Stanford Law School

What *Does* Matter? The Case for Killing the Trolley Problem (Or Letting it Die)

Barbara H. Fried
Stanford Law School

Research Paper No. 1781102

Stanford Public Law and
Legal Theory
Working Paper Series

Stanford Law School
Crown Quadrangle
Stanford, California 94305

This paper can be downloaded without charge from the
Social Science Research Network Electronic Paper Collection
http://ssrn.com/abstract=1781102

Electronic copy available at: http://ssrn.com/abstract=1781102

Barbara H. Fried, “What *Does* Matter? The Case for Killing the Trolley Problem (Or Letting it Die)”

ABSTRACT

For the past forty years, a significant portion of nonconsequentialist moral philosophy has been devoted to refining our moral intuitions about the harms to others we may or may not causally bring about through our acts or omissions. Discussion has focused almost exclusively on trolley-type hypotheticals that share the following features: The consequences of the available choices are stipulated to be known with certainty ex ante; the agents are all individuals (as opposed to institutions); the would-be victims are identifiable individuals in close proximity to the agents; and the agents face a one-off decision about how to act (that is to say, readers are not invited to consider whether the moral principles by which the immediate dilemma is resolved can be scaled up to a large number of (or largenumber) cases).

Derek Parfit’s forthcoming book, 'On What Matters', is the most recent addition to the trolley-problem oeuvre. In this review essay, I argue that the tyranny of trolley-type problems in philosophical thought has yielded moral principles that (whatever their moral virtues) cannot be applied beyond trolley-type cases. In particular, they cannot help resolve the permissibility of the sort of conduct that accounts for virtually all harm to others outside of the criminal context: socially useful conduct that poses some risk of harm to as yet unidentified others. If nonconsequentialist principles about permissible harm to others cannot shed light on the problem of risk, they are doomed to at best marginal significance.

Electronic copy available at: http://ssrn.com/abstract=1781102

Revised draft March 23, 2011

What *Does* Matter? The case for killing the trolley problem (or letting it die)

- Barbara H. Fried
William W. and Gertrude Saunders Professor of Law
Stanford University
bfried@stanford.edu

1. Introduction

For the past forty years, a significant portion of nonconsequentialist moral philosophy has been devoted to refining our moral intuitions about the harms to others we may or may not causally bring about through our acts or omissions. The literature has mined two veins. The first concerns when we may actively inflict harm on some to save others from harm. The second concerns the scope of our affirmative duty to rescue. Forcible transplant cases, Bernard Williams's Jim and the Indians, trolley problems, munitions bombing cases and other hypotheticals raising the doctrine of double effect, are standard examples of the first. Examples of the second have traditionally involved isolated heroic rescues (e.g., our good Samaritan duty to save a drowning man); at least for the moment most nonconsequentialists show no inclination to extend the duty of rescue beyond such isolated events (a 'Duty of Easy Rescue,' in Scanlon's terminology), to (for example) structural and widespread calamities created by poverty or disaster (starvation, disease, etc.).

The "duty not to harm" and conventional "duty of easy rescue" hypotheticals typically share the following three features: The consequences of the available choices are stipulated to be known with certainty ex ante; the agents are all

Electronic copy available at: http://ssrn.com/abstract=1781102

individuals (as opposed to institutions); and 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). In addition, agents 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 sacrificing some to benefit others vary, for ease of exposition, I will follow Prof. Allen Wood's lead and refer to all hypotheticals that share these features as "trolley problems."[1]

Derek Parfit's just-published *On What Matters* is representative in this respect. In the course of hundreds of pages, Parfit offers and interrogates dozens of hypotheticals to tease out our 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 known consequences to identified, close at hand, fictitious persons[2]-- that is, everyone is (in my sense of the term) a trolley problem.

The tyranny of the trolley problem has, in my view, shaped the nonconsequentialist literature on harm to others in a number of unfortunate ways.

---

[1] Allen Wood, "Humanity as an End in Itself," in On What Matters (Oxford University Press, 2011), vol 2, pp. 58-82.

[2] 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 10 years later by a five-year-old child (vol. 2, pp. 217-28). 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 recognize 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 figure out how to deal with the as-yet nonexistent victim.

First, and most directly, it has resulted in nonconsequentialists' devoting the bulk of their attention to a scenario that is significant in the realm of intentional wrongdoing by others (criminal activity, warfare), while largely ignoring conduct that in the civil context accounts for virtually all harm to others: conduct that is socially productive but carries some uncertain risk of harm to (generally unidentified) others.[3] Various moral principles have emerged from the now four-decades-long preoccupation with trolley problems, but none of them can handle the problem of garden-variety risk. Nonconsequentialists who acknowledge this limitation have nonetheless defended the moral intelligibility of trolley-generated moral principles by arguing that harming others and imposing a risk of harm on others are different in kind, and hence appropriately resolved by different moral principles. I believe that view is indefensible as both a descriptive and normative matter. But even if it were not, it saves trolleyology at the high price of ceding virtually all harm-producing conduct outside of the context of war and criminal activities to welfarism.

Second, and relatedly, the hermetic focus on trolley cases has led nonconsequentialists to misdiagnose what is going on in trolley cases themselves. The only way to get purchase on the question of whether our moral intuitions are

[3] Another way of making the point that trolleyologists are engaged in what is generally a moral side show is to put their argument in legal terms. From a legal perspective, the trolley literature is addressing a minor issue in criminal law— whether one can *justify (vel non)* conduct that is prima facie legally wrongful and hence prima facie illegal (i.e., acting in a manner intended to cause death or serious injury to another party)—while failing to address the major issue on which it is resting: what sort of potentially harmful conduct *is* prima facie wrongful (e.g., when is it unreasonable to engage in every-day conduct that poses some risk of harm to others). These latter issue is not a moral sideshow in the context of 'just warfare' problems, precisely because such problems typically involve conduct that everyone understands to be prima facie wrongful (deliberately torturing or killing another); the only question on the table is whether such acts may be justified or excused under the extraordinary circumstances posed by terrorist threats or warfare.

responding, principally, to the fact that I have harmed (failed to rescue) you or instead to secondary features that are peculiar to trolley problems is to examine how intuitions change when we relax each of those of those features. The trolley literature has meticulously tracked how our intuitions change when we vary factual assumptions from one trolley problem to another. What it has not done is to see which of these intuitions survive if we relax factual assumptions that are common to *all* trolley problems—in particular, the assumptions that we are dealing with one-off cases in which the outcomes of available course of action are known with certainty ex ante.

Third, trolleyologists have generally treated the factual posture in which trolley problems 'arise' —and in particular the knowledge that the hypothetical actors are taken to possess about the consequences *to them* of adopting different courses of action-- as exogenous variables that define the moral problem to be solved but are themselves morally neutral. In my view, that position is difficult if not impossible to defend. It has led to no end of analytical confusion in the literature. It has also opened the entire trolley enterprise to the charge that it is begging the question. To put the point in the strongest terms, if nonconsequentialists decided on further thought 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* if that principle is adopted, then virtually the entire trolley literature becomes morally irrelevant, and nonconsequentialists will likely end up with criteria for determining the permissibility of potentially harmful conduct that

diverge little if at all from standard aggregative techniques.[4]

Given that Parfit's ultimate objective in *On What Matters* is to show that all roads—in particular Kantian and Contractualist-- lead to something very close to standard aggregation, it is something of a mystery why he didn't help himself to this straightforward path from nonaggregative premises to aggregative conclusions, rather than assuming the premises of trolleyology and then showing he could get to aggregation even from that inhospitable starting point. But since he has chosen the latter course, I would like to use the publication of *On What Matters* as an occasion to make the case for nonconsequentialists' killing the trolley problem, or at the very least letting it languish for awhile, while they turn their attention to the much more important task of figuring out what nonconsequentialist principles have to say about the problem of risk.

2. What Has the Trolley Problem Wrought?

2.1 Minimizing the prevalence of tragic choices (or, what happened to risk?)

Over half of Allen Wood's commentary on Parfit's Tanner lectures, reprinted in On What Matters, is devoted to an extended attack on "trolley problems." As I

[4] An example here is the ex ante contractualist solution to Taurek's famous dilemma. See, e.g., Jussi Suikkanen, 'What We Owe to Many,' Social Theory and Practice 30 (2004): 485. Suikkanen suggests that we can get to what most people take to be the obviously right answer—save the five on one rock rather than the one on the other—by imagining what all six people would have chosen if we had asked them at an earlier moment in time, when each of them had no reason to think that their odds of being one of the five were anything but five times greater than their odds of being the one. But as Larry Alexander has observed, the argument proves much more than deontologists are going to want to admit, as "ex ante contractarian arguments tend to efface all deontological constraints and lead to pure consequentialism." "Lesser Evils: A Closer Look at Paradigmatic Justification," *Law and Philosophy,* 24:6 *(Nov., 2005*), pp. 611-643 at 617 n.23. For further discussion of the respects in which one might expect the conclusions of ex ante contractualism to diverge from conventional aggregation, see Barbara H. Fried, "Can Contractualism Save Us From Aggregation?" http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1781092.

had been thinking along the same lines myself, I was delighted to come upon an ally, particularly one who (unlike me) has unimpeachable credentials as a Kantian moral philosopher. But as I read further, it became clear that Prof. Wood's principal objection to trolley problems was very different from—indeed, opposite to-- mine.

Before turning to our differences, I should say that I agree with many of Prof. Wood's objections: The fantastical nature of the dilemmas trolley problems ask us to resolve; the demand that we choose whether to kill one person or let five people die without the sort of contextual information that in real life changes the moral complexion of a choice (e.g., was one of the parties' presence on the tracks more negligent than others?); and most of all, the stipulation that the outcomes of all available choices are known with certainty ex ante (a condition, Prof. Wood rightly points out, that is never met in real life). Prof. Wood's critique is worth quoting at length:

> Most of the situations described in trolley problems are highly unlikely to occur in real life and the situations are described in ways that are so impoverished as to be downright cartoonish…. The deceptiveness in trolley problems is indirectly related to their cartoonishness… in that it consists at least partly in the fact that we are usually deprived of morally relevant facts that we would often have in real life, and often just as significantly, that we are required to stipulate that we are certain about some matters which in real life could never be certain. The result is that we are subtly encouraged to ignore some moral principles (as irrelevant or inoperative, since their applicability has been stipulated away). And in their place, we are incited to invoke (or even invent) quite other principles, and even to regard these principles as morally fundamental, when in real life such principles could seldom come into play, or even if they did, they would never seem to us as compelling as they do in the situation described in the trolley problem. (draft ms. 400-401)

My disagreement with Prof. Wood concerns the ways in which the collective preoccupation with trolley problems has (mis)shaped nonconsequentialist thought. Prof. Wood believes that obsessing on trolley problems has led nonconsequentialists to overestimate the real-life necessity of making tragic choices 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 effects can—and I believe do—occur simultaneously. But the latter is, to my mind, the far more troubling one from a policy perspective.

Wood's argument that trolley problems lead us to overestimate the necessity of tragic choices is straightforward: By stipulating that we face a tragic choice and focusing all attention on how we ought to choose-- whether to kill one or let five die, whether to rescue the five on one rock or the one on another-- trolley problems encourage readers to take as given that such tragic choices are 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." (vol. 2, p. 74). By pushing those earlier decisions offstage, trolley problems have 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.

Prof. 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 (vol. 2, p. 79). 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." (vol. 2, p. 80).

I agree with Prof. Wood that many of the tragedies that occur in the developed world are easily preventable. Indeed, a case could be made that the failure to invest prudently in prevention-- using "prevention" broadly to refer to all ex ante investments to improve ex post outcomes-- is the most serious problem facing contemporary American society. But in my view the fixation on trolley problems doesn't play a significant role in that systemic failure, except indirectly, by leading people to misunderstand the nature of the choices we face when we *do* invest proactively in prevention.

In contrast, I believe that nonconsequentialists' 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 problem literature has

inadvertently led both authors and consumers of the literature to regard tragic choices *themselves* as rare and exceptional, indeed freakish, in nature. But they are neither of these things. They are ubiquitous and for the most part quotidian, and typically result not, as Prof. Wood puts it, from "human vulnerability to nature, and … human wickedness" (vol 2, p. 81), but from the finite nature of the resources we depend on to realize 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 meet all competing demands for it.[5] In this sense of the word, most 'goods' in society are necessarily scarce, either because the material resources necessary 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 towards treating all treatable diseases just by allocating our current health care budget more sensibly. But the needs on the prevention end are virtually limitless (diabetes,

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

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 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 and groups." (408) Moreover, the way we make those tradeoffs typically invites the one outcome that almost all nonaggregationists agree is morally impermissible: inflicting serious harm or death on (failing to rescue) one person in order to realize trivial benefits for others, no matter how numerous those others are. [6]

---

[6] Scanlon provides the central articulation of this principle, in what Parfit has dubbed Scanlon's "Greater Burden Claim" (GBC). Vol. 2, ch. 21, pp. 191-93.

The act/omission distinction is lurking troublesomely in the background here. Under at least some circumstances, ll trolleyologists would (all other things being equal) impose a higher duty on people not to harm others by their actions than to save them from impending harm. That asymmetry requires one to qualify, in complicated ways, any broad principle giving trumping power to the person facing the greatest burden under a proposed principle. Where the choice is whether to inflict harm on A or B, Scanlon's GBC applies in a relatively straightforward manner. Ditto where the choice is whether to rescue A or B. The asymmetry matters only when the choice is between inflicting lesser harm on some to save (rescue) others from a greater harm. In that case, depending on the nature of the harm and how it is inflicted, the former group may prevail. The classic hypothetical here is the forcible harvesting of one person's organs (at the cost of suffering or a trivially shortened life) in order to save others from death.

I have deliberately left ambiguous here whether trolleyologists would consider the government or the trolley company's investment, or failure to invest, in safety measures an act, an omission, or neither. Working through this and other complications inherent in operationalizing the act/omission distinction requires a much longer discussion than is appropriate here. For current purposes, I hope at least one of the following observations will suffice to convince trolleyologists that the argument I advance here is not one they can dispose of simply by noting that investing in preventive measures falls under our duty to save rather than duty not to harm:

(i) Investing or failing to invest in safety measures can be considered an act, either directly, or indirectly because it determines the consequences of subsequent conduct that is unambiguously an

To make this concrete, consider the problem of trolley safety. Suppose we decide that we should 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 nonconsequentialist would disagree with that, and all would draw the line considerably short of $500 billion. But wherever we draw it, we will be knowingly choosing to increase the number of preventable deaths 'merely' to save some money (to be used for some other generally offstage purpose.)

---

act. (Prof. Wood seems to have the latter in mind, when he imagines we (the government? The trolley company?) have a duty to "prevent[] anyone from being in places where they might be killed or injured by a runaway train or trolley." (vol. 2, p. 74);

(ii) investing (or failing to invest in) safety measures is not an act for these purposes, but may well fall under our duty to rescue (see, e.g., Frances Kamm's hypothetical of the headache cure, *Intricate Ethics*, at pp. 35-37. While Kamm inclines to the view that "macro decisions whether to invest in research to cure a disease that will kill a few people or in research to cure a disease that will only wither an arm in many" are not morally equivalent to decisions to save many arms in the here and now rather than a few lives, the distinction in her mind seems to be based on the distinction between certain and uncertain harms, *not* on a judgment that such macro decisions wouldn't be considered a form of rescue to begin with.)

(iii) Investing or failing to invest in safety measures doesn't itself fall under the duty not to harm/duty to rescue, but we can fold exactly the same requirements into what is unambiguously an act, by describing the duty of the trolley driver, or the company for which he is an agent, to be to "drive with adequate safety precautions," including making sure that proper signage is posted along the tracks, proper safety gates installed, etc. And if all else fails…

(iv) substitute a different example involving trading off lives for lesser goods that doesn't raise any issue under the act/omission distinct. E.g,: At what speed is it permissible for me to drive through the downtown area, knowing that for every incremental 5 MPH over zero, the risks killing or maiming someone double, starting (at 5 MPH) at one person for every 50,000 trips? Why not forbid everyone from driving at all through downtown, at minor—or even major-- inconvenience to them all, if it will save (say) one statistical life over a five year period?

If that choice is immoral, almost everything we do is immoral, as almost every action imposes some risk (however slight) of serious harm on some, typically to secure much less weighty benefits for many. 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 that choice is not immoral (as I assume all nonconsequentialists would agree), why? 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 permit one death for the mere convenience of millions of others cease to be tragic, just because we don't yet know the identity of the victim or because her 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 we decide when a risk is 'too great' or the benefits to others too small to permit it?[7]

This is, of course, the garden-variety problem of risk. Once we set to the side criminal acts and other forms of conduct that generate no social benefits (or at least none we are willing to count) and hence pose no interpersonal tradeoffs between agent and victim that we care about,[8] almost all harm inflicted on others through human agency is accidental harm that results from conduct that is prima facie

---

[7] For an astute analysis of a real-life example (railway safety in England) of the public's inability to think rationally about these decisions— our inability, that is, to accept we could deliberately stop short of the utmost we could possibly do to reduce one isolated risk-- see J. Wolff, Risk, Fear, Blame and Shame: The Regulation of Public Safety," *Economics and Philosophy*, 22:3, pp. 409-427 (2006).

[8] They do involve other kinds of interpersonal tradeoffs that we do care about and that are also unavoidable-- e.g., the tradeoff between false positives and false negatives in conviction rates, the tradeoff between investing more money in lowering the crime rate and investing it in other functions of government that will disproportionately benefit a different group of people.

permissible but imposes some risks of harm on generally unidentified others. That is to say, the problem of unintentional (in the strong sense of undesired) harms to others *is* the problem of risk. 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 nonconsequentialist philosophers' obsession with trolley problems is to obscure this truth-- to encourage people to pay no attention to the man behind the curtain, as it were (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 unremarked-on in the nonconsequentialist literature. But in On What Matters, Parfit offers a rare, explicit defense of the choice to focus just on cases of 'certain' harms:

> 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. (vol. 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 the five would

have seen the trolley in time and moved out of the way. A fortiori, what is true of intentionally 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.[9]

In what sense then does risk present a "less fundamental" problem? Parfit supplies one possible explanation in the next sentence: "These [that is, problems of risk] are not the questions about which different people, and different moral theories, most deeply disagree." A second, related, explanation suggested in the literature is that the problem of risk is isomorphic to trolley problems, but with the stakes lowered on all sides. As a result, if we can resolve the right way to handle trolley problems, we will, a fortiori, have resolved the problem of risk as well. In either case, the claim is not that risk is an unimportant problem but rather that is an easy problem to solve, at least as compared to the problem of certain harm. On that assumption, it makes perfect sense to set the big guns to work on trolley problems, and either leave the problem of risk to others or ignore it entirely.

Both versions of the 'risk is easy' claim, in my view, get things exactly backwards. If classic trolley problems present a difficult choice for nonconsequentialists, garden-variety risk presents an impossible one. And the moral principles nonconsequentialists have extracted from their encounter with

[9] What distinguishes intentional from accidental harms is not uncertainty per se, but two other factors: (i) acts intended to harm others are *more likely* to result in harm than ones not intended to harm; or (ii) acts intended to cause harm to others are generally prima facie impermissible, whether or not they succeed; that prima facie assumption can be defeated only by showing they justified or excused by extraordinary circumstances. I doubt trolleyologists want to rest on the former distinction, since it concedes at the start that we are dealing with differences in degree, not kind. The latter distinction seems to me key, but generally has nothing to with the probability we will succeed in what we intend to do—that is, harm others.

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.

Given how few choices in life 'read' to the philosophical (and lay) mind as trolley problems, if nonaggregationists stuck to their guns in every one of them-- we may not kill one even to save 100 from death or 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. (Parfit departs from the conventional Kantian answer even in most of these cases, but that's another story.)[10] In this very practical sense, nonaggregative solutions to trolley problems are plausible, whether or not one ultimately finds them morally persuasive.

Not so in the case of garden-variety risks. If the numbers do not count in determining permissible risks, 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. In saying there is no 'deep disagreement' about how to handle risk, Parfit may mean to acknowledge this— to acknowledge, that is, that when it comes to risk, we are all aggregationists. A number of nonconsequentialists have explicitly reached that conclusion; many more, I think, have done so implicitly.[11] If that is indeed what Parfit means, it

---

[10] As noted in the formal commentaries included in *On What Matters,* Parfit appears to get the convergence he seeks between contractualism, Kantianism and consequentialism by assuming it—or more precisely, by assuming that the only "principles that everyone could will to be universal laws" are "optimific" [that is, Rule Consequentialist] principles, (vol. 1, p. 411), a conclusion that rests on the normative assumption that "car[ing] as much about some … things [other than our own well-being], such as the well-being of others," is objectively rational. (vol. 1, p. 358)

[11] For further discussion, see Barbara Fried, "Is There a Coherent Alternative to Cost-Benefit Analysis?" (draft ms. 2011, pp. 27-29). For representative statements ceding the problem of risk to

justifies the decision to ignore risk, but at a high cost to the nonaggregationist project. If the domain of nonaggregative principles is limited to harms that are ex ante certain to occur to known victims as a consequence of our acts (or, in the case of the duty of easy rescue, failures to act), given the rarity of such dilemmas outside of the context of warfare, can we possibly justify the amount of attention that has been given to them? Who cares whether or not we should flip the switch on the trolley tracks, thereby certainly saving five lives at the certain cost of one, if we will never face that choice in real life, and if the 'right' answer to the hypothetical, by nonconsequentialists' own acknowledgment, has no bearing on the kinds of choices we do face daily (whether it is permissible to undertake act X, knowing that it poses some risk of harming some number of as yet unidentified people)?

Alternatively, Parfit may mean to say that there *is* a viable, generally applicable, solution to the problem of risk that does not rely on aggregation, but as all nonconsequentialists are in agreement as to what it is, we need not belabor it. If that is what he means, once again I think he is wrong that such a solution exists. But I do not wish to argue the point here. Instead, I want to urge that whether or not it exists matters a lot more to nonconsequentialist philosophy and to social policy than the 'right' answer to trolley problems.

2.2 Misspecifying the appropriate domain of general principles on harm to others.

By their own account, trolleyologists are seeking general principles to guide

aggregative techniques, see Jules Coleman, Risks and Wrongs, at 210; Joel Feinberg, Harm to Others, 190-93; Thomas Scanlon, What We Owe to Each Other, at 204, 205.

our duty not to harm (duty to rescue) others, not rules for resolving particular cases. But the common-law method deployed in the literature gets at the general through the lens of the particular: Focus on a particular trolley problem and identify our moral intuitions about the 'right' answer; formulate a general principle that, when applied to that particular case, gets us to the 'right' answer and survives general moral reflection; test the principle against a second trolley problem to see if it gets us the intuitively right answer in that case as well; if it doesn't, revise the principle to generate the 'right' result in both cases, and so forth.

Whatever its other virtues, the method poses a couple of dangers, both arising from the domain of particular cases considered.

2.2.1 Limiting the domain of cases to trolley problems.

In one important respect foreshadowed above, the domain is too narrow to support the general principles that have emerged concerning the scope of our duties not to harm others (or to save them from harm). I suggested above that the failure to take risk seriously has left a gaping hole in the nonconsequentialist analysis of both duties. I want to suggest here that the obsessive focus on trolley problems has also done no favors to nonconsequentialists' understanding of trolley problems themselves. Once again, the claim may seem paradoxical, but I don't think it is.

While it is a fool's errand to try to generalize the normative conclusions drawn from trolley problems, I think it is fair to say that the following three

factors are identified more often than any others as morally salient: Whether we are a necessary cause of harm to others; the exact mechanism by which we cause that harm (chiefly the act/omission distinction),[12] and whether the harm we cause is much less serious than whatever other benefits we thereby secure.

But all three factors produce the 'wrong' answer if applied to garden-variety acts that impose risk of harm on others (e.g., putting up a building, knowing that there will be some irreducible risk of death to innocent passersby from falling debris). If serious injury to a passerby does result, the party constructing the building will be the active causal agent of it, and the harm it causes to the passerby will almost always be much more serious than the harm thereby prevented (e.g., inconvenience to those who would have occupied the building were it built). Notwithstanding that all three factors suggest that it would be wrong to put up the building as long as there is a non-zero chance of serious injury to passersby, no one-- trolleyologists most certainly included—thinks this is the right answer.[13]

This suggests not only that the general principles extracted from the trolley literature are useless in handling risk, but that they misidentify what is driving

---

[12] In the classic trolley problems where we are deciding whether it is permissible to divert the trolley, thereby "killing" one person to save several others, "acts" are often further subdivided into those that are 'upstream' of the harm, meaning a necessary means to achieve the desired beneficial result, and those that are downstream of the harm, meaning the foreseeable and unfortunate consequence of achieving it. The standard hypos in the literature illustrating the difference are the Fat Man (may we throw a fat man on the tracks *in order* to stop the trolley and thereby save five) versus the original trolley problem (may we divert the trolley away from the five, knowing that the unfortunate result will be to kill the one.) See Kamm, Intricate Ethics, at .

[13] For Scanlon's acknowledgement of that apparent inconsistency and his effort to justify it, see What We Owe to Each Other, pp. [ ] .

our intuitions in trolley cases as well. What then is driving them? Likely culprits include the other factors identified at the outset that are common to trolley problems: that the consequences of the alternatives are deemed to be known with certainty at the moment the agent must commit to a course of action; that the would-be victims are identified and close to hand; and that the causal chain is short, direct, and often violent. Parfit acknowledges as much, in explaining why statistical deaths are not the same as certain, known deaths:

> When we know that the lives of certain people are in danger, as would be true, for example, if some group of miners are trapped underground, we have reasons to want great efforts to be made to save these people's lives. Some economists point out that we would do more to increase people's life-expectancy if, rather than spending huge sums on trying to save known particular people in such emergencies, we spent this money on more cost-effective safety measures that would prevent a greater number of statistically predictable future deaths. But we could reasonably deny that this fact is morally decisive. We have strong reasons to want great efforts to be made to save the lives of known particular people who are in danger. By making or supporting such efforts, for example, we reaffirm and express our solidarity with, and concern for, everyone in our community. That is less true of acts that merely prevent the statistically predictable future deaths of unknown people.
>
> We have similar reasons to want it to be true that no one would be hunted down and have their organs removed by force. And though such acts would be done to save the lives of certain known particular people, these acts would also produce much anxiety, conflict, and mistrust. (vol. 2, p. 211)

Parfit locates our intuition that statistical deaths are a wholly different thing from certain deaths in the moral sphere. Perhaps it is; perhaps it is better explained by emotional or psychological factors.[14] Whatever the source, if the intuition is recalcitrant enough, it may require deference on purely pragmatic

[14] 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 in Kantian and other deontological frameworks.

grounds, for reasons nicely captured by the Rawlsian 'strains of commitment.' But the failure to identify correctly the features of trolley problems that our intuitions are snagging on has serious consequences for the nonconsequentialist project. It means that attempts to generalize the nonconsequentialist principles extracted from our encounters with trolley problems to other types of problems are likely to misfire. That is not Parfit's problem, since he has managed to extract consequentialist principles from his encounter with trolley problems. But to the extent others dismiss risk as an 'easy' case because they believe it is isomorphic to trolley-type problems but with lower stakes all around, they have gone astray in just this way.

2.2.2. Including all trolley problems in the domain of relevant cases.

In another important respect, the domain of cases used to develop and test general principles about harm to others is much too broad. Trolleyologists have proceeded, de facto, on the assumption that it is the job of sound general principles to make sense of our intuitive responses to all trolley problems, whatever factual posture they come up in, and that all other things being equal, the 'best' general principle will be the one that can accommodate (make cohere) our intuitions about the largest number of trolley scenarios. To put it another way, every trolley problem is equally appropriate grist for trolleyologists' mill. I believe this assumption cannot be defended, and has resulted in no end of confusion in the trolley literature, leaving trolleyologists grappling with nonexistent problems and settling for gerry-rigged solutions to real ones.

Consider for example the question of what knowledge people are permitted

to have about their own ex post fate if a given principle is adopted.[15] If we ask whether person X could, based on her own subjective preferences, reasonably object to permitting ambulances to speed, knowing that five patient lives will thereby be saved for every one pedestrian killed, we will get one answer if X knows she will be one of the ones killed, and we will get another if she doesn't know whether she will turn out to be the one, or one of the five, or (in almost all cases) neither.

Which answer counts for normative purposes? The de facto response from trolleyologists is both: If person X 'happens' to know she is the pedestrian about to be mowed down, we count her as reasonably objecting to allowing ambulances to speed. If she 'happens' to know nothing about her own particular situation that would differentiate her odds from anyone else's, we count her as in favor of letting ambulances speed. And if, having been counted as voting in favor of speeding from a position of ignorance, she subsequently learns that a speeding ambulance is about to mow *her* down, we let her switch her vote (assuming she has the time to do so after she learns about her impending fate and before it is sealed).[16]

Parfit implicitly takes on board the view that whatever information we happen to have about our ex post fate under a given principle is a "relevant,

---

[15] For a more extended discussion of this issue, see Barbara Fried, "Can Contractualism Save Us From Aggregation?" http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1781092

[16] The examples are based on Frances Kamm's "Ambulances" cases in Intricate Ethics, pp. [] . For further discussion of the widespread view among nonconsequentialists that one has a right to renege on prior agreements once one learns one's own fate under them, see B. Fried, "Can Contractualism Save Us From Aggregation?" http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1781092

reason-giving fact" that must be taken as given in assessing what it would be reasonable for us to accept or reject. (vol. 1, p. 356). A number of trolleyologists have defended that position explicitly, on the grounds that only by doing so do we show each individual the respect she or he is owed as a unique individual.[17]

Treating whatever information we happen to have about ex post outcomes as part of our natural endowment has resulted in something of an analytical and normative train (trolley?) wreck in the literature. Consider, for example, the variant on Taurek's famous dilemma taken up by Parfit and Scanlon in *On What Matters*.[18] In Taurek's original hypothetical, a would-be rescuer 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 on the table is: if the would-be rescuer (Green) chooses to save the 5 rather than the 1 based solely on numbers, will she thereby violate Orange's Kantian right to equal respect?

In Lifeboat, the variant Parfit and Scanlon consider, Orange is Green's child. As a consequence, Green has a strong personal preference to save Orange rather than the five. Now what? Parfit takes it to be the self-evidently optimific solution to permit Green to save his own child:

> 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

[17] For further discussion, see B. Fried, "Can Contractualism Save Us From Aggregation?," TAN 24.
[18] John Taurek, "Should the Numbers Count?" Philosophy and Public Affairs 6 (1977).

children's well-being. (vol. 1, p. 385)

Scanlon reaches the same conclusion under his RCM:

> Rather than appealing to the idea of the best outcome—what everyone has impartial reason to prefer— my argument was based on what each individual has reason to want for him or herself. 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. (vol. 2, p. 133)

But it is an easy case for both Parfit and Scanlon only because they have implicitly assumed that we should assess the reasonableness of a candidate principle to govern this choice *not* from the perspective of the particular Green, White, Blue, Yellow, Red, Black and Orange who find themselves in the situation described in the hypothetical, but rather from the perspective of representative persons who do not know whether they will ever find themselves in the role of any of these fictive persons or (much more likely) none. Given that every representative person is operating from a position of equal ignorance about his or her own ex post fate and all are presumed to share a preference for partiality, we get the straightforward answer, under conventional aggregation or RCM, that partiality trumps.

But suppose that the epistemological perspective from which representative persons are to judge the reasonableness of a general principle is taken to be just whatever knowledge the characters in the trolley hypothetical happen to possess at the moment we come upon them. A now self-identified Green must decide whether to save a now self-identified White, Blue, Yellow, Red and Black or

instead save his own child, Orange.

Under this revised hypothetical (call it Lifeboat II), we have a problem getting the 'right' answer under optimific principles, because White, Blue, Yellow, Red and Black will all strongly prefer a rule that requires Green to save them rather than Orange. This is so, even if each of them, answering from a position of equal ignorance about the probabilities they will ever find themselves in any of these roles, would have chosen to let partiality trump. It is also true even if White, Blue, Yellow, Red and Black all realize at the moment of choice that there is non-zero chance that the fate of their own child might someday be at stake, should their surviving spouse have the exceedingly bad luck to find herself in Green's position in some future Lifeboat II; and that their surviving spouses will have to live with the emotional strains of that possibility, whether or not it comes to pass. However strongly White, Blue, Yellow, Red and Black would prefer a rule permitting their respective spouses to be partial to their child should that second Lifeboat II ever come to pass, rational self-interest will lead them strongly to prefer a rule that requires Green to save them rather than Orange now, as long as the 'certain' death each of the five would face under a rule of partiality is valued at its full disvalue to them, and the chance that the tables will be turned at some future time is discounted by its exceedingly low probability.[19] The only way to get the 'right' answer under optimific principles

[19] The foregoing analysis limits the persons whose 'individual reasons' are counted to the onstage players in the hypothetical—the de facto restriction followed in most trolley problems. Were we also to weigh the individual reasons that the offstage loved ones of these six might have to favor one result over the other, those added reasons would only strengthen the case against partiality.

when all the players know their own fates is the one ultimately taken by Parfit: to ditch subjective preferences for objective reasons, and then stipulate that all of the relevant parties have objective reasons to give more weight to Green's and Orange's suffering than to their own collective suffering.

How Scanlon would have us resolve this revised hypothetical under RCM is not clear, at least to me. If we don't count Green's agent-relative preference to save Orange, the five and the one are each facing death if things don't go their way, giving each group complaints of equal strength. At that point, Scanlon's tiebreaker principle would kick in to let the numbers count, once again yielding the 'wrong' answer. Adding in Green's preference would seem to change this result only if we are allowed to aggregate Green's reasons to save Orange and Orange's reasons to want to be saved, and count the sum as *one* individual's reason.[20] How exactly we could justify aggregating Green's and Orange's reasons under a Scanlonian 'individual reasons' restriction is not obvious.

How *should* nonaggregationists approach a problem like Lifeboat? I think we have to begin by acknowledging that there are two very different sorts of problems at play in Lifeboat I and Lifeboat II that implicate very different moral issues and that cannot be resolved by the same principles. [21]

---

[20] This analysis ignores the preferences of all the off-stage loved ones of the five. Were we to count those preferences as well, it might well change the result under RCM. If we assume that at least one of those off-stage loved ones has as strong a preference to have their family member saved as Green has to save his own child, then we have a tie, which invokes Scanlon's tiebreaker principle, once again yielding the wrong answer.

[21] For a fuller discussion of the issues raised in the discussion that follows, see Barbara Fried, "Can Contractualism Save Us From Aggregation?" http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1781092

Insofar as we are seeking general principles to live by—the ostensible project of trolleyologists-- I think that the *only* morally defensible epistemological point of view from which to ask what principles everyone would have reasons to accept (or not reject) is *before* they know how things will turn out for them ex post under a given principle. It is also the only point of view that can yield anything approximating general agreement (unless, like Parfit, we smuggle in altruistic motivations that override our clear personal interests). Trolleyologists have concluded otherwise-- that everyone is entitled to whatever information they just happen to possess when we just happen to solicit their views on general principles—, I believe, because they have confused the question whether people are entitled to know their own preferences, aptitudes, and general situation in life (yes, if one rejects the thick Rawlsian veil of ignorance for 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.[22] The latter is orthogonal to whether we are committed to thick selves: Green is Green before he knows that he will be the one who has to choose between the one and the five and his own child, and he is Green after he discovers it. The question is,

[22] Parfit makes the same mistake, I believe, when defending the superiority of the "full information" condition in Kantian contractualism to Rawls's veil of ignorance. Noting that the central problem with the Rawlsian veil of ignorance is that it achieves impartiality and unanimity not by resolving the disagreements between different people but by suppressing them (vol. 1, p. 356), Parfit argues that Kantian contractualism avoids that problem by endowing us with "full information" concerning the "relevant, reason-giving facts." 'Full information' for these purposes, Parfit argues, includes 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 (vol. 1, p. 356).

But general knowledge about our selves and our circumstances in life alone is sufficient to surface genuine disagreements between real individuals. Whether individuals should be endowed in addition with knowledge about ex post outcomes is a separate question that has to be resolved on other grounds.

which Green's views on the 'right' choice is it morally appropriate for us to solicit? If it is the latter, that proposition has to be established on some grounds other than the commitment to 'thick' selves.

But what if, having agreed on general principles of action without knowing the outcome of any particular acts taken pursuant to them, we subsequently find out, while there is still time to change our minds, who the big losers will be if an agreed-on general principle is adhered to in a particular case? It is one thing for everyone to agree 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, most people feel, for a driver to adhere to that agreement when Pedestrian Smith walks in front of the ambulance and there is time to avoid hitting him by applying the brakes.

I agree. The question is, what kind of thing is it, how should we respond to it, and (most importantly for my purposes here) what relationship if any does the right response to trolley problems have to do with the right way to go about choosing general principles of action?

For utilitarians, the answer to the first two questions is simple. We should respond to each problem in the fashion that we believe will maximize welfare. Whether the optimal response to the first problem will look anything like the optimal response to the second turns purely on empirical assumptions.

For nonconsequentialists, the answer to the first two questions is more complicated. (For more details, see the trolley literature.) My immediate interest is in the last of these questions: what light if any does the right answer to

trolley problems shed on the problem of formulating general principles of action from a position of ex ante uncertainty about outcomes? In my view, the answer is none. The factors driving our nonconsequentialist moral intuitions in the former case are absent in the latter, and we are able to accommodate them in the former case only because such cases rarely arise.

Because nonconsequentialists have said so little about the appropriate principles to govern decisionmaking under conditions of uncertainty, it is hard for me to guess how they would describe the relationship between the two problems. 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 it is the wrong choice, and hope nonconsequentialists will at least consider letting trolley-type problems languish for awhile while they turn their attention to the problem of risk.

3. Hard cases make bad law.

The obsession with trolley-type problems is emblematic, I think, of a more general problem in philosophical thought: the tendency to anchor arguments in extreme examples. There is an adage in the legal world that "hard cases make bad law," by which lawyers mean that cases that present exceptional facts are likely to provoke exceptional responses that do not generalize well to more typical cases. If laws of general application are what you seek, you are much better off starting with

the most common case they must resolve, not the rarest.

The same danger presents itself in the moral realm. If slavery is the example that anchors an exploration of coercion, or the right not to have your kidney yanked out in the middle of the night the example that anchors an exploration of the concept of self-ownership, conclusions reached about the nature and implications of coercion or self-ownership are likely not to generalize well.

Structural features of our legal system make it difficult for courts to avoid dwelling disproportionately on hard cases, because those are the ones that tend to get litigated to final judgment. Courts have many tools at their disposal to handle this problem, including (at the extreme) declaring that a case has no precedential value (that is, does not extend beyond its facts).[23]

Philosophical inquiry doesn't labor under similar constraints; philosophers are free to choose the examples they want to dwell on. At a minimum, however, if extreme cases remain the point of departure, philosophers need to ask themselves, what is the next example of coercion *after* slavery, of self-ownership *after* forcible transplant, of the tragic choices we face between harming some and benefiting others *after* trolley-type problems? If the principles articulated in the context of the extreme cases can't explain our intuitions in those less extreme cases, then we very likely have an instance of hard cases making bad moral law.

[23] This is the technique that the majority in Bush v. Gore helped themselves to, in order to ensure that a hard case made no law at all.