses the expected decrease in altruism. I would also speculate that the sort of people who engage in rational thought for the express purpose of helping others as much as they possibly can will be among the least susceptible to having their motivation desiccated by reasoning processes. Perhaps the tradeoff between effectiveness and altruism is not such a problem outside the lab. While this is still an open question, the available evidence suggests that rational thought is essential for effective altruism.

CONCLUSION

Today, most people are ineffective altruists. We perform actions for the sake of helping others, but we do so in such a way that gives less help to fewer people than we otherwise could. Most of our motivation for donating to charity comes from a desire to feel good about ourselves and score reputation points. And most of what determines our choice of which charities to donate to is a collection of cognitive biases. As a result, millions of people and non-human animals will continue to suffer unnecessarily.

Effective altruism is the antidote to this miserable state of affairs. Concisely put, effective altruism is about "aiming to do the most good that one can". I have offered a more precise explanation of what this means, and shown it to be a much milder and more intuitive philosophy than utilitarianism. We do not have to accept the re-

6. Rand et. al. 2012 shows a 15% decrease in contribution to public goods games; Loewenstein et. al. 2006 shows a 50% decrease in charitable contribution to a statistical victim versus an identifiable victim.

pugnant conclusion or consider ourselves morally blameworthy for actions which are not bad for anyone in order to be effective altruists. Nor do we have to relegate considerations of deontological values like justice and fairness to a role of merely instrumental importance. All we have to believe is that when we act altruistically, it is preferable to give more help to more people, rather than less help to fewer people, all else being equal.

If we are to live in accordance with this preference, we need to revolutionize how we think about altruism. In addition to thinking intuitively, we need to think rationally. I suggest that we reconceive of altruism in economic terms, whereby we view acts of charity as an investment in the well-being of valuable creatures. And we should demand nothing less of ourselves than to see our investment yield maximum returns. Making even the simple decision to donate to effective charities can increase our impact by orders of magnitude. Faced with these facts, it should be evident by the light of our own values that it is no longer acceptable to just make the world a *better* place. This is too modest a goal. Instead, we should endeavor to improve the lives of as many people as possible by as much as possible, and use our altruism to do the most good we can.

Acknowledgements: I would like to thank the Oxford Uehiro Centre for organizing this essay contest, my friends from Crockett Lab - especially Lucius Caviola, Andreas Kappes, and Molly Crockett - for their enormous influence on my thinking about moral psychology, my friends from the Centre for Effective Altruism - especially William MacAskill and Hauke Hillebrandt - for introducing me to a wonderful new way of thinking about moral philosophy, and my advisor Professor Daniel Dennett to whom I owe a greater debt of intellectual gratitude than I could ever hope to repay.

REFERENCES

Andreoni, J. (1989). Giving with impure altruism: applications to charity and Ricardian equivalence. *The Journal of Political Economy*, 1447-1458.

Caviola, L., Faulmüller, N., Everett, J. A., Savulescu, J., & Kahane, G. (2014). The evaluability bias in charitable giving: Saving administration costs or saving lives?. *Judgment and Decision Making*, 9(4), 303.

Crumpler, H., & Grossman, P. J. (2008). An experimental test of warm glow giving. *Journal of Public Economics*, 92(5), 1011-1021.

Gallup. (2013, December 13). Most Americans Practice Charitable Giving, Volunteerism. Retrieved June 15, 2016, from http://www.gallup.com/poll/166250/americans-practice-charitable-giving-volunteerism.aspx .

GiveWell. (2014, November). Against Malaria Foundation (AMF) | *GiveWell*. Retrieved July 24, 2015, from http://www.givewell.org/international/top-charities/amf.

Giving USA. (2015, June 29). Giving USA: Americans Donated an Estimated $358.38 Billion to Charity in 2014; Highest Total in Report's 60-year History. Retrieved June 15, 2016, from http://givingusa.org/giving-usa-2015-press-release-giving-usa-americans-donated-an-estimated-358-38-billion-to-charity-in-2014-highest-total-in-reports-60-year-history/.

Green, L., Fry, A. F., & Myerson, J. (1994). Discounting of delayed rewards: A life-span comparison. *Psychological Science, 5*(1), 33-36.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment.*Neuron*, 44(2), 389-400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105-2108.

Greene, J. D. (2013). *Moral tribes: Emotion, Reason, and the Gap Between Us and Them*. New York: Penguin.

Guide Dogs of America. (2015). Donating FAQ. Retrieved June 15, 2016, from http://www.guidedogsofamerica.org/1/help/donate/.

Hope Consulting. (2011, November). Money for Good II: Driving Dollars to the Highest Performing Non-Profits. Retrieved June 15, 2016, from https://www.guidestar.org/ViewCmsFile.aspx?ContentID=4038.

Karlan, D., & McConnell, M. A. (2014). Hey look at me: The effect of giving circles on giving. *Journal of Economic Behavior & Organization*, 106, 402-412.

Lacetera, N., & Macis, M. (2010). Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior & Organization*, 76(2), 225-237.

Loewenstein, G., & Deborah, A. Small, and Jeff Strnad. 2006. Statistical, Identifiable, and Iconic Victims. *Behavioral Public Finance*, 32-46.

Madden, G. J., Begotka, A. M., Raiff, B. R., & Kastern, L. L. (2003). Delay discounting of real and hypothetical rewards. *Experimental and Clinical Psychopharmacology*, 11(2), 139.

Make a Wish Foundation. (2014, February 04). Combined Financial Statements. Retrieved July 24, 2015, from http://wish.org/~/media/100-000/About%20Us/Making%20a%20Difference/Managing%20Our%20Funds/Documents/FY2013/FY13%20MAWFA%20Combined%20FS_Final%202_04_14.ashx?la=en.

Make-A-Wish America. (2011). Wish Impact & Facts. Retrieved June 15, 2016, from http://wish.org/wishes/wish-impact#sm.00000r9h9eqoladpqvbz65boolroh.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards.*Science*, 306(5695), 503-507.

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: dynamics of willpower. *Psychological Review*, 106(1), 3.

Modigliani, F. (1966). The life cycle hypothesis of saving, the demand for wealth and the supply of capital. *Social Research*, 160-217.

Ord, T. (2013, March). The moral imperative toward cost-effectiveness in global health. *Center for Global Development*, 1-12. Retrieved June 15, 2016, from http://www.cgdev.org/sites/default/files/1427016_file_moral_imperative_cost_effectiveness.pdf.

Raihani, N. J., & Smith, S. (2015). Competitive helping in online giving.*Current Biology*, 25(9), 1183-1186.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences, 17*(8), 413-425.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416), 427-430.

Singer, P., & MacAskill, Will. (2015). Introduction. In R. Carey, *Effective Altruism Handbook.* (viii-xvii). Oxford: Centre for Effective Altruism.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs*, 229-243.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict.*The Social Psychology of Intergroup Relations*, 33(47), 74.

US Census Bureau. (2014). Selected Income Characteristics. Retrieved June 15, 2016, from http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_5YR_DP03.

World Bank. (2015). Poverty Overview. Retrieved July 24, 2015, from http://www.worldbank.org/en/topic/poverty/overview.

Going Viral: Vaccines, Free Speech, and the Harm Principle

OXFORD UEHIRO PRIZE IN PRACTICAL ETHICS 2014-15

MILES UNTERREINER

University of Oxford

ABSTRACT

This paper analyzes the case of public anti-vaccine campaigns and examines whether there may be a normative case for placing limitations on public speech of this type on harm principle grounds. It suggests that there is such a case; outlines a framework for when this case applies; and considers seven objections to the case for limitation. While not definitive, the case that some limitation should be placed on empirically false and harmful speech is stronger than it at first appears.

In December 2014, an outbreak of measles erupted at Disneyland theme park in Anaheim, California, United States of America. By mid-January, the virus had spread north to San Francisco, infecting (thus far) at least 70 people across the state. (Chang 2015).

Measles is a highly contagious airborne disease that typically manifests itself in a red splotchy rash that covers the entire body and is often accompanied by a fever and cough. In certain cases, however, measles is much more dangerous. Persons with weakened immune systems—such as those afflicted with HIV or AIDS—are much more susceptible to the disease, and the measles mortality rate is significantly higher in developing nations. According to the World Health Organization, approximately 145,700 persons died from measles worldwide in 2013. (WHO 2014)

According to public health officials, the best way to stop the spread of measles is to receive a vaccine shot—an inert sample of the virus that effectively trains the body's immune system to resist the real thing. Prior to the start of the United States' national measles vaccination program in 1963, that country reported between 3 and 4 million cases of measles annually. Of persons infected each year, between 400 and 500 died and approximately 48,000 were hospitalized. Thanks to intensive national vaccination efforts, however, the measles virus has been considered eradicated in the United States since the year 2000. (Centre for Disease Control and Prevention 2015)

So why is measles back?

The answer can largely be traced not to a new or mutant form of the virus, but to the spread of something much more difficult to combat: false information.

Antivaccination campaigns now pose a threat to public health efforts around the globe. Such campaigns are sometimes grounded in objections based on religious, philosophical or ethical grounds. Frequently, however, they are based upon the distribution of incorrect empirical information about vaccines themselves.

In 1998, research in *The Lancet*, a British medical journal, appeared to demonstrate a link between the MMR vaccine—the vaccine most frequently given to children to prevent measles—and increased autism rates in children. The editor-in-chief of the British Medical Journal (BMJ) announced in 2011 that this research had been found to be fraudulent, (Godlee et al 2011) and the paper's lead author was found guilty of professional misconduct and barred from practicing medicine in the United Kingdom.

By then, however, it was too late. MMR vaccination rates dropped significantly in the United Kingdom after the fraudulent *Lancet* article was published, from 91 percent in 1998 to 80 percent in 2003. The number of new measles cases rose accordingly, from 56 in 1998 to 1,370 in 2008. (Flaherty 2011) By 2008, the disease was endemic to the UK, a country in which it had once been eradicated. (Batty 2009) Professor Dennis K. Flaherty of the University of West Virginia has called the vaccine-autism scare perhaps "the most damaging medical hoax of the last 100 years." (Flaherty 2011)

> *"Despite the overwhelming evidence of the safety and effectiveness of the MMR vaccine,"* continues Professor Flaherty, *"the vaccine-autism connection gained traction on the Internet and was perpetuated by print and television media eager for increased circulation or higher ratings. Entertainment shows contributed to the controversy by offering vaccine-autism connection proponents a platform to make*

> *their case, largely unchallenged. By 2009, 1 of 5 parents in the US believed that vaccines cause autism in otherwise healthy children. Moreover, 10% of parents in a study published in 2010 were refusing 1 or more newer vaccines for their children." (Ibid)*

What may the state do about all this, if anything? Certainly, few people doubt that the state may permissibly spread correct information as widely as possible, or fund public health vaccination programs to encourage wider uptake of vaccines.

But what if these strategies prove to be insufficient in convincing the public of the safety of vaccination, and hence ensuring that a sufficient number of persons are in fact vaccinated? May the state justifiably limit the free speech of anti-vaccination campaigners, and if so, on what grounds might it do so?

Anti-vaccination advocates have already begun to stake out a principled free-speech argument in favor of their cause. When every venue in Australia at which she had hoped to speak canceled her invitations in January 2015, anti-vaccine campaigner Sherri Tenpenny's organization replied that this tactic amounted to "bullying by vested interests who do not believe in informed consent, free speech and respect for other's rights, and who appear to support censorship of thought and science." (Medew 2015)

Persons conditioned to believe in the inherent value of free speech—myself included—are often inclined to agree with Tenpenny that limiting speech in this way is not permissible. I believe the public has in mind here some form of Mill's harm principle—that actions which do not harm others should not be limited by the state—combined with some sort of belief that speech acts do not "cause harm" in a morally relevant way.

In this paper, I will argue in favor of the following proposal: even under a conventional and philosophically libertarian version of the harm principle—a version that restricts state action to limit the liberty of individuals to cases in which the exercise of that liberty causes suitably *direct physical* (not emotional or psychological) harm—the state has sufficient normative grounds to limit the free speech of anti-vaccine campaigners who spread empirically false information. I use the vaccine case only as a currently relevant and clear real-world example; I do not claim that there is anything normatively unique about anti-vaccine campaigns specifically, and I of course extend the argument to any persons who engage in normatively equivalent acts of speech. Importantly, I do not consider whether the state might limit speech on the

grounds that it causes psychological offense or constitutes "hate speech," although there is a significant literature regarding this question and employing it would make my argument easier.

In the course of stating this normative case for state action to limit speech under certain circumstances, I will consider and reject several arguments to the contrary, as follows.

Objection A: Speech acts cannot cause harm.

Objection B: Speech acts can cause harm, but the harm is too indirect to warrant state interference.

Objection C: The presence of human intermediaries in the causal chain that leads from information distribution to the harm caused requires us to place responsibility for the harm caused with the human intermediaries located proximately closest to the harm, not with the original information distributors.

Objection D: The argument ignores important normative discrepancies between the real-world parallels given to help justify the argument and the case of empirical information distribution to the general public. Specifically, the argument ignores 1) the sincerity of anti-vaccine campaigners and 2) the alternative and equally accessible information, also open to the public, in favor of vaccination, and 3) the lack of a certain type of special relationship between persons that would lead to a reasonable expectation of empirical accuracy in information.

Objection E: The argument is too broad. Such a claim would justify limiting all types of speech acts that might potentially lead to physical harm in any way, and this is an unacceptable proposition.

Objection F: The proposal is not practical.

Objection G: We cannot know the scientific truth to a sufficiently rigorous degree to justify limiting speech that appears to contradict such truth.

WHAT IS DIFFERENT ABOUT SPEECH?

Most libertarians agree that it is the state's job to prevent individuals from harming other individuals, even if it is not the state's job to do much of anything else. If I steal your car, hit you with a baseball bat, or roll a large boulder down a hill onto your property and destroy your home, most libertarian philosophers will agree that the state should prevent me from doing this, or punish me after I have done it. This is because I have done you or your property harm, and the harm principle allows the state to act to protect some individuals from harm caused by others.

Are speech acts qualitatively different from the above types of act, and if so, how? The answer to this question is important. If we answer in the negative, then our task is over: speech may be regulated like any other sufficiently harmful action. It is only if speech is different in some important and relevant way that there remains more work to do.

One could argue that speech acts are different in one important way: they cannot cause harm because the only morally relevant way to cause harm under the harm principle is by an act of physical force or movement. Mill himself did not appear to accept this idea: in the third chapter of *On Liberty*, he argued that "even opinions lose their immunity [protection from state interference], when the circumstances in which they are expressed are such as to constitute their expression a positive instigation to some mischievous act. An opinion that corn-dealers are starvers of the poor, or that private property is robbery... may justly incur punishment when delivered orally to an excited mob assembled before the house of a corn-dealer, or when handed about among the same mob in the form of a placard." (Mill 1869) In other words, Mill thought that spoken or written opinions which (sufficiently directly) instigated *other persons* to cause harm could be limited under the harm principle as well.

But even more basically than that, it is not entirely clear that speech acts are really qualitatively different from other types of action. Consider the case of John, who thinks that vaccines don't work and has decided to host a public reading of a popular anti-vaccine pamphlet. In order to reach the maximum number of people, John purchases a loudspeaker and advertises the reading widely online. But imagine that John, his loudspeaker, and the assembled crowd all gather in a remote mountain village where the sound waves generated by loud noises are known to trigger deadly avalanches. John boldly asserts that his right to free speech outweighs the harm to others that is likely to follow, and begins to read the anti-vaccine pamphlet aloud into

his megaphone. The avalanche that follows kills five people and injures 500. I think it is clear that John has harmed these people, and harmed them in the same sense that he would have harmed them had he stood on top of the mountain and drilled away at the snow with a sledgehammer. (One could imagine plenty of other examples of this type: a person in a room full of otherwise silent people who knows that a spoken word will trigger a noise-sensitive bomb; less similarly but also less absurdly, a person who transmits vital security information to a terrorist group and leads to the death of thousands of civilians in a terrorist attack.)

I think we can therefore reject *Objection A: Speech acts cannot cause harm*. Speech acts can cause harm, and sometimes in precisely the same sense that other, more conventionally recognized types of action do.

I think we can also reject *Objection B: Speech acts can cause harm, but the harm is too indirect to warrant state interference*. This objection does not hold much theoretical weight if by "indirect" we mean "having many causal steps in between the original action and the harm eventually caused." If a line of causation is sufficiently clear and certain, the presence and number of intermediary steps is irrelevant to assigning responsibility for the harm done. If I push a large rock off a cliff, the rock lands on a fuel tank and causes it to explode, and the resulting fire from the fuel tank burns down the town at the bottom of the cliff, I am still responsible for causing the town to burn down.

INTERMEDIATING PERSONS

More difficult and much more realistic, however—and primarily at issue here—are cases in which speech might be said (in some sense) to lead to harm, but the thoughts and beliefs of other persons intermediate between speech and the harm it could be said to cause. What happens, in other words, when other people become part of a causal chain leading from a speech act X to some harm Y? This lies at the heart of *Objection C* and it is the general phenomenon of which anti-vaccine speech is one specific example. Let us consider two test cases.

Smoking: Barbara is the CEO of a major cigarette company, circa 1952. Although the company's scientists inform her that smoking cigarettes over a long period of time causes cancer, she directs her advertising division to market the cigarettes as safe and fun to consumers. Successfully duped by the advertising, millions of people buy cigarettes and later suffer significant negative health effects.

Theatre: Clyde goes to a popular movie, and every seat of every row is full. Just for fun, halfway through the film, Clyde yells "FIRE!," although there is in fact no fire at all. In the stampede that follows as panicked theatregoers flee the room in droves, seven people are trampled and suffer significant injuries.

In both *Smoking* and *Theatre,* false information was distributed (the advertisements or Clyde's false warning); persons responded to the false information with physical action (buying cigarettes or attempting to flee the theatre); and their response to the information caused either themselves or others (or both) harm (cancer or injuries due to the stampede). This leaves us with the following questions: (a) Can Barbara and/or Clyde be said to have caused the harm by distributing false information; (b) should the state hold them liable for the harm thus caused; and (c) can this type of state action logically extend to the type of speech promoted by anti-vaccine campaigners? I propose that we accept the answer to both (a) and (b) to be a clear "yes" (as they were in fact answered in the real world) and to use these answers to help us explore the solution to question (c).

To help answer (c), it may be helpful to begin by asking what we mean when we say that one action "causes" another in more generally accepted cases of wrongful harm causation. Let us consider a clear example containing two sub-examples:

Car: John is walking along the street, doing nothing wrong, when I drive off the road and hit him with my automobile.

The case is clearest if I have done this

> *Intentionally:* It is reasonable to say that I am morally liable for his injury because of the following components of my action: 1) My action (driving a car toward John at high velocity) is one that can be reasonably expected by a reasonable person, possessed of full information, to result in his injury. 2) I intended this to occur. 3) It is not reasonable to expect that he could have avoided my car by acting differently, as he could not have possessed the information necessary to do so (i.e. that he should walk on a different street or be prepared to jump out of the way of my car).

The case is less clear but I think still indicative if I have done this:

> *Unintentionally*: If, say, I were texting while driving or were drunk. 1) still holds, although less directly; my action (texting or drinking) can be reasonably expected to seriously increase the chances of injury to other persons, although it does not make such injury certain. 3) still holds just as in the intentional case. It is for this reason (among others) that the law also restricts those who cause physical harm to other persons unintentionally.

Let us consider a less clear example, but one more pertinent to the issue we are attempting to solve.

Minefield: Say that a dangerous explosive has been left outside John's home during the night, a fact of which he is unaware. I know that if I tell him it is raining, he will most likely go outside to the tool shed to retrieve his umbrella, placing him very near the explosive. Although it is not certain that John will step on the mine in the process, he does so and is injured.

Does this example cohere with the straightforward cases of wrongful causation given previously? I think that it does. 1) Walking into the minefield could be reasonably expected by a reasonable person possessed of complete information to cause likely injury. 2) I intended John to walk outside after hearing the information I gave him (*Intentionally* only applies here). 3) John could not have known that he should act otherwise, and it was perfectly reasonable of him (given the level of information he could be expected to possess) to go to the tool shed in order to retrieve his umbrella. I think it is therefore fair to say that I have caused him to be injured just as I would have caused him to be injured by my car in the straightforward case of moral liability given earlier. It does not matter morally that one step of the causal chain—John deciding to go outside and retrieve his umbrella—involved the action of an independent human moral agent.

POSSIBLE FACTORS CONTRIBUTING TO MORAL RESPONSIBILITY FOR HARM CAUSED BY SPEECH

I now want to emphasize a few important elements of *Minefield* and to connect them with the real-world *Theatre* and *Smoking* cases given previously. In the course of doing so, I hope to delineate the senses in which information distributors can be morally responsible for the indirect effects of their speech when interpreted and acted upon by information-receiving agents. The ideas are as follows:

Two key elements in assigning moral liability for harm caused by speech seem to be be the *level of information* (LI) possessed by each agent and the *reasonableness of the*

response (RR) by each agent. In *Minefield*, I knew something about the level of danger that John didn't (LI), and he could not have been expected to act otherwise given the information he possessed (RR), which led to his injury. In *Theatre*, it would have been theoretically possible for the theatregoers to ignore Clyde's warning and not to have suffered injuries in the resulting stampede, but it would not be reasonable to expect them to have done so; fleeing a fire is the response that would be reasonably expected given the circumstances (RR), and they could not have known (given the short time frame and level of danger involved in waiting around to find out) that there was in fact no fire (LI).

In *Smoking*, Barbara should be held liable for the injuries sustained by the smokers her company deceived (and her company should not have been allowed to say that smoking is safe in the first place) because consumers at the time could not have been reasonably expected to have possessed the information necessary to convince them that smoking is not safe (LI), and engaging in an activity one believes to be perfectly safe (walking, drinking water, eating dinner) is a perfectly reasonable thing to do (RR). It is for these reasons that I believe it is right to extend liability for harm to, as many jurisdictions do, companies who market unsafe products as safe, doctors who give patients empirically bad treatment information, or car salesmen who sell defective cars. The causal agent distributing the false information, unlike the agent receiving the information, knew or should have known that the information was false and harmful (LI), and the consumer or patient could not have been reasonably expected not to have bought the product, followed the doctor's advice, or bought the apparently safe car (RR).

When taken together, we can see LI and RR to be normatively important, collectively and jointly, because they both strongly influence *the certainty that information will be acted upon in a certain way* (C). C is important because if it is unlikely or uncertain, due to LI or RR or some other factor(s), that false and/or harmful information will be acted upon in a certain way by a reasonable moral agent, it is difficult to assign moral blame to the agent distributing the information. For example, if I advise you to jump off a tall cliff with rocks at the bottom—informing you that contrary to popular belief, doing so would be quite safe—and you do so, it would be wrong to say that I am to blame for your action; the LI of a normal person, combined with the low RR of jumping off the cliff in response to a mere suggestion by a stranger, ought to lead us to lay the blame at your feet rather than mine.

A third contributory factor to C and an important factor in assigning moral li-

ability for harm caused by speech is *status* (S): the level of responsibility, command, or authority assumed by the information distributor. In *Cliff* above, it is difficult to assign blame to me, a random stranger, for your decision to jump off the cliff. But if I were your commanding officer in a military unit, and if I ordered you to jump off the cliff and told you to trust me rather than merely suggesting that you do so as a disinterested bystander, it becomes more plausible to shift the blame to me for the harm done to you as a result of my speech act (the order). I mention S in more detail under the discussion of Objection D2 below.

A fourth and final factor in assigning responsibility for harm caused by speech may be the mental state (M) of the person listening to the speech act.[1] This is particularly relevant in cases where the party receiving the information is in a mental state M which is abnormal and renders them particularly vulnerable to speech that is false and/or causative of harmful results. This is of especial importance in cases like *Theatre*, where the party receiving the information is likely to act in a certain way due to reasonably acquired fear of danger or harm. It is also of importance because it takes into account the reduced decision-making capability of those persons with non-standard mental processing capacities, such as children or the cognitively disabled. It is reasonable to expect adults with standard levels of decision-making ability to sort through and balance the alternative avenues of possible action when given information to process; it would not be reasonable to expect a child to know right from wrong, or a clever choice of action from a foolish one, in quite the same way.

Again, M is normatively important because it is a contributory factor to C—the certainty of an agent responding to information in a certain way—and C is normatively important because with a sufficiently low C, it becomes very difficult to assign causal responsibility for any action, including speech.

According to one set of powerful objections, however, the examples and reasoning given above are not adequate to prove the case I am attempting to show. I now want to explain why this so and how these objections might be addressed.

ANOTHER SET OF OBJECTIONS

I want to now move on to:

Objection D: The examples given and reasoning applied in the argument do not apply

1. (I say "may be" because it seems plausible to merely subsume M as a subset of RR rather than maintaining it as its own separate factor.)

to the vaccine case (or other normatively equivalent cases) because of important disanalogies in reasoning.

There are indeed several discrepancies or gaps between the examples and argumentation given thus far and the real-world anti-vaccine case (and cases like it). I read these to be the following:

> *Objection D1: Unlike in* Minefield, Smoking, *and* Theatre, *the alleged wrongdoers in the vaccine case almost certainly believe the information they are distributing to be empirically correct and, furthermore, that by distributing it they are in fact helping the people with whom they speak. It would therefore be wrong to hold them morally responsible for the harm caused by their speech in the same way we ought to hold the alleged wrongdoers in the other cases given responsible. This is also true of the other examples mentioned: doctors prescribing bad treatments, companies selling impure food, and car salesmen selling defective cars all do so deliberately.*

> *Objection D2: At least some of the cases mentioned thus far derive their force from a special relationship that exists between the party distributing information and the party receiving it. In the case of companies selling products or salesmen selling cars, the consumer enters a de facto contract with the seller that the seller breaks by falsely advertising the content of his, her, or its products. In the case of doctors prescribing improper treatment, the patient has entered into a special contract that requires a higher level of moral responsibility from the doctor than would be expected of the general public. There is no such special relationship between anti-vaccine campaigners and the public; they are private persons acting in a private capacity and the analogy is therefore flawed.*

> *Objection D3: Unlike in the cases mentioned thus far, in which there could have been no reasonable expectation that the parties receiving information would act differently in response to the informational stimuli given to them, there* is *a reasonable expectation that members of the general public, as mature moral agents, possess a meaningful choice as to whether or not to follow the advice of, and listen to information provided by, anti-vaccine campaigners. Instead of laying the blame at the feet of anti-vaccine campaigners, we should therefore place the blame for any harmful consequences of non-vaccination with the persons located closest causally to the harm done: those people who listen to, and act upon, the information provided.*

I consider each of these objections now.

D1: Intent

Should a lack of malicious intent matter in assigning moral liability for harm caused by speech? I think that it should, but that as with other cases of liability due to negligence, it renders the person in question merely less responsible, rather than entirely blameless, for the negative consequences of his or her actions. If one intends to mislead, falsify, or gain monetarily from a harm done to others, then moral liability is much easier to establish.

But even in cases wherein malicious intent is absent—where the actor distributing information does not actually believe the information to be false or harmful—I think moral liability is easier to assign than might be otherwise thought. This is so in the case when an agent with status S_1 conveys information to an agent S_2, where 1 is significantly higher than 2, and agents with status S_1 are reasonably believed or expected to possess a significantly higher LI with regard to the information conveyed than agents with status S_2.

In simpler terms, it is also fair to assign moral blame for harm unintentionally caused by incorrect or misleading speech when the person speaking ought to know what they're talking about, but they don't. I turn to that now in more detail.

D2: The Capacity in Which One Acts

Special relationships are an important factor in assigning moral blame for harm caused by acts of communication primarily because of contributory factor S (status). S is normatively important, to recapitulate, because (like LI and RR) it influences the certainty of action based upon the information given (C).

High relative S factors—which lead to trust and hence to a higher degree of certainty of action C—can be conveyed informationally in a variety of ways. S can be conveyed explicitly: by licensure (of doctors, lawyers, engineers); by the adoption of a contract—such as that between a buyer and a seller in a marketplace; by a special relationship between parents or guardians and their children; or by order hierarchies (in an organization with structured top-down power relationships). When one is in possession of a high status S with regard to another person, information conveyed from the high-status individual to the other person attains a special quality which

renders the high-status person specially liable for that speech's consequences. These status-differential relationships are often important and necessary because they allow persons without expertise in a field to engage productively with those who have it; in other words, such relationships enable persons to trust complex or difficult information which they would otherwise be incapable of processing themselves. Without the existence of such relationships, asymmetrical information problems would prevent contracts and discussion between persons of drastically different talents and skill sets. Often, when the importance of acting upon correct information is sufficiently important for the physical safety of some or many persons, such relationships are legally codified to the extent that information about certain subjects cannot be conveyed in certain ways *except* by the proper individuals; it is not legal to offer certain types of medical or engineering advice without a license, for instance.

In these types of relationship, the information-receiving party has been reasonably led to believe that the advice they are given from the information-distributing party is reliable, even though the complexity of such information renders a definitive independent judgment on this question difficult or impossible for the information-receiving party. As such, it is reasonable for the information-receiving party to act upon the information as suggested, leading to a high certainty of action factor C, to an increased directness of causation from a speech act to the relevant physical harm, and to an increased moral liability for that harm on the part of the information-distributing agent.

The extent to which anti-vaccine campaigners (and persons like them) fall under this special category of relationship, which ought to place an additional burden of liability upon these persons acting jointly to distribute information, depends to the extent to which and ways in which the campaign is professionalized, institutionalized, and branded. If the organization attains a quality sufficiently similar to the others mentioned—companies in a marketplace, medical or legal professionals offering advice, and so on—anti-vaccine campaigners place themselves in a moral position that ought to render them increasingly liable to censure if matters go wrong.

D3: Alternative Information Sources

One final possible problem with the analogies and examples given is that they do not adequately take into account the possibility of an information-receiving agent weighing alternative sources of information when considering how to act in the real

world. The more and better sources of information on a subject that are available, the reasoning might go, the less any individual agent promoting false or dangerous information ought to be held liable for that information's consequences. In the particular case under consideration, there appear to be many doctors, scientists, and other professionals offering advice that would contradict the dangerous information distributed by anti-vaccine campaigners; ought not we then hold the person who listens to the anti-vaccine campaigners responsible, as independent moral agents, for making that choice, rather than the campaigners themselves?

Perhaps it ought to be noted that this reasoning does not seem to apply to the other types of speech mentioned thus far; the possibility of individuals encountering some other information sources indicating that smoking is not safe does not appear to relieve cigarette companies of the responsibility to market their products only in a certain way clearly noting that smoking is not in fact safe. The possibility that some sort of private consumer organization might distribute information that some doctors are better or safer than others does not relieve doctors of the responsibility to offer correct treatment advice. Again, the presence of alternative information sources counts against, but does not seem to completely eliminate, moral liability for harm caused by speech when the combination of the relevant *level of information* (including the complexity of the information), *reasonableness of response, status,* and/or *mental state* of the two parties are such that an information-receiving agent can be expected to act with reasonable certainty in a certain way C in response to the speech of the information-distributing party.

Objections E and F: Over generalisability of the Proposal

Is the proposal—that anti-vaccine campaigns may sometimes be limited by state action—too broad and therefore dangerous or impractical? Would it license state limitation of any and all types of speech that might lead to physical harm, however indirectly, to persons who take such speech seriously?

It does not necessarily do so. The set of entities to which such limitations would apply would be seriously circumscribed by the criteria for judging moral liability given in the discussion above.

Limitation of any kind would not be applicable to purely private entities acting individually under such a model for speech limitation because the *status* relationship between two private persons is not the type of relationship that would lead

to the type of special trust discussed above; because the *level of information* reasonably expected to be possessed by private persons is not high or apparently high; and because the *reasonableness of response* to private individuals is generally the responsibility of the information receiver. The state would not be properly allowed to fine my Aunt Muriel for offering her opinion to friends or acquaintances that vaccines cause autism, for instance; no one reasonably perceives Aunt Muriel to be an authority on the matter, no one ought reasonably to believe that she possesses special expertise that would allow her to make such a judgment; and no one has entered any sort of differential-status relationship with her that would make her in any way liable for the consequences of her speech.

The reasoning given here would, however, be more closely applicable to anti-vaccine organizations (or any organization distributing harmful information) that are sufficiently institutionalized, professionalized, organized, and advertised to the public to meet the criteria outlined above. The more closely such organizations or groups of persons approximate the characteristics that render companies, professionals, and salespeople peculiarly liable for their speech acts, the more closely they ought to be subject to scrutiny on the grounds that they may in fact cause harm through speech.

Objection G: Doubtful Certainty of Scientific Truth

Can we ever know with sufficient certainty that vaccines, for instance, do not cause autism—or that smoking causes cancer, or that eating too much fat is bad for you—for the state to justify making a definitive judgment on the matter as reflected in the type of speech advocacy it allows or encourages?

The answer is very rarely yes. The level of certainty about the question under consideration ought to be extraordinarily high, and although a precise percentage seems difficult to suggest, 99 percent (as measured, perhaps, by expert consensus) may be a reasonable number with which to begin. Furthermore, the level of certainty about a subject ought not only be high; the consequences of the correct course of action *not* being taken must be sufficiently harmful to justify some form of state intervention. Although it is not true that drinking water will give one supernatural powers, it is also not harmful to drink water; it would then seem wrong for the state to seek to limit those who encourage drinking water in order to attain supernatural powers.

One Final Note: Methods of State Limitation

Limiting speech on public safety grounds need not (and certainly ought not) take the form of men in black masks kicking down doors in the night to take away those with whom we disagree. It may be, depending on the nature and proximity of the harm caused, something more like mandating a warning label to keep consumers fully informed of risks, or requiring that advertisements or promotional materials contain a fixed text containing a basic set of correct empirical informational notes. It is not practical, possible, or desirable to apply such methods to private individuals—such methods might only be applied to organizations or institutions. The precise form speech restrictions take, however, may vary; it is the basic ethical question of whether the state may impose *some* form of limitation that must first be answered.

I do not pretend to have answered it completely. This question deserves far more normative and empirical analysis than I am capable of giving here. But whether anti-vaccination campaigns (for instance) should qualify for free speech protection is a question that may quite literally determine whether some people live or die, and it is at the very least a question worthy of further consideration.

REFERENCES

Batty, David, "Record Number of Measles Cases Sparks Fear of Epidemic," *The Guardian*, 9 January 2009.

Centers for Disease Control and Prevention, "Frequently Asked Questions About Measles in the U.S.," 21 January 2015. [Available at: http://www.cdc.gov/measles/about/faqs.html . Accessed 22 January 2015]

Chang, Alicia. "Measles Outbreak Casts Spotlight on Anti-Vaccine Movement," *The San Francisco Chronicle*, 23 January 2015. [Available at: http://www.sfgate.com/news/medical/article/Disney-parks-linked-measles-outbreak-grows-to-70-6031882.php . Accessed 23 January 2015]

Flaherty, Dennis K., "The Vaccine-Autism Connection: A Public Health Crisis Caused by Unethical Medical Practices and Fraudulent Science." *Annals of Pharmacotherapy* October 2011 Vol. 45 No. 10. [Available at: http://aop.sagepub.com/content/45/10/1302.full . Accessed 22 January 2015]

Godlee, Fiona, Jane Smith and Harvey Marcovitch, "Wakefield's article linking MMR vaccine and autism was fraudulent," *BMJ* 2011; 342:c7452. [Available at: http://www.bmj.com/content/342/bmj.c7452.full . Accessed 22 January 2015]

Medew, Julia, "Anti-Vaccination Campaigner Sherri Tenpenny's Tour in Jeopardy," *The Sydney Morning Herald*, 20 January 2015.

Mill, John Stuart. *On Liberty.* London: Longman, Roberts, & Co., 1869.

World Health Organization, "Measles: Fact Sheet No. 286," November 2014. [Available at: http://www.who.int/mediacentre/factsheets/fs286/en/. Accessed 22 January 2015]