Trolleyology as First Philosophy:
A Puzzle-Centered Approach to Introducing the Discipline*

Vaughn Bryan Baltzly
Texas State University

*** To appear in *Teaching Philosophy* 44(2) ***

(10,132 words exclusive of Appendix; 19,308(!!) with Appendix)

Abstract: Though sometimes maligned, "trolleyology" offers an effective means of opening and framing, not only classes in ethics, but indeed *any* introductory philosophy course taking a broadly "puzzle-based" approach. When properly sequenced, a subset of the thought experiments that are trolleyology's stock-in-trade can generate a series of puzzles illustrating the shortcomings of our untutored moral intuitions, and which thus motivate the very enterprise of *moral theorizing*. Students can be engaged in the attempt to resolve said puzzles, inasmuch as they're accessible and compelling, and their resolutions generally easy to achieve. Once thus engaged, students can be directed to the fact that (perhaps unbeknownst to them) they had already rolled up their sleeves and begun "doing philosophy." In this way, engagement with trolleyological puzzles can serve as a "microcosm" for philosophy more broadly, illustrating the processes of critical thinking that are likewise the stock-in-trade of philosophers across many different domains of inquiry.

PHILOSOPHERS HAVE RECENTLY ADOPTED the term "trolleyology" to refer to the substantial body of literature concerning issues raised by Philippa Foot's famed "trolley problem."[1] Though the term perhaps originated as a pejorative, it is probably fair to say that many philosophers have come to rather embrace the label—if not as a substantive research program in its own right, then at least as a neutral descriptor of a certain way of engaging students and readers in the fundamental elements of moral philosophy.[2] Whatever its merits as a research program, it seems clear (at least on the basis of anecdotal evidence available to this author) that trolleyology is widely valued for at least its *pedagogical* potential—as a device for classroom instruction in introductory moral philosophy courses.

However, it is less clear whether philosophy instructors have completely appreciated the full range of pedagogical resources that trolleyology has to offer—not only with respect to moral philosophy, but with respect to philosophical methodology more widely. In part, this is due to the fact that philosophy instructors too often wish to *begin* with the so-called "trolley problem," when really Foot's (and Judith Jarvis Thomson's) trolley scenarios ought not to be deployed except as late-stage entries in a series of thought experiments centered on moral dilemmas. Carefully deploying a "trolleyological" approach can direct students' attention to a set of puzzles generated by the juxtaposition of common intuitive reactions to a carefully-curated sequence of thought experiments—thought experiments that are the stock-in-trade of trolleyology, broadly construed. In this way, trolleyology can serve to illustrate the processes

[1] See Foot [1967]. Other early canonical discussions of the problem include Thomson [1976] and [1985] and Costa [1986] and [1987]. "Trolleyology" is a neologism coined (at least in print) by Appiah [2008].

[2] See, for example, Cathcart [2013] Edmonds [2014], and Kamm [2015], three instructive examples of writers who take Trolleyology seriously.

* I would like to thank Jesse Ehnert, Andrew Kania, and two anonymous reviewers from *Teaching Philosophy* for helpful comments on earlier versions of this paper. Jesse's influence goes far beyond his role in reviewing a draft, however, as he has been my principal interlocutor on all matters trolleyological for more than 20 years.

of critical thinking and puzzle resolution that are likewise the stock-in-trade of all philosophers, across many domains. Trolleyology, that is, serves as a natural vehicle for a *puzzle-centered* approach to introducing our discipline—one that makes "intra-personal dissensus" visible to students in ways that enable them to learn how to productively resolve both intra- *and* inter-personal dissensus.[3]

In this paper, I share a trolleyological approach I have developed and deployed to good effect as the opening module, not only in introductory courses in ethics and applied ethics, but also in general introductory courses in philosophy and critical thinking. As just mentioned, this approach involves presenting a sequence of five moral-dilemma thought experiments, each one structured so as elicit common (but by no means unanimous) intuitive reactions as to the proper resolution of the dilemma. But the intuitions thus elicited do not—at least initially—appear to cohere well with one another. I thus generate a series of four puzzles—each one arising from the juxtaposition of two seemingly-contradictory (but widely-held) intuitions. Initial appearance can prove deceiving, however: often it turns out that, despite the seeming contradiction, the apparent inconsistency between intuitions can be resolved. Specifically, the inconsistency can be eliminated via the articulation (or the discovery) of a deeper, heretofore-unnoticed principle or distinction that demonstrates how and why the pair of intuitive reactions can be consistently maintained after all. In section 1 of this paper I sketch this "puzzle-generating" approach to trolleyology. This sets up the pedagogical point I develop in section 2, the heart of the paper. I argue there that, in confronting these puzzles—and in noticing subtle disanalogies between seemingly parallel cases, and in articulating principles that traffic in these now salient (if subtle) distinctions—students are engaged in what we might term "critical thinking," and are already philosophizing. Furthermore, this activity of thinking critically in response to various puzzles (puzzles that arise from the casual deployment of our pre-theoretical, unreflective, intuitive moral beliefs) generalizes beyond moral philosophy: this "puzzle-driven approach" can be a fruitful paradigm for understanding philosophy (and/or critical thinking) more broadly. I thus sketch some of these more general applications in section 3. Section 4 responds to a number of potential concerns regarding this approach; principal among them is the worry that it might unintentionally engender moral skepticism. Finally, an Appendix provides, for any readers who may be interested, more detail about the moral-dilemma thought-experiments (and correspondingly-generated puzzles) that I briefly sketched in section 1.

1. Five "Moral Dilemma Thought Experiments" and Four Related Puzzles

I first establish a baseline by presenting a simple thought experiment regarding a moral dilemma. By design, this dilemma is such that students can be expected to unanimously agree on a decision. Further, students have no problem articulating a rationale for *why* they chose this horn of the dilemma over the other—they have no trouble, that is, articulating a *principle* that *grounds* or *justifies* their intuitive reaction. Once this baseline has been established, the fun begins. For I then confront students with a sequence of four further moral-dilemma-based thought experiments—each of which appears structurally similar to a case that precedes it, but which is such that (for many students, at least) the intuitive reaction to that case seemingly

[3] I am indebted to an anonymous reviewer from *Teaching Philosophy* for noticing that this is perhaps the principal thesis of my essay, and in particular for this way of phrasing that thesis.

clashes with a principle articulated in an analysis of a preceding case. The juxtaposition of these cases, along with the seemingly-inconsistent (but natural and widespread) responses to them, raises (at least for those students who share the requisite contradictory reactions) a series of puzzles. Let's look at how this unfolds in action.

I begin by asking students to imagine a lifeguard who must choose between rescuing one drowning swimmer and rescuing five drowning swimmers, but who cannot save all six. Students are typically unanimous in the judgment that one should "go with the numbers" in this case—no surprise here. I then revise the scenario so that the lifeguard's saving five drowning swimmers is now made contingent on their killing one innocent party—say, by having to drive over a hiker (pinned beneath a fallen tree limb) in order to reach the drowning swimmers in time. Unsurprisingly, answers shift here. While there is often a small contingent expressing willingness to kill one to save five, usually large majorities demure at this prospect. I term the two variants of this case "Rescue I" and "Rescue II," respectively, and their juxtaposition yields our first puzzle.[4] I call this the "Numbers Puzzle," and articulate it something as followings: "Given that the rationale for saving the five in Rescue I was an appeal to the numbers, whence the widespread reluctance to similarly 'go with the numbers' in Rescue II?" Students do not have much difficulty resolving the Numbers Puzzle, however. With little or no prompting, they are able to articulate a version of the doing/allowing distinction that, for ease of future reference, I denominate "KLD": short for "the Killing/Letting Die Distinction."

I next ask students to imagine a variant of Thomson's famous "Transplant" case.[5] However, instead of contemplating the killing of one otherwise-healthy hospital patient (as transpires in the original version), I ask students to consider *letting one person* die in order to do so. This is effected by imagining the protagonist to be a "bystander" of sorts—someone who (unbeknownst to anyone but herself) is in sole possession of an antidote that could save the life of a poisoned and thus rapidly-dying patient in a nearby hospital. This same hospital hosts five other (not-*quite*-so-rapidly) dying patients, all of whom are in need of life-saving transplants of different organs. The poisoned patient (who will expire before the other five do, and whose poisoning will not tarnish any of the organs needed by the other five patients) shares the same rare blood condition that afflicts the five potential organ recipients, making him uniquely suited to serve as their organ donor. (Organs donated from virtually anyone else would be rejected by their new hosts.) These facts, too, are known to our bystander. The question then naturally arises: should this person reveal to the hospital staff the fact that she possesses the antidote that would save the first patient's life? If she produces the antidote, this patient will live, though the other five will assuredly die. If instead she keeps mum about her antidote, the first patient will die—but will then become a veritable "organ farm," whereby the other five patients' lives may be saved.

I call this the "Antidote" case, and for any students answering "yes" to the question "Should the bystander produce the antidote?" (and there are typically many who do so), our second puzzle arises. Deemed the "Letting Die Puzzle," it goes roughly as follows: "Given the

[4] With very minor modifications, these cases—together with their (non-modified) titles—are borrowed from Foot [2002: 81-2].

[5] This case originates with Thomson [1976], though it has antecedents in Foot's discussion of the "Serum" case at her [1967: 63]. See also the discussion at Fischer and Ravizza [1992: 2].

willingness to let one person die in Rescue I for the sake of saving five other lives—and given that, by withholding your antidote, you don't seem to be *killing* the one poisoned patient (as you plausibly had to kill the one hiker to save five drowning swimmers in Rescue II)—why not similarly let one die to save five in Antidote?" With perhaps just a bit of light prompting, students can in due course easily resolve this puzzle as well. This they do by articulating a principle which, like KLD before it, is able to account for a salient disanalogy between two cases which initially appear to be analogous. Specifically, students recognize that, unlike the lifeguard in Rescue I, the bystander in Antidote *intends* the poisoned patient's death. That is, they discover a version of the intend/foresee distinction that, for ease of future reference, I denominate "INH": short for "the Intend-No-Harm Principle."[6]

It is only now, having considered these three cases (and the two puzzles generated by their various juxtapositions), that introduction of the actual Trolley scenarios is appropriate. I start by introducing two standard variations on Foot's [1967] original puzzle, which I call "Trolley I" and "Trolley II." The former is the familiar "spur case," wherein a bystander has the opportunity to divert a runaway train from a track whereupon it will kill five workers to a spur whereupon it will kill one. The latter is the familiar "girthsome man" case, wherein a bystander has the opportunity to deposit a large individual onto the path of an oncoming runaway train, thereby stopping the train before it can kill five workers stuck in the tunnel ahead.[7] Reliably, there is a contingent of students willing to sacrifice the one for the five in the first case but not the second. (To ensure as much symmetry between the cases as possible, I arrange things so that the bystander takes the same course of action in either case: pulling a lever. In Trolley I, this lever alters a switch in the track, diverting the train down the spur; in Trolley II, this lever opens a trapdoor on a pedestrian footbridge overlooking the track—a trapdoor whereupon the girthsome stranger stands.) These students are then forced to confront our third puzzle. The "Trolley Puzzle" can be expressed as follows: "Given that you were willing to throw the lever in Trolley I—and given that this action seems tantamount to *killing* the lone worker on the spur—why don't you similarly throw the lever to kill the girthsome man to save the five workers in Trolley II?"

Owing to reasons explored more fully in the Appendix, certain considerations may complicate the attempt to resolve this Puzzle in similar fashion to the preceding one—*viz.*, by invoking INH and pointing out that, while foreseen, the death of the lone spur(n)ed worker in Trolley I is unintended. I thus push students to search for other ways of resolving this puzzle. Typically, this pushing will eventually lead to the spontaneous identification of a quasi-Kantian injunction on utilizing persons as mere means to our ends (in the manner that the girthsome man, though not the lone worker, is used in the Trolley scenarios). For ease of future reference, I denominate this the "Impermissibility of Using Persons as Mere Means to Your Ends" principle—or the "Means/Ends" principle for short; "M/E" for shorter.

We have now encountered all five of the cases that jointly constitute my term-opening trolleyological exercise, but one more puzzle remains. The hardest of the four (because it does

[6] Rachels [1975] provides a classic expression of the position that in all such cases, it is only ever intent, rather than a (what he regards as spurious) distinction between *doing* and *allowing*, that is morally salient.

[7] These two now-standard variations on the Trolley case were first introduced by Thomson [1985]. In Foot's original version, the agent at the center of the drama is the driver of the trolley itself, rather than a bystander who can throw a lever to divert the train.

not, in my assessment, offer any *clear* or *clean* resolution), the "Killing Puzzle" arises from juxtaposing Trolley I with one of the earliest cases in the sequence: Rescue II. It confronts any students willing to throw the lever in the former case but unwilling to drive over the hiker in the latter case, and I articulate it roughly as follows: "Given that you were willing to throw the lever in Trolley I, for the sake of saving the five workers, why aren't you similarly willing to drive over the hiker in Rescue II, for the sake of saving the five drowning swimmers? Conversely: given your unwillingness to kill one to save five in Rescue II, whence your willingness to do just this in Trolley I?"

This is the hardest puzzle of the bunch! Whereas the first three admit of fairly straightforward solution via the invocation of basic distinctions and principles, this one offers no clear resolution. Owing to reasons explored more fully in the Appendix, various considerations complicate attempts to resolve this Puzzle by invoking any of the three previously-cited principles (KLD, INH, or M/E). The best that can be done to resolve the Killing Puzzle, students often conclude (after a considerable amount of productive class discussion), is to assign complete parity to the two actions constituting the horns of the dilemma in Trolley I. In other words, students eventually move to resist the initial characterization in terms of "killing the lone worker on the spur, vs. doing nothing and letting the five workers on the main track die." Though students won't typically express it quite this way, one might also borrow the terms Thomson uses in her [1976: 76], and speak of there being a pre-existing lethal agent, which in some sense is threatening all six workers equally. (That lever could very well have been in the other position, in which case the "default" scenario would have been that the lone worker down the spur stood in the path of the runaway train.) Seen under that guise, by throwing the lever you are merely deflecting that pre-existing threat to where it will do the least harm. In that fashion, perhaps, one can draw a principled distinction between Trolley I and Rescue II, such that the killing/letting-die distinction does not apply to Trolley I, and once again the only relevant considerations become maximizing benefit and minimizing harm. (Indeed, construed in this fashion, the case comes to look almost parallel to the one with which we started, Rescue I, except that the contemplated "letting-dies" are replaced with "killings." In other words, we're faced with the unhappy fact that we must either kill one, or kill five—"tempered," perhaps, by the consolation that at least we cannot kill all six—leaving "the numbers" as once again the only relevant dimension of moral analysis.)

But let's not digress too far. Readers interested in reflecting further on the Killing Puzzle are invited to consult the relevant portion of the Appendix. The point at this juncture is not to make advances in extant trolleyological theory; the point is merely to sketch one distinct and useful version of trolleyological pedagogy. The "payoff" of the pedagogy just sketched is this: trolleyology has established in students' minds two observations—one pessimistic, and one optimistic. The *pessimistic* observation is that (for many of us, at least) our everyday, ordinary, "pre-theoretical" moral intuitions do not always appear to hang together very well. Juxtaposing common, widespread reactions to some famous moral-dilemma thought experiments reveals some interesting puzzles: pairings of intuitions which, on the surface, appear to clash. Why are we willing to "kill one to save five" in the first Trolley case, for example, but not in the second? However, students' ability to effect resolutions of (most of) these puzzles illustrates the second, and more *optimistic*, observation. This is the recognition that we possess the capacity to employ moral reasoning to discover, identify, and articulate principles and distinctions that can render cohesive that which initially appeared incohesive—to resolve apparent contradictions among our moral intuitions. Our brief, term-opening foray into trolleyology demonstrates to students *that* our moral judgments are improvable, and even more importantly it demonstrates *how* we might make progress in improving them. In short, it shows us how to do moral theory—and, for that matter, how to *philosophize*. Students can be surprised and impressed to learn that, unbeknownst to them, they had already rolled up their sleeves and begun doing philosophy, in the very opening days of the class.

This general approach to trolleyology, and to moral theorizing, can serve as a foundation for a whole term's course organized around this "puzzle-based" methodology. In this paper's third section, we will take a look at how this approach might extend to other domains of philosophical inquiry. But first, let's turn our attention to a handful of other pedagogical opportunities that the foregoing trolleyological framework raises vis-à-vis the teaching of moral philosophy specifically.

2. General Lessons for Motivating and Teaching Moral Philosophy

Once the trolleyological exercises described above have been concluded, I try to convey to students what I take to be the "*moral* of the story" (pun intended) with respect to (i) the status of our moral intuitions and the nature of moral disagreement, (ii) our prospects for improving upon our moral intuitions and (at least partially) ameliorating this disagreement, and (iii) the nature, role, and content of moral theorizing. Let us discuss each of these in turn.

2.1 Moral (Intuitions and Moral) Disagreement

This particular approach to trolleyology serves to illustrate some interesting features of our everyday, ordinary, pre-theoretical moral intuitions. The first and most obvious of these features is *dissensus*—or what I call the "Fact of Moral Disagreement" (FMD). The FMD captures two distinct phenomena. The first, as the name casually suggests, is *inter-personal* disagreement. From the early stages of our trolleyological exercises, we see dissensus emerge within the classroom: the "numbers theorists" early on form a (sometimes vocal and proud; usually confident) minority, willing to buck the prevailing aversion to killing one for the sake of saving five. Later, further cleavages emerge even among the non-numbers theorists: some are willing to kill the lone worker in Trolley I for the sake of saving the five; others remain resolute in their aversion to maximizing lives saved at the expense of violating KLD. Of course, no one needs to enroll in a philosophy class to learn that interpersonal moral disagreement exists: this phenomenon is pervasive and (sadly) all too familiar. But a second form of moral disagreement also lurks here and, in many ways, it is more salient for our purposes. Indeed, it is the phenomenon thrown into highest relief by the sequence of four puzzles generated by our five cases—for the FMD also encompasses what I call *intra-personal* disagreement. That is: disagreement, not just across different persons, but "internally," "within" a single person. For many students will find that their initial, untutored reactions to these cases do not cohere very well. Or, at any rate, they will find that some of their intuitions appear *at least initially* to be inconsistent with one another. Those initial appearances were not all decisive, of course. Much of the point of the sequence of thought experiments is to generate puzzles—but puzzles that are (fairly) easily resolvable. The fact that the puzzles were amenable to resolution demonstrates that (at least some of) our moral disagreement, both inter- and intra-personal, is ameliorable. We'll have more to say about the nature of this puzzle-resolution and disagreement-amelioration later. Before discussing this, though, I like to have students explore more fully the nature of moral disagreement by comparing it with another relevantly (dis)similar form of disagreement: *gustatory* disagreement.

Like our reactions to these moral-dilemma thought experiments, our reactions to various tastes and flavors diverge as well.[8] In other words, just as we have the Fact of Moral Disagreement, we also have what we might term the "Fact of Gustatory Disagreement" (FGD). Just as we disagree about whether or not to throw the lever in Trolley I, and about the relative merits of the Numbers Principle vs. KLD, so also do we disagree as to whether vanilla ice cream tastes better than chocolate, and about the relative merits of raspberries versus strawberries. Like moral disagreement, gustatory disagreement can also come in both inter- and intra-personal flavors. (Since some students seem surprised by this latter claim, I often share here some autobiographical detail regarding my ice cream preferences.[9]) Yet philosophers (let alone laypersons) tend not to argue about these issues, and (curricularly speaking) we don't have introductory classes called "the Philosophy of Gustatory Judgments" or "Contemporary Gustatory Controversies." Why is this? What's relevantly different about the FGD that distinguishes it from the FMD in this regard?

I have in mind three dimensions along which moral disagreement and gustatory disagreement differ, and I find that students are able to identify each of them with very little prompting. The first of these is what I call the *objectivity* dimension. Simply stated, our moral intuitions strike us as being objective—as being (in at least a wide range of central cases) right or wrong, correct or incorrect; moral judgments strike us as being truth-apt.[10] When two people disagree about a moral matter, it seems that (at least) one of them must be mistaken. By contrast, our gustatory disagreements simply resolve into matters of divergent tastes or preferences.

[8] By "reactions to various tastes and flavors," here I'm talking only of our evaluative responses *to* gustatory sensations—not the gustatory sensations themselves. (Though it is perhaps the case that we exhibit considerable interpersonal variation with respect to these very sensations too; on this matter, see e.g. Baltzly [2020].) For the latter probably share important features in common with moral intuitions, and thus do not differ from moral intuitions along those dimensions to be developed shortly below. (For example, gustatory sensations probably *are* best characterized as being truth-apt—this herb you're tasting either is, or is not, cilantro—and these gustatory sensations probably *are* improvable—by developing your palate, you can come to discriminate more and more subtle notes and flavors in a wine or a cheese, etc.)

[9] OK, since you asked: Relatively early in my adulthood, I had determined that my two favorite flavors of ice cream were coffee and (regular, non-mint) chocolate chip. But it was to be a number of years before I was to first discover *coffee-chocolate chip* ice cream. Upon first encountering it, I was very excited: my love of its two component ice cream flavors—together with my intuitive gustatory sense that these are two great flavors that would (unlike, say, chocolate and *pickles*) go great together—led me to expect that I would love *this* flavor of ice cream above all others. You can imagine my surprise, then, when I found myself *disappointed* with the experience. I didn't actually appreciate the flavors' combination—even though it had seemed to me for all the world that I ought to. I was encountering, in other words, a case of *intrapersonal inconsistency* in my gustatory judgments. Or so one might argue, at least.

[10] Of course, I do not wish to prejudge matters in favor of some form of (as various authors variously formulate it) *rationalism*, or *objectivism*, or *realism*, or *cognitivism* about morality, as opposed to (e.g.) *expressivism*, or *emotivism*, or *subjectivism*, or *anti-realism*, or *non-cognitivism*. This is why it is helpful to hedge here by saying that moral intuitions *seem* to be truth-*apt*. In class, I offer (but do not dwell upon) the possibility that this appearance may be only that—an appearance, at odds with the underlying reality. (I also return to this issue later in the course, for purposes of illustrating the sorts of questions asked, and the kinds of theorizing engaged in, when one is doing *meta-ethical* inquiry, as opposed to straightforward normative ethics of the kind conducted thus far.) But as even the most committed anti-realists are likely to allow, our moral judgments bear a *prima facie* appearance of objectivity—and that is all we need for present purposes.

Gustatory "intuitions" (if one can even use that phrase) are merely subjective—their "truth" (if that notion even gets a grip here) is in the eye (or perhaps the tongue) of the beholder.

The second dimension is what I call *gravity*: not only do our moral intuitions seem to be truth-apt, but it also seems to truly *matter* that our moral judgments be true. When we encounter someone who disagrees with us on an important moral matter, we find ourselves troubled by this fact (particularly if it is someone we know well and care for), and it strikes us as important that this person be shown the error of his or her ways. Similarly with respect to our attitudes toward our own moral intuitions: once (if) we recognize and take seriously the possibility that some (or many) of our moral intuitions may be in error, it becomes a compelling project to endeavor to expunge as much error as possible from our web of moral belief and commitment. Gustatory disagreement, by contrast, elicits no such urgency, with respect either to our own or to others' gustatory positions. Gustatory disagreement just doesn't matter that much; it doesn't, that is, have practical *weight*. We can better appreciate the conjunction of *objectivity* and *gravity*, a conjunction characteristic of moral intuitions, when we consider another relevant and instructive comparison in this neighborhood: a comparison with a case where we seemingly observe objectivity *without* gravity. Students will concur that mathematical judgments are, like moral ones, highly (perhaps fully) objective—even though mathematical judgments, like moral ones, appear to be non-empirical. However, disagreement about objective mathematical "facts" (if that is the proper word to use here) does not strike us as quite as weighty as disagreement about moral "facts." Disagreement about the status of, e.g., Hilbert's Problems simply doesn't get the same grip on us as does disagreement about, e.g., the morality of euthanasia or abortion.

When we *do* encounter persons who disagree with our deeply-cherished moral views, and especially when we furthermore find that we care considerably about the fact that they err, a further consequence follows quite naturally: we find ourselves moved to correct these persons' mistaken convictions. (Likewise, we would want to be shown the error of our own moral commitments by those positioned to recognize them as errors.) What's more, we do not find ourselves moved to employ just *any* means of changing their minds. Our first reaction is not (typically) to plead, or to pester, or to harass, or to bully or intimidate, or to coerce them into revising their moral views. Rather, we usually find it most promising to *reason* with them—to offer facts, arguments, counter-examples, parity-claims, and the like. Similarly when we encounter internal inconsistencies plaguing our own webs of moral intuition: we seek to reason our way through matters in an effort to arrive at a more rational, consistent, and coherent moral outlook. These reactions point to, and would not be possible without, the third differentiation between moral and gustatory judgments: the *improvability* dimension. In short, our moral intuitions seem *susceptible to improvement* in a way that our gustatory reactions do not; moral disagreement is *ameliorable*, in a way that gustatory disagreement is not. Much more can be said about this dimension,[11] so let's afford it its own sub-section.

[11] One note here about my talk of "dimensions" along which gustatory and moral disagreement differ (rather than speaking simply of, e.g., "three respects" in which the FMD and the FGD vary). The reason is that, in a writing assignment later in the semester, I will ask students where they locate *aesthetic* intuitions and judgments along these three dimensions as well. Employing the language of "dimensions" here facilitates students' ease of expression of various subtle positions. For example, they may locate aesthetic judgments as occupying some intermediate position on a continuum of objectivity—with moral judgments being largely (or maybe wholly) objective, and gustatory reactions being solely subjective. (Likewise with respect to improvability. However, students tend to be fairly uniform in characterizing aesthetic judgments as being just as "low-gravity" as gustatory

2.2 Improvability

Rather than resting content with the general observations that we will seek to minimize intra-personal dissensus via procedures of reasoning, and that we will seek to ameliorate inter-personal dissensus via procedures of rational discussion and argumentation, I offer students some remarks on one concrete way that these procedures of reasoning and argumentation might be conducted. In other words, I describe a methodology by which we might seek to improve our improvable moral intuitions. To do this, I simply elaborate on a procedure students have by now encountered in their assigned reading: Elliot Sober's notion of *reciprocal illumination*.[12] In essence, reciprocal illumination is the procedure that many philosophers (somewhat sloppily) call "reflective equilibrium." Strictly speaking, *reflective equilibrium* is a state—a goal, an aspiration—but not a process. One common and important procedure by which one might *aim* at reflective equilibrium often gets called "reflective equilibrium" as well, but this is to confuse means with ends. Sober is quite correct, in my judgment, to isolate this one important procedure for effecting equilibrium, and to give it its own proper name and characterization—so I follow him in doing so.

I begin my explanation of reciprocal illumination by introducing a pair of distinctions. The first is that between moral *intuitions* and moral *judgments*. Students usually have little trouble guessing what I'm after with this distinction: the former term refers to our initial, instinctive, unreflective reactions to certain morally-salient matters (e.g., descriptions of (actual or hypothetical) moral dilemmas; candidate moral principles); the latter refers to our more considered, reflective, "final answers" on such matters. The second distinction applies to either intuitions or judgments (in our just-introduced, stipulative senses of those words), and is that between *particular* and *general*. The meaning of this distinction is a little less intuitive, but it runs as follows: the former refers to one's moral assessments (whether "intuitive" or "judgmental") of specific cases, or decisions, or actions; the latter to one's assessment of more abstract *principles* of (more or less) wide applicability. (So, for example, one's initial instinct that one would not run over the trapped hiker in Rescue II is a *particular intuition*, while one's considered, reflective acceptance of the killing/letting-die distinction is a *general judgment*.) Armed with this pair of distinctions, we can now give pithy expression to the technique of reciprocal illumination sketched in Sober's text: it is the examination of our *particular* intuitions in light of (hence the "illumination" metaphor) our *general* intuitions, and vice versa. The idea is to search for cases of "moral cognitive dissonance": to find instances where one of our pre-theoretical particular intuitions is inconsistent with some general principle we find intuitively attractive, and to find general principles which, despite their intuitive attractiveness, might clash with some of our intuitions about particular cases. When such inconsistencies are discovered (as they were as many as four times during the trolleyological exercises described in section 1), we compare the relative strengths of our commitments to each member of the pair of inconsistent intuitions, in an effort to "illuminate" to ourselves which commitments seem strongest. In ways that are familiar to readers acquainted with our profession's more casual use of the term "reflective equilibrium," we then decide which intuitions to retain, and

ones.) Having set up the "FMD-vs.-FGD" discussion in terms of "dimensions" prepares the way for this later philosophical exercise. If I weren't planning to use this essay prompt later in the term, however, I might dispense with the dimensions-talk, and just employ the simpler language of "three ways the FMD and the FGD differ."

[12] Sober's description of reciprocal illumination in his introductory text *Core Questions in Philosophy* can be found at pp. 342-43 (6th Edition), in his discussion of Mill's Utilitarianism.

which to reject—which revisions to our web of moral belief we're most prepared to make, in order to patch up the integrity of the whole. The reactions (both particular and general) that survive this comprehensive practice of reciprocal illumination, then, merit the honorific of "judgments."

In order to illustrate the technique of reciprocal illumination in action, I ask students to consider a tactic that might be used to resolve the Trolley Puzzle. Many students find that their particular intuition in Trolley I—that it's acceptable to divert the train towards the lone worker—clashes with their general intuitive sense that killing is worse than letting die. To resolve this conflict between clashing intuitions, it is not uncommon for some students to revise their particular intuition in Trolley I—to announce that they've changed their mind, and that now they do *not* believe it permissible to throw the lever and switch tracks. The result is that a pair of inconsistent intuitions—one particular and one general—has been partially revised and replaced with a pair of consistent judgments.

The ultimate goal of this *process* of reciprocal illumination is to arrive at a *state* of reflective equilibrium—and it is this (slightly unconventional) sense of the phrase that I introduce to students. I stress to students that they should understand reflective equilibrium only as an aspiration—a goal of moral theorizing, but not a state that many of us should ever hope to achieve. Nevertheless, the vision of one's particular and general moral judgments hanging together in harmonious balance, with no detectable traces of incoherence among them, serves as an attractive and helpful regulative ideal. And we can certainly work our way ever closer to this ideal—bringing our judgments into ever-greater harmony, even if we may never obtain perfect equilibrium. I'm also always careful to stress the difference between reflective and *un*-reflective equilibrium. Moral turbulence, like any form of cognitive dissonance, can be avoided on the cheap if, like an ostrich, one simply refuses to confront (or even to acknowledge) one's intra-personal moral inconsistency. (I shall presume that the reality of *inter*-personal moral inconsistency is impossible to ignore, however.)

2.3 Moral Theory & Moral Theorizing

This method of reciprocal illumination, aimed at reflective equilibrium, suggests a mechanical, methodical, ground-up approach to moral theorizing: an approach wherein one grinds through a bunch of carefully-constructed cases in an attempt to tease out the inconsistencies in our pre-theoretical intuitions, and wherein we try our best to discover deeper principles and distinctions that might resolve these inconsistencies. Once students have come to appreciate the possibility of such a "bottom-up" moral methodology, I introduce the notion of a *moral theory* as the kind of desired output of such a process—moral theory as something like a regulative ideal for (the process of) moral theorizing. And the toy case I use to illustrate this is the Doctrine of Double Effect (DDE). I had actually first introduced the DDE near the completion of our trolleyological pursuits, as an example of a sort of "compound moral principle"—a general moral prescription, like that found in INH, albeit one that incorporates several sub-conditions.[13] I now offer up the DDE, not so much as a compound moral

[13] The version of the DDE I employ in class goes roughly as follows: "In cases where an action brings about both good and bad outcomes, it is permissible for you to perform this action as long as the following three conditions obtain: (1) the value of the good outcome sufficiently outweighs the bad; (2) you do not intend the bad outcome (even if you do foresee it); and (3) the bad outcome is not a means to effecting the good." As thus formulated, conditions (1)-(3) can be seen as incorporating (or at least corresponding to) each of the distinctions

principle (one with multiple moving parts), but rather as something like a "proto moral *theory*" that one might assemble using this sort of bottom-up approach to moral theorizing. After all, simple though it is, the DDE seems to have the requisite features of a theory: a fairly broad (but well-defined) domain of applicability, internal structure, and the like.

However plausible it may or may not be to accord the DDE the status of "moral theory," no one is likely to mistake it for a *complete* moral theory. For one thing, its range of application is still somewhat limited: it's formulated only so as to provide guidance for moral decision-making in cases of "double effect": cases where our actions will have two (or more) outcomes, (at least) one good but the other(s) bad.[14] So we might desire a moral theory with wider scope than this. I ask students to contemplate what, ultimately, we as moral theorists might aspire to. I suggest that two cardinal virtues of any theory (in morality, as elsewhere) are *generality* and *simplicity*. That is, theories should have the widest possible scope of application, and the most elegant or parsimonious possible formulation. Consideration of these two theoretical virtues suggests that the pinnacle of moral theorizing is to discover or formulate what Immanuel Kant called "the Supreme Principle of Morality" (SPM): a *single* moral principle that (in contrast with the DDE, e.g.) applies to *every* moral decision. (And indeed, Kant is one important philosopher who claimed to have identified such a Supreme Principle.) It is important to acknowledge, of course, that many philosophers deny that such Principle is on offer. Nevertheless, there is considerable value in examining the views of two famous and influential philosophers who did defend such a principle. One is Kant, and the other is John Stuart Mill.

At this point I segue into a brief examination (something like a week each) of Mill's Utilitarianism and Kant's deontological ethics—treatments I shan't summarize here. But before making this transition, I offer just one further remark about the nature of moral theorizing. I note to students that, in our search for this SPM, we needn't restrict ourselves to the aforementioned bottom-up approach. I point out that if this were a semester-length course in moral theory, we would examine Kant's and Mill's "top-down" philosophical arguments for (their favored candidates for) the SPM—their approaches to defending their respective principles on their broader merits, and with reference to wider considerations. These approaches go beyond merely relying on repeated applications of the method of reciprocal illumination. For students interested in acquiring a taste of Mill's reasoning, I refer them to Sober's treatment of Mill's analogical argument concerning *visibility* and *desirability*.[15] Likewise, I suggest a broader picture of Kant's moral theorizing by explaining that his overall practical philosophy (of which his formulation(s) of the Categorical Imperative was merely a portion) was arguably an effort to understand the nature of genuine human *freedom*. For present purposes, though, we can only look at the "punch lines" of Kant's and Mill's proposals—we'll examine Kant's Categorical Imperative and Mill's Greatest Happiness Principle as candidates for SPM. As our brief examinations of these two principles quickly

and principles we'd articulated during trolleyology: (1) KLD as a sort of "weighting principle"; (2) INH; and (3) M/E. This particular rendering of the DDE is, I believe, not implausible—even if it may not be completely faithful to important historic formulations of the doctrine, like that of St. Thomas Aquinas.

[14] The phrase "double effect" is, I believe, most often taken to refer to the duality of *intended* and *foreseen-yet-unintended* effects of our actions. Nevertheless, I find that it's helpful to present the phrase to students as capturing the *good consequences/bad consequences* duality. The latter distinction is more intuitive and familiar, and so stressing it can serve to effectively "lock in" the DDE's main thrust in students' minds and memories.

[15] Sober's treatment of this can be found at pp. 341-2.

reveal, Mill's approach (which we study first) can be plausibly regarded as a systematization of the "numbers theory," while Kant's SPM (at least in the guise of the "Humanity Formulation" of the Categorical Imperative) can be plausibly regarded as a systematization of the principle we'd termed "M/E."

3. Applications Beyond Moral Philosophy

At this point, some readers may say, "So far, so good—if you happen to be teaching a class focused on morality." And it's true: the foregoing description of my trolleyological approach (from section 1), together with the lessons I draw from it regarding, e.g., the nature of moral disagreement and moral theorizing (described in section 2), clearly lend themselves to the opening stages of the introductory applied ethics courses ubiquitous in (at least American) universities' core curricula. But my claim here is that trolleyology is suitable, not only for such courses, but for *any* introductory philosophy courses—and perhaps for a host of more advanced courses as well. (For example, I have also used this semester-opening motif to good effect in a graduate-level public policy course called "Moral Dimensions of Public Policy.")

I make this claim because I believe that this particular approach to trolleyology provides a useful template for a "puzzle-based" approach to introducing our discipline—a framework that is useful for engaging a wide range of philosophical topics that one might expect to cover in introductory classes in philosophy and critical thinking. Not only do the four moral-dilemma-generated puzzles from section 1 provide a fruitful template for posing other central philosophical problems, but the means students employed to try to resolve these puzzles can be deployed in reasoning and arguing through other important philosophical problems as well. Significantly, students find that the forms and styles of rational inquiry, reasoned debate, and argumentation employed during our trolleyological exercises came quite naturally to them, with little in the way of explicit guidance or direction from the instructor. Students, that is, are able to discover—without being told how, without being told what they are doing, and without at first even realizing that this is what they are doing—their innate ability to roll up their sleeves and "do philosophy," in the very first days of their very first philosophy class. Armed with this discovery, and the concomitant confidence it breeds, student are (I have found) well-equipped and enthusiastic to launch into a term's worth of investigation of other fundamental philosophical problems. In the remainder of this section, I suggest some ways this general approach can be adapted so as to frame classroom discussion of a few other central philosophical subjects commonly treated in introductory classes.

The subject to which this framework might most profitably be employed is the *problem of evil.* By casting this problem as an inconsistent triad—and not, as do some authors (including Sober, in his introductory textbook), as an argument—its puzzle-like nature becomes clear. The riddle becomes, "Which of these three statements[16] should we abandon?" And then, as student discussion (perhaps with the aid of some light shepherding from the instructor) develops and permits ascending levels of sophistication, the puzzle may evolve into, "How might all three of these statements be true at once (notwithstanding the surface appearance of

[16] In one common formulation, the three statements constituting the putatively inconsistent triad are as follows: (i) "God is omnipotent" (where omnipotence is understood to include omniscience); (ii) "God is omnibenevolent"; and (iii) "Evil and/or suffering exists."

mutual inconsistency)?"[17] This puzzle-centered approach informing students' examination of the problem may take a variety of forms. It can be tacit—a background framework known explicitly only (at first) to the instructor, whose vision of it serves to guide student discussion first toward a recognition of either or both of the riddles mentioned above, and from there on to a discovery of some of the riddles' possible solutions. Or, the puzzle-based formulations can be an explicit part of the presentation of the problem itself (as would be appropriate and necessary in a large lecture class, e.g.). Other hybrid variants surely are possible as well.

Another subject which lends itself naturally to this puzzle-centered strategy is the issue of *freedom and determinism*. One might formulate the philosophical worries about free will in the form of either or both of the following puzzles. (Other formulations are no doubt possible too.) First, there is what we might term the "causal necessity puzzle": if our brains (or at least our central nervous systems) are the causes of all our behaviors, and if our brains are purely physical systems, whose component parts (neurons, neurotransmitters, etc.) are fully governed by the relevant and causally-necessitous laws of nature (laws of neurophysiology, laws of biochemistry and organic chemistry, laws of physics, etc.), then how can any of our behaviors truly be free? Are not our behaviors as physically and causally determined as those of any other physical system? Is not the difference between us and a computer—or, for that matter, between us and the balls on a billiards table—only a difference of degree, and not of kind? Second, we have what we might call the "divine foreknowledge" worry: if there is a God, and if s/he is truly omniscient, then doesn't God know the future—including *our* futures? And if God knows for certain what grade we will get in this class, whether and when we will graduate (and what our G.P.A. will be), what our first job will be, whether and when we will marry and reproduce, and the exact time and manner of our death … then how can any of our future decisions be taken freely? Is not God's foreknowledge incompatible with our having genuine free will? Either of these formulations can be expressed in even more dramatic puzzle-like form, in fact, by offering them up as inconsistent triads in the manner of the problem of evil.[18] To illustrate, taking just the causal necessity puzzle: (i) *Freedom:* It is true of some of our actions that we could have acted differently. (ii) *Causalism:* Our actions (like everything else) are wholly governed by causal laws of nature operating on the way we happen to be. (iii) *Incompatibilism:* If the thesis of freedom is true, then the thesis of causalism is false.

Introductions to still other areas of philosophy can be made to fit this puzzle-centered mold, even if (unlike the problem of evil and the freedom/determinism issues) they are not as naturally or obviously formulated as riddles. For example, I have had some success in motivating *theory of knowledge* by constructing interesting and compelling cases (or as interesting and compelling as I can manage to make them, from students' perspectives) involving pairs of individuals, one of whom "merely" has a true belief (but intuitively not *knowledge*), and the other of whom seemingly does have knowledge. I then ask of these individuals, "What does

[17] Candidate answers here, which you might wish to discuss with your students, include of course the "free will defense" (as it's popularly called), and either of several versions of (what might be called) the "good-requires-evil defense." Examples of this latter strategy include what we might term "epistemic" versions of the argument—*viz.,* that God permits the existence of some evil, for otherwise our limited minds would be unable to recognize and appreciate his created goodness—and "metaphysical" versions—*viz.*, that the existence of (comparatively more valuable) "second-order goods" like courage and compassion is logically dependent upon the existence of certain "first-order evils" like danger and suffering.

[18] I am indebted to Andrew Kania for suggesting this to me, and for the particular formulation of the inconsistent triad that follows.

the latter have, which the former lacks?" In this way, students are able to arrive themselves at an appreciation of the notion of *justification* or *warrant* as it has influenced the field of epistemology, and are thus supplied with a somewhat higher degree of motivation to tackle some of the main debates and positions in this field. (At least, they display a higher degree of motivation relative to the degree of enthusiasm I'd formerly observed, before re-crafting my opening epistemology session so as to fit this puzzle-based mold.) Of course, the problem of *skepticism* also provides a compelling *entrée* into epistemology—and it, too, can be presented in an essentially puzzle-centered format. For instance, you might pitch theory of knowledge as being, in part, an attempt to counter the skeptic—the person who responds to the following inconsistent triad by abandoning its first claim: (i) It is possible to have knowledge of the "external world." That is, there are some propositions about the external world *P*, such that some subjects *S* have justified true beliefs that *P*. (ii) If *S* knows *P*, then *S* cannot possibly be in doubt as to *P*. (iii) For any *P*—at least, any *P* about the "external world"—it is possible to be in doubt as to *P*. (Obviously, this skepticism-centric approach to introducing epistemology works especially well if students are reading Descartes's *Meditations* at this juncture of the course.)

Finally: the *mind/body problem* too can be made to fit this mold. One way to do so may be by juxtaposing "mind talk" with "brain talk," in something like the following fashion. On the one hand, we often deploy *mind talk* when engaged in various explanations and descriptions of behavior. "Mind talk" refers to our customary habit of invoking our (and other persons') minds and their contents—held to consist of beliefs, desires, attitudes, aptitudes, and the like. On the other hand, we sometimes deploy "brain talk"—language that invokes various medical or biological terms ("grey matter," "neurons," "hemispheres," and the like)—to describe that which occupies the space between our ears. And quite often, we find that brain talk can substitute for mind talk. For example, we might say (rather literally) that "I have a lot going on in my brain" (rather than "a lot on my mind") if we're presently preoccupied with worry. Similarly, we might say (rather metaphorically) "My brain hurts" after taking a calculus exam. But in other cases, brain talk and mind talk are *not* so clearly interchangeable. Rarely if ever would we say that our minds weigh approximately three pounds. Likewise, even the most physicalist physicist would demure at amending Stephen Hawking's characterization of the search for the (grander, more unified) Theory of Everything as the effort to know the "Brain of God." So what gives? What do we make of this fact that brain talk sometimes is, and sometimes is not, interchangeable with mind talk? Are the mind and brain the same thing, or not?

4. Objections and Replies

So much for the trolleyological approach I favor, and for the template it offers for introducing not only moral philosophy, but also a host of other central philosophical topics. At this point we should pause to consider some concerns that might arise in instructors' minds as they contemplate employing such an approach. Specifically, in this section I respond to three worries that might be troubling readers: a worry about the wider applicability of specific methodological tools I've described; a worry about over-simplifying or even trivializing various subject matters; and lastly a worry about the possibility of engendering moral skepticism. Let us take each in turn.

4.1 Does Reciprocal Illumination Generalize to Other Philosophical Domains?

As shown in section 3, the puzzle-centered methodology (developed in section 1) of my trolleyological approach can be adapted for the purpose of introducing other domains of philosophy. Many perennial philosophical topics, that is, can be introduced by way of puzzles or riddles. However, it is less clear whether all the specific methodological lessons (developed in section 2) can be similarly applied across the board. This is particularly the case with the technique of reciprocal illumination. It is simply unclear how our intuitive reactions to hypothetical thought-experiments concerning theodicy, determinism, epistemic justification, and the mind/body problem could be characterized according to categories like "particular" and "general," in the way presupposed by my presentation of Sober's technique. I grant this difficulty. Nevertheless, the related notion of reflective equilibrium may get just as much traction in any of these cases as it does in the domain of moral theory. Furthermore, I would argue that the *spirit* of reciprocal illumination applies in each of these domains, even when the specific technique of comparing intuitive reactions regarding principles and cases has no purchase. For the animating spirit of Sober's technique is to run *toward* (conceptual or doxastic) conflict, rather than away from it—to seek out, rather than to avoid, cases of cognitive dissonance; to revel in intellectual disequilibrium, rather than to shy away from it; and to bravely welcome it as an opportunity to shore up one's various commitments. By encouraging students to search for the most economical resolutions of cognitive dissonance (that is: those minimally-disruptive revisions or modifications of their antecedent webs of belief), instructors still impart the spirit (if not the specifics) of reciprocal illumination.

4.2 Oversimplifying or Trivializing?

Section 2.3 described a methodology-centered lead-in to moral theory—a "bottom-up" approach to moral theorizing as an entrée to the idea of fully-fledged moral theories, such as one finds in consequentialism and Kantianism. But one might reasonably worry that that style of introducing students to the notion of a moral theory could prime them to read Kant's and Mill's theories in an over-simplified or simplistic light. This may be of particular concern for those teaching semester-long courses in ethical theory, who might wish to give these authors and positions careful examination. Mightn't "alumni" of the trolleyological approach described above be tempted to shoehorn all of Kant and Mill into glorified versions of "INH" or "M/E" or the "Numbers theory"? I grant the legitimacy of this concern as well. However, I believe that, with due care, instructors can effectively pre-empt this simplifying tendency. By explicitly characterizing the bottom-up methodology of section 2.3 as an "intellectual *hors d'oeuvres*"—designed to whet students' appetites for moral theorizing, and to motivate students' interest in the subject—instructors can ensure that trolleyology serves as an entrée *into* moral theory, but not as an intellectual entrée in its own right. ("So much for our intellectual appetizer; now for our main course.") Instructors might find that further leveraging the "bottom-up/top-down" terminology is helpful here. "Top-down" moral theorizing (of the sort exemplified in the canonical texts likely to serve as centerpieces of any term-length course in ethical theory) might even be framed as a sort of *shortcut* to arriving at an adequately general, adequately simple moral theory—an alternative to a bottom-up approach that might drag out indefinitely. But however precisely this segue is effected, instructors of semester-length courses in ethics can surely find a way of pre-empting in students' minds any temptation to conflate trolleyology with moral theorizing more broadly—or to sever that incipient link, if the term's opening weeks have caused it to begin to take hold.

4.3 Engendering Moral Skepticism?

Finally, readers may worry that this trolleyological approach could have the unwitting effect of encouraging moral skepticism. Students might draw the lesson that, for every putatively general moral principle, there is a counterexample. They may further conclude that there are no answers to be found in moral thinking, and that our intuitions to the contrary are erroneous. (They may conclude, in other words, that the FMD and the FGD pick out relevantly similar forms of dissensus after all.) They are liable to come to endorse an error theory of morality, that is, precisely because trolleyology tempts them to conclude that, when there is a disagreement, both people can be correct and that, moreover, no reasons can be nor need to be given; our morals (they will now think) are determined simply by our arbitrary, capricious feelings.[19]

This is a natural concern. However, I believe that closer attention to the full resources of trolleyology largely dispels it. True, trolleyology does traffic in *puzzles*, and many of these puzzles betoken seeming counter-examples to principles that were *themselves* previously proffered as resolutions to other puzzles. And true, this could potentially effect the impression that the sequence of "principle → counter-example/puzzle → new principle → new counter-example/puzzle → …" could be extended indefinitely. But the true spirit of trolleyology is to demonstrate that, often enough—and often to the surprise of students in whom they may have initially generated a sense of befuddlement—these puzzles *do* have resolutions. The spirit of trolleyology, that is, is the spirit of (*re*)*solving* puzzles, not of generating them. I deploy my trolleyological approach to demonstrate that, in cases of moral conflict (or at very least, in many cases of intra-personal dissensus), principled rational resolution of puzzles is possible. True, on my favored approach, we *do* eventually reach the "Killing Puzzle," whose resolution proves more challenging. But even here, the difficulty encountered in the attempt to resolve this fourth puzzle will typically function as an enticing invitation to further work. In fact, there is pedagogical value in ending the introductory trolleyological exercise with a difficult and compelling problem. Otherwise one runs the risk that all this "puzzle-mongering" might come to seem like little more than a parlor game. (That is: by introducing at least one puzzle that appears soluble—but which nevertheless remains somewhat unsatisfactorily unresolved—one reinforces the impression that this is in fact serious business that we're up to here, and not simply a more conceptual version of a crossword puzzle.) At any rate: at some point—and this is the case with *any* method for introducing moral philosophy—you've just got to take your chances and recognize that some students may emerge out the other end with a higher degree of moral skepticism than they started with … just as there may be other students who experience the opposite dynamic. For my part, I'm fairly confident that this trolleyological approach still yields a "net positive" in terms of converting students to and from moral skepticism. For every student who *wouldn't* have been a skeptic, save for the Killing Puzzle, there's probably more than one student who *would* have been a skeptic, but who instead found herself "converted." And these conversions transpire when students are sufficiently impressed by the possibility of real moral progress—a possibility revealed, among other places, in the resolutions of the first three puzzles.

[19] I am indebted for another anonymous reviewer from *Teaching Philosophy* for pressing me on this concern, and in particular for this way of expressing it.

5. Conclusion

Trolleyology is famous and familiar. However, I worry that it's not often done properly—that its full potential often remains unexploited. This is seen most commonly in its portrayal in popular culture—most famously, on the recent NBC situation comedy *The Good Place*, where "the Trolley Problem" is portrayed simply as the question of whether or not to divert a runaway trolley away from five workers and towards another worker (or some variant thereof). Philosophers, in my experience, tend to be a bit more sophisticated, and generally understand that the *Problem* per se is not the simple question as to whether or not to throw a lever, saving five lives at the expense of one; philosophers typically understand that the eponymous "problem" is the puzzle that arises from the juxtaposition of common responses to a relevant range of alternative trolley scenarios. (For instance: the juxtaposition of a pair of common intuitions in (what above we termed) "Trolley I" and "Trolley II"—our "Trolley Puzzle.") Furthermore, the general use of trolley-like thought experiments has come in for its fair share of justifiable criticism—including in the pages of this very journal. (See Martena [2018].) However, this paper demonstrates that, by adopting the more expansive notion of "trolleyology" that encompasses not only runaway-train-based scenarios, but also the various scenarios comprised by the Rescue and Antidote cases (and other similar cases besides, perhaps), philosophy instructors can provide students an attractive, accessible—and, dare I say, *fun?*—template for "doing philosophy": *all* of philosophy, not only moral philosophy.

APPENDIX: Teaching Trolleyology: Into the Weeds

This appendix supplies further detail about many aspects of the five-moral-dilemma/four-puzzle sequence briefly sketched in section 1, with an eye toward highlighting pedagogical possibilities that run somewhat tangential to the main thrust of that section. It supplies further detail, not only about each of the cases and puzzles described above, but also about one additional case, plus a handful of other philosophical topics one might pursue in the course of employing this approach.

To briefly recap the description of this enterprise provided at the beginning of section 1 above: Recall that I confront students with a sequence of moral-dilemma-based thought experiments—each of which appears structurally similar to a case that precedes it, but which is such that (for many students, at least) the intuitive reaction to that case seemingly clashes with a principle articulated in an analysis of a preceding case. The juxtaposition of these cases, along with the seemingly-inconsistent (but natural and widespread) responses to them, raises (at least for those students who share the requisite contradictory reactions) a series of puzzles. In order to generate these puzzles cleanly, it helps (as will become clearer as we proceed) to analyze each thought experiment in terms of the following tripartite framework: for each experiment, we try to articulate precisely (i) the *dilemma* we confront; (ii) the *decision* we would make in the face of this dilemma; and (iii) the *rationale* or *principle* we would invoke to explain or justify this decision.

<u>A.1 Rescue I:</u> Our first pair of cases involves what we might term, not "lifeboat ethics," but "life*guard* ethics." I typically canvas the room for lifeguards (there's often one or two students who have worked in this capacity), and ask the class to imagine them to be the "star" in the following scenario:

> **Rescue I:** You are a lifeguard charged with guarding two beaches: West Beach and East Beach. One day, from your perch equipoised between the two beaches, you notice six swimmers drowning, all at once: one on West Beach, and five on East. From your ample lifeguarding experience, you know beyond doubt that you'd be able to save the one swimmer's life if you went West, and that you'd be able to save all five swimmers' lives if you went East; however, you also know you simply won't have time to save all six. In other words: your options are, unfortunately, to save the one drowning swimmer, or to save the five. Which do you choose?

Students' near-unanimous response here—indeed, one might say, the only reasonable response—is to go to the East Beach and save five, allowing the lone swimmer on the West Beach to drown.[20] When I ask students why they opted for the East Beach over the West Beach, the answer (once students are satisfied that I'm not posing a trick question) is swiftly forthcoming, and unanimously assented to: you save the five swimmers, because you "go with the numbers."[21] Expressed in terms of our tripartite template, our analysis of Rescue I runs

[20] In most of my classes, this result has been unanimous. However, on a few occasions I have had a student argue that the most reasonable course of action is to simply choose at random, as by flipping a coin. Though we never have time to sufficiently pursue this reaction when it arises, from cursory discussion of the issue it often appears that this intuition is driven by considerations similar to those that influence Taurek [1977].

[21] By this point, I've typically canvased the classroom, not only for lifeguards, but for journalism majors; from among them, I will solicit a volunteer to "co-star" in the role of "reporter." I'll ask the reporter to simulate the

as follows: the dilemma is "to save the one, or to save the five?"; the decision is (almost always unanimously) to "save the five"; and the reason offered is "the numbers!".

A.2 Rescue II: This case is a variant on Rescue I; the most important difference here concerns a change in (or important further specification of) the (literal) landscape on which the drama unfolds:

> **Rescue II:** Imagine yourself again to be charged with guarding two beaches, but this time no one is drowning at West: only the five are drowning at East. However, in order to get to East Beach, you must *drive*: specifically, you must drive your standard lifeguard-issue Land Rover from your perch, poised high above beaches, down a very narrow mountain road, with steep rocky sides. The catch is this: mid-way down this road, there is a hiker pinned beneath a fallen tree limb in the middle of the road. She's[22] basically safe (perhaps she—like the tree—has suffered a broken limb, but nothing more); however, she is trapped. She's used her cell phone to call a tree surgeon, who will arrive in an hour (after the five swimmers have already perished) to cut away the (tree) limb to free her. Your Land Rover is capable of driving *over* the felled limb—but in doing so, it would crush the poor hiker pinned beneath, and she would die. So your choice is this: do you drive over the hiker, killing her, in order to get to the five drowning swimmers (whom you are certain you will save)? Or do you refrain from driving over her, allowing the five swimmers to drown?[23]

Many people "switch their answer" here—opting *not* to go with the numbers, but rather to refrain from driving over the hiker—thereby allowing the five swimmers to drown. There is always a small (but typically confident and well-spoken) minority of students, though, who "stick with the numbers" here, and opt to drive over the hiker in order to save the five drowning swimmers. Such students have not (yet) exposed themselves to the charge of

portion of his or her interview with the lifeguard (they are to imagine that the reporter is planning to write a splashy feature story for the local paper) wherein the lifeguard is asked to explain his or her reasoning in deciding to save the five Eastern swimmers rather than the lone Western one. I'll then ask the wider classroom whether they concur with the interviewee's explanation. This reporter can continue to play this role—the antagonist to the lifeguard's protagonist role—at several more points in the subsequent discussion: this is a helpful role-playing exercise whereby you can elicit students' thoughts regarding the rationales or principles they take to undergird their decisions.

[22] Caveat: I've not yet worked out a satisfactory approach to the "gendering" of the subjects of my thought experiments. In general, I just do my best to randomize their genders, any time references to them via gendered pronouns seems unavoidable, and hope that the resultant gendering of the various parties does not distort students' reactions to the cases. (I'd once considered assigning gender-neutral names to these characters—e.g., Alex, Chris, Kelly, Kim, Pat—but concluded that doing so would serve to "humanize" *these* characters relative to the (otherwise anonymous) persons who co-star with them in their respective thought experiments. It seems best, all things considered, to treat all subjects of all thought-experiments as symmetrically "abstract"—unnamed, and referred to only via the use of pronouns.) In what follows, I adopt the same approach: when convenience dictates the use of pronouns to refer to these hypothetical swimmers, hikers, etc., I will simply alternate pronoun genders. (I prefer this approach to the use of the gender-neutral pronoun "they" because—attractive thought it might be to utilize a non-gendered pronoun—given the nature of the cases, there are numerous occasions when the use of "they" would invite confusion as to whether it's being used in the singular or the plural.)

[23] At this point, I often find it helpful to expand students' deliberative capacities in the following fashion: "If you have difficulty imagining yourself performing this action, or any similarly gruesome actions contemplated by these moral-dilemma thought experiments, you may simply switch from the first person to the third person. That is, you might consider, not what *you* might (be able to bring yourself to) do in this situation, but rather how you would morally evaluate some other identically-situated person who faced this same decision."

inconsistency. For everyone else, however, a puzzle arises at this point—the first in our sequence of four:

> **Numbers Puzzle:** Given that your rationale for saving the Eastern swimmers in Rescue I was that you wanted to save the greater number of lives, why don't you similarly "go with the numbers" in Rescue II? For, numbers-wise, Rescue I and Rescue II appear symmetrical: it's five alive, one dead, in each case. So why the switch?

Students do not have much difficulty resolving the Numbers Puzzle, however: with little or no prompting, they are able to articulate what philosophers typically call the "Killing/Letting Die Distinction" (KLD). Though students will not typically include the *ceteris paribus* clause, they will offer something like this formulation: "All else equal, *killing* someone is morally worse than (merely) *letting someone die*."[24] Thus, expressed in terms of our tripartite template, and formulated in such fashion as to incorporate the resolution to the Numbers Puzzle, our encounter with Rescue II can be expressed as follows: the dilemma is "do we or do we not kill the one hiker, in order to save the five drowning swimmers?"; the decision is (overwhelmingly, but not uniformly) to refrain from killing (even if this means letting five die); and the rationale offered (on behalf of the majority) is KLD. (The vocal minority, of course, continues to offer "the numbers!" as their rationale for driving over the hiker in order to save the five swimmers.)

A.2.1 Interlude: "Solving for *X*" (or: "Absolute vs. Threshold-Sensitive Constraints"):

I've found it useful to pause at this point to gauge students' intuitions regarding *Absolutist* versus *Thresholdist* interpretations of KLD (and by proxy, perhaps, of such "deontological constraints" more generally). I do this by asking how many of the "proto-deontologists" (those who invoked KLD as their rationale for not killing Rescue II's hiker) would similarly demur at the prospect of allowing *six* swimmers to die, or *ten*, or *twenty*, or *one hundred*, or … rather than to drive over the one hiker. I sometimes make this polling more systematic by asking them to "solve for *X*" in the following formulation: "I would be willing to drive over (and kill) one hiker in order to save *X* drowning swimmers."[25] I ask all the deontologists (I have not yet introduced this term to them) to raise their hands, and then to put down their hands once *X* is large enough that they are now prepared to kill the hiker: "*ten* swimmers? … twenty? … *ninety?* …", etc.) I usually find that most hands are down once we're much past the twenties; the remaining hands can be quickly adduced ("*one-hundred* drowning swimmers? *A thousand?!*") as belonging to Absolutists about the KLD principle. According to such Absolutists, there is no threshold above which the badness of allowing *X* persons to drown outweighs the badness of killing one hiker.

At this point, the Thresholdists' incredulity—aimed at the Absolutists' refusal to save even a *million* lives at the cost of killing one person—is palpable. But I try to disabuse the

[24] This is, of course, a special application of the familiar "doing/allowing" distinction, widely discussed in the literature. (See, e.g., the papers collected in Part III of Fischer and Ravizza [1992].)

[25] To elicit their full participation, I often have to assure them that there are contrivances by which we might ensure that we can intelligibly contemplate the saving of *X* persons, for progressively larger values of *X*. For instance: perhaps we need to drive over the hiker in order to reach our speedboat, which we will then use to motor out to a capsizing ship of panicking party-goers, which we alone have the skill and calm presence of mind to right.

Thresholdists of their belief that the Absolutists' is the only incredible position in the room. Soliciting one Thresholdist volunteer's position for special examination, I'll point out to him or her (and thus the entire class) that it appears no less reasonable to suppose that there's a firm line between saving, say, *seventeen* lives and saving *eighteen* lives. Nor does it seem any more reasonable to assert (as some students at this point will) that there's perhaps no *exact* threshold, but that nevertheless there's a range within which the moral permissibility changes: that "it's definitely somewhere between five and twenty," for example. (However, I do let these students off the hook to some degree, by pointing out that these sorts of "sorites challenges," while admittedly puzzling, are nevertheless ubiquitous. Clearly there is a distinction between *daytime* and *nighttime*, even if there is no precise moment at which daytime ends and nighttime begins. I console Thresholdists that it is no more or less puzzling to affirm that, likewise, there are values of X that are too small to justify killing one person to save X lives, and there are values of X that are large enough to justify killing one person to save X lives, even if there is no precise delineation between the two. But—and this is a key point I wish to convey—there *is* something genuinely puzzling about the sorites-like nature of this distinction. Furthermore, this puzzle at very least counsels some humility in one's consideration of rival (and putatively "unreasonable") views, such as those of the Absolutist, or the strict (as I like to call them at this point) "Numbers Theorist.")

It's also worth "solving for X" in the opposite direction too. That is: it is not safe to assume that everyone who's willing to drive over Rescue II's hiker in order to save five drowning swimmers *is* a strict Numbers Theorist. Some of them may actually be Thresholdists for whom the Threshold at which it's worth killing one to prevent X drownings is actually *lower* than five. To investigate this possibility, I ask the handful of Numbers Theorists to raise their hands, and then to put down their hand once X becomes *low* enough that they're not willing to kill one to prevent X drownings: "four swimmers? three? *two*? *ONE*?!" Even the most devoted Numbers Theorists, though, admit that they are not indifferent between *killing one* and *letting one die*—a fact betokened by their reluctance to kill one to save one from dying. What this shows, I point out to the class, is that seemingly *everybody* accepts the Killing/Letting Die distinction—it's just that we all have different thresholds. (Excepting, perhaps, the Absolutists, who do not admit the existence of any threshold.) For most Numbers Theorists, though, the threshold is exceeded already by the time X reaches 2.

A.3 Antidote: Returning now to our sequence of thought experiments: for the next case, it shan't be necessary to elicit from the classroom a volunteer who can draw upon his or her previous employment experience to role-play as the drama's protagonist—and for an important variant of this case, which I will also introduce, it shan't be *possible*. (Unless you happen to have among your students any former medical professionals taking an introductory philosophy class purely for purposes of edification.) In the former instance, this is because the imagined case is (like each of the subsequent cases) even "purer" than the Rescue cases: the protagonist is simply a *bystander*, and serves in no special, (quasi-)professional role (as with the lifeguard in the previous two cases). In the latter instance, this is because the protagonist *does* occupy a special, professional role, which few (if any) of your undergraduate students will have previously occupied—viz., that of a physician or surgeon. As we shall see, this difference between these two cases, vis-à-vis the professional obligations of their protagonists, is a crucial reason why I've opted to modify the latter case (a more widely-known thought experiment, perhaps, but one that poses its own distinctive challenges) into the former case. But so much for general characterization; let's describe the first case, together with its associated analysis, before going on to present the second, more famous case from which it's derived (along with a diagnosis of the complications that have prompted me to modify it). Thus, Antidote:

> **Antidote:** Imagine yourself a resident of a small, remote town (in, say, the upper Yukon Territory), whose lone medical center suddenly finds itself confronting quite a fantastical circumstance. For the hospital currently houses *six* dying patients. In the hospital's west ward lies a lone patient who has ingested a rare poison called *Misology*—one that kills the brain but leaves one's other organs intact. Left untreated, this patient will die within the day. In the hospital's east ward are five dying patients: each in need of a life-saving vital organ transplant, and each needing a different organ. (Specifically: two need lungs, two need kidneys, and one needs a heart.) Furthermore, each of these five patients has the same rare blood condition, such that they require a donor with the same rare condition. Without transplants, each of these patients will die in two days. Essentially, this means that these five patients are each fated for death: even if a donated organ with the correct blood condition were to become available *right now*, in all likelihood it would take longer than 48 hours for it to be properly prepared and transported to this remote medical outpost. However, the westward patient right next door just happens to have the very same rare blood condition; he just so happens to be a registered organ donor too. Furthermore—this being an isolated, remote town—word of this state of affairs quickly gets around. It gets to *you*, at any rate, in time for you to recall a curious conversation you'd had with your grandmother just before she'd died. She'd bequeathed to you the family's stash of an antidote—an antidote you never expected to need, for a poison you'd never heard of it. But once word arrives of the hospital's fantastical situation, you suddenly remember the name your grandmother spoke when describing this poison: "Misology!" Since this years-ago exchange with your late grandmother was so peculiar, and since you were rather young at the time, you'd never told anyone about it. Thus, no one in the village (not even your own family) knows that you are custodian of the town's (indeed, the entire region's) only stash of antidote; further, the fact that you are *uniquely* a possessor of this antidote is known to you. Now fully apprised of the situation, and of your potential role in it, you recognize the choice that faces you. On the one hand, you could stride immediately to the hospital and offer the chief of medicine your supply of the antidote—thereby ensuring that the lone westward patient lives. But you also realize that if you fail to come forth and share some of your antidote stockpile, the westward patient will die, and he will become a veritable organ farm: his organs can be harvested for purposes of performing five life-saving operations. And furthermore, *no one will ever be the wiser*. Your dilemma now, as this bystander citizen, is whether or not to come forward (thereby bringing about the saving of one patient's life), or to keep mum about your antidote (thereby effectively bringing about the saving of five lives). What do you do?

Student responses will likely be mixed here; there may be a significant number of students in both camps. But for any who fall into the "produce the antidote and save the westward patient" camp (call students who share this intuition the "Lifesavers"), a second puzzle arises:

> **Letting-Die Puzzle:** Given that you were willing to let one person die in Rescue I for the sake of saving five other lives—and given that, by withholding your antidote, you don't seem to be *killing* the one poisoned patient (as you plausibly had to kill the one hiker to save five drowning swimmers in Rescue II)—why

> don't you similarly let one die to save five in Antidote? For—both in terms of numbers, and in terms of the KLD distinction—Rescue I and Antidote appear symmetrical. So why the divergent intuitions in these two cases?

Initially, some Lifesavers may deny the symmetry I attribute to the two cases with this formulation of the Letting-Die Puzzle: they will claim that—notwithstanding initial appearances to the contrary—we *are* "killing" the westward patient with our decision to keep mum about our life-saving antidote. Leave these reactions aside for now; we will return to them when we discuss the related case of "Transplant" below.

With perhaps just a bit of light prompting, though, students can in due course easily resolve this puzzle in a different (and likely more satisfying) way: by articulating a principle which, like KLD before it, is able to account for a salient and morally-relevant disanalogy between two cases which initially appear to be analogous. Specifically, students recognize that, unlike the lifeguard in Rescue I, the bystander in Antidote *intends* the westward patient's death. There is a morally relevant difference, many students recognize, between *foreseeing* that one's action (e.g., rushing to the East Beach to save five drowning swimmers there) will allow one person to die, and *intending* that one person die. (Even if that intention is manifested by refraining from a certain action—e.g., refraining from producing a life-saving antidote—such that you are "merely" letting that person die.) And in Antidote, the "keep mum" response is an instance of the latter—even if it's an instance of "letting die," and thus cannot be condemned by invoking the KLD principle. Its condemnation, then, must rest upon something like (what I introduce as) the "Intended No Harm" (INH) principle. When prompted, students have little difficulty formulating such a principle roughly as follows: "All else equal, *intending* harm to someone is impermissible, whether or not that harm results from an action on your part, or an inaction." INH in some ways echoes the language that opens a well-known formulation of the Hippocratic Oath: "First, *Do No Harm* …" My reasons for thus echoing the Hippocratic formulation will become evident from the subsequent Interlude discussing another similar, and similarly well-known, case from the literature—an Interlude that merits inclusion for independent reasons.

<u>A.3.1 Interlude: The "Transplant" Case:</u> Many readers will recognize Antidote as a variant of another famous case from the trolleyological canon: "Transplant." (In Thomson's version of the Transplant case, however, a surgeon contemplates *killing* a healthy patient in order to carve him up and redistribute his organs to five patients who share the same rare blood condition and are each in need of a life-saving transplant (and each in need of a different organ). Accordingly, for our purposes, this version of Transplant is actually structurally more akin to Rescue II.) And indeed, for a number of years, I would present my own variation on the traditional Transplant case at this juncture in the dialectic (a variation that, as we will shortly see, has the surgeon simply *allowing* the patient to swiftly die). However, I found that even my modified case posed its own peculiar challenges. So, over the years, I developed the Antidote case, first as a supplement for, and more recently as an alternative to, my original version of Transplant. Because these challenges are themselves instructive, however—and because the traditional Transplant case (at least as formulated by Thomson) is so widely known—I think it is worthwhile to include here an Interlude discussing this case.

First, a description of Transplant as I had been wont to present it:

> **Transplant:** Similar to the scenario in Antidote, here in the west ward of our remote hospital lies a lone patient suffering from a fatal malady—one that kills the brain but leaves one's other organs intact (a swiftly-growing tumor, perhaps)—whose death can be prevented with a straightforward surgical procedure. If performed, the town's (very fine and accomplished) surgeon has every confidence the patient will live, and will suffer no lasting effects; if not performed, the surgeon is just as confident that the patient will die within the day. Furthermore, this patient is (and is known to the surgeon to be) a hermit: no one (other than the surgeon and one or two (trustworthy) members of the hospital's small staff) knows that she checked into the hospital yesterday, complaining of headaches and dizziness before swiftly lapsing into unconsciousness upon being admitted. You are the surgeon, and you suddenly realize you face a choice: if you decline to operate on the hermit she will die within the day, and no one will ever know you easily could have saved her. (Being a hermit, she never apprised anyone as to her hospitalization, or even as to her condition.) With the hermit's death, then, you'd have a veritable organ farm: you'd be able to carve her up and redistribute her organs to the five dying patients in the east ward, each of whose lives would therefore be saved. What do you do?

Compared to Antidote, students' responses here were less varied. While consensus rarely emerged, there was always a large contingent—often a majority—who favored operating on the hermit and saving her life, even while foreseeing that five other patients would die. That is, with Transplant, the Lifesaver intuition did seem to predominate. (Transplant—as is the case with Antidote—tended also to be the first case in which there was a significant contingent unable (or unwilling?) to endorse one of the options, or at least willing to admit "I don't know!" as their response.) With Transplant, though, I'd found that students' reactions—whichever camp they occupied—were to a considerable extent driven by features of the case that are largely irrelevant for my purposes. For example, many students who affirmed the "keep mum and let the hermit die" intuition explained that the westward patient's status as a *hermit* influenced their reactions: *she* will not be missed if she dies today; her absence may not even be noticed (by anyone beyond a handful of hospital personnel). Whereas presumably each of the five patients in the east ward has families and friends who will miss them should they die, and who will be delighted should they live. Additionally, complications arose with respect to the Lifesavers' intuitions owing to considerations regarding the professional obligations of surgeons and doctors: unlike lifeguards, who stand under no professional (or perhaps even legal) obligations regulating their conduct with regard to their "patients," surgeons and other doctors *have* taken the Hippocratic Oath. This Oath (at least as many students understand it) obliges doctors to take all reasonable available steps to heal, serve the health of, or save the lives of their patients. A plausible and commonly-offered interpretation of this obligation, as it bears on the present circumstances, is that the surgeon ought to do everything possible with the means available to save whatever lives can be saved—which means that the surgeon ought to perform the operation on the westward patient, rather than to create a fund of healthy organs available for life-saving transplants by allowing the hermit to succumb to a lethal ailment. *Not* performing the operation, according to many students, is more aptly described as killing the hermit, rather than merely letting her die. (This, in fact, is the context in which one is most likely to observe the resistance, first noted above, to construing this case as analogous to Rescue I. Insofar as the surgeon is seen as actually *killing* the hermit in Transplant, the case is more analogous to Rescue II.)

After experiencing this dynamic for several semesters, and having diagnosed the foregoing features as the confounding factors, I began to augment the classroom discussion of Transplant by asking students to consider the case anew, once subjected to modifications suitable to ameliorate these complication—*viz.*, the case I eventually came to term (and which I describe above) as Antidote. Using this "sequential strategy"—first considering Transplant, then considering its modified form as Antidote—posed its own challenges, however. Most troublingly, there is reason to suspect that well-known "ordering effects" come into play here. Few students seemed prepared to switch to the "let one die to save five" intuition in Antidote, in circumstances where students considered that case only *after* they'd considered Transplant. It seems plausible to hypothesize that this results from the fact that students' "you've just *got* to save the westward patient" intuition in Antidote is already primed by their prior consideration of Transplant, with its concomitant considerations of professional medical ethics. More recently, then, I have begun to experiment with an approach that dispenses with Transplant altogether, and proceeds immediate from Rescue II to Antidote. Preliminary observations indicate that this tactic *does* result in greater student willingness to let the westward patient die for the sake of saving the five eastward patients—though by no means does it seem that this position now predominates.[26]

Of course, Transplant may also merit discussion in its own right. For one thing, it has an admirable pedigree: it figures prominently in several famous philosophical discussions (as cited above). So for this reason alone, it may be worth presenting to introductory philosophy students, who may be likely to encounter the case later in their studies—perhaps even later in that very semester, if the class in question is a semester-long treatment of normative ethics, which might include, e.g., Foot [1967], Thomson [1976], and/or Thomson [1985] among its assigned readings. Furthermore, the case may prove useful precisely *because* it raises some of the matters I've here characterized as "sideline distraction" issues—the special moral obligations of medical professionals; the phenomenon of "special obligations" or "role ethics" more generally; the relevance of the lone patient's hermitism; etc. And so the posing of these "distractions" may be worthwhile for its own sake. However, it remains the case that, for purposes of the distinctive take on trolleyology-as-first-philosophy developed here, the distractions that accompany Transplant are likely more trouble than they're worth.

* * * * * * * * *

Returning to the main thread, we can formulate our analysis of Antidote in terms of our tripartite template as follows: the dilemma is "do we reveal our stash of antidote, thereby enabling the saving of the westward patient's life, or do we keep mum about it, thereby enabling five life-saving organ transplants?"; the decision is mixed—though there are likely to be a number of Lifesavers; and the rationale these Lifesavers can offer (particularly once confronted with the Letting-Die Puzzle) is "INH." (Contrariwise, students who wish to keep mum can appeal to the numbers and interpret this case on the model of Rescue I, as letting one die to save five.)

[26] There is a further reason for wishing to replace Transplant with Antidote: doing so ensures that INH is far more likely to be invoked, as the solution to the Letting-Die Puzzle, than is the Hippocratic Oath. Why is this important? I want to get INH on the table because I'm planning to put it to work later in our conversation as well. Whereas invoking the Oath would be one-off, a sort of *deus ex machina*—it wouldn't do any subsequent work for us.

A.4 "True" Trolleyology (Finally!): Only at this point in our sequence, I maintain, does it make sense to introduce the traditional Trolley Problem. I start by introducing two standard (and Thomson-ian) variations on Foot's [1967] original puzzle, which I call "Trolley I" and "Trolley II":

> **Trolley I:** Metro service was supposed to stop at Socrates Station today, as there's track work on both spurs just south of this station (the spur heading on out to the west, along the Plato Line, and the spur branching eastwards along the Xenophon line). Both spurs contain tunnels, and there is currently one worker laboring in the west tunnel and five workers laboring in the east tunnel. Now, the trains are *supposed* to stop at Socrates; however, *this* train is a runaway train—no brakes! You are a bystander, fully apprised of the situation, and you just so happen to find yourself next to a lever. By throwing the lever, you can divert the train from its current course (where it's hurtling towards the five workers in the east tunnel, each of whom will die if the train continues to plunge in their direction) over to the other spur—where it will now be on a collision course with the lone worker in the west tunnel. (Being narrow and shallow, these tunnels afford no means of escape. To occupy the same tunnel as a runaway train, then, is to be condemned to certain death.) What do you do?

Many students opt to throw the lever—diverting the train away from the five workers and down the track where it will kill just the lone worker. Typically, this intuition is more widely shared than either of the two intuitions in Antidote (or in Transplant), but less common than either of the prevailing intuitions in Rescues I and II ("rescue the five swimmers" and "*don't* run over the one hiker," respectively). Now for Trolley II:

> **Trolley II:** Like Trolley I, as far as runaway trains go—except now there's no spur: just a runaway train (or tram—the sort of vehicle that the Brits[27] might call a "trolley") heading towards five workers stuck in a tunnel. And now, the lever you're standing next to is connected, not to a switch that will divert the train down a different track, but to a trapdoor on a pedestrian footbridge that crosses over the train track. There is presently a gentleman standing just atop this trapdoor, also watching this scene unfold—and you just so happen to notice that he is … well, *girthsome* enough that his body could stop the train. (Your background in mechanical engineering makes you certain of this. It also leaves you certain that if *you* were to run down and throw yourself onto the tracks, it would avail you nothing: you're too slender to stop the train, so you would be crushed right along with all five workers.) Question: do you throw the lever, dropping the large man from the bridge into the path of the oncoming runaway Trolley, thereby killing him—but stopping the train and saving

[27] Foot, the originator of the trolley problem, was British. Interestingly, though, she was the granddaughter of an American President—a President who even happened to share his name with a (then-)great American city, no less. (Make Cleveland great again!) (I tease because I love: as a native Ohioan, and a one-time resident of the Cleveland metropolitan area, I have nothing but respect for this once- and still-great city.) Edmonds [2014: chapter 3] provides these (and many other) wonderful biographical details and background information about Foot and the other "founding mothers" of trolleyology: G.E.M. Anscombe and Iris Murdoch. I often tell students the tale of trolleyology's founding mothers—a habit many readers might wish to adopt, particularly if they're keen to foreground for students women's important contributions to philosophy.

> the five workers' lives? Or do you refrain from doing so—allowing the five workers to be killed by the train?[28]

Most students opt *not* to throw the lever in this case. (Though the "Numbers Theorists"—by this point a small but reliably consistent contingent—will opt to sacrifice the bystander for the sake of the five workers … just as they sacrificed the lone worker in Trolley I, and the westward patient in Antidote, and the hiker in Rescue II.) In any event, it is often easy to identify a plurality of the students who simultaneously affirm the "throw the lever" reaction to Trolley I and the "don't throw" reaction in Trolley II. But such students must confront the third puzzle:

> **Trolley Puzzle:** Given that you were willing to throw the lever in Trolley I—and given that this action seems tantamount to *killing* the lone worker in the other tunnel—why don't you similarly throw the lever to kill the girthsome man to save the five workers in Trolley II? For—both in terms of numbers, and in terms of the KLD distinction—Trolley I and Trolley II appear symmetrical: you're killing one to save five in either case. So why the divergent intuitions?

Students may initially attempt to resolve this puzzle in the same manner used to resolve the previous one—by invoking INH and arguing that in Trolley II we intend the bystander's death, whereas in Trolley I we merely foresee the one worker's death. Admittedly, this response may work. But for purposes of exploring the conceptual terrain more fully, I push the dialectic along by playing devil's advocate and attempting to convince the students that plausibly, you *don't* intend the girthsome man's death in Trolley II. Instead, you merely intend for his girth to intercept the trolley's momentum, and to bring it to a halt before it reaches (and kills) the five workers. You may foresee that this will almost certainly result in the girthsome man's death. But that is just as certainly not your intent! If, having brought the train to a halt, he were to get up, dust himself off, and walk away, this would be just fine with you; in fact, it would be your preferred outcome. But no such counterfactual obtains in Antidote—the case where, in arguable contrast to Trolley II, it *is* unavoidably clear that you *do* intend a death. (This, by the way, is an opportune moment to introduce your students to the notion of a *counterfactual*.) There, if you had had reason to believe that the westward patient might very well live, even without the benefit of your antidote, you'd actually have *less* reason to withhold your antidote. But not so in Trolley II: to the extent that you have reason to believe that the bystander might survive his encounter with the runaway train, you would have *additional* reason

[28] Thomson [1985] is the original source for the "girthsome-man-on-a-bridge" version of Trolley. However, her original case had the agent at the center of the drama *pushing* the bystander off of the bridge and onto the track below. As has been subsequently widely recognized, this "hands-on" aspect of Thomson's scenario poses unnecessarily confounding factors. (For a summary, see Greene [2013] chapter 9.) It also feels a tad … *politically incorrect*. Subsequent commentators have therefore sought to eliminate this political incorrectness (and, perhaps, sought to pre-empt the possibility that "size-ism" may be distorting people's intuitions here) by replacing the famed "fat man" with various contrivances: artifices involving, e.g., *fat suits* or *large backpacks*. The most elegant solution, though, belongs to Parfit [2011: 218], who simply imagines the trap door to be underneath an individual whose size is left undescribed; all we know is that his or her collision with the train would trigger its automatic brake. However, owing to the fat man's outsized role in popular conceptions of trolley-based thought experiments (*vide* the titles of Cathcart [2013] and Edmonds [2014]), I will here employ this version, notwithstanding my mild discomfort.

to throw the lever. Or so, at any rate, I argue as devil's advocate: it's at least not obvious that an appeal to INH is able to resolve the Trolley Puzzle.[29]

But there's another asymmetry between Trolley I and Trolley II: even if one might plausibly maintain that one doesn't intend the bystander's death in Trolley II, one is still using him as a *means* of saving the five workers. (In the same fashion, you (literally!) used the westward patient as a means to your end of saving the five patients, by carving him up and redistributing his organs.) His girth, his person—his very being—plays a necessary causal role in your plan to save the five workers, and accordingly you are treating him as a tool, a mere means. Call the principle (familiar to anyone acquainted with the relevant formulation of Kant's Categorical Imperative) the "Impermissibility of Using Persons as Mere Means to Your Ends" principle, or (not quite as clunkily) "M/E" for short.

If this appeal is successful, it reveals that the Lifesaver intuition in Antidote is overdetermined: it can be supported, not only by invoking the INH principle (as we saw above), but also by invoking M/E. And indeed, you will likely have had students who also invoked something like M/E during the initial discussion of Antidote. I will freely admit in this forum (though I do not admit it in the classroom) that, during our discussion of Antidote, I usually employ a bit of professorial subterfuge in order to preserve the desired dialectical shape of my trolleyology. I'll do this either by deliberately minimizing students' (sometimes latent) appeals to M/E during discussion of Antidote, and/or by deliberately conflating considerations of *means/ends* with considerations of *intent* (when summing up students' reactions to Antidote, or drawing out the common themes expressed during their discussion), and then concluding that the students identified only INH in their exploration of possible rationales for saving the one life in Antidote. This may strike readers as slightly manipulative. I have some sympathy for this worry; nevertheless, I think this tactic is warranted on pedagogical grounds—it is worth "holding M/E in reserve" until such time as it can make a sudden, dramatic appearance as the solution to the Trolley Puzzle.

Thus, expressed in terms of our tripartite template, we might characterize the majority's reaction to Trolley I as follows: the dilemma is whether or not to throw the lever and divert the runaway train away from its present course (where it will kill five workers) and down an alternate track (where it will kill only one worker). The decision (for the substantial majority of students) is to throw the lever, thereby saving the five but killing the one. And the most common rationale offered—a rationale made necessary by consideration of the Trolley Puzzle—is that, unlike our use of the girthsome man in Trolley II, our decision to deflect the train toward the lone worker in Trolley I does not in any objectionable sense *use* that one worker as a mere means to an end. (It does not run afoul of the M/E principle, that is.) And for Trolley II, our template yields the following: the dilemma is whether or not to throw the lever, dropping the girthsome man into the path of the oncoming train; the decision is (for all

[29] Pro tip: though students often begin by greeting this devilishly-advocated casuistry with considerable incredulity, I've found there's a way to instantaneously convert much of that into credulity. I do this by reminding them of an opening scene in Pixar's 2004 film *The Incredibles*, wherein Mr. Incredible narrowly forestalls a commuter train derailment by fully absorbing the train's momentum with nothing but his own body. It's almost as if students recognize that, having already suspended disbelief in precisely this regard in a fictional encounter once previously in their lives, they're now obliged to do the same here. That is: the devil's advocacy here strikes them as incredible, until they realize that they'd once credited it in *The Incredibles*—at which point it suddenly becomes credible.

but the Numbers Theorists) to not throw the lever; and the rationale—invoking the same principle used as justification in Trolley I—is that our throwing the lever would be an impermissible *use* of the man on the footbridge (whereas no such using-as-mere-means is contemplated in Trolley I).[30]

Of course, the foregoing tripartite analysis of Trolley I is formulated so as to be suitable for the circumstances at hand: responding to the Trolley Puzzle. But for many students, it cannot, at the end of the day, serve as a wholly adequate analysis. And that's because it is incapable of grounding a (to them) satisfactory response to a fourth puzzle—one they are now about to confront:

> **Killing Puzzle:** Given that you were willing to throw the lever in Trolley I, for the sake of saving the five workers, why aren't you similarly willing to drive over the hiker in Rescue II, for the sake of saving the five drowning swimmers? Conversely: given your unwillingness to kill one to save five in Rescue II, whence your willingness to do just this in Trolley I? In short: why these divergent intuitions?

This is the hardest puzzle of the bunch! Whereas the first three admit of fairly straightforward solution via the invocation of basic distinctions and principles (KLD, INH, and M/E, respectively), this one offers no clear resolution. The very wording of the Puzzle ("… given your unwillingness to kill one to save five in Rescue II, whence your willingness to do just this in Trolley I?") is designed to rule an appeal to KLD out of court. Students will therefore often begin by trying to invoke INH, arguing that in running over the hiker in Rescue II, you intend her death, and that adherence to INH therefore precludes your saving the five drowning swimmers in this case. Initially, this seems plausible, but considering the parallels with Trolley I may serve to undermine our confidence in this strategy. After all, if we can seriously maintain that we don't intend the lone worker's death in Trolley I (even though we foresee it), can't we maintain with equal justification that we don't intend the hiker's death? Consider the fact that, were driving over the tree limb beneath which she's pinned to have no lethal effects, we would certainly still do it. And even if we did anticipate lethal effects: should it turn out she survived being driven over by a Land Rover unscathed, we would be *delighted*—not disappointed! (Compare this counterfactual to the relevantly similar counterfactuals in Antidote: not only would we not welcome the "non-lethal outcome" for the westward patient, vis-à-vis our decision to keep mum and hold onto our stash of antidote, but such an outcome would utterly thwart our intentions.) In comparison with our paradigm cases of *intended* and *unintended-but-foreseen* deaths (Antidote and Trolley I, respectively), Rescue II seems more similar to the latter than to the former.

[30] One note on the tripartite "dilemma/decision/rationale" template: I do not introduce this three-part analytic template on the first pass. On the first day of trolleyology (which, if you're doing it correctly, is probably also the first day of the term; I prefer to introduce at least the first few thought experiments even before I've gone over the syllabus), I just present the cases in sequence, and the puzzles as they arise. It is only on day two, when we are "reconstructing" the previous class period's discussion, that I introduce the tripartite framework as a way of rigorously capturing and comparing our reactions. (My "day two" happens to fall on our third meeting, however: after having a spontaneous, only-lightly-structured discussion of the various moral-dilemma thought experiments and puzzles on the first day of term, I reserve the second meeting for the typical "day one" mechanics: reviewing the syllabus and course requirements, participant introductions, and the like.)

Students may then try to invoke the M/E principle to ground their diverging judgments in Rescue II and Trolley I. They may argue that the prohibited action in Rescue II requires impermissibly using the hiker as a means to your end, whereas nothing of the sort seems to be going on in Trolley I. This seems plausible initially. But once again, comparison with one of our paradigm cases here—namely, Trolley II—casts some doubt on this analysis. The contrast with Trolley II, where it seems utterly evident that the bystander is being (putatively impermissibly) treated as a mere means to our end, seems palpable. For it is far from "utterly evident" that we are similarly treating the hiker. After all, if she were not pinned under that tree limb, we would still drive over it to get to the beach to save the five drowning swimmers. Likewise, if there were any alternative means of reaching the beach whatsoever (provided it still got us there in time to save the five swimmers), we would take that route rather than run over the hiker. Consideration of these (and other similar) relevant counterfactuals seems to reveal that our (certainly lethal) interaction with the hiker is in no wise a part of our life-saving action, and that she is not serving as a means to our end of saving the five drowning swimmers. Contrariwise, there are seemingly no relevantly similar counterfactuals in Trolley II where we wind up saving the five workers yet sparing the life (or at least the bodily integrity) of the bystander. (For example: it's not like we can say that we would have thrown the lever "even if there were no one standing on that trap door," or "even if there were a slender-er person standing atop the trap door," etc.[31])

The best that can be done to resolve the Killing Puzzle here, I believe, is to assign complete parity to the two actions constituting the horns of the dilemma in Trolley I; that is, to resist the initial characterization in terms of "killing the lone worker in the tunnel down the spur, vs. doing nothing and letting the five workers on the main track die." As students are wont to express the matter here: "Either way you're making a choice, you're taking an action; you can't evade responsibility for *one* of the outcomes, but not the other, on grounds that not throwing the lever is 'mere inaction,' a 'declining to intervene.' Either decision is on a par; at the end of the day, you'll have acted either to save five, or to save one." Though students won't typically express it this way, one might also borrow the terms Thomson uses in her [1976: 76], and speak of there being a pre-existing lethal agent, which in some sense is threatening all six workers equally (that lever could well have been in the other position, in which case the "default" scenario would have been that the lone worker down the spur stood in the path of the runaway train), and which the agent in question is merely "redistributing." Seen under that guise, by throwing the lever you are merely deflecting that pre-existing threat to where it will do the least harm. In that fashion, perhaps, one can draw a principled distinction between Trolley I and Rescue II, such that the killing/letting-die distinction does not apply to Trolley I, and once again the only relevant considerations become maximizing benefit and minimizing harm. As mentioned in section 1: construed in this fashion, the case comes to look almost parallel to the one with which we started, Rescue I. Here, we're faced with the unhappy fact that we must either kill one, or kill five, leaving "the numbers" as once again the only relevant dimension of moral analysis.[32]

[31] There are, of course, "counterfactual scenarios" in which we save the five without killing the one, but they appear distant enough that they no longer seem to be counterfactual versions *of* Trolley II. Trolley I, for example, is one such counterfactual scenario that seems to inhabit too distant a possible world to remain a relevant comparison.

[32] It's worth noting at this juncture that, somewhere in your discussion of the Trolley scenarios, you'll likely have some students tempted by an appeal to *fate* as a reason for not throwing the lever in Trolley I. "Sure, that lever 'could very well have been in the other position,' such that the runaway train was hurtling towards the one rather

But I digress: the point here is not to make advances in extant trolleyological theory; the point is merely to sketch one distinct and useful version of trolleyological pedagogy, and resolving the Killing Puzzle is unnecessary for our purposes. If anything—and as suggested in section 4.3 above—there is perhaps pedagogical *value* in leaving the Killing Puzzle tantalizingly unresolved. For doing so may serve as enticement to the very sorts of further philosophizing described in sections 2-4 of this essay.

than the five. *But it wasn't*—it was positioned such that the train was headed towards the five. Thus, it would appear that it's the *fate* of the five to die, and of the one to live—and who am I to interfere with fate??" If their classmates don't spontaneously do so first, you should of course disabuse such students of the misconception that such appeals to fate could possibly prove dispositive. You might do so gently, starting off with an acknowledgement that such appeals to fate are powerful—though in fact, they're *too* powerful. "For it seems that one could argue with equal force that it was the *bystander's* 'fate' to be standing next to the lever, and to be thus positioned to save five lives." Such appeals are unhelpful, precisely because they can cut both ways. (You might further point out that the students' initial fate-ful appeal also seemingly proscribes most *any* form of intervention in the world—seeking medical treatment for life-threatening injuries, say.)

References

Appiah, Anthony Kwame [2008]. *Experiments in Ethics.* Cambridge, MA: Harvard University Press.

Baltzly, Vaughn Bryan [2020]. "The Interpersonal Variability of Gustatory Sensations and the Prospects of an Alimentary Aesthetics." *intervalla: platform for intellectual exchange* volume 7. https://www.fus.edu/intervalla/volume-7-questions-of-taste.

Cathcart, Thomas [2013]. *The Trolley Problem, or Would You Throw the Fat Guy Off the Bridge?: A Philosophical Conundrum.* New York: Workman.

Costa, Michael J. [1986]. "The Trolley Problem Revisited." *Southern Journal of Philosophy* 24(4): 437-49.
______ [1987]. "Another Trip on the Trolley." *Southern Journal of Philosophy* 25(4): 461-66.

Edmonds, Dave [2014]. *Would You Kill the Fat Man?* Princeton, NJ: Princeton University Press.

Fischer, John Martin, and Mark Ravizza [1992]. *Ethics: Problems & Principles.* Fort Worth: Harcourt Brace Jovanovich.

Foot, Philippa [1967]. "The Problem of Abortion and the Doctrine of Double Effect." *Oxford Review* 5. Reprinted in Fischer and Ravizza [1992]: 60-7

_______ [2002]. "Killing and Letting Die." Chapter 6 of her *Moral Dilemmas: and Other Topics in Moral Philosophy.* Oxford University Press. (Originally published in Garfield, J. (ed.), *Abortion: Moral and Legal Perspectives,* University of Massachusetts Press, 1985.)

Greene, Joshua [2013]. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them.* New York: Penguin.

Kamm, Frances [2015]. *The Trolley Problem Mysteries.* New York: Oxford University Press.

Martena, Laura [2018]. "Thinking Inside the Box: Concerns about Trolley Problems in the Ethics Classroom." *Teaching Philosophy* 41(4): 381-406.

Parfit, Derek [2011]. *On What Matters.* New York: Oxford University Press.

Rachels, James [1975]. "Active and Passive Euthanasia." *The New England Journal of Medicine* 292: 78-80.

Sober, Elliot [2013]. *Core Questions in Philosophy: A Text with Readings* (6th Edition). Boston: Pearson.

Taurek, John M. [1977]. "Should the Numbers Count?" *Philosophy and Public Affairs* 6(4): 293-316.

Thomson, Judith Jarvis [1976]. "Killing, Letting Die, and the Trolley Problem." *The Monist* 59(2): 204-17. Reprinted in Fischer and Ravizza [1992]: 69-77

________ [1985]. "The Trolley Problem." *The Yale Law Journal* 94(6): 1395-1415. Reprinted in *Fischer and Ravizza* [1992]: 280-92.