Introduction

Hallvard Lillehammer

I.1 The context

The trolley problem is one of the most notorious puzzles in contemporary moral philosophy. Over little more than fifty years, the problem has become a reference point for systematic reflection on moral theory; medical ethics; the ethics of war; the ethics of automation; neuroscience; social psychology; intercultural moral comparisons, and more. Beyond academia, the problem has been the topic of popular books; short films; TV series, and online games. In short, the trolley problem has the rare distinction in philosophy of having become something of a cultural phenomenon. This fact is itself a topic of controversy. There are those who consider the problem a conduit to discovering the nature of morality. Yet there are also those who consider it an example of academic theorizing at its most pointless. The essays in this volume critically address this dispute by discussing the main questions that have been at issue in the growing literature on this topic.

I.2 The problem

Sometimes the trolley problem is introduced by giving a single example of a tragic choice, such as that of a train driver who can save five people on one track by killing a different person on another track. Yet as standardly understood in the academic literature, the trolley problem proper arises when we compare two examples, in both of which the resulting harms are the same, yet most people judge the two cases differently. Thus, in what is normally considered its first appearance in the literature, the problem was introduced by comparing the case of the driver of a train who can save five people on a track by switching to another track and thereby killing one person to the case of a judge who can save five people by having one innocent person executed (Foot 1967/2002). Thus understood, the problem is to explain why in the one case the saving of the greater number should be thought of as permissible while in the other case it is not.

Although most contributors to the literature take as their starting point either two or more examples to which people are expected to respond in different ways, they differ importantly in how narrowly they understand the problem and the range of examples taken to instantiate it. Judith Jarvis Thomson, who gave the problem its name, gave two different specifications of what the problem is. In her first formulation, Thomson asked: "[W]hy is it that Edward [a train driver] may turn the trolley to save his five, but David [a hospital surgeon] may not cut up his healthy specimen to save his five" (Thomson 1976, 206). When returning to the issue a decade later, she defined the problem thus: "Why is it that the bystander [who, unlike the driver has had no role in previously directing the trolley] may turn his trolley, though the surgeon may not remove the young man's lungs, kidneys and heart" (Thomson 1985, 1401). Although Thomson was later to "regret having muddied the water by giving the same name to both problems," she did not think this "water-muddying" had done any real harm. After all, the explanatory challenge these two problems raised were, if not exactly the same, then very similar (Thomson 2016, 115–116).

A wider definition of the problem is provided by Frances Kamm, whose formulation abstracts away from the substance of the examples to focus on their structural features. According to Kamm, "we could see the trolley problem . . . as presenting a challenge to nonconsequentialists who. . . think there is what is called a side-constraint on harming non-threatening people to produce greater goods" (Kamm 2016, 12). The challenge is then to "explain exactly what the side-constraint on harming amounts to, and what its form is. . ." (13). On this definition, the problem is understood as a question about "why it is sometimes permissible to kill, even rather than let die, when we come to kill in some ways but not others" (47; 183; 195; c.f. Thomson 2016, 224). This problem also applies to cases not involving trolleys that have a similar structure, and solving the problem requires explaining the moral differences among different ways of causing death.

On a third interpretation, the label "trolley problem" picks out a way of thinking about moral questions, namely, one that invokes a set of schematic thought experiments in which a small number of actors are supposed to choose between a small number of actions the outcomes of which are normally assumed to be fixed. This is the understanding of "trolley problem" invoked by Barbara S. Fried when she complains about "the intellectual hegemony of the trolley problem" in recent moral philosophy (Fried 2012a, 2). Fried's definition includes the understanding of the trolley problem put forward by Thomson and Kamm. Yet it goes beyond their understanding

to include a wider range of moral problems that either may or may not share the structural features that distinguish the problem defined by Thomson or Kamm. Although some of the chapters in this volume operate within the constraints employed by Thomson or Kamm's definition, other chapters extend their use of examples beyond those constraints to develop an argument with wider theoretical ramifications for what has come to be widely known as "trolleyology".

I.3 The chapters

This volume contains twelve original essays on the trolley problem written by some of the most influential contributors to the study of that problem; some of whom have either defined or established the research agenda in their field.

The first chapter traces the history of the problem from its initial appearance in the work of Philippa Foot, through its formal definition by Thomson, to its further elaboration by Kamm. It then confronts the problem with three skeptical responses to its claim to intellectual significance, as articulated by Fried. The historical period covered spans around five decades (from 1967 to 2012), by which time the literature on the trolley problem was experiencing an explosive growth, in ways addressed in the chapters that follow.

William J. FitzPatrick's chapter departs from Thomson's objection to Foot's original discussion. Thomson's challenge was to explain why a bystander (as opposed to a trolley driver) should be permitted to switch the trolley in spite of the fact that she would thereby be killing one as opposed to letting five die (as opposed to the driver who would be killing one as opposed to killing five). Thomson later changed her mind about this case, going on to deny that the bystander is permitted to switch. FitzPatrick argues that Thomson ought not to have changed her mind about the bystander. In doing so, he argues that there is an important challenge posed by the problem as originally defined. Moreover, this challenge can be met by giving an account of reasonable norms of shared risk to the loss of innocent life in public spaces.

Much discussion of Thomson's work has focused on whether it is *permissible* or *impermissible* for the bystander to switch. Yet it has also been argued that not only is it permissible for the bystander to switch, it is actually *required*. In his chapter, Peter A. Graham responds to a recent argument by Helen Frowe in favor of this conclusion. According to Graham, Frowe's argument depends on the premise that people have a duty to prevent harm to others when they can do so without violating anyone's rights or bearing an unreasonable personal cost. Graham argues that Frowe's premise is

implausible. In addition, he presents a direct argument for the claim that it would be permissible for the bystander *not* to switch the trolley, inspired by previous work by Kamm.

Frances Kamm agrees that it is permissible for the bystander to switch. In her chapter, Kamm explains her reasons for her agreement and how those reasons derive from the non-consequentialist idea that goods brought about improperly are not morally justified. She calls this The Doctrine of Productive Purity. During the course of her chapter, Kamm explores the details of this doctrine by applying it to a range of cases discussed in the recent literature. She also discusses other conditions that need to be met in order to make the switching of the trolley permissible; the relation of those conditions to the distinction between killing and letting die; and general ideas about persons and their relations to each other that explain the significance of non-consequentialist constraints on permissible action.

In their chapter, Dana Nelkin and Samuel C. Rickless address the challenge that standard formulations of the trolley problem fail to take account of the ethical significance of risk because it considers the harms caused by the available options available as fixed. Nelkin and Rickless argue that the imposition of risk should itself be understood as a kind of harm alongside actual injury or death, and therefore something that innocent people can be considered as having a defeasible right against. By drawing on a distinction between direct and indirect harmful agency, according to which the former kind of agency is harder to justify – all other things being equal – than the latter, Nelkin and Rickless argue that insofar as a version of the Doctrine of Double Effect can be justified as one moral principle among others, so can a probabilistic version of that doctrine. It follows that the distinction between direct and indirect harm can be extended to account for other cases involving risk.

In her chapter, Fiona Woollard discusses the relationship between the trolley problem and the distinction between doing and allowing harm. According to Woollard, the distinction between killing and letting die that has been at issue in the trolley problem literature is an application of the doing/allowing harm distinction to the specific harm of death. According to Woollard, the distinction between doing and allowing should be understood as a distinction between different burdens of moral justification. In short, and other things being equal, doing harm is harder to justify than merely allowing harm. She also argues that the lesson to draw from trolley cases is not that there is something wrong with the doing/allowing distinction, but rather that there is a need for additional deontological distinctions to make sense of the full range of such cases.

Liezl van Zyl's chapter approaches the trolley problem from the perspective of virtue ethics. On the one hand, Van Zyl argues that there are reasons to be skeptical about the theoretical significance and practical relevance of the trolley problem. On the other hand, she argues that a virtue ethical approach can provide a plausible diagnosis of what distinguishes the acceptable act of a bystander saving five by turning a switch from the unacceptable act of saving five by pushing someone off a bridge. She also argues that a virtue ethical approach can account for what goes on in Thomson's infamous Loop case by noting that diverting the trolley does not necessarily involve viewing the sole workman as an object or as a means to an end, even if the death of the one is causally necessary for the saving of the five.

In their chapter, Guy Kahane and Jim Everett describe how in recent decades the trolley problem has become a central focus of empirical research in moral psychology. Much of this research has been framed in terms of a contrast between "deontological" and "utilitarian" approaches to morality. Kahane and Everett argue that this framing is misleading in two ways. First, some of the lay responses to trolley cases which psychologists have classified as "utilitarian" have only a tenuous relation to what philosophers have traditionally meant by this term. Second, even when what underlies lay responses to trolley cases echoes aspects of utilitarianism, this doesn't generalize to other aspects of moral thought. Kahane and Everett conclude that while trolley cases have a useful role to play in psychological research, the centrality of such cases in recent moral psychology is theoretically problematic.

Joshua D. Greene's chapter shows how, for more than two decades, the empirical study of trolley dilemmas has substantially improved the scientific understanding of the psychological mechanics of moral judgment. In doing so, Greene picks up on the methodological challenges raised in Kahane and Everett's chapter. By drawing on his own path-breaking work in experimental psychology and cognitive neuroscience, Greene draws attention to how the cognitive processes engaged by trolley dilemmas pose explanatory and normative challenges for traditional ways of describing and evaluating moral judgments. In particular, Greene argues that the study of trolley dilemmas has exposed the evidential or justificatory limitations of moral intuitions by bringing to light patterns in moral thought that fail to withstand critical scrutiny. During his critical survey of recent work on this topic, Greene brings out a number of potential misunderstandings that can arise from attempts to translate the results of empirical work in moral psychology into orthodox philosophical categories, such as "deontological" and "utilitarian," or vice versa.

Natalie Gold's chapter provides a critical survey of empirical studies that compare the responses to trolley cases across different cultures. Gold draws two conclusions from this survey. First, the amount of evidence available is limited. Second, with respect to statistical significance, the picture that emerges from the evidence is mixed. Gold suggests that there is some support for the claim that cross-cultural differences in moral judgments on trolley cases do exist and that those differences are shaped by the particularities of the culture in which they are embedded. Gold also considers the potential implications of these cultural differences for the epistemology and metaphysics of moral judgment. She agrees that the mere fact of cultural difference and disagreement is consistent with the claim that some cultures have a better grasp of moral facts than others. Yet all things considered, she is more inclined to consider the differences in question as evidence for a form of ethical constructivism.

The chapters by Sven Nyholm and Ezio di Nucci discuss the trolley problem in the context of applied ethics, public policy, and law. Nyholm's chapter concerns the relationship between the trolley problem and the ethics of autonomous vehicles. Nyholm argues that it can be useful to compare crashes involving self-driving cars and a range of trolley cases. It can be *directly* useful because the trolley problem brings to light ethical issues of immediate importance for the ethics of self-driving cars. It can be *indirectly* useful because by highlighting the differences between the ethics of self-driving cars and ethics of trolley cases, it is possible to clarify what really matters in the ethics of self-driving cars.

In his chapter, di Nucci considers the trolley problem in the context of healthcare, more specifically in connection with the global health crisis caused by COVID-19. The hypothesis evaluated in di Nucci's chapter is that the trolley problem can be used to distinguish between the alleged permissibility of lifting "lockdowns" and the alleged impermissibility of pursuing "herd immunity." After careful consideration, di Nucci rejects this hypothesis. In doing so, he makes a number of observations about the extent to which thinking about the trolley problem could be useful in thinking about the ethics of pandemics, on the one hand, and the extent to which thinking about the ethics of pandemics could be useful in thinking about the trolley problem, on the other hand.

1 Keeping track of your trolleys

Origins and destinations

Hallvard Lillehammer

1.1 Introduction

This chapter presents a selective overview of the emergence of the trolley problem as a theoretical concern in moral philosophy. It does so by giving a snapshot of the problem as understood by three of the most influential contributors to discussions of that problem, namely, Philippa Foot, Judith Jarvis Thomson, and Frances Kamm. The chapter ends with a review of three criticisms that have been leveled at the trolley problem when considered as an instance of a certain way of thinking about moral philosophy. It does so by considering the arguments of one vocal critic of this way of thinking about moral philosophy, namely Barbara H. Fried.

1.2 Origins

1.2.1 Foot

It all started with a thought experiment involving a tram driver, introduced by Philippa Foot in order to draw an analogy during the course of an argument about the ethics of abortion. The argument first occurred in a paper published by *The Oxford Review* in 1967 where it was responded to by her colleague and friend Elizabeth Anscombe (Anscombe 1967; Foot 1967/2002a).

Both the timing of the argument and the identity of the commentator are worthy of comment. It was the year 1967 in which abortion was legalized and regulated in the United Kingdom (UK) (with the exception of Northern Ireland) through an act of parliament. The topic of discussion was therefore one of contemporary public interest. Foot's commentator, Anscombe, was a committed practitioner of Roman Catholicism, historically associated with the endorsement of the moral distinction between intention and foresight, as embodied in the claim that an act of killing that would normally be sinful or wrong might nevertheless be permissible if committed as an act of self-defense, for example, as a side effect of saving one's own life. Some would

argue that this doctrine, widely known as "the doctrine of double effect," serves to specify the conditions when an otherwise impermissible act of abortion is permissible. It was this doctrine that Foot subjected to critical scrutiny in her paper.

Although Foot did not reject outright the moral significance of the distinction between intention and foresight in her paper, she did argue that in a wide range of cases, including the case of an unfortunate tram driver who has the choice between either killing five people or killing one, the operative moral distinction is that between acting contrary to a positive duty to aid people versus acting contrary to a negative duty not to harm people, where a negative duty will normally trump a positive duty. This alternative distinction, or something very much like it, is one that has been thought to support another widely recognized moral distinction, namely, that between killing and letting die.

The case of the tram driver made only a brief appearance in Foot's paper. This is how Foot introduced it:

> Suppose that a judge or magistrate is faced with rioters demanding that a culprit be found for a certain crime and threatening otherwise to take their own bloody revenge on a particular section of the community. The real culprit being unknown, the judge sees himself as able to prevent the bloodshed only by framing some innocent person and having him executed. Beside this example is placed another in which a pilot whose aeroplane is about to crash is deciding whether to steer from a more to a less inhabited area. To make the parallel as close as possible it may rather be supposed that he is the driver of a runaway tram which he can only steer from one narrow track to another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed. In the case of the riots the mob have five hostages, so that in both the exchange is supposed to be one man's life for the life of five. The question is why we should say, without hesitation, that the driver should steer for the less occupied track, while most of us would be appalled at the idea that the innocent man could be framed (Foot 1967/2002a, 24).[1]

[1] Later in the same article, Foot appeals to the example of a doctor who can either kill or let die an innocent person to save five dying patients. This example (later known as "Transplant") would soon replace the case of the judge as the standard contrast case to that of the driver, until Thomson abstracted away from such role-based complexities with her "Fat Man" example (later known as "Footbridge"), in which the five people on the track can be saved by pushing or otherwise making a very large person fall on the track in front of the five (see Thomson 1976, 207–208).

The basic elements of the puzzle that Thomson would later call "the trolley problem" are clearly visible in this passage. Some individual agent (a judge, a pilot, or a driver) is faced with a choice between a set of serious harms, the magnitude of which is quantifiably different on either side (i.e., five deaths versus one). In some cases (e.g., that of the driver), it arguably seems right to minimize the harms, whereas in other cases (e.g., that of the judge) it does not. The puzzle is to work out what, if anything, could plausibly explain the difference.[2] As already noted, Foot's proposed solution to this puzzle was to appeal to the distinction between negative and positive duties.

Foot's article is notable for its theoretical modesty. During the course of the article, she writes: "In many cases we find it very hard to know what to say, and I have not been arguing for any general conclusion" (Foot 1967/2002a, 30). Instead, she has been "trying to discern some of the currents that are pulling us back and forth" (ibid. 32). Whereas her main aim in the paper is "to show that even if we reject the doctrine of the double effect we are not forced to the conclusion that the size of the evil must always be our guide" (ibid. 31), she says that she has "not, of course, argued that there are no other principles" (ibid. 30). She describes some of her more outlandish thought experiments as having been "introduced for light relief" (ibid. 22) and apologizes for their "levity" (ibid. 32). Yet she does not think this makes them frivolous, trivial, or un-illuminating. For example, she asserts that her case of a doctor who can kill an innocent person to save five dying patients by redistributing that innocent person's organs is "not over-fanciful considering present controversies about prolonging the life of mortally ill patients whose eyes or kidneys are to be used for others" (ibid. 25). What Foot was aiming to produce by constructing her thought experiments involving the judge, the doctor, and the driver was a series of imperfect but illuminating analogies. As with all arguments by analogy, there are inevitably some respects in which the cases on either side of the analogy will differ.[3] The interesting question is whether the cases in question are sufficiently similar in morally relevant respects.

[2] Strict act consequentialists deny that there is a morally basic difference between these cases and so would recommend that we minimize consequent harms, all else being equal. For this reason, the trolley problem is often introduced against the background of a non-consequentialist theoretical project that consists in the articulation of so-called side-constraints on the promotion of the good (see, e.g., Kamm 2007).

[3] Foot's case of the pilot might be thought to fail the test of relevant similarity in comparison with the cases of the driver, the judge, or the doctor, given that the pilot may reasonably be assumed to end up dead themselves. Yet given the way in which the range of trolley cases expanded over the years, even that could be reasonably contested.

1.2.2 Thomson

In the United States (US), the legal counterpart of the UK Abortion Act is the Roe vs. Wade case decided by the Supreme Court in 1973. As is well known, Thomson made a seminal intervention in the philosophical debate about abortion in the run-up to that decision in her paper, "A Defense of Abortion," published in the inaugural volume of *Philosophy and Public Affairs* in 1971 (Thomson 1971). There is no tram driver in evidence in Thomson's paper on abortion. Nor is there is any reference to Foot's paper. On the other hand, there is plenty of "light relief"; most famously the example of a person who wakes up to find herself having been connected to a famous violinist by the "Society of Music Lovers" for the purpose of providing a nine month course of dialysis for said violinist. In other words, the practice of constructing imaginative thought experiments for the purposes of arguing by analogy is one in which Thomson was an expert practitioner well before she turned her attention to the trolley problem.

It was in her 1976 paper "Killing, Letting Die and the Trolley Problem" that Thomson gave the name to what she there describes as "a lovely, nasty difficulty," namely, "why is it that Edward [the driver] may turn the trolley to save his five, but David [the doctor] may not cut up his healthy specimen to save his five?" Thomson labeled this difficulty "the trolley problem, in honor of Mrs. Foot's example" (Thomson 1976, 206). As with Foot before her, the theoretical context of Thomson discussion was her interest in the moral distinction between killing and letting die. One of the main conclusions of Thomson's paper was that this distinction "cannot be used in any mechanical way in order to yield conclusions about abortion, euthanasia, and the distribution of scarce medical resources. The cases have to be looked at individually" (ibid. 217).

To show this, Thomson introduced the Passenger case, in which the driver in Foot's example is replaced by an innocent passenger whose choice is that between killing one and letting five die, given that the passenger (unlike the driver) never set the train in motion. A decade later, in Thomson (1985), the passenger was replaced with a now more famous bystander in order to avoid any ambiguity on this score. The point of the Passenger case (and later the Bystander case) is that *if* switching the trolley from one track to another is at least permissible in this case (which Thomson at this point asserts that it is), *then* there are structurally equivalent cases to that of the Driver case in which negative duties do not trump positive duties, so Foot's distinction does not after all explain the moral difference between the driver on the one hand and