Silicon at the Switch:
A Computational Moral Grammar for Solving Trolley Problems

Luke Swartz

Abstract

A popular set of moral dilemmas in philosophical literature are "trolley problems"—choices between killing (or letting die) different groups of people, named after the most famous example, involving a runaway trolley. Dr. John Mikhail has suggested a generative moral grammar for solving such problems; this paper critically examines this grammar, especially through a partial implementation of that grammar in Prolog. Extensions, amendments, and alternatives to both the grammar and its computational implementation are discussed.

Overview of the Trolley Problem

Philippa Foot first suggested the following moral dilemma in a 1967 essay, "The Problem of Abortion and the Doctrine of the Double Effect":

> [T]he driver of a runaway tram…can only steer from one narrow track on to another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed (p. 270).

The article, and this problem in particular, spurred Judith Jarvis Thomson to write "Killing, Letting Die, and the Trolley Problem," in which she elaborated upon Foot's earlier example and contrasted it with a seemingly similar scenario:

> David is a great transplant surgeon. Five of his patients need new parts—one needs a heart, the others need, respectively, liver, stomach, spleen, and spinal cord—but all are of the same, relatively rare, blood-type. By chance, David learns of a healthy specimen with that very blood-type. David can take the healthy specimen's parts, killing him, and install them in his patients, saving them. Or he can refrain from taking the healthy specimen's parts, letting his patients die (p. 80).

Most people's moral intuition is that in the first case (which we shall call, in keeping with the literature, "Driver") it is permissible for the driver to steer towards the one man, while in the

second case ("Transplant"), it is not permissible for David to harvest the healthy specimen's organs. The problem is that, on strict consequentialist grounds, it is impossible to arrive at the appropriate moral intuition: in each case, the results are one dead and five saved. What, then, separates these cases?

A vast literature has arisen to answer just that question, with widely varying solutions. While the situations may seem artificial at first, there are some cases of similar moral dilemmas in the real world. For example, in World War II, both the English and Germans either considered or implemented a policy of directing enemy bombers toward smaller villages in order to spare large cities (Clark, 1995, p. 198 endnote 1). Admittedly, some cases in the literature are so convoluted that they border on absurd (thus prompting parodies, e.g. Patton, 1988). Nevertheless, "trolley problems" (as we shall refer to this class of dilemma) illuminate interesting distinctions in the ways people cognitively process moral situations.

Mikhail's Grammar Solution to Moral Intuition

John Mikhail recently posited a unique solution to trolley problems, drawing upon an analogy of John Rawl's between permissible acts and grammatical sentences, suggesting that a "moral grammar" could be developed that distinguishes between permissible and non-permissible acts. That is, just as native speakers of a language can instinctively tell whether a sentence is grammatical or not based on some mental linguistic grammar, perhaps our intuitions about moral problems such as these are based on a mental moral grammar. Before we analyze Mikhail's solution in detail, we should make clear the purpose of such a grammar.

Mikhail's analysis is based on empirical data from psychological experiments in which participants were given various scenarios and asked whether the actors' action is morally permissible or not (Mihail, Sorrentino and Spelke, 1998; Mihail and Sorrentino 1999). In particular, seven variations on the classic trolley problem were presented, such as the following:

> Hank is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Hank sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Fortunately, Hank is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the men. Unfortunately, there is a man standing on the side track with his back turned. Hank can throw the switch, killing him; or he can refrain from doing this, letting the five die.
>
> Is it morally permissible for Hank to throw the switch? Y or N (Mikhail, 2000, p. 125).

This situation is closest to a situation initially proposed by Thomson, called Bystander (as in this case Hank is a bystander on the track as opposed to the driver of the trolley as in Driver). The other six cases subtly change the parameters for characters named Ian, Karl, Luke, Ned, Oscar, and Paul (see Mikhail, 2000, p. 125-7). In these experiments, participants consistently rated Hank, Luke, and Oscar's actions as permissible, Ian, Karl, and Ned's actions as forbidden; and Paul's action as obligatory (p. 128). Furthermore, these intuitions were spontaneous (unintentional and nonvoluntary), immediate (made quickly), stable (not changing if presented at a later time), and stringent (participants could not be easily convinced to think otherwise) (p. 101).

Our aim, then, will be to describe what cognitive processes might result in these decisions. Note that this aim makes the unfamiliarity of the various Trolley Problem examples unimportant; the very fact that people consistently make the same permissibility judgments about the cases is enough—just as people's ability to judge the grammaticality of unfamiliar sentences still explains syntactic intuitions, despite their unfamiliarity (Mikhail, pp. 116-8). In this paper and in Mikhail's analysis, we are *not* interested in an *explanatory* account of moral intuition (for example, investigating whether these intuitions are a result of nature or nurture), nor in a *prescriptive* moral theory (that is, critiquing these moral intuitions). It is another question entirely as to whether or not our moral intuitions are *right*; assuming that they are would result in a prescriptive conventionalism (Sheng, 1995, p. 213). Furthermore, while people's explanations of their moral intuitions may be interesting, this introspection can't adequately solve the descriptive problem, because people's explanations of their moral judgements were inconsistent from one participant to another, despite the similarity between the judgements *themselves*

(Mikhail, 2000, p. 102). Finally, we shall not seek to apply these descriptive intuitions to contentious contemporary ethical issues (such as abortion or euthanasia), which have so often motivated philosophical debate on trolley problems.

A Computational Moral Grammar

How, then, could one construct a moral grammar to describe most people's intuitions about these trolley problems? Mikhail suggests that such a grammar would involve a series of postulates regarding causation, good and bad effects, and moral laws. It should first be noted that his notation takes into account the idea that events are ordered in time and that they depend on the circumstances; in this brief summary, we shall assume these features for brevity and clarity.

In regards to causation, Mikhail posits two kinds of morally relevant causation: K-generation and I-generation. K-generation refers to an actor knowingly committing some action *V* that has an effect *W*; I-generation is K-generation but with intention—that is, *V* I-generates *W* if and only if *W* is either the actor's goal or a means to getting that goal (that is, that *W* I-generates the goal) (p. 154). Likewise, he posits four Features of the Residual Permission Principle (the idea that what is not forbidden is permissible and vice-versa): F_1 states that if action *V* I-generates homicide, then *V*-ing is forbidden (that is, intentional homicide is forbidden). F_2 states that if action *V* I-generates battery, then *V*-ing is forbidden (intentional battery is forbidden). F_3 states that if, in short, if action *V* results in prevention of a bad effect, with no other bad (side) effects of its own, then it is obligatory (note that under Mikhail's grammar, the only bad effect is death). F_4 is a formulation of the Doctrine of Double Effect (henceforth, DDE), which we shall examine in some depth later; essentially it allows one to K-generate a bad effect so long as one does not I-generate that bad effect. Together with Conversion Rules for converting a situation into Action Plans, these rules can generate an algorithm for each trolley problem, allowing one to determine permissibility.

Mikhail's hypothesis is interesting and his grammar is convincing, but is it possible to make such a grammar computational—that is, could one model a subset of this grammar on a computer? An implementation of just that has been achieved in the Prolog language (see attached code). This program takes as its input a list of causational facts about the situation; that is, the output of the first few stages of Mikhail's Conversion Rules theory (p. 134). That is, a human being must first linguistically process the description of a scenario, evaluate its temporal structure, and posit a causal structure on those temporal items. Second, the program requires as its input the goal of the user's action; this is "cheating" to some degree (as one is not told the goal of the actors in each situation), but this will be justified later.

The program begins with some recursive notions of "causes" and "i-causes," which roughly correspond to Mikhail's K- and I-generation (here we assume that each actor knows the effects of all his or her actions). This is combined with a notion of prevention ("prevents"), which can be considered (roughly) the opposite of causes; in particular, if event *X* prevents *Y*, and *Y* would have caused *Z*, then *X* prevents *Z*. Note that there are also more basic notions of "directly_causes" and "directly_prevents," which are introduced merely to have a "base case" for the recursion; as we shall examine later, "no one agrees about how to *count* events" and thus the adverb "directly" should be taken with a very large grain of salt (Costa, 1986, p. 441).

Permissibility is determined based on the assumption that all actions are permissible unless deemed forbidden. Thus, in keeping with Mikhail's F_1 and F_2, all intentionally-caused "bad" effects are forbidden. ("Bad" is defined as killing or hitting any number of people.) However, this does not solve the case of Karl, or, as I like to call it, the Trolley Problem in Reverse: Karl is faced with a train headed directly for one person, but can switch it to a side track with five people. This should be judged as impermissible, but with only a stricture against intentional battery and homicide, it is permissible (since he would be turning the trolley with the intention of saving the one, not killing the five). Thus, there must be some notion of utilitarian consequences also embedded in the theory—that is, while examining the consequences isn't *sufficient* to determine permissibility, it is *necessary* for permissibility that the good effect outweigh the bad.

This is defined in the program in terms of the function "worse," which assumes that killing or hitting (as an effect or consequence) is worse if done to more people (thus assuming that all lives are worth the same), and that killing any number of people is worse than hitting any other number of people (that is, that homicide is worse than battery, regardless of how many people are battered). Clearly, in the real world these assumptions do not hold: we may value someone because of their importance (say, if he/she is a doctor about to save people in a hospital), youth (e.g. saving a child over a octegenarian), health, etc.; likewise, it seems that at least in the extreme cases battery can be worse than homicide (say, if one had the choice between paralyzing 1,000 people or killing one).

Another addition to the theory has to do with another variation of Bystander, where the agent has the choice between letting the train hit one person or switching the train to another track, hitting another person. In this scenario, which I dub One on Each Track, my intuition is that it seems less permissible to throw the switch. This is essentially a question about the distinction between killing and letting die, which has been much discussed in the literature. Michael Tooley famously contended that there is no difference whatsoever between killing and letting die, suggesting an example in which two children, John and Mary, are in a room; if one pushes a button, John will die, while otherwise Mary will die (1994, p. 104).

Heidi Malm argues strongly that there is a difference between killing John and letting Mary die; the acts would be morally equivalent only if "an agent is at liberty to change who lives and who dies when other things are equal" (1989, p. 245). She suggests that if one has to justify either pressing the button or not, it seems much easier to justify inaction than action (p. 246). Thus, Malm argues that "it is morally impermissible to substitute one person's death for another's (change who lives and who dies) without a good reason for doing so" (p. 248). Likewise, Richard Trammel asserts that "in some cases we are under greater obligation to avoid taking a life than to save a life, even though effort and motivation are constants" (1975, p. 291). Motivated by this dialog and my own personal intuitions (ideally these intuitions would be tested in a controlled experiment with other people), the computational grammar incorporates a

preference for inaction in cases where the effects are equally bad (i.e. the number of people killed or hit is equal). Put another way, all things being equal, killing is (perceived as) worse than letting die.

Finally, both Mikhail's grammar and my implementation of it have a strict deontology—that is, things are only defined as permissible and not-permissible (and obligatory). There is no accounting for *degrees* of permissibility, largely because the psychological experiments merely asked whether an action was permissible or not—it did not ask how strongly the subjects felt either way. Likewise, there is no notion of what *should* be done by a morally upright actor (that is, supererogatory actions), which largely differs from merely permissible actions when the actor has to make a sacrifice for another.

Possible Extensions/Improvements

Obligation

One particularly weak aspect of the current program is that it defines "obligatory" merely as actions which prevent a bad effect but which have no bad effects of their own. While this is similar to F_3, ideally an action *X* should be defined as obligatory when *failing* to do *X* is not permissible—thus keeping permissibility as a basic deontological construct (see Mikhail, 2000, pp. 141-3).

However, this introduces a number of problems. First, it assumes that the actor in question is the only actor in the given situation: for if there are two actors who can, say, throw the lever, one's failing to pull the lever will not necessarily entail that the lever will not be pulled. Second, it introduces questions of special obligation, such as the distinction between a child dying of starvation because its parents failed to feed it and a homeless person dying of starvation because someone failed to feed him/her: parents have a special obligation to feed their children, while

unless one is a Franciscan one does not have a special obligation to feed the poor. We could make the assumption that there is but one actor who has no special obligation towards anyone, but the problem would still require a second feature—namely, goal determination. If one fails to do action *X*, one needs to know the goal of that decision in order to evaluate its permissibility.

Goal Determination

The current program requires one to explicitly state the goal of the actor; thus, the program cannot determine on its own the goal of any action other than the main one. Nevertheless, this does have one benefit: if one changes the goal to be evil in intent (e.g. saying that Hank throws the lever with the intention of killing the one person—that five people will be saved is of no consequence to him), then, appropriately, the deontic status of the action changes to not-permissible. However, both because human moral cognition requires determining a subject's goals and for utility in deciding obligation, some automatic ability to discern a subject's goal would be useful.

I attempted to detect goals by appealing to the notion of a final-effect—that is, an effect which itself is not the cause of any other effect. In most of the trolley problem cases, the final effects are either killing or saving people's lives, which correspond to the goals. Nevertheless, this is not always the case: for example, there could be a similar situation in which, as a result of saving the five, they would throw a party for the actor. While this is an added benefit to throwing the switch, the desire to have a party in one's honor does not seem an adequate goal to justify killing someone. Thus, one would probably have to traverse each causational tree and determine the effect which produces the greatest good (or, in these trolley problem cases where the only effects are homicide and battery, the effect which is the prevention of the most harm). Assuming—as it seems all the participants in the psychological experiment did—that each actor has the highest aims in mind, one would then assign this good effect as the goal.

A similar improvement would be eliminating the current notion of "prevention" from the input. A more natural action description would describe the options available to the actor, listing the consequences of those actions; one could then note, for example, that Hank's not throwing the switch would cause the death of the five, while Hank's throwing the switch does not; therefore, one could *conclude* algorithmically that throwing the switch prevents the death of the five. However, it would not be clear exactly *how* throwing the switch prevents the death of the five—which is a much more difficult problem, and essential to determining the permissibility of Hank's action.

Conflating Terms

One final difficulty is that the program does not let users make a distinction between terms such as killing and letting die in the input. Likewise, "hitting" refers both personal battery as well as by a distance (i.e it doesn't matter whether it's the trolley or the actor doing the hitting). This makes for input that is often forced an unnatural, as we don't always place every kind of death under the label "killing."

Perhaps it would have been better to stick with "murder" and "battery," which at least have technical legal meanings (and which are defined rather precisely by Mikhail); nevertheless, these words are charged with so much meaning that it seemed counter-productive. Ideally, the program would accept different names for the same idea, and translate ("lexicalize," in Mikhail's terminology) those terms into more technical notions of murder and battery.

<u>On the Doctrine of Double Effect</u>

The careful reader may note that while F_1, F_2, and F_3 each have analogues in the computational grammar, F_4 does not. That is, I don't actually implement the explicit Doctrine of Double Effect

in my program. The program merely assumes that anything that is not forbidden is permissible (in keeping with the Residual Permission Principle), and since nonintentional murder/battery are not defined as forbidden, then they are permissible, given that the consequences are better from a utilitarian standpoint and given a preference for inaction if all other things are equal.

Thus, an explicit DDE does not seem necessary to describe the trolley problems; by focusing on the intentionality of the murder or homicide, in a way the DDE is assumed. However, F_4 is also omitted because it doesn't seem to jibe with Mikhail's description of the DDE—that is, his prose description of it makes sense, but his symbolic formalism of it doesn't seem to jibe with that prose.

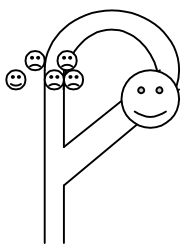
Under Mikhail's interpretation, the DDE can explain how otherwise impermissible actions are permissible, but it is "not itself a test of whether an action is right or wrong" (p. 162). It "states the necessary conditions that must hold for a presumptively wrong action to be justified" (p. 162). He further argues that legal notions of battery and murder (and the prohibition against intentionally committing such acts) explain deontic status on their own, as his view of the DDE *cannot* explain why an act is impermissible (p. 162-3). As one philosopher describs it, the DDE only says that "*sometimes* the end justifies the means" (Costa, 1986, p. 447). Nevertheless, in (my reading of) F_4, he states that if an actor K-generates battery while I-generating a good effect, at the same time K-generating a bad effect which is the same as the good effect, that action is forbidden. I must admit that I'm not sure what the description means, but it seems that stating that an action is forbidden is tantamount to "a test of whether an action is right or wrong."

In any case, the DDE has come under attack as a solution to trolley problems, because it (at least seems to) suggest the wrong deontic status for a variety of moral dilemmas:

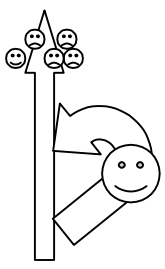
In the most famous case ("Gas Leak"), provided by Foot in the essay that spurred this entire discourse, there are "five patients in a hospital whose lives could be saved by the manufacture of a certain gas, but that this inevitably releases lethal fumes into the room of another patient whom

for some reason we are unable to move" (1967, p. 276). She suggests that the relatives of the gassed patient could successfully sue the hospital, and that manufacturing the gas seems intuitively non-permissible. However, the DDE would suggest that, since the killing of the one is not intended (rather, it is a regrettable foreseen consequence of producing the gas), the action is permissible.

Another case, presented first by Thomson, is the "Loop Variant" of the classic trolley problem: the track straight ahead has five workmen, and the track on the right has one (very fat) workman, but the tracks connect together such that if one steers the train straight, it will hit first the five men and then the one, but if it hits the one first he will stop the train (1986b, p. 102). She argues that this is intuitively permissible, but impermissible under the DDE (as one would intend the trolley to hit the fat worker as a means to save the five); nevertheless, it seems clear that on utilitarian grounds one could discard this example—the choice is, after all, between killing all the workers or one of them. One could conceive, however, of an alternate example, where the loop only goes one way—that is, it is a clear choice between using the fat man as a trolley stop or letting the five die (see figure). This is similar to the situation described for Ned in Mikhail's study, which was judged as impermissible (2000, p. 126). Thomson and others, however, see switching the trolley as permissible (Clark, 1995, pp. 191-2; Kamm, 1991, p. 573-4; Hallborg, 1997, pp. 299-300). Indeed, both Hallborg and Thomson admit some reservations about classifying the turning of the trolley as permissible; "Some people feel more discomfort at the idea of turning the trolley in the loop variant than in the original…" (Thomson, 1986b, p. 102). I am inclined to agree with the Mikhail study, a much more controlled and scientific assessment than a philosopher probing a handful of colleagues; nevertheless, it seems that this case is at least controversial in regards to its intuitive deontic status.



The Loop Variant



Alternative (Ned)

Likewise, another counterexample seems to be that "one may not drive over someone standing on the road in one's attempt to get to a hospital with five dying patients, though one only foresees the death one causes and does not intend it" (Kamm, 1991, p. 573). This situation, dubbed "Rescue II" by Foot (1984, p. 283), could admittedly be interpreted so as to coincide with the DDE, by saying that one does intend to run over the one as a means of getting to the hospital, but this seems to blur the distinction between intended and forseen consequences—after all, if the one were not in one's way, one can still achieve one's overall goal of saving the five.

It seems, then, that while the DDE is a useful theory for explaining aspects of trolley problems, it falls apart in certain cases. It would be worthwhile to consider additional or alternative explanations for moral intuition.

Alternate Solutions to Trolley Problems

Summarizing the various attempts to solve it over about 30 years, Michael Clark concludes, "[The trolley problem] has proved remarkably intractable" (Clark, 1995, p. 189). Indeed, almost any attempt to explain (or, in many cases, critique) people's intuitive moral judgements seems to have some flaws (such as those pointed out in the DDE above). Is there some alternative to the DDE which would produce better results, and is that alternative computationally feasible in the context of a moral grammar?

Commensurate and Incommensurate Evils

One relatively recent solution proposed by Robert Hallborg rests on the distinction between commensurate and incommensurate evils (taken from Joseph Raz). This means that one must compare the types of harm threatening the various participants—considering the severity, likelihood, extent, etc. of the evils in a "multiplicity of significant criteria" (1997, p. 308). Hallborg argues that in the case that the evils are commensurate (i.e. of roughly equivalent severity, likelihood, etc.) then one is permitted to choose the lesser (in utilitarian terms) evil; otherwise, one may not take the action (transplanting the organs, throwing the switch, etc.). This analysis is interesting and compelling, as it considers various aspects of the harm threatening a group; however, on a closer look it is decidedly circular. Hallborg describes actions' harms as being commensurate in the case that the actions themselves are permissible, which merely begs the question. Perhaps a more stringent notion of "commensurate" would produce a more plausible solution; it is certainly an interesting paradigm to work within even if it is not currently well-developed so that humans (let alone computers) can use it to predict moral intuition.

Deflecting and Originating Harm

First developed by Foot in her original essay in response to the DDE is the idea of deflecting versus originating harm. While this theory has been formulated in different ways, nearly every version draws the distinction between various trolley problems based on the origin of the threat or harm facing the (potential) victims. It is argued that while one may change an already present harm which threatens two groups from one group to another, one cannot originate a new harm to threaten one of the groups. A facet of the kill/let die distinction, this theory is related to the bifurcation of act and omission, doing and allowing, or positive and negative rights.

While earlier versions of this theory (mainly from Foot) were easily rebutted with counterexamples (e.g. Thomson's Bystander example, which suggests that letting die can often

result in the same moral intuition as killing), newer formulations seem to explain many problems quite well. For example, Thomson explains why Gas Leak is not permissible, explaining that the fumes are a new threat to the one patient, which is quite different from the threat facing the five (i.e. their disease); by contrast, she argues, if lethal fumes were coming up through the heating system, it would be permissible to deflect the fumes toward the one, because both parties are threatened with the same harm (1986b, 107).

Montmarquet rephrased this theory by explaining that the distinction "depends on that factor's being present at that time in lethal form" (1982, p. 448). He goes on to define this notion in more precision:

> *X* is threatened with death by factor *Y* at time *t* if and only if either
> (i) *Y* will kill *X* in due course if both are left undisturbed after *t*, or
> (ii) at *t*, *Y* can still be directed into a position in which it will kill *X* in due course if both are henceforth left undisturbed (Ibid.).

Perhaps the best formulation of this sort of theory is Clark's idea of directing misfortune; here, he appeals to "a fundamental principle of tort law: that loss and misfortune should lie with the victim unless there are compelling reasons to shift it to others" (Clark, 1995, p. 195). Thus, in Transplant, those needing organs are already under the misfortune of missing organs, while in Bystander, the workmen's misfortune has yet to hit them (quite literally). Most interestingly, he suggests a combination of Bystander and Transplant: the actor didn't switch trolley, so the five are now missing organs and the one from the side track goes to the hospital with them. While perhaps the actor *should* have switched the track, it is still not permissible to cut up the one to save the five, as the misfortune has already been visited upon them (p. 196).

The trouble with all of these formulations of the "source of harm" theory is that they rely on fuzzy definitions of what constitutes "originating" harm. After all, the one workman on the side track in Bystander is in some sense not threatened at all by the oncoming train—that is, until the bystander chooses to switch the track. Gorr asserts, "Montmarquet is just mistaken in claiming that in those cases [he examines] the members of the smaller group are within the range of whatever threatens the larger group!" For example, in Bystander, "the workman on the side track

is not *at this time* in any danger whatsoever. Granted, he could be *placed* in danger" but he isn't (Gorr, 1990,p. 98). Even if one posits that the same *kind* of threat has to face both parties, one could change Transplant to have similar features to Bystander in terms of the harms threatening the various groups: For example, let us take a bizarre situation I shall call "Organ Sucker." A new, twisted device that removes organs, the Organ Sucker, has been used to rob five people of their vital organs. Is it permissible to use the Organ Sucker on one healthy, compatible patient? Clearly not, although the threat is exactly the same. One could appeal to Clark's notion that the five are already under misfortune, but Clark himself notes that there is some ambiguity about when misfortune hits (Clark, 1995, p. 195). Again, if one cannot determine a reasonable definition of "origination" of harm or misfortune for humans to use, clearly our computational grammar cannot take advantage of this theory.

Avoidable and Unavoidable Violation of Rights or Autonomy

Another interesting although much more easily disposed theory is that of avoidable and unavoidable violation of rights (or, in Don Locke's case, of autonomy). This theory, proposed by Margery Naylor, states that while in Bystander we must decide the fate of the one workman with or without his consent, in Transplant we could ask the donor for consent; thus, the first constitutes an unavoidable violation of the right to control over one's body, while the second constitutes an avoidable violation. Many, including Clark, have critiqued this theory, but Alastair Norcross has perhaps the best refutation; he reduces the theory to absurdity by positing the moral dilemma "Toaster":

> You are hard at work mending an electric toaster. I see the cord of the toaster hanging over the edge of the table with the plug in fairly close proximity to an electrical outlet, which would almost certainly harm you, nor do I ask you whether you would like me to do so, which would be, to say the least, a silly question (1989, p. 717).

This can hardly be considered a non-permissible avoidable violation of rights. Similarly, one can argue that if one had a line of communication with the workman in Bystander, or if the healthy patient in Transplant was in a temporary coma, this wouldn't change the situations' deontic

status. The interesting notion of this theory, however, is its appeal to consent, which can be more fully developed into a more robust solution.

Probabalistic/Utilitiarian/Contractarian Hypothetical Consent

Thomson's latest attempt at the trolley problem discarded her earlier notion of "claims" to rights and origin of harm, by positing some notion of hypothetical consent. While she accedes that "hypothetical consent is not sufficient for permissibility" (1990, p. 188) it can be shown that while people would rationally give consent to switch the trolley in Bystander even if they knew they would (later) be on one of the tracks, it is irrational for healthy people to give consent to operate in Transplant, as the likelihood of being one of those needing organs is less (p. 195). Eric Rakowski posits a similar theory (1993).

Similarly, Alexander Rosenberg suggests that along "contractarian" lines, "a rational individual would bargain for social institutions that did not minimize the risks of being a 'trolley' victim, but did minimize the risks of being a 'transplant' victim" (1992, p. 89). That is, they would consent to such a system because it would give the maximal benefit, given the "costs and benefits of exposure to the risks" (p. 90).

However, even if one could explain some decisions through the maximization of benefit in a hypothetical consent situation, many of our moral intuitions simply don't coincide with the maximally beneficial choice: The best example of this is John Harris' "Survival Lottery," in which a computer picks people at random to be harvested for organ donation. Harris explains that the lottery would be of maximal benefit to everyone, even considering the psychological stress of being under threat of death, for "as we have seen, the chances of actually being called upon to make the ultimate sacrifice might be smaller than is the present risk of being killed on the roads, and most of us do not lie trembling a-bed, appalled at the prospect of being dispatched on the morrow" (1975, p. 261). Interestingly, Harris suggests another solution to Transplant

which does not coincide directly with the maximization of benefit theory but which does seem to poke a hole in the dilemma itself: why not, at random, harvest one of the five's organs for the other four? (p. 263). Nevertheless, most people see the idea of a Survival Lottery as being completely repugnant—despite the fact that a rational person acting solely for maximum benefit would join it. Indeed, "So strong is our intuitive repugnance for the transplant sacrifice that it extends to accepting the heroic offer of a patient who is prepared to be sacrificed" (Clark, 1995, p. 194).

Combination

"Perhaps there *is* no formula, no principle or small set of principles, which will do the job satisfactorily" to explain our moral intuitions about trolley problems (Clark, 1995, p. 195). Indeed, throwing up their hands, Fischer and Ravizza argue that our moral intuitions are wrong and nothing really distinguishes them (1992b, p. 80). Postow also argues that "we feel safest in relying on intuitions in concrete cases in which the rightness or wrongness of an action is apt to be overdetermined. The artificial [trolley problem] examples…are not only highly uncommon—they are also free of overdetermination" and thus we cannot trust our intuitions (1989, p. 536).

Even if this is so, it is still interesting and worthwhile to explain how commonsense moral intuitions are cognitive processed—that is, to develop an accurate *descriptive* moral intuition theory. We might take these philosophers' argument to heart, however, in their assertion that there are "middle cases" in which people cannot or do not agree on the permissibility of an action. Clark notes that if a series of principles "are to explain our intuitive judgements, then this [ambiguity] is just as it should be, since the intuitions in many of the intermediate cases are weaker, less constant and less widely shared than those about the paradigms" (Clark, 1995, p. 198).

I am convinced that some combination of the various theories will result in a more accurate, if perhaps more complicated, descriptive moral grammar. Certainly a more developed distinction between creating and diverting harm could be integrated into the grammar (just as a rudimentary notion of killing being worse than letting die if all other things are equal is already part of the grammar). The DDE seems very useful in certain classes of problems, but it appears that it cannot explain all our intuitions on its own. Likely if one changes the grammatical deontology to a gradient, between permissible and forbidden, one could posit that certain factors add to or subtract from an action's permissibility—there may not be one, definitive, necessary and sufficient framework. Such a change would require more empirical studies and fairly fundamental changes in the moral grammar, but the result would be a more accurate description of mental intuition.

Of course, it may be possible that "moral knowledge is stored in the form of examples and stories" and thus our symbolic, linguistic-derived approach may be inappropriate (Stitch, p. 225). Nevertheless, at least a small subset of human moral cognition can be successfully modeled in a computational grammar, which suggests that we are on the right track. It is possible, also, that one could construct a perfectly workable descriptive theory which does not map to the way that moral cognition actually occurs—still, merely understanding what moral knowledge *is* seems to be the first, necessary step before one can investigate the implementation of that knowledge.

Works Cited/Consulted:

Bica, C. C. (1999). Another perspective on the doctrine of double effect. *Public Affairs Quarterly*, 13(2), 131-139.

Boyle, J. M., Jr. (1980). Toward understanding the principle of double effect. *Ethics*, 90(4), 527-538.

Clark, M. (1995). Sacrificing one to save many. *Journal of Applied Philosophy*, 12(2), 189-200.

Costa, M. J. (1986). The trolley problem revisited. *Southern Journal of Philosophy*, 24(4), 437-449.

——— (1987). Another trip on the trolley. *Southern Journal of Philosophy*, 25, 461-466.

Davis, N. A. (1994). The priority of avoiding harm. In *Killing and Letting Die*, 298-354

Fischer, J. M. (1991). Tooley and the trolley. *Philosophical Studies*, 62, 93-100.

Fischer, J. M. and Ravizza, M. (1992a). Quinn on doing and allowing. *The Philosophical Review*, 101(2), 343-352.

——— (1992b). Thomson and the trolley. *Journal of Social Philosophy*, 23(2), 64-87.

Foot, Philippa (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, 5. Reprinted in *Killing and Letting Die*, 266- 279.

——— (1984). Killing and letting die. *Abortion and Legal Perspectives*, Jay L. Garfield and Patricia Henessey (Eds.). Amherst: University of Massachusetts Press. Reprinted in *Killing and Letting Die*, 280-289.

Gorr, M. (1990). Thomson and the trolley problem. *Philosophical Studies*, 59, 91-100.

Hallborg, R., Jr. (1997). Comparing harms: the lesser-evil defense and the trolley problem. *Legal Theory*, 3, 291-316.

Hanna, R. (1993). Participants and bystanders. *Journal of Social Philosophy*, 24(3), 161-169.

Harris, J. (1975). The survival lottery. *Philosophy*, 50 (191), 81-87. Reprinted in *Killing and Letting Die*, 257-265.

Kamm, F. M. (1991). The doctrine of double effect: reflections on theoretical and practical issues. *The Journal of Medicine and Philosophy*, 16, 571-585.

——— (1992). Non-consequentialism, the person as an end-in-itself, and the significance of status. *Philosophy and Public Affairs*, 21(4), 354-389.

Locke, D. (1982). The choice between lives. *Philosophy*, 57, 453-475.

Malm, H. M. (1989). Killing, letting die, and simple conflicts. *Philosophy and Public Affairs*, 18(3), 238-258.

Mikhail, J. (2000). *Rawls' Linguistic Analogy: A Study of the "Generative Grammar" Model of Moral Theory Described by John Rawls in* A Theory of Justice. Cornell University, PhD Dissertation.

Mikhail, J. and Sorrentino, C. (1999). Toward a fragment of moral grammar: knowledge of the principle of double effect in children ages 8-12. Poster presented to the Society for Research in Child Development, Albuequerque, N.M.

Mikhail, J., Sorentino, C., and Spelke, E. (1998). Toward a universal moral grammar, in Morton Ann Gernsbacher and Sharon J. Derry, eds. *Proceedings, Twentieth Annual Conference of the Cognitive Science Society*, Mahwah, New Jersey: Lawrence Erlbaum Associates, 1250.

Montmarquet, J. A. (1982). On doing good: the right and the wrong way. *The Journal of Philosophy*, 79(8), 439-455.

Naylor, M. B. (1988). The moral of the trolley problem. *Philosophy and Phenomenological Research*, 48 (4), 711-722.

Norcross, A. (1989). A reply to Margery Naylor. *Philosophy and Phenomenological Research*, 49 (4), 715-719.

Patton, M. F., Jr. (1988). Tissues in the Profession: Can Bad Men Make Good Brains Do Bad Things? *Proceedings and Addresses of the American Philosophical Association*, 61-3. Available online at http://www.mindspring.com/~mfpatton/Tissues.htm

Postow, B. C. (1989). Thomson and the trolley problem. *The Southern Journal of Philosophy*, 27(4), 529-537.

Quinn, W. (1989). Actions, intentions, and consequences: the doctrine of doing and allowing. *Philosophical Review*, 98, 287-312. Reprinted in *Morality and Action* (1993). New York: Cambridge University Press, 149-174.

Rakowski, E. (1993). Taking and saving lives. *Columbia Law Review*, 93(5), 1063-1156.

Rosenberg, A. (1992). Contracarianism and the “trolley” problem. *Journal of Social Philosophy*, 23(2), 88-104.

Sheng, C. L. (1995). A suggested solution to the trolley problem. *Journal of Social Philosophy*, 25(1), 203-217.

Steinbock B. and Norcross, A. (eds.). (1994). *Killing and Letting Die*. New York: Fordham University Press.

Stitch, S. P. Moral philosophy and mental representation. In *The Origin of Values*, M. Hechter, L. Nadel, and R. E. Michod (eds.), Hawthorne, NY: Aldine de Gruyter.

Thomson, J. J. (1986a). Killing, letting die, and the trolley problem. *Rights, Restitution, and Risk*. Cambridge, Massachusetts: Harvard University Press, 78-93.

——— (1986b). The trolley problem. *Rights, Restitution, and Risk*. Cambridge, Massachusetts: Harvard University Press, 94-116.

——— (1990). *The Realm of Rights*. Cambridge, Massachusetts: Harvard University Press.

Tooley, M. (1994). An irrelevant consideration: killing versus letting die. In *Killing and Letting Die*, 103-111.

Trammell, R. (1975). Saving life and taking life. *Journal of Philosophy*, 72(5), 131-37. Reprinted in Killing and Letting Die, 290-297.