MORAL JUDGMENTS ABOUT ALTRUISTIC SELF-SACRIFICE: WHEN PHILOSOPHICAL AND FOLK INTUITIONS CLASH

Bryce Huebner
Department of Philosophy
Georgetown University

Marc D. Hauser
Departments of Psychology and Human Evolutionary Biology
Harvard University

On 1 April 2008, it was announced that the family of a Navy SEAL named Michael Monsoor would be presented with a Congressional Medal of Honor (Tyson, 2008); two months later a Medal of Honor was presented to the family of Pfc. Ross McGinnis (Hohmann, 2008). Each of these soldiers engaged in a selflessly heroic act; they both jumped onto a live grenade to save the lives of others who were nearby. Soldiers are, of course, trained to do whatever they can to escape the blast of a grenade. In military training, no one ever suggests that a soldier *should* jump onto a grenade. However, these acts of altruistic self-sacrifice occur more often than one might expect. Those soldiers who engage in such acts are seen as heroes who have expressed courage and altruism beyond what could ever be *expected* of a person. Yet, while most of us praise such acts of self-sacrifice, it seems implausible to claim that anyone ever has a moral obligation to throw herself on top of a grenade, even if doing so would save her closest friends.

Altruistic self-sacrifice is rare, supererogatory, and not to be *expected* of any rational agent. The possibility of such actions has, however, played an important role in moral theorizing. J.O. Urmson (1958) famously argued that such 'saintly' and 'heroic' actions should lead us to revise our classification of morally significant actions. He claimed that such actions were praiseworthy, not merely permitted, and yet they could neither be classified as forbidden or obligatory. Susan Wolf (1982, p. 419), by contrast, suggests that "moral saintliness, does not constitute a model of personal well-being toward which it would be particularly rational or good or desirable for a human being to strive."

Stepping away from the views advanced by moral theorists, moral psychologists have also casually reported that when lay people are approached on the street and asked to make judgments about moral dilemmas, some of them ask why the normatively preferable option—sacrificing oneself—has been left out of the experimental design. For example, in a previous experiment where we asked people in a public park how many lives would have to be saved for it to be obligatory to push someone into harm's way, we found that some participants thought that the only viable option was an act of altruistic sacrifice. They made claims like "Eric should not sacrifice someone else to save others. He should jump in himself" and "I'd jump in front, I weigh 220". The data that would tell us precisely how often such thoughts are expressed is not always collected or recorded. However, anecdotal evidence at least suggests that a small, though reliably present proportion of people will ask why the person described in a moral dilemma doesn't sacrifice herself for the good of some number of anonymous strangers.

The fact that people offer self-sacrifice as a response to an artificial moral dilemma like the trolley problem is surprising. While there is no doubt that human beings express compassion toward one another, it is unclear why anyone would ever treat a radical act of altruism such as throwing oneself in front of an oncoming trolley *to save the lives of a small number of unknown people* as the right thing to do.[1] Unlike those cases where a soldier sacrifices his life to save his platoon-mates, the trolley scenarios that populate philosophy papers and psychology experiments tend to strip away all identifying information from those who are threatened (cf., Foot, 1967, 1985; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Hauser, Young, & Cushman, 2007; Kamm, 2005; Mikhail, 2007; Moore, Clark, & Kane, 2008; Petrinovich, O'Neill, & Jorgensen, 1993; Pizarro, Uhlmann, Tannenbaum, & Ditto, in prep; Sinnott-Armstrong, Mallon, Hull, & McCoy, forthcoming; Unger, 1996; Valdesolo & DeSteno, 2006; Waldmann & Dietrich, 2007). There are *five people* down the track, not your five children, and there is *one person* on the side track, not *one orphan*. While considerations of kinship may drive a mother to sacrifice herself for the good of her child, and while there is reason to think that such considerations play a prominent role in folk morality (Petrinovich et al., 1993), the unconditional altruism required to jump in front of a runaway trolley to save a small number of anonymous strangers requires motivation of an entirely different order. The person who is inclined toward such radical acts of altruistic self-sacrifice seems to think very little of her own life—and this is why many people have seen such saintly actions as morally unattractive.[2]

Perhaps the suggestion that a person should throw herself onto the tracks points to a deeper concern with the plausibility of trolley dilemmas and the intuitions that yield moral judgments about such apocryphal cases. Imagining turning the trolley onto an unsuspecting person may evoke a sense of discomfort as well as a sense that something has gone wrong with the set-up. When participants in psychology studies are asked if it is permissible to sacrifice the life of one person to save five others, they may think that the scenario is absurd, and that they can point out what has gone wrong by instead suggesting an absurd act of radical altruistic self-sacrifice. If this is the case, the response may tell us that there is something deeply wrong with our intuitions about trolley scenarios and all other related dilemmas.

1. THOMSON'S CRITIQUE OF THE TROLLEY PROBLEM

In a recent paper, Judith Jarvis Thomson (2008) has argued that our intuitions about altruistic self-sacrifice do, in fact, suggest that something has gone wrong in philosophical debates over the trolley problem (Foot, 1967; Kamm, 2005; Thomson, 1985). Thomson asks us to consider a case in which a bystander sees a runaway trolley coming down the tracks, she knows that there are five people on the main track, there is one person on a side track to the right, and the bystander is on a side track to the left. The bystander realizes that she has only three viable options:

[1] This is not, of course, to claim that the judgment that one ought to engage in such a radical act of self-sacrifice is utterly unintelligible. In fact, there may some cases where such an act would likely be prescribed by a radical form of act utilitarianism. We only wish to claim that it would be surprising if this form of act utilitarianism (or any other sophisticated philosophical position) were to play a prominent role in less reflective judgments about right and wrong.

[2] Again, there may be more reflective positions from which it would make sense to sacrifice oneself for the good of unknown others. Having adopted a moral theory that required a thoroughly impersonal point of view, a person could carry out a calculation that would lead her to judge that her own life had no more value than the lives of these unknown others. It is a further question whether such a 'moral point of view' is psychologically plausible; but even on the assumption that it is, it is not obvious that this is the sort of moral theory that we should prefer.

1) She can do nothing and let the trolley kill the five people on the main track;
2) She can flip a switch to the right, killing the one anonymous person but saving the lives of the five people on the main track; or
3) She can flip a switch to the left, killing herself but saving the lives of the five people on the main track.

Although Thomson (2008, pp. 364-365) concedes that it would be a 'good deed' if the bystander were to give up her own life to save the five people on the main track, she argues that it would be morally reprehensible if the bystander were to force an anonymous person to pay the cost of this good deed because she was unwilling to do so herself. On Thomson's understanding of the case, the bystander is unlikely to sacrifice herself to save five unknown people, but there is no obvious justification for sacrificing the life of another to save the people on the main track. After all, any reasons that can be employed in deciding against altruistic self-sacrifice will be equally applicable to the unknown stranger on the other track: put briefly, the bystander doesn't want to die, but neither does the anonymous stranger. Since the bystander has no obligation to sacrifice herself to save five anonymous strangers (barring a desire to take her own life), the only morally permissible option is to let the trolley continue along the main track, killing the five people.

With this case in hand, Thomson argues that it is impermissible to divert the trolley onto an anonymous person in the more familiar two-track bystander case (where the only options are flipping the switch and killing one person, or letting the five die). She argues that even if the bystander in this case *would* be willing to sacrifice her own life for the good of five unknown people, she cannot assume that the anonymous person on the side track would be equally altruistic. More importantly, the bystander has no reason to treat this person as having an obligation to be sacrificed, and since the person on the side track has not consented to sacrifice her life, the bystander must adopt the clearly permissible action of doing nothing, letting the trolley kill the five people on the main track.

Thomson clearly recognizes that her claim that it is impermissible to flip the switch in the standard bystander case is highly counterintuitive. So, she also attempts to explain why it has seemed to many people—including philosophers, psychologists, and lay people—that it is permissible to flip the switch in the two-track bystander case. People are " moved by the fact that more people will live if the bystander turns the trolley than if he doesn't" (Thomson, 2008, p. 373). But, the fact that ordinary people are impressed by the number of lives that are saved shows only that our *recognition* of the duty not to harm a person without her consent is sensitive to the way in which this harm is brought about. In a case where one must only turn a trolley, the negative duty not to harm another person is just less salient than it is in cases where bringing about the desired end requires direct physical contact—as when a large person must be pushed in front of an oncoming trolley.

Our *primary* goal in this paper is to use Thomson's philosophical analysis of the trolley dilemma and her new trilemma cases as a starting point from which to explore folk intuitions about morally permissible actions. Specifically, we examine the way that including the option to altruistic self-sacrifice influences the decision to save the lives of anonymous others. Secondarily, we also address some of the theoretical problems with Thomson's argument, and in particular, the idea that prior philosophical discussions of the bystander case are misguided because the trilemma version eliminates the moral permissibility of killing the one on the side track to save the five on the main track. We begin with this secondary task in order to draw out the theoretical framework for developing our experimental methods. We conclude by briefly discussing the theoretical

issues that are raised when philosophical and folk intuitions conflict, and what this means for thinking about the role of intuitions in both philosophy and cognitive science. While we recognize that folk intuitions about these cases may be dismissed by some philosophers as irrelevant for answering questions about whether one should ever sacrifice her own life to save the lives of anonymous others, we suggest that folk-moral judgments about this case shed interesting light on the nature of moral judgments independent of this significant prescriptive issue.

2. WHEN THE TROLLEY GOES OFF THE TRACK

As noted above, Thomson believes that her intuition about the three-track bystander case sheds light on a previously unnoticed difficulty with the more familiar two-track bystander case. Unfortunately, the introduction of this third morally significant option yields a more striking modification of the two-track case than Thomson appears to have realized. After all, if Thomson is to demonstrate that an intuition about the three-track bystander case can be extended *without revision* to the familiar two-track bystander case, it must be established that no additional, morally significant factors are introduced by adding this option. Things would, of course, be much easier in ethics if we could assume that the overall context of an action was never morally significant. But, as the shape of the space in which a decision must be made is changed, this is likely to introduce a variety of (possibly morally significant) considerations, a point made repeatedly in the philosophical work dealing with precisely the kinds of cases that concern Thomson (Kamm, 2005). In fact, there is good reason to think that the space of options with which we are presented is likely to have an influence on our decisions. For example, experimental studies have shown that decisions are influenced not only by the details of our various options, but also by the number of options that we happen to have (cf., Schwartz, 2004). Moreover, both philosophical arguments (e.g., Kamm, 2005) and psychological results (e.g., Cushman, 2008; Cushman, Young, & Hauser, 2006; Greene et al., In Press; Moore et al., 2008; Waldmann & Dietrich, 2007) suggest that it is difficult to seamlessly introduce new factors into moral dilemmas without affecting the intuitions about the case at hand. In short, it is difficult to guarantee that every morally significant consideration has been held constant in moving between apparently similar cases. Our goal in this section is to show that establishing that there are no normatively significant differences between the three-track case and the two-track case requires that Thomson assume that only considerations of consent are normatively significant in these two cases. This means that establishing a strong analogy between these two cases requires the adoption of theoretical assumptions that are likely to be shared by only those who accept Thomson's preferred deontological theory.

The key to Thomson's argument in the three-track case is that if there is nothing to tell in favor of turning the trolley to the left or to the right, the only viable option is to let the trolley continue along the main track. At first blush, this assumption is quite plausible. A similar all-or-nothing principle of non-interference arises in one interpretation of Jewish ethics (Elster, 2007, p. 82). Where a decision must be made that will selectively disadvantage some group of people, and where there is nothing that obviously tells in favor of disadvantaging one group over another, it seems that the normatively preferable option is just not to act. However, as this principle is typically understood, the prohibition is "not a ban on causing a person to be killed to save others, but on selecting who it shall be" (Elster, 2007, p. 82). This principle, however, leads to an important difference between the two-track and three-track bystander cases. Indeed, it is

of primary importance that the reasons that can be brought to bear in deciding what ought to be done diverge radically in these two cases.

In the two-track case, the people on the tracks are typically described in a way that makes them completely anonymous. Describing the bystander case in this way allows us to abstract away from possibly irrelevant differences between the people on tracks, and this allows us to focus on the question of whether it is ever permissible to sacrifice one person to save five others. There are, of course worries about abstracting away from the intricate details of particular cases, but it is a move that is strongly endorsed in a variety of philosophical works on our moral psychology (though this view is perhaps most pronounced in Kamm's (2005) work, it is a common approach to moral philosophy). In evaluating familiar sorts of moral dilemmas, we can justifiably rely only on the agent-neutral reasons that are at play when we evaluate the proposed action. Regardless of whether the bystander's intentions, the outcomes, or the victim's consent provide the most important constraint on morally acceptable action in this case, each of these factors can be evaluated without regard to the facts about the particular agents whose lives are at stake. However, the reasons that must be evaluated in the three-track bystander case are more complex. As Sidgwick (1907) convincingly argued, even the staunchest consequentialist must draw a distinction between reasons of self-interest and reasons of general benevolence. Consequently, introducing considerations of altruistic self-sacrifice in the case of the three-track bystander appears to also introduce a range of agent-relative, or more precisely, self-regarding reasons for acting (cf., Nagel, 1986; Parfit, 1984; Pettit, 1987) that are not present in the original bystander case. This is not to say that, *all things considered,* these self-regarding reasons make a difference in deciding what it is right to do in this case. However, the introduction of these considerations at least modifies the decision-procedure that one must go through in evaluating the two cases. Thus, to show that an intuition about the three-track case can be extended *without revision* to the two-track bystander case, it would be necessary to show either that these self-regarding reasons are not sufficient to modulate the decision procedure in the three-track case, or at least that they are just as important in the decision procedure that must be carried out in the two-track case as they are in the three-track case.

To clarify our concern with Thomson's argument, consider the thesis that only agent-neutral reasons are morally significant. There are at least two ways in which this idea might be developed, one along broadly consequentialist lines and one along the broadly deontological lines adopted by Thomson. On the consequentialist version of this argument, one could argue that only the outcomes matter in this case. Here, there is an obvious answer to what ought to be done in the two-track case (sacrifice the one to save the five); but the answer is far less obvious in the three-track case: while it is clear that the five ought to be saved, it is not immediately obvious whether to flip the switch to the left or to the right. For such a consequentialist perspective, there is thus an obvious and normatively significant difference between the two-track and three-track bystander cases. The consequentialist seems to have a strong reason to flip the switch, but there is no reason to flip the switch to the left rather than the right as either choice is equally arbitrary. The deontological version of the argument requires the recognition that the negative duty not to harm another person without her consent is the weightiest of moral principles. This is the view that is adopted by Thomson in her interpretation of the three-track case, and it gives a principled reason for thinking that the only viable option in this case is to let the trolley continue along the main track.

We are inclined to think, as Mike Ridge (2001) has argued in another context, that the assumptions that the deontologist is forced to make in such cases look slightly self-indulgent. But, there are even deeper problems with this move. Thomson's argument that

it is obviously permissible to 'let the trolley go' in both of these cases, regardless of what the alternative options are, assumes that letting the trolley go is a clear omission, and is thus obviously permissible (Thomson, 2008, p. 369).[3] Thomson's argument that there are no morally sufficient reasons that tell in favor of turning the trolley thus presumes that the bystander is responsible for the outcome if she flips the switch, but that she is not responsible for the outcome if she just lets the trolley continue along the main track. However, the fact that the two-track bystander case tends to strike most people as a genuine moral dilemma tells strongly against this assumption. Philosophers and non-philosophers alike tend to see the bystander as faced with a tragic choice between two unappealing options. But, it doesn't follow from the fact that this choice is tragic that there are no reasons that speak in favor of choosing one of these unappealing options. More importantly, the bystander cases make it clear that all of the relevant choices and outcomes have been made transparent to the actor. But this takes away much of the force from the claim that letting the trolley go is obviously permissible. When the actor is completely aware of all of the choices and outcomes, her refusal to intervene is plausibly treated as a deliberate and intentional inaction, and hence as something that the actor *does* (Quinn, 1989). Here, the loss of five lives is the result of the bystander's intentional inaction, and it is in no way obvious that such intentional inactions are always permissible.[4]

Finally, it seems as though the person who turns the trolley onto herself must recognize that her death is an inevitable consequence of her action; but, if she is to override the motivation of self-preservation, she must also form an intention to bring about her death as the only possible means to save these people's lives. Building on an argument developed by Kamm (2005), it seems that the agent who diverts the trolley onto a side track in this case does so in order to save her life and the lives of the people further down the track. So, her reason for redirecting the trolley is best explained by her intention to save herself and the five people on the main track rather than by appeal to an intention to bring about the death of the person on the side track. In this case, the death of the one person on the side track is, no doubt, a tragic consequence of her action. However, intending to bring about this outcome would require a desire (the death of the one person) that is utterly lacking in the bystander, at least as the case is described. To preserve the parity required for Thomson's argument, the bystander would have to intend to sacrifice the anonymous person's life in precisely the same way. But, this has not been

[3] Thomson is, of course, likely to be unimpressed by the response in this paragraph. She has previously argued that actions must be understood as types of events; events in turn are to be understood as having clear spatio-temporal properties. Since it is unclear, on her view, when the relevant omission begins and ends, omissions can never be counted as actions. However, even on the assumption that there is a clear boundary between actions and omissions, this does not preclude a distinction between intentional and unintentional omissions. And it seems reasonable to suppose that an agent's reason for not acting is, at least sometimes, to be viewed as morally salient in evaluating the moral status of her inaction.

[4] Thomson (2008, p. 370) acknowledges that such considerations may cause trouble for her argument. In discussing the case of the Driver's Three Options, she argues that the driver of a car can plausibly be seen as killing five people in the road in front of her if she takes her hands off of the wheel (in a refusal to make a decision). Thomson claims that taking one's hands off of the wheel must be understood in the context of the fact that *she started her car.* She therefore claims that because driving the car is an intentional action, and it doesn't stop being intentional when the brakes fail, the driver kills the five people in the road if she 'does nothing'. This is, it seems, a special case of the distinction between those actions that an agent *initiates* or *keeps going* and those that she *allows to run their course* (Foot, 1985). We do not wish to deny that Thomson's argument establishes one important different between the driver and the bystander. However, our claim is that the mental state of the bystander, and the reasons for action that are currently available to her, also make it implausible to say that she can merely 'let the trolley go'. Since the bystander *can* redirect the threat, and since she *intends* to let it go, the death of the five people on the main track can plausibly be seen as occurring as a result of the bystander's inaction (Quinn, 1989).

established, except on the condition that relevant intentions are bound up with considerations of consent.

Although the strong deontological premises that are required for Thomson's argument entail that the bystander case isn't a genuine moral dilemma, they also force a theoretical perspective that takes consent to be the sole, determining variable for moral reasoning. Since the legitimacy of Thomson's preferred deontological theory is precisely what is at issue in the bystander case, she has no room to rely on her preferred deontological intuition as a means to dispense with the bystander dilemma. For Thomson, the third track is a third-wheel, offering no additional traction on difficult theoretical issues that continue to preoccupy moral philosophers.

Even if we put to the side the aforementioned problems with Thomson's argument against the bystander case, there is a further methodological question about how the introduction of a third option modifies the space of moral dilemmas, including trolley cases. Thomson assumes, based on her own intuitions, that the introduction of altruistic self-sacrifice should lead others to see the hypocrisy inherent in sacrificing the person on the side track and to judge that the bystander should let the trolley continue along the main track. But, it is unclear how such considerations will be evaluated when people who do not share Thomson's commitment to a deontological theory are presented with such cases. It is on this point that we come to a further untested assumption that is raised by Thomson's thought experiment: how does the introduction of a third track affect judgments from the standpoint of folk-morality.

3. FOLK INTUITIONS ABOUT SELF-SACRIFICE IN TROLLEY CASES

Before we present our experimental data, we would like to make it clear why we believe that it matters how the introduction of a third track is processed from the standpoint of folk-morality. Our interest in this issue arose from a claim that Thomson makes in her paper. Having offered her account of why the bystander *may not* turn the trolley, Thomson also attempts to address the question of why it seems to so many people (read: folk morality) that she *should* turn the trolley. In considering this question, Thomson (2008, p. 373) makes the following claim: "Ninety-three percent of the seniors at South Regional High School in Dayton, Ohio, say that the bystander may turn the trolley in Bystander's Two Options!" It seems that there are two closely related reasons why folk-moral judgments such as these might matter. First, it is genuinely interesting to know whether there is a difference between folk-moral judgments about difficult cases and more tutored philosophical intuitions about these cases. Where there are such differences, there is a further question about which of these intuitions (if any) ought to be taken seriously in the construction of an adequate moral theory, and how tutoring affects our native intuitions. Thomson suggests that this 'result' is likely to derive from two considerations: 1) these people are overly impressed by the numbers of lives that are at stake, and 2) the structure of the case prevents people from noticing that they infringe on a negative duty in turning the trolley onto the one person.

While Thomson may be right that this is the case, it is unclear why she felt compelled to invent a hypothetical result when there are real data available from numerous studies, with literally thousands of participants offering judgments about different dilemmas including the bystander case (Cushman et al., 2006; Greene et al., In Press; Hauser, Cushman et al., 2007; Mikhail, 2007). If there is an interest in describing the judgments that people make in such cases, and in using these patterns of responses to come to a better understanding of which features are morally relevant, then there is no possible cost in consulting this literature. For example, a recent study revealed that eighty-five percent

of participants judged that it was morally permissible to flip the switch in the two-track bystander case (Hauser, Cushman et al., 2007). However, there is no existing empirical data on the sorts of moral *trilemmas* that are at issue here. Thus, rather than inventing seemingly plausible results, we see great merit in collecting real data to generate real results that speak to the nature of folk moral intuitions, recognizing that these results may not impact upon our prescriptive theories.

Second, there is a question about how different factors are evaluated from the standpoint of folk-morality. Thomson's philosophical intuition is that considering the option of altruistic self-sacrifice makes the negative duty not to harm another person more salient. If this is the case, then perhaps the introduction of such an option will also have a significant effect on the proportion of people who will judge that it is permissible to flip the switch in a trolley case. We suggested above that the three-track bystander case does not provide a compelling theoretical reason for thinking that the bystander should not turn the trolley. But, perhaps the folk-moral judgment that it is permissible to turn the trolley results from being overly impressed by the numbers where these have been made most salient. This is where empirically driven moral psychology can make a crisp connection with Thomson's intuitions. Stated differently, though we think Thomson's philosophical argument can be critiqued on philosophical grounds, we also think that her argument has raised a fascinating thesis that calls for empirical testing.

As we see it, someone who was willing to adopt Thomson's argument should also be willing to make the following clear empirical prediction. In a three-track case where the option of altruistic self-sacrifice is included, we should expect to find a significant increase in the proportion of people who are unwilling to flip a switch to divert a trolley onto another anonymous person when the comparison class is flipping the switch to engage in an act of altruistic self-sacrifice. Considerations of altruistic self-sacrifice should make the negative duty not to harm another person more salient then they would have been were only numerical considerations transparent. If this is the case, it will provide evidence for the claim that the folk-moral judgment in the two-track bystander case results from focusing (*pace* Thomson) on numbers as opposed to negative duties. An alternative prediction is suggested by the all-or-nothing principle of non-interference from Jewish ethics that we mentioned above. If something like this principle is operative in folk-morality, then where there is nothing to tell in favor of turning the trolley to the left or to the right, the option of letting the trolley go will seem more plausible. Here, we would expect that considerations of self-sacrifice will introduce additional considerations such as an interest in self-preservation that *can* tell in favor of one of the options; so, there should not be a significant increase in the number of people who judge that it is permissible to let the trolley go.

In investigating such hypotheses, experimental psychologists and philosophers alike have to be careful to control the differences between scenarios (as Kamm has frequently argued, there are many cases in which they aren't). In the following experiment, we examine three of the factors that might evoke differences in folk-moral judgments in a three-track case: 1) the inclusion of the option of altruistic self sacrifice; 2) the presence of self-regarding reasons for action; and 3) the bare inclusion of a third choice.

Between 25 November 2008 and 10 March 2009, 3266 participants voluntarily signed on to the Moral Sense Test website (http://moral.wjh.harvard.edu) and responded to one of four conditions modeled on the thought experiment proposed by Thomson (2008). In each condition, the text of a familiar trolley dilemma was modified to include a boxcar rather than a trolley (to avoid the plausible assumption that additional lives may be at stake if the trolley is carrying passengers) and to include a situation where a runaway boxcar was coming down the tracks and two side tracks branched off from the

main track at a switch point. In Condition 1, participants responded to a moral trilemma in which the protagonist, Jesse, can either engage in an act of altruistic self-sacrifice to save the lives of five anonymous people, turn the boxcar onto a track where there is one anonymous person, or let the boxcar continue along its course (1-5-Jesse). Condition 2 included a trilemma that was identical to the one used in Condition 1 except that the person who was reading the scenario was asked what she or he should do rather than being asked what Jesse should do (1-5-You). Condition 3 proposed a moral trilemma in which Jesse could either divert the boxcar to the left or the right track (each with one anonymous person) or let the boxcar continue along its course (1-5-1). And finally, in Condition 4, participants were presented with a standard bystander case with an additional empty side track (1-5-0).[5]

In Condition 1, participants read the following moral trilemma and were asked what Jesse should do.

> Jesse is standing near the railroad tracks and notices an empty boxcar coming down the tracks, moving fast enough to kill anyone that it hits. If Jesse does nothing, the boxcar will continue along the main track, killing five people who are walking down the main track. There is a switch nearby that Jesse can use to divert the boxcar onto either of two side tracks that split off from the main track in opposite directions. There is one person walking along the right-side track. So, if Jesse flips the switch to the right, the boxcar will hit and kill this person. Jesse's foot is stuck in the track on the left-side track. So if Jesse flips the switch to the left, he will be hit and killed by the trolley himself. What should Jesse do?

Faced with this case, the majority of the participants (43%) judged that Jesse should flip the switch to the right and a surprisingly large proportion of our participants (38.3%) judged that Jesse should engage in an act of altruistic self-sacrifice to save the five people on the main track. Thus, when people were presented with the opportunity to engage in an act of altruistic sacrifice, the vast majority (81.3%) judged that they should take some sort of action, and only a small proportion of participants (18.7%) judged that Jesse should allow the trolley to proceed along the main track so as to kill the five unknown people (see Figure 1 for the results of Conditions 1-3). However, in line with the hypothesis that the introduction of altruistic self-sacrifice would have an effect on people's willingness to divert the trolley off of the main track, the proportion of participants who judged that they should do nothing was slightly larger than the proportion predicted on the basis of the results obtained in the familiar 2-track bystander case (15%).[6] However, while this is a significant increase, it is important to note that it is an increase of just 3.7% of the population. There is room for debate over the importance of this small difference, and we return to discuss this issue in the concluding section of this paper.

Condition 2 introduced a modification of the wording that was used in the previous trilemma, replacing the name 'Jesse' with the pronoun 'you' to allow for a closer examination of the way in which considerations of altruistic self-sacrifice are interpreted

[5] All participants were asked to complete the test without interruption, to read the scenarios and associated question carefully, and to answer the questions solely on the basis of the information provided. Previous research presenting moral dilemmas of this kind has demonstrated that there is no substantive difference between responses obtained using Web-based questionnaires and more traditional pen-and-paper questionnaires (Hauser, Young et al., 2007). All procedures were conducted in accordance with the Institutional Review Board of Harvard University, and followed the testing procedures of other web-based research (e.g., Nosek, Banaji, & Greenwald, 2002).

[6] χ^2 (1, N=845) = 9.06, p=.003, w=.037.

from the standpoint of commonsense psychology. Although a surprisingly large proportion of people judged that Jesse should engage in an act of altruistic self-sacrifice to save the lives of five anonymous people, we predicted that when participants were asked what *they* should do, they would be less likely to prefer this option. That is, although a large number of participants were willing to say that another person should engage in an act of altruistic self-sacrifice, we expected that they would be less willing to say the same of themselves. In considering this case, two hypotheses present themselves. First, while the distinction between 'you' and 'Jesse' may make no moral difference, people may rely on different psychological strategies in interpreting these two cases, and so display a different pattern of judgments. That is, people may see another person as having *more reason* to sacrifice their lives to save some anonymous people than they would have themselves. Second, if people rely on strategies of mental simulation in interpreting moral scenarios such as this, there should be no significant difference between the willingness to engage in self-sacrifice where the second-person pronoun is used instead of a proper name. After all, if people imagine themselves having to make the choice when they are faced with a moral scenario, then being told to explicitly imagine that they were in this situation should not make much of a difference.

In line with this latter hypothesis, the results obtained in Condition 2 paralleled the results that we obtained in Condition 1. The majority of participants (48%) once again judged that they should flip the switch to the right; however, much to our surprise, a substantial number of participants (33.7%) continued to judge that they should flip the switch to the left, supporting an act of altruistic self-sacrifice. Replicating our initial results, the vast majority of participants (81.7%) who were presented with the opportunity to engage in an act of altruistic self-sacrifice judged that they should engage in some action, and only a small proportion of participants (18.2%) judged that they should do nothing, allowing the trolley to proceed along the main track and kill the five unknown people. Strikingly, there was no significant difference between the proportion of people who chose to do nothing in this case as opposed to the case in Condition 1.[7] Once again, however, there was a slight increase in the proportion of people who judged that they should do nothing when compared to the values expected on the basis of the familiar two-track bystander case.[8]

According to the hypothesis suggested by Thomson's argument, those people who were presented with a question about what a third party bystander should do might have continued to be swayed by the numbers because they were not really entertaining the possibility of altruistic self-sacrifice. However, this should not have been the case where a participant was explicitly presented with the option of diverting the oncoming trolley onto *herself*. In this case the hypocrisy of diverting the trolley onto an anonymous individual should have been more pronounced, making the impermissibility of turning the trolley far more transparent. However, although there was a slight increase in the proportion of people who were willing to divert the trolley onto the anonymous person, there was no significant difference in the proportion of people who were willing to let the trolley go when they were asked what they should do rather than being asked what Jesse should do. Further, slightly fewer people were willing to engage in altruistic self-sacrifice when their own hypothetical life was on the track, so to speak, evidence that makes less plausible the suggestion that people always engage in mental simulation when they are faced with a moral scenario, (see Figure 1).[9]

[7] χ^2 (1, N=943) = .13, p=.717, w=.004.
[8] χ^2 (1, N=943) = 7.76, p=.005, w=.033.
[9] χ^2 (2, N=943) = 10.83, p=.004, w=.048.

There is surely something right about the claim that we often rely on strategies of perspective taking when we make moral judgments. However, the fact that people tend to engage in mental simulation when presented with the third-person case cannot *by itself* explain why people are only slightly less likely to judge that they should sacrifice their own lives to save anonymous people than to judge that someone else should sacrifice herself to save some anonymous people. The truly striking fact is that so many people in these two conditions judged that they should engage in altruistic self-sacrifice. Approximately *one-third* of the participants who were presented with this option in each of these two conditions offered such a judgment. However, it would be surprising if the people who offered this response in the context of answering a moral dilemma would actually engage in such radical acts of altruistic self-sacrifice in the real world. Social organisms like ourselves are likely to face strong selection pressures against the motivation to engage in such unbounded acts of radical altruism (Dawkins, 1976; Maynard Smith, 1964). We are not, after all, sacrificial lemmings who always act for the good of the group! Of course, there may be strong social and evolutionary pressures that lead individuals to engage in costly altruistic actions when they benefit close genetic kin, or when there are likely to be reciprocated. However, pure altruism for the sake of altruism is unlikely to have evolved (or remained stable) in our species. Moreover, and perhaps more importantly from the standpoint of normative ethics, Thomson (2008) and Wolf (1982) have offered compelling reasons for thinking that there is little that is worthy of praise in such radically altruistic actions as sacrificing one's life for the good of a small number of anonymous people. A person who engages in such a sacrifice shows such a lack of concern for herself (unless, perhaps, she is a calculating and thoroughly rational act utilitarian) that it would be hard to make sense of, let alone praise, her motivation. This raises the question: why would participants ever judge that radical acts of altruistic self-sacrifice are normatively preferable?

Although biological pressures are likely to militate against engaging in acts of radical altruism, there may be cultural pressures that lead people to offer this judgment, thinking it is the morally appropriate answer. For example, the prominence of discussions of martyrdom and radical altruistic self-sacrifice across a number of religious traditions suggest that religious beliefs may play a crucial role in driving the judgment that a person *should* be willing to sacrifice her life for the greater good. To examine this possibility, we compared the responses of participants who reported some religious affiliation with the responses of participants who reported that they had no religious affiliation. In line with this hypothesis, participants who reported some religious affiliation were substantially more likely to judge that Jesse should engage in altruistic self-sacrifice, and they were less likely to judge that Jesse should let the boxcar continue along the main track.[10] Similarly, participants who reported some religious affiliation were significantly more likely to judge that they should engage in altruistic self-sacrifice themselves.[11]

We also considered the possibility that taking a course in moral philosophy would have an effect on judgments about what should be done in this case, and that

[10] Of the participants who reported *no religious affiliation* 47.1% judged that Jesse should turn the trolley to the left, 31.5% opted for self-sacrifice, and 21.5% judged that Jesse should let the trolley go. Strikingly, participants who reported *some religious affiliation* were less likely to turn the trolley to the left (39.4%), more likely to opt for self-sacrifice (44.3%), and less likely to judge that Jesse should let the trolley go (16.3%). This comparison was highly significant, χ^2 (2, N=845) = 14.87, p=.001, w=.112.

[11] Of the participants who reported *no religious affiliation* 53.1 judged that Jesse should turn the trolley to the left, 28.9% opted for self-sacrifice, and 18.0% judged that they should let the trolley go. Participants who reported *some religious affiliation*, were less likely to turn the trolley to the left (44.5%), more likely to opt for self-sacrifice (37.1%), and *slightly* more likely to let the trolley go (18.4%). Again, this comparison was significant, χ^2 (2, N=943) = 8.13, p=.017, w=.084.

backgrounds in religion and morality might dissociate in some way. Further, it is plausible to think that some moral philosophy courses (or readings in this area) would present some version of the two-track-trolley problem. Thus, when faced with the three-track case, participants who had seen the bystander case previously may be more likely to perceive what more is at stake in the three-track case. However, this analysis failed to reveal any significant difference between the judgments of those participants who had taken a course in moral philosophy, and those participants who had not.[12] On the basis of these results, we suggest that participants who identify with some religion are more likely to judge that they should engage in an act of radical altruistic self-sacrifice. However, training in moral philosophy seems to have little effect on the pattern of judgments that people provide in response to these sorts of dilemmas.

To understand precisely what role such considerations are playing in these judgments, it is also necessary to examine the role that they play in cases where altruistic self-sacrifice is not at issue. So, having examined the role of altruistic self-sacrifice in Conditions 1 and 2, we turned in Condition 3 to another question about the psychological processes that might be in play when people make judgments about moral trilemmas. Here, we examined a 'double-bystander' case in which Jesse had to make a decision about whether to redirect a runaway boxcar to a left side track or a right side track, where a single anonymous person was walking along each of these tracks (once again, there were five people walking along the main track). This case provides a strong test of the hypothesis that when there is nothing to tell in favor of turning the trolley to the left or to the right, the option of letting the trolley go will seem more plausible. The scenario read as follows:

> Jesse is standing near the railroad tracks and notices an empty boxcar coming down the tracks, moving fast enough to kill anyone that it hits. If Jesse does nothing, the boxcar will continue along the main track, killing five people who are walking down the main track. There is a switch nearby that Jesse can use to divert the boxcar onto either of two side tracks that split off from the main track in opposite directions. There is one person walking along the right-side track. So, if Jesse flips the switch to the right, the boxcar will hit and kill this person. There is one person walking along the left-side track. So, if Jesse flips the switch to the right, the boxcar will hit and kill this person. What should Jesse do?

We chose to examine this case because we were interested in the effect of introducing a third, morally significant option on participants' moral judgments, especially in a case where neither of the two proposed options was clearly preferable to the other. Building on the case from Jewish ethics that we raised above, we hypothesized that when people are faced with two equally unappealing choices, a greater proportion will be unwilling to turn the trolley. Since there is no reason to chose one track over the other in a case where only agent-neutral considerations are at play, people who are presented with this case should be less willing to make an arbitrary choice about who is to live and who is to die.

---

[12] In Condition1, the majority of participants in each group judged that Jesse should turn the trolley to the left (no moral course, 42.0%; at least one moral course, 45.6%); many participants opted for self-sacrifice herself (no moral course, 38.3%; at least one moral course, 38.5%); and, some judged that Jesse should let the trolley go (no moral course, 19.7%; at least one moral course, 15.9%); there was no significant difference between these groups, χ^2 (2, N=845) = 1.76, p=.415, w=.037. Similarly, in Condition 2, the majority of participants in each group judged that they should turn the trolley to the left (no moral course, 48.1%; at least one moral course, 47.8%); many participants opted for self-sacrifice (no moral course, 32.9%; at least one moral course, 36.1%); and, some judged that they should let the trolley go (no moral course, 19.0%; at least one moral course, 16.1%); again, there was no significant difference between these groups, χ^2 (2, N=943) = 1.47, p=.479, w=.031.

Although the majority of our participants continued to judge that they should take some action in this case (40.9% judged that Jesse should flip the switch to the left; 29.5% judged that Jesse should flip the switch to the right), a significantly greater proportion of participants (29.6%) judged that Jesse should let the trolley continue along the main track than would be predicted on the basis of the two-track bystander case.[13] Moreover, this was a significantly greater increase than the one that we found in Condition 1 (see Figure 1).[14] However, unlike Conditions 1 and 2, where considerations of altruistic self sacrifice were in play, there were no significant differences between the judgments that were offered by people who reported some religious affiliation and participants who reported no religious affiliation.[15] Similarly, having taken a course in moral philosophy had no significant effect on the judgments that were offered.[16] These data provide some evidence for the claim that the introduction of an *arbitrary* third option leads to an increased proportion of participants who are unwilling to flip the switch, but with potentially relevant cultural background (i.e., religion, moral coursework) playing no role.

In Condition 4 we returned to the question of whether the mere presence of a third option for action would have a significant effect on folk-moral judgment. To examine this question, we developed a final scenario that introduced only a slight modification to the wording used in Condition 3. Rather than presenting two *occupied* side tracks, this trilemma included the clearly preferable option of diverting a runaway boxcar onto an unoccupied track, essentially reducing the scenario to a bystander case with an easy, no-cost alternative. Unsurprisingly, almost everyone who read this case judged that Jesse should flip the switch to the left (93.2%), with only a small minority (3.8%) refusing to intervene, or judging that Jesse should flip the switch onto the occupied sidetrack (3.0%).[17]

4. CLASHES OF INTUITION

Thomson's (2008) purported refutation of the trolley problem builds on a principle that has formed the core of contemporary moral theory: "To reason in accord with the dictates of morality is to view oneself as unexceptional" (Gendler, 2007, p. 80). Thomson (Thomson, 2008, p. 364) hopes that by laying bare the reasoning that would lead the bystander to judge it permissible to turn the trolley she can show that such actions are transparently immoral. After all, although it would be good to save the people on the main track, a person who intentionally kills another because she is unwilling to pay the cost of this good has clearly done something wrong. The data reported in the previous section minimally suggests that Thomson's intuition is radically at odds with the commonsense intuition of what a person should do when she is faced with a difficult moral trilemma such as the three-track bystander case. If Thomson were right, we would expect it to be far more *transparent* that it is immoral to turn the trolley, and this increased transparency should be reflected, at least to some extent, in the folk-moral

[13] χ^2 (1, N=709) = 118.85, p<.0001, w=.146.

[14] χ^2 (1, N=709) = 55.60, p<.0001, w=.109.

[15] χ^2 (2, N=709) =.331, p=.848, w=.084.

[16] χ^2 (2, N=709) = .864, p=.650, w=.032.

[17] The proportion of people who judged that they should do nothing in this case was much smaller than the number who judge that they should do nothing in the familiar two-track bystander case, χ^2 (1, N=769) = 112.73, p<.0001, w=.116. There was no significant effect of religion [χ^2 (1, N=769) = 4.53, p=.104, w=.003] nor taking a moral course [χ^2 (1, N=769) =.34, p=.843, w=.008] on this judgment. However, 23 participants did judge that they should flip the switch onto the occupied side track. We suppose that these people failed to read the scenario carefully, made a mistake, or for some other reason made a choice that they would not reflectively avow.

judgments that are offered in response to this case. Even if not everyone was sensitive to the hypocrisy of turning the trolley, we would expect a striking increase in the proportion of people who judge that one should just let the trolley go in such a case. In stark contrast to this hypothesis, we found no such pattern in folk-moral judgments about the three-track bystander case. Across both conditions in which altruistic self-sacrifice was at issue, only a small minority of participants (approximately 18%) judged that they should let the trolley go, suggesting that even when people are explicitly presented with the option of altruistic self-sacrifice, the inclination to save the five people on the main track continues to dominate their intuitive judgments.

It is important to note, however, that including the option of altruistic self-sacrifice *did* elicit a slight increase in the proportion of people who judged that they should let the trolley go. However, this shift was remarkably small (3.3-3.7%, $w<.04$) and not of obvious theoretical import. Perhaps a small proportion of people reason as Thomson expects, focusing on the negative duty not to harm another person without her consent when presented with the option of altruistic self-sacrifice. Alternatively, this shift could indicate a performance error that arises as a result of the increased complexity of the case relative to the two-track bystander. That is, some participants may not attend to all of the relevant options and may thus make judgments that they would not avow if they thought more carefully through the case. This interpretation is bolstered by the fact that approximately three-percent of participants presented with the control scenario also judged that they should sacrifice the life of one person when they had a cost-free alternative. Regardless of which interpretation is correct, this small but significant shift in judgments may have important pragmatic implications. If a jury were faced with a case where one person harms another as a foreseen consequence of her action, and where she could have sacrificed herself instead, the presence of a single person whose intuition diverged in this way could make the difference between a unanimous decision and a hung jury. Given the pragmatic import of such data, we suggest that further studies on the folk-morality of altruistic self-sacrifice should be carried out to examine the mechanisms at play in folk-moral judgments about altruistic self-sacrifice. However, it is clear from our data that this small shift in judgments is not sufficient to justify the claim that it is *transparently* obvious that one should let the trolley go in the three-track bystander case.

Turning to the responses that were offered in the three-track 'double-bystander' case lead us to an even clearer understanding of why the principles that are operative in folk-morality yield a pattern of responses that conflicts with the predictions of Thomson's theory. Here, there was a theoretically and statistically significant increase in the proportion of people who judged that they should just let the trolley go. In fact, the proportion of people who offered this judgment was twice as large as was predicted on the basis of the familiar two-track bystander case. This result suggests that the extent to which self-regarding reasons come into play in evaluating a proposed action plays an important role in determining whether people will judge that they should let the trolley continue along the main track. Where there is nothing to speak in favor of adopting one course of action over another, people are more likely to judge that they should just let the trolley go. However, where some consideration speaks in favor of adopting one option rather than another (e.g., the self-regarding reasons that are generated by the impulse toward self-preservation), the choice to divert the trolley onto an anonymous person seems far more justifiable. This also sits well with our earlier suggestion that Thomson's intuition in the three-track bystander case goes through only on the assumption that there are no morally significant self-regarding reasons available in that case.

Our data also raise an important difficulty for using psychological results as evidence against philosophical moral theories. Where there was no morally salient feature that

could tell in favor of adopting one option over another, people relied on a morally irrelevant feature of the case in making their judgment, revealing a pronounced preference for the first option that was presented. This preference for the first option presented is a massively robust effect in research on judgment and decision making, including studies of rats, pigeons, monkeys and humans (Atkinson & Shiffrin, 1968; Insko, 1954; Modigliani & Hedges, 1987; Pineño & Miller, 2005; Waugh & Norman, 1965; Wright, Santiago, Sands, Kendrick, & Cook, 1985). It is well known that people sometimes rely on morally irrelevant factors in making moral judgments, and this fact has played a prominent role in recent philosophical thinking about the evidential status of folk-moral judgments. So, the fact that we find this pattern of judgments in our data raises a truly difficult question: when does a dominant folk-moral intuition provide a reason to modify or abandon a conscious, deliberative philosophical moral theory. Faced with the results of interdisciplinary research on moral psychology, there are two extreme responses that immediately spring to mind—both of which are ultimately untenable.

First, it might be suggested that the presence of a dominant folk-moral intuition that conflicts with a moral theory *always* provides a reason for revising or abandoning that theory, even where it is driven by some morally irrelevant factor. This response is obviously untenable because it requires that we assume that folk-morality cannot be wrong. While there may be ways of arguing for the claim that there are no moral facts-of-the-matter, we believe that this conclusion should not be the starting point for evaluating the status of folk-moral judgments.

Second, and more plausibly, it might be suggested that the presence of a dominant folk-moral intuition that conflicts with a moral theory *never* provides a reason for revising or abandoning that theory. Thomson, for example, is likely to see the self-regarding reasons evoked by considerations of self-interest as morally irrelevant (at least in the three-track bystander case), and therefore, as normatively on a par with the decision to flip the switch to the left because it was the first option presented. This response underwrites much of the mainstream skepticism about experimental philosophy and the interdisciplinary study of empirical moral psychology. It is often suggested that responses collected in moral psychology experiments are of an entirely different sort than the intuitions required to support a philosophical theory. Philosophical theories rely on more 'robust intuitions' that require philosophical dialogue and reflection (Kauppinen, 2007). But should we rest content with the claim that reflective philosophical intuitions count as evidence for the truth of a moral theory? Perhaps some folk-moral intuitions are epistemically suspect because they are the result of selection pressures that would not lead us to make moral judgments that we would reflectively avow. However, it is also possible that reflective philosophical theorizing can lead us to assume the truth of dominant ideological prejudices rather than uncovering genuinely philosophical truths. We thus should take care to examine the theoretical principles that are operative in producing reflective philosophical intuitions as well.

The principles that lead to Thomson's reflective philosophical view about the three-track bystander case are made clear when she turns to a second thought experiment. She asks her readers to imagine a case where she doesn't want to send her own money to Oxfam, so she steals someone else's money to send instead. Thomson rightly concludes that this action is 'pretty bad', and we assume that empirical tests of folk intuition would provide resounding confirmation. But, since this thought experiment is intended to show that her position on the bystander case is intuitively obvious she also argues that "if the bystander proceeds to turn the trolley onto the one on the right-hand track in Bystander's Three Options, then what he does is markedly worse, because the cost in Bystander's Three Options isn't money, it is life" (Thomson 2008, p.365). Unfortunately, the analogy

between these two cases is far from apparent. Unlike her intuition in the Oxfam case, Thomson's considered position on the bystander requires that she abandon a number of theoretical principles that we have antecedent reason to accept. It requires, for example, that she ignore self-regarding reasons that may be at play in this case. It requires that she believe that even when a person has full knowledge of outcomes, and can easily act to divert a threat, deciding not to act should not be seen as something that an agent has done intentionally. Moreover, her view makes it difficult to draw a distinction between intentionally harming someone as a means to some greater good and harming someone as a foreseen side effect of bringing about a greater good. But, this distinction garners much support from *both* theoretical arguments and folk-moral intuitions. As Tamar Gendler (2007) has argued, there is a strong tendency to read philosophical thought experiments in light of antecedent commitments that may or may not have bearing on the case that is under consideration. For this reason, we worry that if Thomson's argument about the three-track bystander succeeds, it is likely to do so because it evokes a response that relies on these elicit assumptions, driving those who adopt a view like Thomson's "(either reflectively or unreflectively) to represent relevant non-thought experimental content in light of the thought experimental conclusion" (Gendler, 2007, p. 69).

In light of the difficulties faced by these extreme views on the relation between reflective philosophical intuitions and folk-moral intuitions, we offer a more moderate response: the presence of a dominant folk-moral intuition that conflicts with a philosophical moral theory *always provides* a defeasible reason for revising or abandoning that theory; but, since folk-morality can itself be radically mistaken, folk-moral intuitions must be triangulated against reflective philosophical intuitions as well as additional empirical evidence about conditions under which we are likely to rely on irrelevant heuristics and biases (Gigerenzer, 2002; Tversky & Kahneman, 1974). In this endeavor:

> It would be no less of a mistake to think that the resources of commonsense psychology will tell us everything we need to know about, say, the concept of rationality than it would be to think that rationality can be completely understood through empirical studies of people's reasoning habits (Bermudez, 2005, p. 10).

This being the case, where a dominant folk-moral intuition contradicts a philosophical theory, this intuition must be explained (or explained away)—and this requires philosophers to take notice of the patterns of judgments that people actually make in response to philosophical thought experiments. The dominance of a folk-moral intuition does not, of course, rule out the possibility that most people are relying on theoretically misguided considerations; however, it does provide defeasible evidence against that theory.

As we argued above, Thomson's intuition derives much of its force from a set of theoretically problematic presuppositions that she has brought to the table when she considers the relevant thought experiment. Thomson could, of course, mitigate this worry by explaining why folk-morality includes this mistaken intuition about the three-track bystander case. However, given that there was no significant difference in the proportion of people who judged that *they* should let the trolley go compared to the proportion of people who judged that Jesse should let the trolley go, the explanation that she has offered for the two-track bystander case is unlikely to apply. Considerations of number are surely at play (cf., Thomson 2008, p. 373), but it does not seem to be the drastic nature of the action that drives people to focus on the number of lives at stake as opposed to some other set of features (*pace* Thomson 2008, p. 374). On the basis of this argument,

we suggest that the trolley problem remains a problem—and it's a problem that is likely to require real interdisciplinary collaboration to solve. To make this point clear, we suggest that although Thomson has not solved the bystander case with the appeal to a three-track alternative to the familiar bystander case, she has raised a serious issue that calls for a careful analysis of both the folk-moral intuitions and philosophical theories that can be deployed in examining a case such as this. Our hope in this paper is to have suggested one way in which this dialog can proceed. A close attention to both the philosophical issues and the empirical data can lead us to a more promising analysis of the reasons for action in a case where altruistic self-sacrifice is seen as a viable option.

5. REFERENCES:

Atkinson, R., & Shiffrin, R. (1968). Human memory: A proposed system and its control processes. In K. Spence & J. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89-105). New York: Academic Press.

Bermudez, J. (2005). *Philosophy of Psychology*. New York: Routledge.

Cushman, F. (2008). Crime and Punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353-380.

Cushman, F., Young, L., & Hauser, M. (2006). The Role of Reasoning and Intuition in Moral Judgments:Testing three principles of harm *Psychological science, 17*.

Dawkins, R. (1976). *The Selfish Gene*. Oxford: Oxford University Press.

Elster, J. (2007). *Explaining social behavior*. Cambridge: Cambridge University Press.

Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review, 5*, 5-15.

Foot, P. (1985). Morality, action, and outcome. In T. Honderich (Ed.), *Morality and objectivity* (pp. 23-38). London: Routledge.

Gendler, T. (2007). Philosophical Thought Experiments, Intuitions, and Cognitive Equilibrium. *Midwest studies in philosophy, 31*, 68-89.

Gigerenzer, G. (2002). *Adaptive thinking*. Oxford: Oxford University Press.

Greene, J., Cushman, F., L., S., Lowenberg, K., Nystrom, L., & Cohen, J. (In Press). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*.

Greene, J., Nystrom, L., Engell, A., Darley, J., & Cohen, J. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(389-400).

Hauser, M., Cushman, F., Young, L., Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind and Language, 22*, 1-21.

Hauser, M., Young, L., & Cushman, F. (2007). Reviving Rawl's linguistic analogy. In W. Sinnott-Armstrong (Ed.), *Moral Psychology* (Vol. 1: The evolution of morality). Cambridge, MA: MIT Press.

Hohmann, J. (2008, 03 June). Soldier who threw himself on grenade to save others is recognized. *Los Angeles Times*,

Insko, C. (1954). Primacy versus recency in persuasion as a function of the timing of arguments and measures. *Journal of Abnormal and Social Psychology, 69*, 381-391.

Kamm, F. (2005). *Intricate Ethics*. Oxford: Oxford University Press.

Kauppinen, A. (2007). The rise and fall of experimental philosophy. *Philosophical explorations, 10*, 119-122.

Maynard Smith, J. (1964). Group Selection and Kin Selection. *Nature, 201*, 1145-1147.

Mikhail, J. (2007). Universal Moral Grammar. *Trends in cognitive science, 11*, 143-152.

Modigliani, V., & Hedges, D. (1987). Distributed rehearsals and the primacy effort in single-trial free recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 13*, 426-436.
Moore, A., Clark, B., & Kane, M. (2008). Who shall not kill? Individual Differences in Working Memory Capacity, Executive Control, and Moral Judgment. *Psychological science, 19*(6), 549-557.
Nagel, T. (1986). *The view from nowhere*. Oxford: Oxford University Press.
Nosek, B., Banaji, M., & Greenwald, A. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group dynamics, 6*(1), 101-115.
Parfit, D. (1984). *Reasons and persons*. Oxford: Clarendon Press.
Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of personality and social psychology, 64*(3), 467-478.
Pettit, P. (1987). Universality Without Utilitarianism. *Mind, 72*, 74-82.
Pineño, O., & Miller, R. (2005). Primacy and recency effects in extinction and latent inhibition: A selective review with implications for models of learning. *Behavioural Processes, 69*, 223-235.
Pizarro, D., Uhlmann, E., Tannenbaum, D., & Ditto, P. (in prep). The motivated use of moral principles.
Quinn, W. (1989). Actions, Intentions, and consequences. *The philosophical review, 48*(3), 287-312.
Ridge, M. (2001). Agent-Neutral Consequentialism From the Inside-Out. *Utilitas, 13*, 236-254.
Schwartz, B. (2004). *Paradox of Choice*. New York: Ecco.
Sidgwick, H. (1907). *The methods of ethics*. London: Macmillan.
Sinnott-Armstrong, W., Mallon, R., Hull, J., & McCoy, T. (forthcoming). Intention, Temporal Order, and Moral Judgments. *Mind and language*.
Thomson, J. (1985). The Trolley Problem. *Yale Law Journal, 94*, 1395-1415.
Thomson, J. (2008). Turning the Trolley. *Philosophy and public affairs, 36*(4), 359-374.
Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. . *Science, 185*, 1124-1130.
Tyson, A. S. (2008, 1 April). SEAL Killed in Iraq To Get Medal of Honor. *The Washington Post,* p. A04,
Unger, P. (1996). *Living high and letting die*. Oxford: Oxford University Press.
Urmson, J. O. (1958). Saints and Heroes. In A. Melden (Ed.), *Essays in Moral Philosophy*. Seattle: University of Washington Press.
Valdesolo, P., & DeSteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. *Psychological science, 17*, 476-477.
Waldmann, M., & Dietrich, J. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological science, 18*(3), 247-253.
Waugh, N., & Norman, D. (1965). Primacy memory. *Psychological Review, 72*, 89-104.
Wolf, S. (1982). Moral saints. *Journal of Philosophy, 79*(8), 419-439.
Wright, A., Santiago, H., Sands, S., Kendrick, D., & Cook, R. (1985). Memory processing of serial lists by pigeons, monkeys, and people. *Science, 229*, 287-289.