The Philosophical Quarterly
The Scots Philosophical Association and the University of St Andrews
The Philosophical Quarterly Vol. 62, No. 248 July 2012
ISSN 0031-8094 doi: 10.1111/j.1467-9213.2012.00061.x

WHAT *DOES* MATTER? THE CASE FOR KILLING THE TROLLEY PROBLEM (OR LETTING IT DIE)

By Barbara H. Fried[1]

I. INTRODUCTION

A significant portion of non-consequentialist thought over the past forty years has been directed to the question of what harms it is permissible for us to cause to others through our action or inaction. The literature has mined two veins.

The first concerns our affirmative duty to rescue others at moderate cost to ourselves (hereinafter referred to as the Duty of Easy Rescue). The general aspiration here has been to carve out a principled exception to what most non-consequentialists take to be our background prerogative *not* to aid others. All other things being equal, the effect of that exception is to push non-consequentialist morality closer to the optimific (that is, aggregate welfarist) solution.

The second concerns when it is impermissible to harm one person or group of persons in order to avoid harm to others (hereinafter referred to as the Duty Not to Harm). In contrast to the Duty of Easy Rescue, the general aspiration here has been to articulate and defend principled limits on the optimific solution, rooted in the inviolability of individuals.

My principal concern here is with the Duty Not to Harm. (The Duty of Easy Rescue presents some related difficulties for non-consequentialists, which I take up briefly below.) For the past forty years, the philosophical literature on the

[1] William W. and Gertrude H. Saunders Professor of Law, Stanford University. I am grateful to the Center for Advanced Study in the Behavioral Sciences, where the initial draft of this essay was written. I would like to thank Joseph Bankman, Mark Kelman, Liam Murphy, Allen Wood and most especially Derek Parfit for their generosity in reading and commenting on prior drafts.

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK, and 350 Main Street, Malden, MA 02148, USA

Duty Not to Harm has developed around a set of (now canonical) hypotheticals requiring us to choose between harms to different people or groups of people: trolley problems, forcible transplants, Bernard Williams's Jim and the Indians, Taurek's dilemma of whether to save the one or save the five, scenarios that implicate the doctrine of double effect (e.g., the munitions bombing cases), etc. The hypotheticals typically share a number of features beyond the basic dilemma of third-party harm/harm tradeoffs. These include that the consequences of the available choices are stipulated to be known with certainty ex ante; that the actors are all individuals (as opposed to institutions); that the would-be victims (of the harm we impose by our actions or allow to occur by our inaction) are generally identifiable individuals in close proximity to the would-be actor(s); and that the causal chain between act and harm is fairly direct and apparent. In addition, actors usually face a one-off decision about how to act. That is to say, readers are typically not invited to consider the consequences of scaling up the moral principle by which the immediate dilemma is resolved to a large number of (or large-number) cases. Although the occasions and mechanisms for making such tragic choices vary, for ease of exposition I will follow Allen Wood's lead and refer to hypotheticals that share most or all of these secondary features as 'trolley problems.'[2]

In my view, the intellectual hegemony of the trolley problem has shaped non-consequentialist thought in a number of unfortunate ways. First, and most directly, it has resulted in non-consequentialists' devoting the bulk of their attention to an oddball set of cases at the margins of human activity, while largely ignoring conduct that (outside of the context of criminal activity and warfare) accounts for virtually all harm to others: conduct that is prima facie permissible (mowing a lawn, fixing your roof, driving a car down a city street) but carries some uncertain risk of accidental harm to generally unidentified others. Of the various moral principles that have emerged from the now four-decades-long preoccupation with trolley problems, none can handle the problem of garden-variety risk. As a result, trolleyology is at best engaged in what amounts to a moral sideshow. [3]

[2] A. Wood, 'Humanity as an End in Itself,' in D. Parfit, *On What Matters* (Oxford UP, 2011), vol 2, pp. 58–82.

[3] Problems of just warfare are arguably the closest real-world analogue to trolley problems. I take no position on the usefulness of trolleyology in thinking through moral dilemmas in this arena. For a skeptical view, see D. Luban, 'Risk Taking and Force Protection,' in I. Benbaji and N. Sussman (eds.) *Reading Walzer* (London: Routledge (forthcoming) (draft available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1855263); D. Luban, 'Unthinking the Time Bomb,' in C.R. Beitz and R.E. Goodin (eds.), *Global Basic Rights*, 181–206 (Oxford UP, 2009). Although criminal conduct often entails harms that, judged ex ante, are close to certain to result, it does not raise the problematic interpersonal tradeoffs that trolleyology is designed to help us think through, because we put no value on one side of the tradeoff: the liberty of individuals to engage in criminalised conduct.

Second, and relatedly, the hermetic focus on trolley problems has led trolleyologists to misdiagnose what is going on in trolley problems themselves. The trolley literature has meticulously analysed how our intuitions about harm to others change as we move from one trolley-type setup to another (life *v* life; life *v* lives, life *v* limbs; act *v* omission; upstream *v* downstream harms). But once we relax the secondary features common to *all* trolley problems (most importantly, that the consequences of stipulated alternative courses of action are known with certainty ex ante, secondarily that any resulting harms will befall *proximate* and *identified* victims), many non-consequentialists have no clear intuitions at all. Of those who do, most conclude that the problem of risk has to be resolved by some form of aggregation, in which the numbers are doing most or all of the work. This suggests that what is driving people's intuitive resistance to aggregation in trolley problems is not the third-party harm/harm tradeoffs *per se* (which are as much a factor in risk cases as in trolley cases) or even the distinction between act and omission, but rather the secondary features peculiar to trolley problems. Maybe our penchant to care much more about certain-to-result harm to identified victims than a risk of harm to as yet unidentified ones is defensible in moral terms. But if so, the terms will have to be very different from those typically invoked to explain the 'right' result in trolley problems. And if it is not, trolleyology more likely belongs in the domain of moral psychology than of non-consequentialist moral philosophy.

Third, trolleyologists generally treat the factual posture in which trolley problems 'happen' to present themselves – and in particular the knowledge that the hypothetical persons in the problem are taken to possess about the consequences *to them* of adopting different available courses of action – as an exogenous variable that defines the moral problem to be solved but is itself morally neutral. I think that view is difficult if not impossible to defend, and has opened the entire trolley enterprise to the charge that it is begging the question. To put the point in the strongest terms, if, on further thought, non-consequentialists conclude that the appropriate epistemological point of view from which representative individuals should formulate their objections to a candidate principle is *before* they know how things will turn out *for them in particular* if that principle is adopted, then virtually the entire trolley literature becomes morally irrelevant, and non-consequentialists will likely end up adopting instead principles that diverge little if at all from standard aggregative techniques.[4]

[4] For an argument that ex ante contractualism leads to aggregation, see L. Alexander, 'Lesser Evils: A Closer Look at Paradigmatic Justification,' *Law and Philosophy*, 24 (2005), pp. 611–43, at 617 n.23. For a discussion of where ex ante contractualism is likely to diverge from conventional aggregation, see B.H. Fried, 'Can Contractualism Save Us From Aggregation?,' *The Journal of Ethics*, 16 (2012), pp. 39–66.

Derek Parfit's 2002 Tanner lectures, just published in the two-volume *On What Matters* along with commentaries by Barbara Herman, Thomas Scanlon, Susan Wolf and Allen Wood, are the most recent contribution to the annals of trolleyology. In the course of hundreds of pages, Parfit offers and interrogates dozens of hypotheticals to tease out non-consequentialists' moral intuitions about when it is permissible or morally required to sacrifice some in order to save others from harm. Every one of those hypotheticals involves choices among consequences deemed certain to befall identified, close-at-hand, fictitious persons – that is to say, every one is (in my sense of the term) a trolley problem.[5]

Parfit himself is hardly a non-consequentialist, and his aim in *On What Matters* is not to defend non-consequentialism in general or trolleyology in particular. It is rather to show that, even starting from the inhospitable premises of trolleyology (inhospitable, that is, to consequentialism), Kantians and Kantian contractualists will be driven to consequentialist conclusions. Parfit is also responsible, more than any other single person, for putting the moral claims of as-yet unidentified victims on the philosophical map, starting with his exploration of the non-identity problem in *Reasons and Persons*.[6] While the moral complexities he surfaces in the non-identity problem are different from those that trolleyology pushes offstage in ignoring uncertainty, the two are not wholly unrelated.

For all these reasons, it would be odd to describe Parfit himself as a trolleyologist, and I do not mean to do so. And yet, even Parfit has fallen prey to the allures of trolleyology, here and elsewhere, in treating decision-making under certainty and decision-making under uncertainty as different moral kinds; treating the former as the 'central part of moral theory' (*Reasons and Persons*, p. 25) and the latter as something more akin to a 'non-moral' dilemma (*On What Matters,* 1, 162); and arguing that an individual's reasons to accept or reject a proposed act should be assessed

[5] Well, not quite. One of the examples Parfit gives of duties we owe to future, as yet unborn, persons involves a piece of glass negligently left in the woods, and stepped on ten years later by a five-year-old child (*On What Matters*, 2, pp. 217-18.) This is clearly a case of risky conduct with uncertain consequences and as yet unidentified potential victims. But Parfit avoids most of the difficulties that arise in regulating risk by choosing conduct that anyone would recognise as negligent and describing it as such. As a result, the 'right' answer — the conduct is impermissible — is essentially stipulated up front. The only remaining moral task is to resolve the non-identity problem – that is, how to deal with the claims of an as-yet non-existent victim.

[6] *Reasons and Persons* (Oxford UP, 1984). Indeed, as discussed below, Parfit relies on the claims of future, as yet unidentified, persons in *On What Matters* to derive consequentialist conclusions from the non-consequentialist premises of trolleyology.

from the ex post epistemological perspective – that is, *after* the individual knows whether *her* life will go better or worse as a result of it.

For Parfit's immediate purposes, the detour through trolleyology may be at worst a waste of time, and at best a tactically shrewd belt-and-suspenders operation to bring along those Kantians who are unpersuaded by ex ante contractualism. But taking a longer-term view, I think it is a tactical error for consequentialists to engage trolleyologists on their own terms (i.e., trolley problems) without registering objections to those terms. As I suggested above, the 'common sense' intuitions that trolleyology traffics in – how could it be right to throw the fat man off the bridge, shoot one innocent person to save five, leave a workman to suffer from electric shocks for an hour so soccer fans worldwide can watch the World Cup uninterrupted? – are arguably not moral at all (in the conventional non-consequentialist understanding of morality) and have at best an extremely limited domain. But the visceral power of those intuitions lends the non-consequentialist principles extracted from trolleyology a surface moral plausibility they have not earned. In short, the right position for consequentialists to take with respect to trolley problems is to just say no to the lot of them.

All of this is to say that my disagreements with Parfit are at most a friendly family quarrel; the real object of criticism here is committed trolleyologists. Nevertheless, *On What Matters* provides a useful text and its appearance an apt occasion to flesh out these concerns, and to make the case for non-consequentialists' killing the trolley problem, or at the very least letting it languish from neglect while they redirect their attention to the far more important problem of risk.

II. WHAT HAS THE TROLLEY PROBLEM WROUGHT?

II.1. *Minimizing the prevalence of tragic choices (or, what happened to risk?)*

Allen Wood devotes a substantial portion of his commentary in *On What Matters* to decrying the outsized role of trolley problems in non-consequentialist philosophical argument. I share many of his objections, including to the fantastical nature of the dilemmas trolley problems pose; the absence of contextual information that in real life changes the moral complexion of tragic choices; and the unrealistic stipulation that the outcomes of all available choices are known with certainty ex ante (*On What Matters,* 2, pp. 69–70). But we have very different – indeed, opposite – worries about the ways in which preoccupation with trolley problems has distorted non-consequentialist thought. That disagreement points to a more profound disagreement about the inevitability of interpersonal

tradeoffs, and concomitantly the limited scope for Kantian notions of individual inviolability in any imaginable social world.

Wood believes that obsessing on trolley problems has led non-consequentialists to *overestimate* the occasions on which we must, unavoidably, choose between one person's life, health or ability to pursue her own projects and another's. I believe that it has led them to underestimate that necessity, dramatically. There is no logical contradiction here. Both of these distortions can – and I believe do – occur simultaneously. But the latter is, to my mind, the far more troubling one.

Wood's argument that obsessing on trolley problems has led us to exaggerate the necessity for tragic choices is straightforward. By stipulating that we face a tragic choice and focusing attention solely on how we ought to choose – whether to flip the switch to kill one or instead let five die, whether to rescue five strangers on one rock or one's own child on another – trolley problems encourage readers to regard such tragic choices as unavoidable. But in real life, Wood argues, most tragic choices could have been avoided if the relevant individuals had taken adequate preventive measures at an earlier moment in time, for example, by building safer trolleys, putting up better signage to warn passersby, erecting fences to 'prevent anyone from being in places where they might be killed or injured by a runaway train or trolley' (*On What Matters,* 2, p. 74). By pushing those earlier decisions offstage, trolleyology has pushed them off the philosophical agenda as well. Out of sight is out of mind in academic discourse as much as in other aspects of life.

Wood acknowledges that some tragic choices are unavoidable. But the ones that are, he suggests, are by and large the province of 'extreme and desperate situations in human life' such as war, anarchy, pestilence, famine or natural disaster. In contrast, when faced with quotidian decisions like allocation of health care services, if we find ourselves deliberating whom to save and whom to let die, it is in all likelihood because we have made a 'voluntary decision… to turn health care, or even human life as a whole, into something horrible and inhuman, something like war, that ought never to exist' (*On What Matters,* 2, pp. 79–80).

Without doubt, many of the tragedies that occur in the developed world are easily preventable. Indeed, a case could be made that the most serious problems facing contemporary American society can be traced to our collective failure to invest prudently in prevention (using 'prevention' broadly to include all ex ante investments to improve ex post outcomes). But in my view the fixation on trolley problems hasn't contributed to that systemic failure, except perhaps indirectly, by leading people to misunderstand the nature of the choices we face when we *do* invest proactively in prevention.

In contrast, non-consequentialists' obsession with trolley problems *has* led many to underestimate radically the occasions in life we are forced to make tragic choices. Given that trolley problems are about nothing but tragic choices, the claim that obsessing on them has led philosophers to underestimate the need to make such choices no doubt seems paradoxical. I don't think it is. By presenting tragic choices only in 'extreme and desperate,' indeed (outside of the context of war) freakish, circumstances, the trolley literature has inadvertently led both authors and consumers of that literature to regard tragic choices *themselves* as rarely occurring and freakish in nature. But they are neither of these things. They are ubiquitous and for the most part quotidian, and typically result not, as Wood suggests, from 'human vulnerability to nature, and … human wickedness' (*On What Matters,* 2, p. 81), but from the finite nature of the resources we depend on to realise our projects in the world.

This is the problem of scarcity, in the sense that economists use the word. It denotes not any absolute level of deprivation, but rather any situation in which the demand for a 'good' exceeds its supply, with the consequence that we cannot satisfy all competing demands for it.[7] In this sense of 'scarce,' most goods in society are necessarily scarce, either because the material resources needed to produce them are finite (e.g., 'goods' like health, product safety, national defense) or the social space required to enjoy them is finite (e.g., any activities we wish to pursue that impose some risk of harm on others).

However wealthy a society is, however many doctors it has trained, however many procedures it underwrites and public health projects it has undertaken, at some point it cannot put any more resources into addressing the health needs of one citizen without leaving unaddressed the health or other pressing needs of other citizens. In a country as rich as the US, we could probably go a long way toward treating all currently treatable diseases just by allocating our existing health care budget more sensibly. But there will always be more we could do for the sick if money were no constraint. And on the prevention side, the needs are virtually limitless (diabetes, obesity, drug use, poor nutrition, poor cardiovascular condition, eradicating infectious diseases, etc.) Long before we reach the point where an additional dollar invested in diabetes prevention would yield no incremental benefit in lives saved, we will (rightly) conclude that investing those additional resources in, say, education, social workers, police, etc. would have a greater positive impact on peoples' lives. And when we

[7] More precisely, it denotes any situation where, if a good were free, people would consume more of it than is available. This qualification is meant to set aside the market solution to scarcity: raise the price of a good until demand no longer exceeds supply.

reach that conclusion, however many resources we have already put into diabetes prevention, someone will eventually die because we did not invest a few dollars more. In that sense, we will unavoidably be making 'tradeoffs between the deepest interests of different people' (*On What Matters,* 2, p. 79). Moreover, the way we make those tradeoffs typically violates the one principle that almost all non-aggregationists agree on: that it is wrong to cause death or serious harm to one person in order to avoid more trivial harm to (realise more trivial benefits for) others, no matter how numerous those others are (hereinafter, the Greater Harms Trump Lesser Harms principle).

To make this concrete, consider the problem of trolley safety. Suppose we decide that we should do as Wood suggests, and invest enough money in trolley safety that we make it the case that no one will ever be faced with this or any other tragic choice created by a runaway trolley. How much money is enough? Suppose that if we invest $5 billion dollars in safety measures, we can reduce expected deaths or serious injuries from trolley accidents from one in every 10 million trolley trips to one in every 12 million trolley trips. Should we (must we?) make that investment? If so, how about $50 billion? $500 billion? In a world of finite resources, we have to draw the line somewhere. I do not imagine any non-consequentialist would disagree with that, and all would draw the line considerably short of $500 billion. But wherever we draw it, we will knowingly be choosing to increase the number of preventable deaths 'merely' to save some others money (to be used for some other generally offstage purposes.)[8]

If that choice is immoral, almost everything we do is immoral, as almost every action poses some risk (however slight) of serious harm to some, typically to secure much less weighty benefits for ourselves or others. However much precaution we use, and however slight that risk is, it can almost always be reduced further by even greater precautions, at the extreme by forbearing from the activity entirely. And if, as virtually all non-consequentialists would agree, that choice is not immoral, why not? What differentiates it, morally speaking, from the identical trade-offs made in the context of trolley-type problems – identical, that is, measured by expected outcomes? Why does the choice to cause one person to die

[8] Some trolleyologists will no doubt be tempted to sidestep the question by supposing that those offstage purposes will themselves prevent deaths, thereby allowing us to let the numbers count as a tie-breaker. For one such effort, see M. Ridge, 'How to Avoid Being Driven to Consequentialism: A Comment on Norcross,' *Philosophy and Public Affairs*, 27 (1998), 50–8. But given the impossibility of proving the negative here, that supposition is always available, raising the concern that trolleyologists will invoke it whenever useful to save trolleyology from clearly unpalatable conclusions and ignore it the rest of the time – a worry that Ridge's example does not allay.

for the mere convenience of millions of others cease to be impermissible (or indeed even tragic, in the sense of an appropriate occasion for deep soul-searching) just because the death is a statistical rather than absolute certainty? And if it is sometimes permissible to impose some risk of death or serious harm on a few for the mere convenience of many others, how do we decide when a risk is 'too great' or the benefits to others too small to permit it?

This is, of course, the garden-variety problem of risk. Outside of the arenas of criminal conduct and warfare, almost all harm inflicted by human agency is accidental harm that results from conduct that is prima facie permissible but imposes some risks on generally unidentified others. Trolley problems, in which all of the consequences of available actions are stipulated with certainty ex ante, are the freaks and sports of human interaction.

The first casualty of the hegemony of trolley problems is to obscure this truth – to encourage non-consequentialists to ignore the garden variety problem of risk and to focus instead on an oddball set of cases at the margins of human life.

Until relatively recently, this lopsided allocation of attention has gone largely unremarked-on in the trolley literature. But in *On What Matters*, Parfit offers a rare, explicit defense of the choice to ignore risk, echoing arguments suggested by others in passing:

> In trying to answer [what acts are right and what wrong], it is best to proceed in two stages. We can first ask which acts would be wrong if we knew all of the morally relevant facts… After answering these questions, we can turn to questions about what we ought morally to do when we don't know all of the relevant facts. These questions are quite different, since they are about how we ought to respond to risks, and to uncertainty. As in the case of non-moral decisions, though these questions have great practical importance, they are less fundamental. These are not the questions about which different people, and different moral theories, most deeply disagree. Given the difference between these two sets of questions, they are best discussed separately. So I shall often suppose that, in my imagined cases, everyone would know all of the relevant facts. We can then ask what we ought to do in the simplest, fact-relative sense. [*On What Matters,* 1, p. 162]

This explanation raises more questions than it answers. First, insofar as what is wrong with (potentially) harmful conduct is the (potential) harm itself, it is not apparent why 'certain' and 'uncertain' harms would raise 'quite different' moral problems. All acts involve consequences that are (ex ante) more or less certain. Trolley problems, in which the consequences are stipulated to be known with certainty ex ante, simply represent the limit case at one extreme. Given that we are dealing with

factually continuous phenomena, why would we think they raise morally discontinuous problems? Why is the difference between (say) 95 percent certainty and 100 percent certainty as to any of the morally relevant facts (the probability that harm will result, the identity of the victim(s), etc.) a difference in kind rather than a difference in degree, and a very slight one at that?

Second, assuming that the two do present 'quite different' moral problems, it is not apparent why we should regard the moral problems raised by uncertain harm as 'less fundamental' than those raised by certain harm. As Parfit acknowledges, it cannot be due to their relative practical importance. While 'certainty' is king in the hypothetical world of trolley problems, in the real world the consequences of our acts are *always* uncertain ex ante. This is true even of harms that are intended (in the strong sense of desired or the weak sense of foreseeable). If I point a loaded gun at your head and pull the trigger, I am overwhelmingly likely to kill or seriously injure you, but I am not certain to do so. The gun could misfire; I could have forgotten to load it; the bullet could be deflected by a metal plate in your skull. If I divert the trolley, I may believe I will thereby save five from certain death at the cost of one life, but I can never be certain, ex ante or ex post. Perhaps diverting the trolley will cause it to tip over before it reaches the one; perhaps if I had done nothing the five would have seen the trolley in time and moved out of the way. *A fortiori*, what is true of knowingly inflicted harms is true of accidental harms. Thus, from an ex ante perspective, the problem of all harm, accidental or not, *is* the problem of risk.

In what sense then does risk present a 'less fundamental' problem? Parfit's comments suggest two possible answers. The first is that the solution to the problem of risk cannot be of fundamental moral interest, because, as of the moment we must choose how to act, we lack a 'morally relevant fact' bearing on the wrongness of the act: what its actual consequences will be. If this is the claim, it is problematic on a number of levels, but I want to focus on just one here. As suggested above, whether the actual consequences of an act *are in fact* morally relevant to its rightness or wrongness needs to be established. From the perspective of a truly ex ante contractualism, they are not. And until it is established, the argument begs the question. (I return to this issue in section II.3 below).

Alternatively, Parfit can be read to suggest not that the problem of risk is unimportant, but that the answer is uncontroversial, at least relative to trolley problems. ('These are not the questions about which different people, and different moral theories, most deeply disagree.') In a related vein, others have suggested that once we resolve the problem of 'certain'

harms, risk will take care of itself, either because the imposition of risk is itself a completed harm or other wrong, or because the wrongness of risk is parasitic on the wrongness of the harm it threatens.[9]

Both versions of the 'risk is unproblematic' claim, in my view, get things exactly backwards. If classic trolley problems present a difficult choice for non-consequentialists, garden-variety risks present an impossible one. And the moral principles non-consequentialists have extracted from their extended engagement with trolley problems, far from solving the problem of risk, are viable only as long as they are *not* extended to garden-variety problems of risk.

While it is a fool's errand to generalise the principles extracted from trolley problems, as noted above most trolleyologists agree, at a minimum, on the Greater Harms Trump Lesser Harms principle: It is prima facie impermissible to inflict serious harm on one person – or indeed, even fail to rescue them from harm – in order to secure qualitatively lesser benefits for others, no matter how numerous those others are.[10] Or as Michael Ridge put it, 'innocent lives always dominate convenience.' [11] But this principle routinely produces the 'wrong' answer when applied to garden-variety conduct that imposes a risk of harm on others. Take for example the permissibility of constructing a sports stadium, knowing that with any conceivable level of safety precautions there will still be some irreducible risk of death to inno-

[9] On imposition of risk as a freestanding wrong, see J. Hampton, 'Correcting Harms versus Righting Wrongs: The Goal of Retribution,' *UCLA Law Review*, 39 (1992), pp. 1659–702, at 1661–6; H. M. Hurd, 'The Deontology of Negligence,' *Boston University Law Review*, 76 (1996), pp. 249–72, at 262; R. Kumar, 'Who Can Be Wronged?,' *Philosophy and Public Affairs*, 31 (2003), pp. 99–118, at 109; C. Finkelstein, 'Is Risk a Harm?,' *University of Pennsylvania Law Review*, 151 (2003), pp. 963–1001; C. Schroeder, 'Corrective Justice and Liability for Increasing Risks,' *UCLA Law Review*, 37 (1990), pp. 439–78; K.W. Simons, 'Corrective Justice and Liability for Risk-Creation: A Comment,' *UCLA Law Review*, 38 (1990), pp. 118–42; C. Schroeder, 'Corrective Justice, Liability for Risks, and Tort Law,' *UCLA Law Review*, 38 (1990), pp. 143–62; R.A. Duff, 'Subjectivism, Objectivism and Criminal Attempts' in A.P. Simester and A.T.H. Smith (eds.), *Harm and Culpability* (Oxford UP, 1996), pp. 19–44, at 37 n. 68; J. Goldberg and B. Zipursky, 'Unrealized Torts,' *Virginia Law Review*, 88 (2002), pp. 1625–719. On the wrongness of risk being parasitic on the wrongness of the harm it threatens, see S. Perry, 'Harm, History, and Counterfactuals,' *San Diego Law Review*, 40 (2003), pp. 1283–314; S. C. Wheeler III, 'Self-Defense: Rights and Coerced Risk-Acceptance,' *Public Affairs Quarterly*, 11 (1997), pp. 431–44; D. McCarthy, 'Liability and Risk,' *Philosophy and Public Affairs*, 25 (1996), pp. 238–62, at 251; Joel Feinberg, *Harm to Others* (Oxford UP, 1984), pp. 190–93.

[10] Most non-consequentialists will qualify this principle when the 'lesser harms' are still serious as a qualitative matter and the numbers of 'lesser harms' are significantly greater (e.g., one life for 1 million limbs). For present purposes, however, that qualification can be set aside.

[11] 'How to Avoid Being Driven to Consequentialism: A Reply to Norcross,' *Philosophy and Public Affairs*, 27 (1998), pp. 50–8, at 55. For other articulations of the principle, see T. Scanlon, *What We Owe to Each Other* (Harvard UP, 1998), p. 235; F. Kamm, *Intricate Ethics* (Harvard UP, 2007), pp. 36–7.

cent passersby from falling debris. If a passerby does die as a result of falling debris, that harm will typically be much more serious than the harms thereby prevented (loss of recreational enjoyment to spectators or revenues to players and management, etc.). Thus, under the Greater Harms Trump Lesser Harms Principle, the project would clearly be ruled out.

But no one, trolleyologists most certainly included, believes it would be wrong to build a stadium in any case in which there is a non-zero chance of serious injury or death to passersby.[12] It is easy to see why. Given how few choices in life 'read' to the philosophical (and lay) mind as trolley problems, if non-aggregationists stuck to their guns in every one of them – we may not kill one even to save 100,000 from the loss of a limb; we must rescue Jones from an hour of extreme pain, even though doing so will deprive a hundred million viewers of the pleasure of watching a World Cup match on TV – life would go on pretty much as before. In this very practical sense, non-aggregative solutions to trolley problems, whether or not they strike most people as morally persuasive, will at least strike them as feasible, for the same reason that the conventional 'Duty of Easy Rescue' strikes most people as feasible: because the circumstances in which anyone will have to alter her daily life to meet that duty are very rare.[13] (To put it another way, 'Fiat justitia ruat caelum' is an easier imperative to live by when we know the heavens are not in fact going to fall.) Not so in the case of garden-variety risks. If the numbers do not count in choosing between greater and lesser harms under conditions of uncertainty, such that the mere possibility of severe harm to even one person precludes actions with expected lesser benefits to millions, life as we know it would pretty much grind to a halt.

Acknowledging the unacceptability of that result, many non-consequentialists have conceded, explicitly or implicitly, that when it comes to risk, we are all aggregationists. This may also be what Parfit has in mind in saying that '[t]hese are not the questions about which different people, and different moral theories, most deeply disagree.'[14] That concession by non-consequentialists, however, comes at a very

[12] For Scanlon's acknowledgment of that apparent inconsistency and his justification for it, see *What We Owe to Each Other*, pp. 235–7. For Frances Kamm's justification, see n. 24 below. For the argument that they are not in fact reconcilable, see A. Norcross, 'Comparing Harms: Headaches and Human Lives,' *Philosophy and Public Affairs*, 26 (1997), pp. 136–67.

[13] For the argument that this undemandingness is precisely what justifies restricting the Duty of Rescue to those 'nearby,' see R. Miller, 'Beneficence, Duty and Distance,' *Philosophy and Public Affairs*, 32 (2004), pp. 357–83, at 379.

[14] For representative statements ceding the problem of risk to aggregative techniques, see J. Coleman, *Risks and Wrongs* (Oxford UP, 2002), p. 210; Feinberg, *Harm to Others*, pp. 190–3; Scanlon, *What We Owe to Each Other*, pp. 204–5.

high cost to their non-aggregationist project. If the domain of non-aggregative principles is limited to tragic choices between ex ante certain outcomes (i.e., trolley problems), such principles will play a minor role at best in governing conduct with the potential to harm others. How much does it matter whether we should save one identified life at the certain cost of 100,000 limbs if we will never face that choice in real life, and if the 'right' answer, by non-consequentialists' own acknowledgment, has no bearing on the kinds of choices we do face daily: whether we may put one statistical life at risk to secure recreational enjoyment for 100,000?

Others have argued that non-aggregative principles do have a role to play in resolving the latter question, but the principles are different from those that apply in the case of trolley problems. But the precise contours of those principles have yet to be spelled out, and it remains to be seen whether any are both viable and clearly distinct from conventional aggregation. I am doubtful, for reasons I have spelled out at length elsewhere.[15] But for now, the only point I wish to urge is that the answer to that question is far more important to the future vitality and relevance of the non-aggregationist project than anything more to be learned from trolleyology.

II.2. *Misdiagnosing trolley problems*

The single-minded focus on trolley problems has not just led trolleyologists to give short shrift to the problem of risk; it has also done no favours to their understanding of trolley problems themselves. Once again, the claim may seem paradoxical, but I don't think it is.

That many of the principles extracted from trolleyology do not produce the 'right' answer if applied beyond trolley cases suggests that the principles have been drawn much too broadly. Take the 'Greater Harms Trump Lesser Harms' Principle. Rather than its being the case that the principle requires that '[i]f one can save a person from serious pain and injury at the cost of inconveniencing others or interfering with their amusement, then one must do so no matter how numerous those others might be,'[16] it requires that one do so *only* if the harms in question are (i) *absolutely certain* to befall (ii) victims who are *identified* and (iii) *generally proximate* to the actor.

[15] Fried, 'Can Contractualism Save Us From Aggregation?'
[16] Scanlon, *What We Owe to Each Other*, p. 235.

Can these limitations be defended? It is beyond dispute that the certainty of harm and the identifiability and proximity of the victim are emotionally and psychologically salient to most people in judging the permissibility of (potentially) harmful conduct. The crucial question for trolleyologists is whether they are *morally* relevant, and if so, why.[17]

Peter Singer's 1972 article, 'Famine, Affluence and Morality,' forced that question with respect to identifiability and proximity, in the context of the Duty of Easy Rescue.[18] The debate he started continues to this day.[19] In Singer's words, if the rationale for the Duty of Easy Rescue is that 'if it is in our power to prevent something very bad from happening, without thereby sacrificing anything morally significant, we ought, morally, to do it,' why should that duty be limited to one-off heroic rescues of proximate victims? (p. 231). Why does it not extend to the obligation of moderately affluent Westerners to make what are, for them, relatively trivial financial sacrifices to help alleviate mass starvation in the developing world? One answer, of course, is that doing so would eliminate one of the chief advantages non-consequentialism is thought to have over utilitarianism: that it does not demand an implausible degree of other-regardingness from each of us in our private lives. But this answer seems to save morality by reducing it to whatever we are willing to do because it costs us little. [20]

Notwithstanding the equally critical role that identifiability, proximity and (most of all) certainty have played in delimiting the Duty Not to Harm, few non-consequentialists have acknowledged that role and even fewer have tried to justify it. While the anomalous status of risk has gotten more attention of late in the non-consequentialist literature, most of that attention is directed at solving the myriad problems that follow from distinguishing categorically between harms 'certain' and 'uncertain' to occur rather than defending the moral significance of the distinction to begin

[17] I don't mean to endorse the distinction between rational and emotional responses. But I take it to be basic in some form to what is meant by rationality or rights in Kantian and deontological frameworks.

[18] *Philosophy and Public Affairs*, 1 (1972), pp. 229–43, at 241.

[19] For trenchant arguments in the Singerian vein, see E. Ashford, 'Utilitarianism, Integrity, and Partiality,' *Journal of Philosophy*, 97 (2000), pp. 421–39; Ashford, 'The Demandingness of Scanlon's Contractualism,' *Ethics*, 113 (2003), pp. 273–302. For various efforts to defend (contra Singer) the moral relevance of identifiability and proximity in formulating the Duty of Easy Rescue, see Parfit, *On What Matters*, 2, p. 211; Sarah Miller, 'Need, Care and Obligation,' *Royal Institute of Philosophy Supplement*, 80 (2006), pp. 137–60, at 141–3; R. Miller, 'Beneficence, Duty and Distance,' at 379; J. Lenman, 'Contractualism and Risk Imposition,' *Politics, Philosophy and Economics*, 7 (2008), pp. 99–122, at 116.

[20] Elizabeth Ashford has pressed this point in critiquing Scanlon's version of the Duty of Easy Rescue. 'The Demandingness of Scanlon's Contractualism,' pp. 273–302.

with. Like Parfit here, most non-consequentialists who have addressed the distinction at all have treated its moral significance as self-evident.[21]

It is not. Given the decisive role that certainty, along with identifiability and proximity, plays in generating the 'right' answer to trolley problems and the critical role that trolley problems in turn play in formulating the Duty Not to Harm, establishing that these factors have a moral and not just psychological or emotional basis is essential to the non-consequentialist project. The first step is to recognise the decisive role that all three factors are *in fact* playing in trolleyology. And that recognition will come only when trolleyologists leave behind the world of trolley problems for the world of uncertain risk of harm to statistical and/or remote victims, and observe systematically which of their intuitions survive the journey.

II.3. *Taking trolley problems as we 'find' them*

An unstated premise of trolleyology is that every trolley problem, whatever its factual setup, is equally appropriate grist for the philosophical mill, and that the 'best' articulation of our Duty Not to Harm is whatever set of principles can accommodate (make cohere) our intuitions about the largest range of possible setups. I believe this assumption cannot be defended, and has left trolleyologists grappling with non-existent problems and settling for gerry-rigged solutions to real ones.

Consider the question of what information people are permitted to have about their own ex post fate under a proposed principle when deciding whether to accept (reject) that principle.[22] Such knowledge swamps in importance all other 'facts' that influence individuals' preferences. If we ask whether hypothetical person X could reasonably agree to ambulances being permitted to speed whenever it is the case that five patient lives will thereby be saved for every one pedestrian killed, we are likely to get one answer (no) if X knows she will be the one killed and a different one (yes) if, as far as she knows, her odds of turning out to be one of the five are five times greater than her odds of turning out to be the one.

Which answer is the relevant one, morally speaking? The de facto response from many non-consequentialists is both: Whatever information we happen to have about our ex post fate under a proposed principle at any given moment is (in Parfit's words) a 'relevant, reason-giving fact' that

[21] The relevant literature here is substantial. For a summary and discussion of it, see Fried, 'Can Contractualism Save us From Aggregation?'

[22] For a more extended discussion of this issue, see Fried, 'Can Contractualism Save Us From Aggregation?'

must be taken into account in assessing whether, as of that moment, we could reasonably accept that principle (*On What Matters*, 1, p. 356). If at that moment person X 'happens' to know nothing about her own particular situation that would differentiate her odds from anyone else's, we count her as in favour of letting ambulances speed (Ambulance I). If at that moment she 'happens' to know that she is the pedestrian about to be mowed down, we count her as against it (Ambulance II). And if, having agreed from the epistemological position of Ambulance I to let the ambulance speed, X subsequently learns that a speeding ambulance is about to mow *her* down (Ambulance II), we let her switch her vote (assuming there is time for her to do so after she learns about her impending fate and before it is sealed).[23] A number of trolleyologists have defended that position explicitly, on the ground that only by allowing X to switch her vote do we respect her right not to be killed against her own wishes (even at the cost of many other lives).[24]

Treating the epistemological point of view of the characters that populate a trolley problem as an exogenous fact that frames moral analysis but is not itself subject to it has produced no end of confusion in the trolley literature. The examples here are legion.[25] Consider one of the variants on Taurek's famous dilemma taken up by both Parfit and Scanlon in *On What Matters*.[26]

[23] The examples are based on Frances Kamm's 'Ambulance' cases in *Intricate Ethics*, pp. 29–30, 274–5. For further discussion of the widespread view among non-consequentialists that one has a right to renege on prior agreements once one learns one's own fate under them, see Fried, 'Can Contractualism Save Us From Aggregation?'

[24] Lenman, 'Contractualism and Risk Imposition,' p. 117; Kamm, *Intricate Ethics*, pp. 36–7, 273. Kamm's rationale for this conclusion seems to rest on a version of the act/omission distinction, in which conduct that ultimately causes harm counts as an 'act' only if we know with certainty, as of the moment we act, both that someone will be harmed and the identity of that person. Her defense of drawing this particular line (as I read it) merely reasserts it: Letting ambulances speed in Ambulance I does not violate anyone's 'right not to be killed' because the decision is made when 'no one does anything to cause harm aside from running vehicles justified for the sake of benefit even when there is a risk of unavoidable harm.' In contrast, letting them speed in Ambulance II 'would sanction doing what will (as a side effect) kill someone when it could easily be avoided (albeit, at the opportunity cost of saving more lives).' *Intricate Ethics*, 275. In one form or another, almost the whole of trolleyology rests on the same intuition that potentially harmful conduct undertaken in conditions of certainty is a different moral kind from the same conduct undertaken in conditions of uncertainty. As discussed below, while Parfit signs on to that distinction as well, in the end he insures it will do almost no work. But of course Parfit's goal, unlike Kantian contractualists and other non-consequentialists, is not to defend trolleyology but to show it will lead us back to consequentialism.

[25] I take up a number of them in 'Can Contractualism Save Us from Aggregation?'; B. Fried, 'The Limits of a Non-consequentialist Approach to Torts,' *Legal Theory* (forthcoming) (draft available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1957467).

[26] John Taurek, 'Should the Numbers Count?' *Philosophy and Public Affairs*, 6 (1977), pp. 293–316.

In Taurek's original hypothetical, a would-be rescuer (Green) is faced with the choice of saving five people (White, Blue, Yellow, Red and Black) on one rock or one person (Orange) on another. The question posed by Taurek is: if Green chooses to save the five rather than the one based solely on the numbers, will she thereby violate Orange's Kantian right to equal respect? Scanlon, like virtually every non-consequentialist (other than Taurek himself) who has considered the hypothetical, concludes the answer is no: At least where one is choosing between identical harms, both of which fall on the same side of the act/omission divide, Scanlon argues we should let the numbers count as a tiebreaker.

In one of the variants that Parfit and Scanlon take up in *On What Matters* (Lifeboat II), Orange is Green's child. As a consequence, Green now has a strong personal preference to save Orange rather than the five. Both Parfit and Scanlon conclude that, contra Lifeboat I, it is now permissible for Green to save Orange at the cost of the five, because over the long run, allowing each of us to be partial to our children in this way would make life go significantly better for almost all of us. In Parfit's words:

> The good effects [of saving 5 rather than 1] would be massively outweighed by the ways in which it would be worse if we all had the motives that such acts would need. For it to be true that we would give no such priority to saving our own children from harm, our love for our children would have to be much weaker. The weakening of such love would both be in itself bad, and have many bad effects. Given these and other similar facts, the optimific principles would in many cases permit us, and in many others require us, to give strong priority to our own children's well-being [*On What Matters,* 1, p. 385].

Fair enough. But how does one get to the optimific solution from non-optimific (Kantian/contractualist) premises? As Parfit notes, the easy answer is the one supplied by ex ante contractualism: count only the preferences people would have before they possessed any individualised knowledge about the odds they would turn out to be Green, Orange or one of the five (*On What Matters,* 1, 349–50). From that position of equal ignorance, if everyone makes the same calculation about the expected value of partiality *in their lives* as Parfit has made with respect to the average person, everyone will prefer a rule that allows Green to be partial to his own child. (This just illustrates the general point that where preferences are more or less uniform, ex ante contractualism converges with the optimific solution.)

From the start, however, Scanlon rejected ex ante contractualism on principle, insisting that representative complaints must be weighted in accordance with the actual (ex post) harms that will befall actual victims, a view that most Scanlonian contractualists have adopted as well:

> Suppose that A is a principle which it would be rational for a self-interested chooser with an equal chance of being in anyone's position to select. Does it follow that no one could reasonably reject A? It seems evident that this does not follow. Suppose that the situation of those *who would fare worst* under A, call them the Losers, is extremely bad, and that there is an alternative to A, call it E, under which no one's situation would be nearly as bad as this. Prima facie, the losers would seem to have a reasonable ground for complaint against A.[27]

But weighting complaints in accordance with actual (ex post) harm produces the 'wrong' answer in Lifeboat II (Green may not choose to save his own child). Ignoring for the moment Green's agent-relative preference to save his own child, the now self-identified White, Blue, Yellow, Red and Black and the now self-identified Orange each face certain death if Green's choice doesn't go their way. This gives each of the six a complaint (certain death) of equal qualitative strength. This is so, even if each of the six, judging the matter from a position of equal ignorance about the consequences *to their lives* of adopting one rule rather than the other, would have chosen to let partiality trump.

At that point, faced with identical complaints from each of the six who are facing certain death, Scanlon's tiebreaker principle would kick in to let the numbers count, once again yielding what Scanlon and Parfit both take to be the 'wrong' answer: Green must save the five rather than his own child, Orange. Adding in Green's preference would seem to change this result only if we are allowed to aggregate Green's reasons to save Orange with Orange's reasons to want to be saved, and count the sum as one individual's reason. How exactly we could justify aggregating Green's and Orange's reasons under a Scanlonian 'individual reasons' restriction is not obvious. Moreover, even if we could justify it, Orange clearly wins the day under Scanlon's Relative Complaint Model only if we ignore the preferences of all the off-stage loved ones of the five who will die as a result. If we assume that at least one of those off-stage loved ones has as strong a preference to have his or her family member saved as Green has to save Orange, then we once again have a tie, relegating us to Scanlon's tiebreaker principle, which again yields the 'wrong' answer.

So instead, Scanlon does what trolleyologists generally do whenever adhering to the ex post perspective generates the 'wrong' answer: He reverts, sub silentio, to ex ante contractualism:

[27] T. Scanlon, 'Contractualism and Utilitarianism', in A. Sen and B. Williams (eds.) *Utilitarianism and Beyond* (Cambridge UP, 1982), pp. 122–23 (italics added). See also Scanlon, *What We Owe to Each Other,* p. 207.

> A principle requiring us always to give the needs of strangers the same weight as those of friends and family members would be one that each of us could reasonably reject, because it would make impossible special relationships that we have strong reasons to want to have [*On What Matters,* 2, p. 133].

Parfit, in contrast, sticks to an ex post perspective in Life Boat II (*On What Matters,* 1, 382, 388–9), but argues that Kantian contractualists can – indeed, must – reach the 'right' (optimific) answer even from that perspective. The argument is complicated, but for immediate purposes the essential moves are these:

(i) In most situations, each of us has sufficient, morally plausible reasons to act either selfishly or altruistically.
(ii) But Kant's Universal Law imposes a further requirement, beyond that we have sufficient reasons to act as we do: We must also act in accordance with principles that everyone could rationally choose, if they were choosing principles that everyone would accept (*On What Matters,* 1, 401–2).
(iii) That further requirement will always be met by the optimific rule, because all of us, while we may have good reasons to act selfishly, by stipulation also have good reasons to choose those principles that will on the whole make things go better for everyone over time (*On What Matters,* 1, pp. 378–379) – that is, to act altruistically.
(iv) In contrast, it will never be met by a principle that is significantly non-optimific, because those who are disadvantaged by the principle more than trivially have neither self-interested nor altruistic reasons to accept it.
(v) Thus, in Lifeboat II, while White, Blue, Yellow, Red and Black have sufficient reason to prefer themselves, they also have sufficient reason to choose the optimific rule (Green may choose to save his own child), even at the cost of their own lives. In contrast, Green has no good reason to accept the non-optimific rule that he must save the five in preference to his own child, because it satisfies neither his altruistic nor his self-interested motivations.

To generalise Parfit's conclusion, whenever self-interest and the optimific rule diverge significantly, Kantians and Kantian contractualists should agree that the optimific rule always wins (*On What Matters,* 1, 378).

But why should we assume that White, Blue, Yellow, Red and Black have sufficient individual reasons to choose the optimific rule, even knowing that their own lives are at stake? As Parfit notes, a desire-based, subjectivist account of reasons will not yield that result, unless we

impute an implausibly high degree of altruistic motivation to individuals (*On What Matters,* 1, 381). At the other extreme, an objectivist account of reasons that holds that 'we always have most reason to do whatever would be impartially best' ('Rational Impartialism') (*On What Matters,* 1, 382) will always yield that result – indeed, it will yield the stronger result that the *only* principle we have sufficient reason to choose is the optimific rule. But it reaches the 'right' (consequentialist) conclusion only by assuming it.

Parfit's actual route to the optimific result in *On What Matters* is different from Rational Impartialism, and subtler. Like Rational Impartialism, Parfit's account of sufficient reasons is objectivist. But unlike Rational Impartialism, his version (a 'wide, value-based objective view') accepts that both partiality and altruism can supply morally sufficient reasons to act, depending on the circumstances. Parfit then argues that one circumstance in which all people would have a morally sufficient reason to sacrifice themselves for others, even if they ultimately choose not to, is when the stakes are high enough – if, for example, they would thereby save millions of other people. (Parfit is not making a claim here about peoples' actual motivations, but it is a plausible description of those as well.) He then converts virtually every trolley-type problem into such a high-stakes choice, by stipulating that 'in the thought-experiment to which the Kantian Formula appeals,' all the on-stage players would weigh in the balance the consequences not just to each other, but to the 'billions of people' who would be governed by a given principle 'both now and in all future centuries' (*On What Matters,* 1, p. 382).

Thus, in Lifeboat I (Taurek's original example) Parfit's Orange would think to himself, if we adopt a rule that says that someone in Green's position should always save the larger number, '[t]hough I would die, my choice would indirectly save at least a million other people…. So even on … more egoistic views, I would have sufficient reasons to give up my life to save these very many other people.' And in Lifeboat II, Parfit's White, Blue, Yellow, Red and Black would think to themselves, if we adopt a rule that allows someone in Green's position to be partial to his own child, we will die, but we will make it possible for billions of people from now through all eternity to be partial to their children in a way that will make all of their lives go significantly better. (Query why the self-sacrificers wouldn't be entitled at the very least to offset the gains to billions of future people from adopting the optimific rule with the aggregate loss to (millions of?) future people whose lives (like theirs) will go better under the non-optimific rule? Doing so will not change the optimific conclusion – by definition, optimific rules will always produce greater overall welfare

than non-optimific ones. But in many cases it might reduce the welfare gap between optimific and non-optimific rules sufficiently to boot the self-interested, non-optimific choice out of the 'high stakes' camp.)

A good case can be (and has been) made that both ex ante contractualism and Parfit's ex post high-stakes altruism produce consequentialist results by smuggling in consequentialist premises. But for present purposes, the point I wish to make is different and more general: The 'right' answer to any trolley problem depends entirely on how the problem is set up, and in particular what information we impute to the choosers at the moment of choice. When Kamm and Scanlon pose the Ambulance and Lifeboat II problems, respectively, from the ex ante epistemological perspective, they get the optimific result. When they pose the identical problems from the ex post perspective, they get the nonoptimific result. So also with Parfit: viewing Lifeboat II initially from the conventional ex post epistemological perspective (count only the preferences of the on-stage players, all of whom know how they will fare under the available choices), he gets the nonoptimific result. But then he in effect smuggles in a partial ex ante perspective, by imputing to the on-stage players knowledge of the aggregate expected costs of the available choices going forward to the end of time, thereby tipping their preferences over to impartiality.

Given the decisive role played by the set-up of a trolley problem, it is imperative for trolleyologists to defend the normative relevance of the set-ups they choose. If they do not – if they take each hypothetical as it just 'happens' to arise – then they will end up with a set of principles that are confused at best, and indefensible at worst. In my view, that pretty much describes the state of trolleyology at the current moment.

What set-up *should* non-aggregationists adopt in analysing the sorts of tragic tradeoffs posed by trolley problems? Insofar as they are seeking general principles to live by – the self-described project of trolleyologists – I think the *only* morally defensible epistemological point of view from which to ask what principles everyone would have reason to accept is *before* they know how things will actually turn out for them ex post, over the course of their lives, if a given principle is accepted. It is also the only point of view that can yield anything approximating general agreement (unless, like Parfit, we allow self-interested motivations to count in theory but ensure they will never carry the day in fact whenever they lead to significantly suboptimal results.)

As I have argued elsewhere, trolleyologists have concluded otherwise – have concluded, that is, that everyone is entitled to take into account whatever information they just happen to possess in whatever hypotheti-

cal factual scenario we just happen to place them – because they have confused the question whether, in choosing principles to live by, people are entitled to know their own preferences, aptitudes, and general situation in life (yes, if one rejects the Rawlsian veil of ignorance in favour of differentiated, 'thick' selves) with the question whether they are also entitled to know how things will actually turn out *for them* if a given principle is adopted.[28] Parfit makes the same mistake in *On What Matters*, in assuming that 'relevant, reason-giving facts' include both general information about who we are and the circumstances we occupy in life *and* knowledge of whether we will turn out to be 'one of the few people on whom … great burdens would be imposed' if a given principle were chosen (*On What Matters*, 1, p. 356).

But general knowledge about ourselves and our circumstances in life is sufficient to generate 'thick,' differentiated representative individuals and surface the genuine disagreements among them. Whether individuals should be endowed as well with knowledge about ex post outcomes is a separate question that is orthogonal to whether we are committed to thick selves: Green is Green before he knows that *he* is the one who will have to choose between saving his own child and saving five, and he is Green after he discovers it. The question is, which Green's preferences is it morally appropriate for us to count? If it is the latter, that proposition has to be established on some grounds other than the commitment to 'thick' selves.

But suppose we agree that Ambulance I is the appropriate epistemological perspective from which to decide whether ambulances should be permitted to speed whenever it is the case that doing so will save more lives, and that from that perspective we conclude everyone would say yes. Now suppose that one of the parties to the agreement in Ambulance I subsequently finds out, while there is still time for the ambulance driver to hit the brakes, that she is the unlucky Pedestrian who is about to be mowed down to save five. It is one thing for everyone to agree in principle that ambulances should be permitted to speed whenever doing so will save five statistical patient lives for every one statistical pedestrian life thereby lost. It is quite another thing, many will feel, for a driver to adhere to that agreement when Pedestrian Smith walks in front of the ambulance and the driver could avoid hitting him by applying the brakes. I agree. The question is, what kind of thing is it, what should we do about it, and how can we reconcile that response with our response in Ambulance I?

[28] Fried, 'Can Contractualism Save Us From Aggregation?'

For utilitarians, the answer to these questions is straightforward, at least in theory. Both choices should be resolved by the same moral criterion: pick the decision rule that will maximise expected welfare. What that rule is and whether we should have different rules depending on whether the decision is being made from the epistemological perspective of Ambulance I or Ambulance II turns purely on empirical assumptions.

In cases that come up in the posture of Ambulance I (ex ante decisions under conditions of uncertainty), presumably the optimal rule will usually be to sacrifice the one to save the five. The optimal rule in Ambulance II is a more complicated calculus. Ambulance II poses a very narrowly defined dilemma: What should we do if we happen to have an opportunity to revisit our earlier decision in Ambulance I *after* we know the identities of the one and the five and *before* the driver irrevocably commits to a course of action? Having that information changes the social meaning of the action for the driver, presumably for the worse; it may also change it for the worse for immediate bystanders and for the larger society, to the extent that each is aware of the nature of the choice that the driver faced and made. If it changes the social meaning enough, it could conceivably tip the optimific calculus, leading a utilitarian to conclude that anyone who ever finds herself in the position of the driver in Ambulance II either may or must hit the brakes, notwithstanding the general agreement in Ambulance I to the contrary. But for utilitarians, that qualification introduces no analytical or moral inconsistency. It simply recognises that in some circumstances, following the 'sacrifice one to save five' rule will occasion enough psychic discomfort and/or political instability that the optimific move is to allow an exception. For utilitarians, the other viscerally powerful factors introduced in the move from Ambulance I to Ambulance II (that harm is no longer merely a statistical probability but now a sure thing; that it will occur up close and personal) raise identical concerns, with the identical solution: if adhering to the general agreement struck in Ambulance I will decrease overall wellbeing, utilitarians would want the driver to defect from that agreement.

My guess is that most non-consequentialists, in the end, would strike roughly the same compromise as utilitarians: Adopt an ex ante decision rule more or less by the numbers, but permit or require individual actors to deviate from the rule in cases in which adhering to it in the face of 'certainty' that harm will result, identification of the victim, or the prospect of inflicting it (or have others witness it) up close and personal seems intuitively and overwhelmingly 'wrong.' This is in fact the compromise they have struck in dividing the universe of potentially harmful conduct into certain harms (trolley problems) and uncertain harms (risk), and

resolving the former by non-consequentialist principles and the latter by aggregative principles that differ little if at all from the optimific solution.

Unlike welfarists, however, it is not clear that non-consequentialists can account for both halves of that compromise by some unifying moral principle.[29] The candidate principle offered to date – that a policy that 'is not acceptable at every time [to every person] is plausibly acceptable at none'– is not going to do the job.[30] At most, it explains why we are drawn to let the pedestrian who finds herself trapped in Ambulance II renege on her assent to the optimific solution in Ambulance I ('Sure, she said that then, but look at her now'). But it does not explain why her assent in Ambulance I was good to begin with, since it was foreseeable from the vantage point of Ambulance I that whoever in that group ultimately turned out to be the pedestrian in some future Ambulance II would live to regret her decision. And it does not explain why we *ought* to let the victim renege in Ambulance II. As much as she now finds unacceptable, from the vantage point of Ambulance II, her agreement to the optimific solution in Ambulance I, she would find equally unacceptable, from the vantage point of Ambulance I, that the future preference of some as-yet unidentified pedestrian in Ambulance II for the non-optimific solution will negate her own preference for the optimific solution.

To put the problem in general terms: Since our preferences with respect to a given policy will always change with different epistemological perspectives, the principle that a policy that 'is not acceptable at every time [to every person] is plausibly acceptable at none' means that no policy will ever be plausibly acceptable. It also means that if a pedestrian's impending death is discovered the moment before the driver can hit the brakes to save her death, her interests are dispositive, but if it is not discovered until the moment after, they are not counted at all. What moral principle could possibly justify this distinction?

Finally, the proposition that every epistemological vantage point is of equal moral relevance is hardly self-evident. For the many reasons spelled out in this article, I think it is difficult if not impossible to defend. But at a minimum, it requires defense.

Other non-consequentalists have responded, in effect, of course the 'right' results in Ambulance I and II cannot be explained by the same moral principle. But why would we think they should be, given that they present very different moral problems? As my discussion in section II.1 suggests, that view seems hard to defend if the fundamental 'bad' in both

[29] For the argument that it cannot, see Norcross, 'Comparing Harms: Headaches and Human Lives.'

[30] Lenman, 'Contractualism and Risk Imposition,' p. 117.

scenarios is the actual harm that could or will befall real people. Why would potential harm to others present a moral problem that is different in kind, depending on whether it is deemed certain or uncertain to come to pass?

If the fundamental 'bad' posed by Ambulance II is not the impending harm itself but rather the secondary features introduced in the move from Ambulance I to Ambulance II (the certainty that harm will result, the identifiability and proximity of the victim, etc.), then there is indeed no reason to assume that we should resolve Ambulance II in accordance with same principles applied in Ambulance I. But in that case, trolleyologists face two further challenges.

The first they share with utilitarians: to show how inconsistent decision rules applied to the same act depending upon the epistemological vantage point from which it is judged can be made to cohere into a socially stable regime. The second they do not: to show that the decision rules they would adopt from the vantage point of Ambulance II have a moral, and not merely psychological or emotional, basis.

Because non-consequentialists have said so little about the appropriate principles to govern decision-making under conditions of uncertainty, it is hard to say more about how those principles can be reconciled with the principles extracted from trolleyology. But they have voted with their pens on the relative philosophical importance of the two, enshrining trolley problems as the paradigmatic case testing the scope of our duty not to harm (duty to rescue) others and risk as a moral sideshow. As is abundantly clear by now, I think that gets things exactly backwards.