Economics Bulletin

Volume 30, Issue 1

A note on the Trolley Problem and Three Weaknesses of Economic Theory

Alessandro Lanteri
Faculty of Political Science, Università del Piemonte Orientale at Alessandria

Abstract

The trolley problem is a moral dilemmas in which human lives are in danger and some, but not all, can be saved by direct intervention of a decision-maker. This article discusses three weaknesses of microeconomics with respect to individual conduct in the trolley problem: (i) it cannot make predictions; (ii) after observing the conduct of participants in an experiment, it cannot explain their decisions; (iii) it cannot suggest policies that ensure the maximization of aggregate welfare, nor can it suggest laws that endorse the prevailing observed conduct.

I am grateful to Sandra Aloia, Mario Cedrini, Chiara Chelini, Caterina Marchionni, Marco Novarese and Salvatore Rizzello for our conversations on the matter of this article and to two anonymous referees for very useful comments. The usual disclaimer applies.
Citation: Alessandro Lanteri, (2010) "A note on the Trolley Problem and Three Weaknesses of Economic Theory ", *Economics Bulletin*, Vol. 30 no.1 pp. 500-507.
Submitted: Oct 13 2009. **Published:** February 09, 2010.

1. Introduction

The trolley problem is a moral dilemma (Foot 1978, Thomson 1985, Unger 1996, Lanteri *et al.* 2008, Chelini *et al.* 2009) in which five human lives are in danger and can only be saved by the direct intervention of a decision-maker. The situation is as follows: a trolley is running towards five people who will be killed if it proceeds on its course. It is possible to save the five people by pulling a lever, which will divert the trolley onto a sidetrack. One person stands on the sidetrack, however, and this person will be killed if the trolley is diverted. So, the decision-maker is faced with a binary choice: action (i.e., pull the lever) or inaction, which result in one or five casualties respectively. The trolley problem is a typical moral dilemma, because the decision-maker has moral reasons both to act and to not act, but doing both is not possible (McConnell 2008).

What will the decision-maker do? Why? Could that decision be changed? How?

In what follows I will argue that mainstream microeconomics cannot answer these simple questions. This article discusses three weaknesses of standard microeconomics with respect to the conduct of the decision-maker in the trolley moral dilemma: (i) it cannot make predictions, (ii) it cannot explain empirical evidence, and (iii) it cannot give advice on policy making.

2. Overview

Economic orthodoxy characterizes individual action as the rational pursuit of the maximal satisfaction of individual preferences. Yet, it remains open as to *which* preferences. To be sure, standard microeconomics accounts for 'private' actions (Sen 1985). This means that these actions are characterized by a strict concern with tending to the agent's consumption (self-centered welfare), disinterest with regard to the welfare of others (self-welfare goal), and a sharp focus on the agent's goals, disregarding whatever allocation others may value (self-goal choice). While 'privateness' does not rule others out altogether, it often leaves them with an instrumental role, what is called 'non-tuism' (Gauthier 1986, p. 87, 311). Non-tuism means that, even when an agent prefers that others behave in a certain way or that something befalls them, such preference is independent from what *they* want, and it only holds insofar as those behavior and happenings serve the agent's satisfaction. It has been suggested, however, that non-tuism is "a feature of particular models and not an assumption that is essentially built into the economic way of thinking" (Pettit 2001, p. 78; Sen 1982). Indeed the three formulations of privateness leave room to significantly different roles for the others (Davis 2007, p. 316):

> *Self-centered welfare concerns only an individual's own satisfaction (or desire fulfillment), but self-welfare goal allows other individual's satisfaction to enter into an individual's satisfaction through sympathy (or antipathy), and self-goal choice allows for non-welfarist goals that are altogether removed from an individual's satisfaction (such as pursuit of social justice).*

Such flexibility, though perhaps praiseworthy under other respects, makes it virtually impossible to ascertain *ex ante* whether in the trolley problem a rational economic agent would or should pull the lever.

3. First Weakness: Prediction

The maximization of aggregate welfare requires that the decision-maker pull the lever, so that five lives get saved. This may be compatible with self-goal choice and with self-welfare goal, but hardly with self-centered welfare. A constituent part of economic rational agency, however, is that "self-regarding desires are generally stronger than […] other-regarding ones" (Pettit 2001, p. 78). Since in the trolley

problem the individual optimum might be different from the social optimum, a self-regarding decision-maker might want to refrain from acting (more on this below), even in the presence of a sympathetic concern for saving lives. So, it is not easy to predict which of these concerns actually prevail, and so which preferences will ultimately be maximized.

One could perhaps try to use one's insight to attempt a prediction. Such an attempt would evoke Max Weber's notion of *Verstehen*, or the "understanding from within by means of intuition and empathy, as opposed to knowledge from without by means of observation and calculation" (Blaug 1980: 43). The notion that introspection granted access to individual motivation was common in the early days of economics (e.g., Machlup 1955). Ever since the 1950's, however, economists have abandoned the 'unscientific' practice of *Verstehen*. According to most economists, nothing discloses the authentic preferences of an agent better than her actual behavior: an agent's actions 'reveal' her preferences. So, we must observe behavior in order to infer preferences, before a prediction is possible.

Microeconomic theory, therefore, is incapable of making pointed predictions as to whether a rational agent would pull the lever or not. Although this is a weakness, accurate prediction is not the only desirable function of a theory of individual behavior. The explanation of observed behavior, too, is desirable and may be sufficient to uphold a theory.

4. Second Weakness: Explanation

As mentioned, we do not know in advance which preferences rational decision-makers are maximally satisfying. We thus ought to observe actual behavior and subsequently make an inference about preferences that triggered it. If we observe that people pull the lever, for example, we can describe their conduct as a manifestation of their preference for aggregate welfare. Otherwise, we account for the data as a manifestation of self-regarding concerns. As seen, either of those would be a plausible economic explanation for each observation respectively.

In an experimental study of the trolley problem, Lanteri *et al.* (2008, p. 795ff.) found that more than 94% of the participants consider pulling the lever acceptable and more than 65% consider doing so morally compelling. Hence, assuming that these respondents did not lie in the questionnaire and that they are not lacking in will power, if they found themselves in a situation of this kind, it is likely that they would pull the lever.[1] Therefore, one would be tempted to infer that they are motivated by a preference for aggregate welfare and so expect that they always act in the pursuit of aggregate welfare.

In a common variant to the standard problem, the same trolley is running towards five people, but this time there are no sidetracks. It is only possible to save the five people by pushing onto the track an overweight stranger, who happens to be standing nearby and whose mass will be sufficient to arrest the trolley. The stranger will of course be killed if pushed on the track. What would a rational economic agent do now?

[1] It would have been impractical – and perhaps altogether impossible – to arrange a direct test of individual behavior in a setting that mimics the main elements of the trolley problem. The data from this 'philosophical experiment' (e.g., Knobe and Nichols 2008) on the acceptability of alternative courses of action nonetheless seem adequate evidence for these preliminary reflections on the ability of economic theory to predict, explain, and modify individual behavior in a moral dilemma. For another use of philosophical experiment in economics, see Cubitt et al. 2009.

As above, we cannot predict but we can explain. Less than 46% consider admissible the pushing of the stranger and less than 3% consider it obligatory to do so. Acting – and so pushing the stranger – is required in order to guarantee the survival of the highest number. However, we now have reasons to believe that, if the respondents faced such decision, they would refrain form pushing the stranger onto the track. They do not seem to pursue the social optimum. Hence, they can no longer be presumed to pursue aggregate welfare.

I suggested that we could not predict, but in truth we could. Since the responses to the two scenarios were prompted in sequence from the same pool of respondents, we could employ the preferences inferred from the first scenario to predict the responses to the second scenario. The respondents' preferences were for aggregate welfare, and so we would have predicted that they push the stranger just like they pulled the lever. Yet, the evidence does not corroborate such prediction.

Though our prediction failed, we can nonetheless try to explain. Our explanation, however, requires that in this second scenario we posit different preferences than those posited to account for the observations from the standard scenario. In order to explain the observations, we must admit that preferences are either volatile or to some extent inconsistent. This is problematic.

We usually infer that a person has certain preferences from seeing the person perform an action corresponding to the preferences, and then explain that action as stimulated by those preferences. The notion of preferences, therefore, has been criticized as being a circular concept that produces an illusion of an explanation while not really explaining anything. In the case of revealed preferences, circularity is avoided by requesting three fundamental properties of preferences: transitivity, completeness, and stability (Camerer *et al.* 2005, p. 10n). When one of the three properties fails, preferences become again circular. Therefore, one must posit the stability of preferences to avoid circularity, but if one requires stable preferences it becomes impossible to capture the evidence in the two treatments of the trolley experiment.

The seeming problem is a failure to appreciate the nuances in the alternative plans one has to enact in order to achieve some results. Assuming the goal of saving five lives, the actions required to obtain the goal are judged only on the grounds of their efficiency. Perhaps the two actions – pulling the lever and pushing the stranger – have different costs. For example, the participants may account for the risk of being charged with murder for pushing the stranger. Conversely, the individual costs and risks associated with operating the lever are lower. Less than 29% believe that pulling the lever amounts to the intentional murder of the one person standing on the sidetrack, but almost 92% consider pushing the stranger a deliberate killing. (Virtually all the respondents also believe that intentionally killing somebody is both morally and legally worse than letting somebody die.) If the cost of pushing the stranger is higher than that of pulling the lever, when both actions grant the same outcome of saving five lives, it is plausible that more agents will pull the lever than push the stranger, though they have the same preferences. It can also be imagined that the participants have preferences which discriminate between the two scenarios and which therefore explain the difference in the responses.[2]

Both these possibilities, however, are questioned by additional evidence. Half of the participants in the study were administered a different, reversed treatment (i.e., they responded to the stranger scenario first, followed by the lever scenario). If the

[2] I owe this remark to an anonymous referee.

explanation sketched above were correct, one should now observe responses very similar to those of the previous treatment. The responses to the stranger scenario are indeed unchanged: 48% deem the push acceptable and 7% deem it compelling, with almost 89% judging it an intentional murder.[3] However, the responses to the lever scenario have become puzzling: in the reversed treatment, less than 78% participants find the pulling acceptable, just above 11% consider it compulsory, and almost 60% now believe that pulling the lever counts as an intentional murder.[4] Both the explanations that could be invoked to make sense for the asymmetries in the two scenarios of the previous treatment now fail to square with the observed responses.

A second weakness of microeconomic theory is thus that it cannot propose a coherent explanation for these empirical observations.[5]

5. Third Weakness: Law and Policy Making

In the trolley problem, although it is not the most common response, inaction is always compatible with rational decision-making. However, since the maximum social welfare obtains when five lives are saved, either by pulling the lever or pushing the stranger, then perhaps some authority may want to draft a law ensuring that this regularly happens. Provided that enough people already pull the lever, achieving such goal requires either an increase in the benefits or a reduction in the costs associated with deliberately killing someone insofar as this ensures the survival of many others. The consequences of such scheme, however, could be dire. Is hunger a problem? Kill the undernourished, so the survivors may dine with gusto. Do you want to reduce the spread of sexually transmitted infections? Kill the people living with STIs and protect the healthy from the threat of contagion….

Obviously, no economist would ever recommend doing so. Since interpersonal utility comparisons are impossible, it is also impossible to establish whether these policies would improve or reduce the overall social welfare. Economists generally agree that the preferable state of the world is that in which everybody is either happier than, or at least as happy as in every alternative state of the world. In other words, if society can unanimously support a policy (i.e., nobody has a reason to veto it), then that policy is justified. The problem, therefore, is to identify those policies that only affect individuals positively (or neutrally), but never damage them. Policies of this kind are called Pareto-efficient. Such is not the case for a policy that invites decision-makers to always act in order to save five lives in the trolley problem, because obviously the stranger and the person standing on the sidetrack are affected negatively by that policy.

Moreover, in a scenario such as the trolley problem there is no room for the most common alternative to the Pareto criterion: a Kaldorian compensation (Kaldor 1939). According to the compensation criterion, although someone is damaged, a policy may nonetheless be commendable. This is the case if, after the new state of the world is achieved, those who benefit from the policy realize gains larger than the losses of those who suffer from it. So that the ‘winners’ can (at least in principle) compensate the ‘losers’ but still report a positive net outcome. In the trolley problem, the losers are killed by the trolley, so that it is arguably impossible to offset their loss.

---

[3] The differences between treatments are not statistically significant.

[4] All statistically significant at the .05 level (see also Lanteri 2009).

[5] This is clearly a weakness of the Pareto principle and not of microeconomic theory *per se*. It counts as a weakness of economics only to the extent that economists subscribe to it. I owe also this remark to an anonymous referee.

A lawmaker striving for popular support may instead try to pass a bill consistent with the prevailing rules of conduct in his polity. Therefore, he would want the law to stipulate as mandatory the pulling of the lever, but forbid the pushing of the stranger, as a representative agent would do anyway. Many real people, however, are not representative agents. In the reversed treatment, for instance, over 22% believe pulling the lever unacceptable and over 7% say pushing the stranger is morally compulsory, so they would presumably break the law. Moreover, Chelini *et al.* (2009) employ the dataset from a different trolley problem experiment to show that indeed the most common pattern of responses is 'pull the lever but don't push the stranger'. Yet, 60% of the participants deviate from the pattern at least once over three repetitions with modest variants of the standard dilemma, so that they, too, would presumably break the law. Even a bill deliberately designed to capture the prevailing behavior, therefore, may send the majority of people before a court.

What one could regard as a third weakness of microeconomic theory with respect to the trolley problem, therefore, is that it does not empower specific policies.

6. Concluding remarks

The trolley problem and its variant are moral dilemmas in which human lives are in danger and some, but not all, can be saved through the direct intervention of a decision-maker. Standard microeconomics reveals three weaknesses with respect to the conduct of the decision-maker in the trolley problem: it cannot predict whether she will act or not; after observing the conduct that may be inferred from the judgments expressed by the participants in an experiment, moreover, it cannot individuate the preferences that explain their decisions; finally, it cannot suggest policies that ensure the maximization of aggregate welfare, nor can it suggest laws that endorse the prevailing conduct.

Taken separately, none of these weaknesses would raise serious concerns, as it is understandable that a theory be successful at addressing some aspects of some phenomena, and not all. Together, however, the three weaknesses question the capacity of economics to contribute to our understanding of human behavior in moral dilemmas. Such failure to address human behavior in moral dilemmas constitutes a problem for economics.

The recognition that reciprocity and fairness often drive individual decisions (e.g., Fehr & Schmidt 1999) and the growing importance of ethical concerns in the economic domain (e.g., consumer boycotts in response to corporate moral violations, the focus on social responsibility and socially responsible investments, …) invite a better understanding of the ways in which the moral judgments of economic agents affect their preferences. A failure to do so would question the ability of economics to claim its title of science of decision-making at large, being instead only applicable to narrow contexts in which material incentives prevail and moral ones are unimportant.

The economic model of decision-making owes much strength to its capacity of being "applicable to all human behavior" (Becker 1976, p. 8, emphasis added), because "*all human behavior* can be viewed as involving participants who maximize their utility from a stable set of preferences and accumulate an optimal amount of information and other inputs in a variety of markets." Denying such generality would force economists to accept a lesser role for their discipline. It does not seem likely that many economists would easily accept such limitation, in a time when the novel strands of pop economics (e.g., Frank 2007, Harford 2008, Landsburg 2007, Levitt and Dubner 2005) herald a new age of 'economics of everything' to side with the traditional academic 'imperialism' of economics (Mäki 2009).

References

Becker, G. (1976) *The Economic Approach to Human Behaviour*, University of Chicago Press: Chicago.

Blaug, M. (1980 [1992]) *The Methodology of Economics: Or How Economists Explain*, Cambridge University Press: Cambridge.

Camerer, C., G. Loewenstein, and D. Prelec (2005) "Neuroeconomics: How Neuroscience can Inform Economics" *Journal of Economic Literature* **43**, 9-64.

Chelini, C., A. Lanteri, and S. Rizzello (2009) "Moral Dilemmas and Decision-Making: An Experimental Trolley Problem" *International Journal of Social Sciences* **4**, 174-182.

Cubitt, R., M. Drouvelis, S. Gaechter and R. Kabalin (2009) "Moral Judgments in Social Dilemmas: How Bad is Free Riding?" University of York Discussion Series in Economics 2009/20.

Davis, J. (2007) "Identity and Commitment: Sen's Fourth Aspect of the Self" in *Rationality and Commitment* by B. Schmidt and F. Peters, Eds., Oxford University Press: Oxford, 313-335.

Fehr, E. and K. Schmidt (1999) "A Theory of Fairness, Competition and Cooperation" *Quarterly Journal of Economics* **114**, 817-868.

Foot, P. (1978) "The Problem of Abortion and the Doctrine of Double Effect", in *Virtues and Vices*, Blackwell: Oxford.

Frank, R.H. (2007) *The Economic Naturalist: In Search of Explanations for Everyday Enigmas*, Perseus Books Group: New York.

Gauthier, D. (1986) *Morals by Agreement*, Clarendon Press: Oxford.

Harford, T. (2008) *The Logic of Life: The Rational Economics of an Irrational World*, Random House: London.

Kaldor, N. (1939) "Welfare Propositions in Economics and Interpersonal Comparisons of Utility" *Economic Journal* **69**, 549-552.

Knobe, J., and S. Nichols (2008) "An Experimental Philosophy Manifesto" in *Experimental philosophy* by J. Knobe and S. Nichols, Eds., Oxford University Press: Oxford, 3-14.

Landsburg, S.E. (2007) *More Sex is Safer Sex*, Free Press: New York.

Lanteri, A. (2009) "Judgments of Intentionality and Moral Worth: Experimental Challenges to Hindriks" *Philosophical Quarterly* **59**, 713-720.

Lanteri, A., C. Chelini, and S. Rizzello (2008) "An Experimental Investigation of Emotions and Reasoning in the Trolley Problem" *Journal of Business Ethics* **83**, 789-804.

Levitt, S.D. and Dubner, S.J. (2005) *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*, William Morrow: New York.

Machlup, F. (1955) "The Problem of Verification in Economics" *Southern Economic Journal* **22**, 1-21.

Mäki, U. (2009) "Economics imperialism: Concept and constraints", *Philosophy of the Social Sciences*, **39**, 351-380.

McConnell, T. (2008) "Moral Dilemmas" in *The Stanford Encyclopedia of Philosophy, 2008 Ed.* by E. N. Zalta, Ed., Online: plato.stanford.edu accessed: October 2009.

Pettit, P. (2001) "The Virtual Reality of *Homo Economicus*" in *The Economic World View: Studies in the Ontology of Economics* by U. Mäki, Ed., Cambridge University Press: Cambridge.

Sen, A. (1982) *Choice, Welfare, and Measurement*, Blackwell: Oxford.

Sen, A. (1985) “Goals, Commitment, and Identity” *Journal of Law, Economics, and Organization* **1**, 341-355.
Thomson, J. (1985) “The Trolley Problem” *Yale Law Journal* **94**, 1395-1415.
Unger, P. (1996) *Living High and Letting Die*, Oxford University Press: Oxford.