Measuring Consequentialism: Trolley Cases and Sample Bias

Amy Lara, Kansas State University [1,2,3]
Michael Smith, Kansas State University [2,3]
Scott Tanona, Kansas State University [2,4]
Bill Schenck-Hamlin, Kansas State University [2]
Ron Downey, Kansas State University [2]
Bruce Glymour, Kansas State University [1,2,3]

[1]Paper writing
[2]Survey design
[3]Data analysis
[4]Paper editing

Contact author: Amy Lara, Philosophy Department, 201 Dickens Hall, Kansas State University, Manhattan, KS, 66506. (785) 532-0357. alara@ksu.edu.

Acknowledgments:
This project was funded by a National Science Foundation grant (0734922). The authors wish to thank David Rintoul, Larry Weaver, Maria Glymour, and Jon Mahoney for valuable assistance.

Measuring Consequentialism: Trolley Cases and Sample Bias

Abstract: We use data from a recent survey on communication norms among scientists to argue that standard Trolley and Sophie's Choice scenarios induce subjects to drop from studies, generating loss of data and sample bias. We explain why these effects make instruments based on such scenarios unreliable measures of the general deontic versus consequentialist commitments of study subjects.

I. Moral Epistemology and Empirical Data.

Since the seminal work by Harman (1999) and Greene et al. (2001), empirical data has been of increasing importance in moral philosophy, and several important studies surveying the range and variation of moral intuitions have been done (Nichols, 2004; Nichols and Mallon, 2006; Hauser et al., 2007). Largely, however, the import of this work has been critical rather than positive. To the extent that the moral intuitions of professional philosophers are not widely shared outside the confines of academic philosophy, the method employed by most of ethics, reflective equilibrium, is suspect, providing at best a check on consistency. Unless the intuitions of professional ethicists can be grounded as especially truth conducive, arguments for moral claims which rely on such intuitions are bad. But for all that this lesson is salutary, it does little to further the aim of moral theory: if the search for reflective equilibria, grounded in the considered moral judgments of ethicists, is bad method, what methods might be better, and to what theoretical ends might they be reliably deployed?

Ethical theory has several aims. Nominally, of course, the central aim is the discovery of a correct ethical theory, or at least the discovery of core features any correct theory must have. But the interest in ethical theory is not merely theoretical—one wants to know the correct moral theory, in part, because one wishes guidance, and that in at least two respects. First, when confronted with an ethical dilemma, one wants a principled resolution, and preferably one that appeals to true moral principles. Secondly, one wants to know how to train others so as to become ethical, i.e. to engage systematically in ethical behavior. Though the latter goal is at best secondary in the modern philosophical cannon, it is the original goal of philosophy itself. As Aristotle says, knowledge of the good is "of great importance for the conduct of our lives" and the goal of ethical study is "action, not knowledge" (*Nicomachean Ethics*, 1094a and 1095a)

Developing methods which allow reliable inference to true moral theory is non-trivial, and, as ever so many have argued (Kant, 1785; Moore, 1903; Frankena, 1939), it may be that the methods of empirical psychology and empirical philosophy simply cannot be brought to bear on questions of theoretical truth in ethics. We take no stand on that issue here. Matters are rather different, however, with respect to training. Whatever the true ethical theory may turn out to be, questions about how beliefs *about* ethical truths can be modified and questions about what behaviors such beliefs influence or can be brought to influence *are* fully empirical. If empirical philosophy and moral psychology are to provide a *positive* rather than merely corrective contribution to moral philosophy, the most likely path to such a contribution is just here. What moral beliefs influence which kinds of behavior? Are the beliefs conscious, or at least can some explicit formulation of them be elicited, or are they unconscious, reflected only in our

dispositions to judge cases in this way rather than that? In either case, what kinds of interventions will change the beliefs in ways that modify behavior? Is better behavior a matter of better reasoning, or reducing ignorance, or changing the most basic principles adopted by students, or changing the local principles they adopt, or changing something else entirely, perhaps their attitudes or the aims they pursue?

To answer these questions using observational data or data gained from survey instruments, we require methods for reliably determining the moral commitments of subjects, whether these commitments take the form of consciously recognized principles or unconscious dispositions. In particular, it is important to know whether the cognitive commitments or judgmental dispositions had by subjects are best accommodated by consequentialist or deontic theories. This is partly a consequence of the need to connect moral psychology with ethical theory. But it is also a concomitant of the desire to intervene to change behavioral dispositions. For example, if dispositions to issue ethical judgments are grounded in conscious inference from moderately high level normative principles, interventions on beliefs about those principles are a likely place to focus training. If, conversely, conscious commitments to normative principles are constructed post-hoc from prior judgments about particular cases (Haidt, 2001), ethical training might better focus on quite different features guiding normative behavior. In either case, it is essential to develop good measures of the highest level moral principles endorsed, explicitly or implicitly, by subjects, for that is a necessary preliminary step in determining the relation between commitments to quite general ethical principles and judgments about particular cases.

Our focus in this essay is the reliability of current methods to measure normative commitments at the most general level. We are specifically interested in the use of Trolley and Sophie's Choice scenarios to elicit judgments about proper behavior from subjects, which judgments are then used to scale the extent to which subjects are more or less committed to deontic or consequentialist theoretical principles. Trolley cases and the kindred Sophie's Choice cases have for many years been the standard test for consequentialist commitments among academic philosophers (Foot, 1967; Thomson, 1976; Quinn, 1989), and their use, both illustrative and diagnostic, in introductory ethics courses is endemic. Several studies have attempted to develop survey instruments employing such scenarios in order to diagnose the most general normative commitments of subjects (Nichols and Mallon, 2006; Hauser et al., 2007). We argue here that such instruments are methodologically unsound. We present statistical evidence that Trolley and Sophie's Choice questions induce loss of data by causing subjects to drop out of studies and we present statistical and anecdotal evidence that those with deontic commitments are more likely to drop out, inducing sample bias. Each difficulty is in itself serious; together they are damning. We argue that alternative methods for assessing commitments to deontic versus consequentialist principles are required.

II. Study Design and Methods.

In 2007 we began a study designed to identify the extent to which scientists' knowledge of various cognitive biases (e.g. confirmation and assimilation bias) influenced their views about the propriety of so-called 'framing' in communications with the general public. As part of that project we constructed a 63-item survey that was

distributed by email to 987 faculty and graduate students at three large state universities in the Midwest, during the spring of 2008. The survey questions covered several demographic variables, and seven content areas: awareness of cognitive biases (we refer to these questions as BIAS variables, 7 instruments), goals of scientific communication (AIM variables, 6 instruments); beliefs about the characteristics of lay audiences (AUD variables, 9 instruments); beliefs about the effectiveness of framing (EFF variables, 12); beliefs about the local norms governing scientific communication (CN variables, 9 instruments); assessments of the appropriateness of specific communication strategies (BEH variables, 13 instruments); and moral predispositions, as judged by responses to Trolley and Sophie's Choice scenarios (MT variables, 5 instruments).[1] Responses to all content instruments were on a seven point Likert scale.

For the MT questions, we used five moral dilemmas, following wording used by Greene et al. (2001).[2] Respondents were asked to evaluate the appropriateness of an action, on a seven point Likert scale, from very inappropriate to very appropriate. MT1 was the classic trolley case involving throwing a large man off of a footbridge in order to stop a runaway trolley that would otherwise kill five people. MT2 was the other classic trolley case, in which one can press a switch to divert a runaway trolley so that it kills only one person on the tracks instead of five. MT3 was a familiar Sophie's Choice case, where one needs to smother a crying baby to prevent a group of villagers from being discovered by enemy soldiers. MT5 asked the respondent to imagine that his/her family is captured, and the only way to prevent the entire family from being killed is to kill one

[1] Responses were given on-line, and were anonymous. Informed consent was obtained. Both the informed consent document and the survey itself were approved by the [redacted for review] IRB. The survey itself can be found on-line at [redacted for review]

[2] See appendix for exact versions of the instruments.

of his/her own children. MT4 asked whether one would find it appropriate to destroy a very valuable sculpture, owned by an art collector, in order to stop a runaway trolley from hitting five people.

MT1 and MT2 together can be used to test for a subject's commitment to the doctrine of double effect (Hauser et al., 2007), which is generally held to be a deontic principle. MT3 and MT5 can each be used to test for a slightly different deontic intuition: the distinction between doing and allowing. MT4 can be used with MT1 to test for personal versus impersonal intuitions such as those concerned with active killing (Greene et al., 2001). The MT variables were distributed as below in Figure 1, and the inter-correlations are given in table 1.

Fig. 1

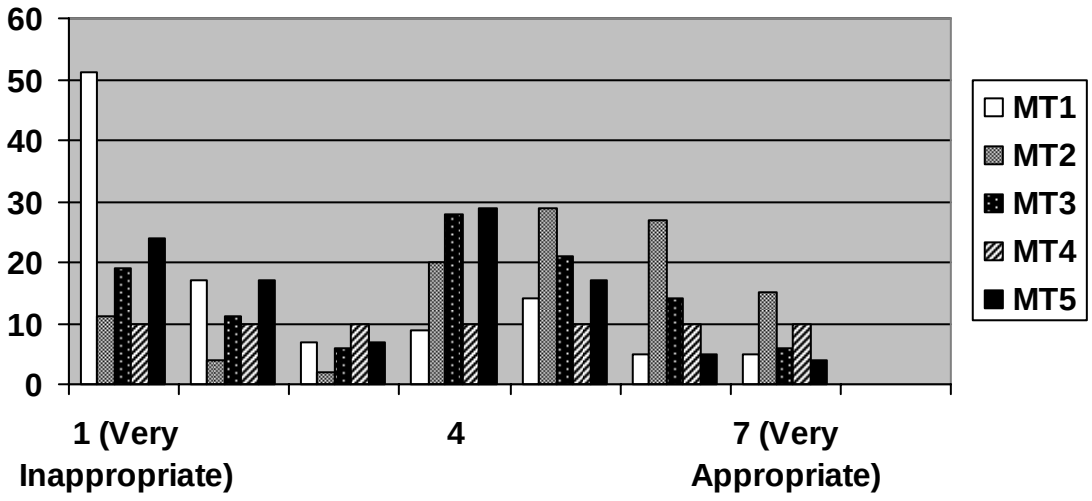

Table 1

Correlations

		MT1	MT2	MT3	MT4	MT5
MT1	Pearson Correlation	1	.382**	.383**	.094	.524**
MT2	Pearson Correlation	.382**	1	.465**	.080	.407**
MT3	Pearson Correlation	.383**	.465**	1	.021	.691**
MT4	Pearson Correlation	.094	.080	.021	1	-.086
MT5	Pearson Correlation	.524**	.407**	.691**	-.086	1

**. Correlation is significant at the 0.01 level (2-tailed).

The response rate was 16.4% (162 of 987). Of the 162 responses, 111 surveys were completed (11 % return rate). Of the 111 completed surveys, a further 8 subjects (7.2%) omitted responses to one or more MT instruments.

III. Analysis of Drop-Out and Sample Bias.

A large number of respondents stopped taking the survey when they came to the moral dilemma questions, or skipped those questions and completed the rest of the survey. Of 162 people who started taking the survey, 51 did not complete it. Of those 51, 20 (39%) stopped taking the survey at the moral theory section. To test the significance of data loss from MT questions, we calculated the relative risk of dropping out at questions in the particular content areas. Each subject who failed to complete the survey began omitting responses at some point in the survey. A question was scored as the 'drop out' point, or point of first omission, for a subject, if a response to that question was omitted, and began a string of two or more omissions in sequence. The relative risk of dropping out at an MT question relative to other questions is 7.4 (chi-square test,

p<<.001), i.e. subjects are 7.4 times more likely to drop at an MT question than to drop at a non-MT question. Drop rates by area are reported in table 3.

Area	MT	BEH	BIAS	CN	EFF	AIM	AUD	Demographic variables
Number Dropping	20	11	4	0	10	3	2	1

The bias in omission rates was also evident among completed surveys. 8 subjects omitted responses to one or more of the 5 MT questions, a significant departure from the average rate of omission. Table 3 reports the relevant statistics for omitted variables, by content area, among completed surveys. The relative risk of omission is 5, and is significant (chi-square, p<<.001).

Table 3

VAR	MT	BEH	BAIS	CN	EFF	AIM	AUD
Pr(no omission in area)	.928	.937	.973	.973	.991	.991	1
PR(at least one omission in area)	.072	.063	.027	.027	.009	.009	0

This loss of data at MT questions, in both completed and uncompleted surveys, cannot be accounted for either by the number of such questions, which was relatively low (5 questions; other areas included 6 to 13 questions), or by the location of the MT questions in the survey—these questions occur early, as questions 11 through 15.

There are likely several reasons that questions about Trolley and Sophie's Choice cases occasion loss of data. For one thing, such cases are far removed from the workaday ethical dilemmas about which subjects are most likely to have clear principles or fixed dispositions. This makes the questions difficult and time consuming to answer. To the extent that this is true, there is doubly good reason to avoid Trolley and Sophie's Choice scenarios. Not only do they induce loss of data, when answered by persevering subjects they generate *misleading* data. Whatever moral commitments, conscious or dispositional, underwrite their judgments and behaviors in the workaday world of interest, those commitments and dispositions do not of themselves resolve the dilemmas posed by Trolley and Sophie's Choice scenarios. Hence, measurements on such resolutions are *confounded* measures of the commitments or dispositions of interest—namely those which influence everyday behavior. Since it is just these commitments or dispositions which we wish to modify by training, it is just these dispositions we wish to measure.

There is also strong anecdotal evidence that Trolley and Sophie's Choice scenarios are simply offensive—subjects don't answer these questions because even thinking about them is morally unpleasant. We quote (with permission, but without attribution to preserve anonymity) from a subject who latter e-mailed us with a complaint:

> "If you must include such horrific hypothetical questions for some legitimate reason, at least give some reasonable choices for answers. Many people when faced with an horrific situation will try to intervene by putting themselves at risk. I can think of no one except a mentally ill person who would be willing to play your numbers game with other people's lives….For those

reasons I refused to answer several of the most egregious questions on that page which I felt allowed no sane answer at all."

The nature of the above quoted complaint suggests that the drop out generated by Sophie's Choice and Trolley cases is non-random. Those most likely to find the dilemmas especially difficult to resolve are also those most likely to be committed to various principles most comfortably at home in deontic theories, e.g. double effect, firm doing/allowing distinctions and strict injunctions against active killing. If data loss is bad empirically for merely pragmatic reasons, sample bias is methodologically dangerous. If the degree to which a subject is deontically oriented influences inclusion in the sample, this can induce associations between *other variables* which causally influence deontic commitments but not each other.

To illustrate, suppose commitments to general principles, and hence the degree to which a subject is assessed as deontic or consequentialist, are constructed post-hoc to justify responses to more local questions about communicative norms (how important is honesty versus effectiveness, for example) and judgments about particular behaviors (how appropriate is this or that particular case of framing?). That is, suppose that an MT variable is caused by a CN variable and also by a BEH variable. If that MT variable *also* causes inclusion or exclusion from the sample, the sample data effectively condition (control for) particular values of the MT variable. But by controlling for a common effect of two otherwise independent causes, we will induce an association between them: in our data, CN and BEH will be associated, and that association will likely underwrite a fallacious inference to the effect that local norms cause assessments (see Spirtes,

Glymour and Scheines, 2000 or Pearl, 2000 for discussion; see Glymour et al., 2008 for discussion of the implications in social science).

To the extent that we are interested in discovering how to intervene on students (i.e. train them) in ways that influence their moral judgments, it is important to learn which kinds of moral commitments cause or are caused by dispositions to judge particular behaviors morally acceptable or morally unacceptable. Those causal inferences require joint measurements of variables tracking general moral commitments, local moral commitments and particular moral judgments. To the extent that instruments for measuring general moral commitments induce data loss or sample bias, they threaten to seriously compromise causal inferences. Survey instruments defined with respect to Trolley and Sophie's Choice cases appear to generate exactly these threats.

IV. Discussion.

Measuring the degree to which subjects are committed, explicitly or implicitly, to deontic or consequentialist principles is important, first in order to adequately represent the kind of moral commitments subjects have, second in order to discover the causal dependencies between various commitments, beliefs, dispositions and behaviors, and third in order to identify variables on which one might intervene to change the behavior, or anyway the considered moral judgments, of students. It has now become fairly standard in the philosophical literature to use Trolley Case scenarios to elicit moral judgments (Greene et. al. 2001, Nichols and Mallon 2006; Hauser et al. 2007). The idea of using such judgments to scale the degree to which subjects are deontic or consequentialist is appealing, and there is some prima facie justification in so doing.

As indicated above, Trolley cases can be used to elicit a standard deontological intuition from respondents: the doctrine of double effect. In particular, it has been found that many people are willing to divert a runaway trolley from a track on which five people are standing onto a track on which only one person is standing; yet, most people would not be willing to throw a large person onto a track to stop a runaway trolley from killing five people (Nichols and Mallon 2006; Hauser et al., 2007). Purely consequentialist reasoning would presumably find both situations morally equivalent. By contrast, deontologists have traditionally drawn a strong moral distinction between killing someone as a means to prevent more killings and performing a life-preserving action that will foreseeably result in another death as a side effect of the action. It is plausible to think that those who draw such a distinction, either consciously or intuitively, are as a consequence more likely to judge these two trolley cases differently.

Another important intuition can be elicited by "Sophie's Choice" cases. Like the Trolley cases, these are moral dilemmas in which one must decide whether to harm someone in order to prevent more harms. The difference, though, is that if one does not commit the harm, the victim will be harmed anyway. A classic case is one in which a group of people are hiding from enemy soldiers, and one's baby begins to cry. If one does not smother and kill the baby, the whole group will be discovered and killed. Here the consequentialist view seems the most reasonable: since the victim is going to die anyway, why not perform the killing oneself and prevent further deaths? Even Bernard Williams, when discussing a similar case in his classic critique of utilitarianism, admits that the consequentialist's answer here is probably right, though he is very critical of the reasoning the consequentialist uses to reach that answer (1963, 117). In order to resist

the consequentialist's answer to this case, one would have to have a strong intuition that corresponds to the deontologist's distinction between doing and allowing: the idea that there is a moral difference between allowing an immoral action to take place, and performing that same immoral action oneself. This distinction is at the root of agent-centered restrictions, which prohibit performing certain kinds of actions, even if performing them is necessary to prevent more of those same types of actions from being performed by others. Not all deontologists endorse these restrictions (McNaughton and Rawling, 2007), but many do (Kant, 1799; Nagel, 1972).

Notwithstanding the appeal of Trolley and Sophie's Choice cases, and their traditional role in philosophical ethics, our survey results strongly indicate use of these moral dilemmas is problematic. There is both empirical and anecdotal reason to think that such scenarios induce data loss and sample bias, and are at best confounded or biased measures of the more local principles implicated in everyday behavior.

A significant number of respondents stopped taking the survey when they came to the moral dilemma questions, or they skipped those questions and completed the rest of the survey. The difference in the rate of first omission among the MT questions and the rate of first omission among other content questions is highly significant. Study results therefore provide good reason to think that Trolley and Sophie's Choice cases induce loss of data.

At the end of the survey, respondents were invited to contact the survey administrators with further questions or comments. One person who completed the survey did so, and she raised very strong objections to the inclusion of the moral dilemmas in the survey. She described the scenarios as "horrific" and argued that there

was “no sane answer” to the dilemmas; thus, she refused to answer the “most egregious” of the dilemmas and said that their inclusion in the survey left her with “a very negative feeling about [the] whole project.” One might wonder why the available response of “very inappropriate” would not count as a sane answer to the proposed actions in these dilemmas. A statement made by this person is telling: “I can think of no one except a mentally ill person who would be willing to play your numbers game with other people’s lives.” In other words, even to ask about the appropriateness of trading lives off against each other is already to commit a moral error.

We have here anecdotal evidence for sample bias. Our respondent’s commitments here are, to a first approximation, deontic. Insofar as those with deontic commitments are differentially likely to leave instruments unanswered, or to drop from the study entirely, there is a great risk of spurious associations appearing in the data. In such a situation, one is effectively conditioning on a common effect of measured variables. This will induce associations between measured variables even when they neither cause one another nor share a common cause. Equally important, even when the causes of a common effect *do* share some causal connection, directly or through some common cause, conditioning on the common effect will lead to misidentification (c.f. Spirtes, Glymour and Scheines, 2000). That is, one will incorrectly estimate the influence of causes on their effects, often quite significantly.

To the extent that the greater stop rate for MT questions is better explained by the fact that they are simply more demanding in that they require respondents to *develop* a principled view that they do not already have, either explicitly or implicitly as a disposition, instruments built around Trolley and Sophie’s Choice cases are confounded

or biased measures of commitments and dispositions of interest. We need to measure the dispositions and or commitments which cause *everyday* behavior. If these workaday dispositions are insufficient to generate responses to Trolley and Sophie's Choice cases, then the responses themselves have a basis in at least some other source. The responses are therefore biased or confounded measures of the commitments or dispositions of interest.

Equally problematic, however, is the effect the moral dilemma questions may have had even on those who did complete the survey. Tetlock (2003) has shown that the mere contemplation of tragic trade-offs can lead to a feeling of contamination, and subsequent "cleansing" actions, such as a willingness to volunteer or make contributions to charity. It is possible that our moral dilemma questions created a similar effect in respondents and influenced the way they answered subsequent questions in the survey, especially questions about normative constraints on communication. These are serious disadvantages to using Trolley cases and Sophie's Choice cases in instruments that attempt to measure the relation between moral attitudes and other behaviors and attitudes.

If Trolley and Sophie's Choice scenarios generate poor measures of the general consequentialist versus deontic commitments of subjects, what alternatives are there? The most centrally troubling feature of Trolley and Sophie's Choice cases is their tragic element. This element is problematic first because it is importantly different from everyday moral dilemmas, both in the magnitude of the consequences and in the fact that these consequences cannot be forestalled. This makes resolving such dilemmas time consuming, and forces subjects to recruit cognitive resources to the resolution which they might well not use at all in resolving more common and less magnitudinous dilemmas.

Finally, imagining killing one's own baby, throwing people off bridges, and running lumbering trolleys over folks is not everyone's cup of tea. Inserting such questions into an instrument that asks mostly sedate questions about familiar topics can be offensive, and may lead to "cleansing" effects. It would be preferable to develop instruments that avoid the tragic element.

Some headway is made on both of these problems in Brady and Wheeler (1996). Their project represents one of the few attempts to develop an instrument for measuring something like consequentialist and deontological predispositions. They developed a number of non-horrific vignettes, and measured both solution-preference and rationale-preference in their subjects. They also developed a parallel instrument that asked short questions about respondents' preferences for certain character traits. There turned out to be high correlation between responses to the two instruments, indicating that the shorter one could be used alone, eliminating the need for presenting long vignettes to subjects. Unfortunately, for various reasons, their methods do not well serve the interests of moral theory.

Any measure of the degree to which subjects are disposed deontically rather than consequentially must capture dispositions that are recognizable to philosophers as deontic or consequentialist, if the measure is to be of use to philosophers themselves. Psychologists have developed a number of reliable measures of moral reasoning (Kohlberg, 1969; Rest, 1979; Forsyth, 1980), but none of them is intended to test for consequentialist versus deontic predispositions. The closest such instrument is that developed by Brady and Wheeler (1996), which measures "formalism" and "utilitarianism." They define formalism as "the human tendency to assess ethical

situations in terms of their consistent conformity to patterns or rules or some other formal features," and utilitarianism as "the tendency to assess ethical situations in terms of their consequences for people. It does not specify kinds of consequences; it does not identify which persons are relevant" (1996, p. 928). These definitions do not capture the philosopher's distinction between deontology and consequentialism. Both deontologists and consequentialists can assess particular actions in terms of their conformity to rules or their consequences for people. A strong aversion to throwing someone off a bridge, for example, may be framed in terms of the harm it will cause to that person, rather than as the breaking of an abstract or formal rule. Likewise, a willingness to throw the person off the bridge could be framed as obedience to an abstract principle of fairness.

What crucially distinguishes deontology from consequentialism is the commitment (or lack of it) to agent-relative restrictions and obligations. Agent-relative restrictions are captured by the doctrine of double effect and the doing/allowing distinction. Agent-relative obligations arise from special relationships that require partiality and favoritism. An instrument that measures deontic versus consequentialist predispositions ought to test for all three of these commitments explicitly.

Further, good measures must be sensitive to two different ways in which subjects may be committed to general abstract principles. It may be that subjects have general commitments about the relative importance of rights and well-being when choosing among policies or assessing behavior. It may also be that they do not have such general commitments, either explicit or dispositional, but instead have commitments or dispositions to consider rights or well-being only when choosing policies in particular contexts. In the latter case, the extent to which subjects have deontic rather than

consequentialist commitments is a matter of both degree within contexts and frequency among contexts. Measures of general theoretical commitments ought to allow one to distinguish between these two cases.

Moreover, such distinctions require that instruments are explicitly formulated to ask about the *relevance* of various features to an assessment of moral propriety, rather than inferring relevance from the assessments themselves. For example, it is important to know whether subjects think that generally it is more (or less) important to consider the effect of policies on well being or to consider the extent to which policies infringe on individual rights. It is similarly important to know whether subjects endorse such principles as the doctrine of double effect or a doing/allowing distinction, in the abstract. It is only with such information that one can find empirical warrant for claims about the extent to which general commitments influence or are influenced by moral assessments of particular behaviors.

Finally, and for the same reasons, it is important to know, e.g., whether subjects think it is more important to consider well being or rights when choosing policies in several contexts, e.g. gun control, health care, communication, public projects and eminent domain, civil defense, and so on. If subjects commonly do not have commitments to consequentialist or deontic principles in the abstract, but do have such commitments in local contexts, it is possible that training regarding abstract deontic and consequentialist principles will change the local commitments subjects have. And of course, it is also possible that local commitments cannot be modified in this way. To know, we need measures of consequentialist versus deontic commitments in local contexts. Finally, it is crucially important that measures of local commitment to deontic

or consequentialist principles consider commonplace contexts—instruments must not confront subjects with scenarios in which all options are horrific, and the scenarios ought to generate dilemmas of a piece with those subjects commonly confront.

We have then four recommendations. Empirical work must not ignore the extent to which subjects have commitments to high-level, abstract deontic or consequentialist principles, whether the commitments are conscious or dispositional. Second, these commitments cannot be well assessed using standard Trolley and Sophie's Choice scenarios. Third, good measures must include instruments that explicitly focus on abstract principles, and query subjects with respect to the relevance of these considerations to moral judgments. Fourth, good measures must also include instruments that focus on specific contexts, but these contexts must be commonplace and must not force choice among horrific options.

V. Summary.

We have presented statistical and anecdotal evidence indicating that Trolley and Sophie's Choice scenarios induce data loss and sample bias in observational studies. Instruments built with respect to such scenarios are therefore unreliable, for at least three reasons. They are pragmatically bad, insofar as they compromise the power of a study by lowering the sample size. Their inclusion risks inducing sample bias, which will compromise both the causal inferences and the estimation of parameters (specification and identification of the correct model). Third, they are arguably irremediably confounded or biased measures of commitment to abstract moral principles, insofar as such commitment is revealed only in local contexts.

We suggest an alternative strategy for measuring general commitment to abstract moral principles. Instruments should focus on the extent to which subjects perceive principles or considerations as *relevant* to moral judgment, both in the abstract and in specific contexts. Contexts should be commonplace and contextualizing scenarios should avoid forcing choices among tragic options.

References

Aristotle. c. 350 B.C./1985. Irwin, T., trans. *Nicomachean Ethics*. Indianapolis: Hackett.

Brady, F. and Wheeler, G. 1996. An Empirical Study of Ethical Predispositions. *Journal of Business Ethics,* 15, 927-940.

Foot, P. 1967. The Problem of Abortion and the Doctrine of the Double Effect. *Oxford Review*, 5. Reprinted in Foot, P. 1978. *Virtues and Vices.* Berkeley: University of California Press, 19-32.

Forsyth, D. 1980. A taxonomy of Ethical Ideologies. *Journal of Personality and Social Psychology,* 39, 175-184.

Frankena, W. 1939. The Naturalistic Fallacy. *Mind*, 48, 464-477.

Glymour, B., Glymour, C., and Glymour, M. 2008. Watching Social Science: The Debate About the Effects of Exposure to Televised Violence on Aggressive Behavior. *American Behavioral Scientist,* 51, 1231-1259.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., and Cohen, J. 2001. An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science,* 293, 2105-2108.

Haidt, J. 2001. The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review,* 108, 814-834.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., and Mikhail, J. 2007. A Dissociation Between Moral Judgments and Justifications. *Mind and Language,* 22:1, 1-21.

Harman, G. 1999. Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error. *Proceedings of the Aristotelian Society*, 99, 315-331.

Kant, I. 1785/1997. Gregor, M., trans. *Groundwork of the Metaphysics of Morals*. Cambridge, UK: Cambridge University Press.

Kant, I. 1799/1993. Ellington, J., trans. On a Supposed Right to Lie because of Philanthropic Concerns. Indianapolis: Hackett.

Kohlberg, L. 1969. Stages and Sequence: The Cognitive-Developmental Approach to Socialization. In Goslin, D., ed., *Handbook of Socialization Theory and Research.* Chicago: Rand McNally, 347-480.

McNaughton, D. and Rawling, P. 2007. Deontology. *Ethics in Practice,* third edition, ed. LaFollette, H. Malden: Blackwell, 31-44.

Moore, G. 1903. *Principia Ethica*. New York: Cambridge University Press.

Nagel, T. 1972. War and Massacre. *Philosophy and Public Affairs*, 1:2, 123-144.

Nichols, S. 2004. After Objectivity: An Empirical Study of Moral Judgment. *Philosophical Psychology*, 17, 5-28.

Nichols, S. and Mallon, R. 2006. Moral Dilemmas and Moral Rules. *Cognition*, 100, 530-542.

Pearl, J. 2000. *Causality*. Cambridge, UK: Cambridge University Press.

Quinn, W. 1989. Actions, Intentions, and Consequences: The Doctrine of Double Effect. *Philosophy and Public Affairs,* 18.

Rest, J. 1979. *Development in Judging Moral Issues.* Minneapolis: University of Minnesota Press.

Spirtes, P., Glymour, C. and Scheines, R. 2000. *Causation, Prediction, and Search.* Cambridge, MA: MIT Press.

Tetlock, P. 2003. Thinking the Unthinkable: Sacred Values and Taboo Cognitions. *TRENDS in Cognitive Sciences,* 7:7, 320-324.

Thomson, J. 1976. Killing, Letting Die, and the Trolley Problem. *The Monist,* 59, 204-217.

Williams, B. and Smart, J.J.C. 1963. *Utilitarianism: For and Against.* Cambridge, UK: Cambridge University Press.

Appendix 1: The MT Instruments.

All Instruments employed a 7 point Likert scale, with 1 being 'Very Inappropriate' and 7 being 'Very Appropriate'.

MT1: A runaway trolley is heading down the tracks toward five workmen who will be killed if the trolley proceeds on its present course.

You are on a footbridge over the tracks, in between the approaching trolley and the five workmen. Next to you on this footbridge is a stranger who happens to be very large. The only way to save the lives of the five workmen is to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if you do this, but the five workmen will be saved.

How appropriate is it for you to push the stranger onto the tracks in order to save the five

workmen?

MT2. You are at the wheel of a runaway trolley quickly approaching a fork in the tracks. On the tracks extending to the left is a group of five railway workmen. On the tracks extending to the right is a single railway workman. If you do nothing, the trolley will proceed to the left, causing the deaths of the five workmen.

The only way to avoid the deaths of these workmen is to hit a switch on your dashboard that will cause the trolley to proceed to the right, causing the death of the single workman.

How appropriate is it for you to hit the switch in order to avoid the deaths of the five workmen?

MT3. Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables.

Your baby begins to cry loudly and you cover his mouth to block the sound. If you remove your hand from his mouth, his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death.

How appropriate is it for you to smother your child in order to save yourself and the other towns

people?

MT4. You are visiting the sculpture garden of a wealthy art collector. The garden overlooks a valley containing a set of train tracks. A railway workman is working on the tracks, and an empty runaway trolley is heading down the tracks toward the workman.

The only way to save the workman's life is to push one of the art collector's prized sculptures down into the valley so that it will roll onto the tracks and block the trolley's passage. Doing this will destroy the sculpture.

How appropriate is it for you to destroy the sculpture in order to save this workman's life?

MT5. You, your husband, and your four children are crossing a mountain range on your return journey to your homeland. You have inadvertently set up camp on a local clan's sacred burial ground.

The leader of the clan says that according to the local laws, you and your family must be put to death. However, he will let you, your husband, and your three other children live if you yourself will kill your oldest son.

How appropriate would it be for you to kill your oldest son in order to save your husband and your other three children?