Journal of Business Ethics (2008) 83:789–804
DOI 10.1007/s10551-008-9665-8

An Experimental Investigation of Emotions and Reasoning in the Trolley Problem

Alessandro Lanteri
Chiara Chelini
Salvatore Rizzello

ABSTRACT. Elaborating on the notions that humans possess different modalities of decision-making and that these are often influenced by moral considerations, we conducted an experimental investigation of the Trolley Problem. We presented the participants with two standard scenarios ('lever' and 'stranger') either in the usual or in reversed order. We observe that responses to the lever scenario, which result from (moral) reasoning, are affected by our manipulation; whereas responses to the stranger scenario, triggered by moral emotions, are unaffected. Furthermore, when asked to express general moral opinions on the themes of the Trolley Problem, about half of the participants reveal some inconsistency with the responses they had previously given.

KEY WORDS: experiments, intuition, moral emotions, moral judgement, moral reasoning, trolley problem

Roughly 30 years of studies of human behaviour and cognition have taught us that decision-making is "*intendedly* rational, but only *limitedly* so" (Simon, 1961, p. xxiv) and that most behaviour is automatic (Bargh and Chartrand, 1999), emotional (Damasio, 1999, 2000; Elster, 1999), instinctive (Camerer et al., 2005; Rubinstein, 2007), or otherwise inspired by some mental 'short-cut' or heuristic (Kahneman, 2003; Simon, 1955, 1978), and not carefully thought over. Simplified procedures of these kinds are largely successful, but occasionally they may fail in ways that are both systematic and predictable (Gilovich et al., 2002). In the last 15 years, we have also witnessed the development of a conspicuous stream of research investigating the role of mental short cuts (and failures thereof) in moral judgements (Baron, 1994, 1995, 1997, 1998; Greene and Haidt, 2002; Haidt, 2007; Haidt and Joseph, 2004; Hauser, 2006; Messick and Schell, 1992; Schelling, 1984; Sunstein, 2004).

Some psychological mechanisms, like emotions and intuition, previously not incorporated in standard choice theory are instead central to our capacity to cope with and thrive in a complex social and natural environment, and they constitute a basis on which an interdisciplinary approach to decision-making could be fruitfully developed. We also believe that, by means of such enrichments, economic theory may aim at a deeper and more nuanced understanding of individual behaviour in the direction of increased realisticness, and that such improved understanding requires, on some occasions, to pay a more than cursory attention to the role played by moral considerations.

In this article, we elaborate on such ideas by addressing two modalities of ethical judgements – moral emotions and moral reasoning – and how do they affect decision-making. We do so with an experimental investigation of the two standard variants of a moral dilemma generally known as the trolley problem.

The trolley problem

The moral philosopher Philippa Foot (1978) was the first to suggest a challenging moral dilemma – the so-called 'trolley problem' – in which a decider is faced with two simple alternatives, both of which result in tragedy. The scenario is roughly as follows: a trolley is running down its track, but nobody is in control. Along the track stand five people who are unavoidably going to die unless the trajectory of the trolley is altered. By flipping a switch it is possible to lead the trolley to a different track, where unfortunately a single person is standing and is then condemned to die. Though a case can be made for or

against both hitting the lever and not hitting it, there seems to be no obviously superior option to choose, whence the dilemmatic dimension of the situation. Nonetheless, most people agree that flipping the switch is permissible.

When it is compared to an experiment with a slightly changed setting, however, the case becomes less straightforward (Thomson, 1976, 1985, 1986). Suppose the same trolley is hurtling down towards five people, but in the absence of sidetracks it is only possible to block it by dropping a heavy weight in front of it. There happens to be a stranger nearby who is considerably overweight: enough for his mass to be sufficient to halt the trolley, if he is pushed on the track.[1] In this case his life would be lost. Most people, even those who tolerated the sacrifice of one person to save five in the previous case, now hesitate.

But why?

Either they are inconsistent, and thus their seeming reliance on intuition proves faulty, or the two situations differ in a morally relevant sense.[2] One difference may be that killing the one is a side-effect to the attempt to save the five in the first case, while in the second case the killing of the stranger is a crucial and deliberate step towards the rescuing of the five. According to the 'doctrine of double effect' (Aquinas, 13th century) the deliberate causation of harm in order to promote some other good is morally inferior to the promotion of some good, whose indirect consequence is to cause harm as a side-effect. This suggestion, however disputable on different grounds (e.g. Kagan, 1989; Unger, 1996), amounts to a trade-off between good and bad deeds. What principle does guide such trade-off?

Since it is impossible to save everybody, the socially maximising conduct is whatever saves five lives. Many philosophers believe this to be the correct source of moral decision-making; these are the advocates of consequentialism, the ethical doctrine that we ought to undertake whatever course of action brings about the greatest benefit to the largest number of people. Most subjects faced with the trolley experiment, nonetheless, refuse to push the overweight person and would feel uneasy if they were told that such is the only moral choice in the trolley problem. This might be taken as a *prima facie* argument against consequentialism; or as demonstration of our failure to apply that principle under certain circumstances.

Following the lead of Immanuel Kant, the adherents to deontic ethics believe instead that we ought to act out of the principles of rational duty: to behave, that is, in a way such that we would want to become a Categorical Imperative for everyone to follow. One such imperative is to never treat anyone (including oneself) as a means, but always as a moral end in himself. The stranger thrown on the tracks is treated a mere means towards the goal of saving someone else and this ought not to be done. However, can we maintain that, in the original scenario, the single person killed by our pulling of the switch is taken as an end in himself?

We want to do good and save the five and we certainly do not want the one to die, so even when the 'special disfavour of fortune' or the 'niggardly provision of a step motherly nature' prevent the actualization of our intended achievements, our good will, ''like a jewel, would still shine by itself, as something that has its full worth in itself'' (Kant, 1784, p. 394). It follows that ''people cannot be morally assessed for what is not their fault, or for what is due to factors beyond their control'' (Nagel, 1993, p. 58). That there is a trolley running towards five people is not our fault. We know, however, that pulling the switch will kill someone, and operating the switch is fully and exclusively under our control. It is too fragile an argument to suggest that it is not our fault that someone dies after we pull the switch. Kant (1784, p. 394) specifies that chance does not affect moral judgement only insofar as good will is not mere wishing, but involves ''the summoning of all means insofar as they are in our control'' (ibid.). Moreover, we would not rationally want 'pull the switch' to become an imperative rule followed by everyone, lest we are prepared to accept our death to be delivered when we least expect it through some hurling trolley that a passer-by diverted towards us to protect someone else. It is therefore not admissible, for deontologists, to pull the switch.

Consequentialists would both pull the switch and push the overweight person. Deontologists would do neither. However, are common people either of those?

Some of us, though perhaps a few, certainly are. It is nonetheless clear to many commentators that ethics should move beyond the logical derivation of

what ought (not) to be done following from some abstract principle and get a better grasp of the psychological background of human decision-making (e.g. Anscombe, 1958).

How do we make sense of the trolley problem? The doctrine of double effect suggests that 'pulling the switch' is admissible under some important conditions. First of all, the real goal of one's action ought to be morally good. Second, the immoral action ought not to be a goal in itself, but merely a side-effect even if its undesirable consequences are known in advance. Lastly, the immoral action ought to be proportionate to the other immoral effects, which would obtain in absence of the original good action. The three conditions are met in the standard version of the problem: pulling the switch aims at saving five lives and it is a good deed, while the death of someone else is unintended and proportionate to the main goal. In the stranger variant, however, the active pushing of someone on the track reverses the situation: a morally bad action – i.e. killing the overweight person – now brings about the good side-effect of saving five people and is not acceptable according to the requirements of the doctrine of double effect.

The distinction between main aim and side-effect in the trolley experiment cannot be easily conflated with deontologism nor consequentialism, and certainly not in a way that is accessible to experimental subjects innocent of moral training. They think it is fine to let one die in the concrete attempt to save five, conversely they would let die five people if saving them requires that one be deliberately sacrificed. They somehow reveal more tolerance of omissions than actions though the consequences might be ultimately worse. Perhaps, therefore, rather than a fully worked-out normative ethical theory, they follow some intuitive hunch. Edward Royzman and Jonathan Baron (2002) indeed uncovered a psychological mechanism in human judgement, which they call 'indirect harm bias', favouring indirectly harmful over directly harmful options both in moral and non-moral issues, irrespective of the associated outcomes, intentions, or self-preservational concerns. In addition, results cannot be fully explained in terms of differences in judgements about which option is more active, more intentional, more likely to cause harm, or more subject to the disapproval of others. It is an intuitive rule of thumb, which is consistently applied, but not easily understood.

Moral intuition and the trolley

Some additional variants to the trolley problem were proposed, in which the trolley in the initial situation can be diverted to a track that loops back to the five people, but a man on this track will actually stop it and die, therefore becoming an active part of the plan to save the five and no longer just a side-casualty. In another version, the trolley can be diverted by means of colliding another trolley into it; both trolleys will then be derailed into the yard of a sleeping man, who will thus be killed. On occasions like these, the doctrine of double effect could provide specific guidance, but people do not seem to follow its prescriptions. Instead, as the moral philosopher Peter Unger (1996, p. 92) suggests, the responses to the new problems are partly dependent on whether the subject has already encountered the standard version, because of a desire to express consistent moral opinions. He also points out how, by means of introducing intermediate alternatives to <do nothing> and <kill one to save five>, the first option may lose appeal and no longer be considered morally superior to the last one. When intermediate options progressively save more lives through increasingly active forms of intervention, Unger's students conclude that <actively kill one to save five> is better than <let one die to save two>, and that <let one die to save two> is better than <do nothing>, therefore, by extension <actively kill one to save five> is considered transitively better than <do nothing>. Rather than saying what they believe to be the right thing, therefore, it seems that subjects express what they believe is right thing, provided that it confirms the rightness of whatever they had said earlier on.[3]

Peter Singer (1999), whence the reflection and the following quote are appropriated from, underlines how adding or deleting intermediate alternatives affect our intuitive judgement of pre-existing options, inducing test subjects to display what behavioural economists term 'preference reversal'. However, at a closer look, it becomes apparent that the intuitive reactions are based on rather odd factors (Unger, 1996, p. 102):

> First, when serious loss will result, it's harder to justify moving a person to, or into, an object than it is to move the object to, or into, the person. Second, when serious loss will result, it's harder to justify changing the speed of a moving object, or changing its rate of motion, than changing the object's direction of motion. Third, when there'll be big loss, it's harder to justify speeding up an object than slowing down an object. Fourth, it's a lot harder to justify taking an object at rest and setting it in motion than to justify taking an object in motion and increasing its speed … [Fifth] it's harder to justify imposing a substantial force on an object than it is to justify allowing a force already present (just about) everywhere, like gravitation, to work on the object.

Are these factors morally relevant? Are our intuitions granted? Perhaps, Singer suggests, these intuitions rely on some proxy of genuine moral factors, but when relentlessly exported to alien contexts they are no longer suitable. Thus, Unger concludes, responses to the trolley problem are rather dependent on psychology than on ethics proper.

Moral intuition and psychology

The mayhem of intuitive feelings is far reaching. If intuition is such a biased process, one may believe it would always be better to sit down and carefully deliberate. Besides the inefficiency of such proposal (because deliberation is very costly, both effort- and time-wise), careful reasoning may even be ineffective. Even thorough analyses often rely on intuitive hunches, and they may prove unable to reach a satisfactory solution.

For one instance, in the 1970s British and Japanese health authorities decided to suspend the provision of DPT vaccines on the basis that they could, as a side-effect, cause the death of a little number of patients, smaller – it should be noted – than the number of patients who would have died in the absence of such vaccine. Similarly, polio vaccine Sabin is more effective than Salk; yet, the first may cause polio in patients. Despite this risk the number of lives saved by Sabin remains significantly higher than Salk. Many specialists, nonetheless, have preferred the less effective treatment on the grounds of an intuitive judgement that procuring harm is worse than not avoiding harm. Omissions are treated with more indulgence than actions (another instantiation of the 'indirect harm bias;' Royzman and Baron, 2002). The consequence is that a number of lives was lost to this line of reasoning. This approach may seem to be in contradiction with the doctrine of double effect (in terms of proportionality), and indeed the delivery of DPT vaccine has been restored, just like Sabin is nowadays once again preferred to Salk. The loss of lives to side-effects (although known with fair statistical certainty) is considered admissible in the attempt of saving a greater number of lives. Cases like these may keep someone struggling to find a solution, but not always there is room for correction.

Amos Tversky and Daniel Kahneman (1981) administered an experiment to two groups of subjects. The setting is a case of a disease expected to kill 600 people. The first group of subjects was faced with alternatives A and B, the second group with C and D.

> If program A is chosen, 200 people will be saved.
>
> If program B is chosen, there is 1/3 probability of saving 600 and 2/3 probability of not saving anyone.
>
> If program C is chosen, 400 people will die.
>
> If program D is chosen, there is 1/3 probability that nobody dies and 2/3 that everybody dies.

This is a case involving human lives and the attempt to save them, we certainly would not want it to be deliberated upon by gut feelings. Since it is evident that program A is identical to C and B to D, we are confident that either A/C is better than B/D or vice versa. Why then 72% of the first group would choose A, while 78% of the second group would choose D? Are such intuitions sound? This is an instantiation of 'framing'. Subjects do not question the real effect of each policy, but take it at its face value. In this case, it is normal to see the granted safety of 200 people as a gain and the certain death of 400 as a loss. As Kahneman and Tversky (1979) also pointed, we are more risk averse with respect to gains than to losses. The words employed in setting the problem to the two groups determine different perceptions, so that in the first case subjects are unwilling to promote the same risky program they support in the second case. It is also noteworthy that none of the responses seems to be particularly

problematic and there seems to be a significant confidence underlying the choice of plan A and plan D. Intuition operates very fast, it is often uncontroversial, and it does not always render due justice to subtle differences.

Moral heuristics

As we shall see in greater detail below, the responses of the participants in our trolley problem experiment do account for the distinctions in the two situations. Given the time constraints imposed on the decision and their lack of formal philosophical training, however, it is rather implausible to imagine that our subjects were rationally considering an actual application of the doctrine of double effect (or any other ethical doctrine). It is more likely that they were employing some sort of heuristic following from that principle or – which in the light of the previous discussion seems more probable – that the doctrine of double effect originates in some measure from such rule of thumb.

Indeed moral judgement does not escape automatic processes: we face situations and promptly deliver a good/bad intuitive evaluation of alternatives as part of our perceptions; explanation comes only afterwards, if at all (Haidt, 2001). There is an expanding body of experimental evidence showing that people make choices for (at least in part) unknown reasons, and then make up reasonable justifications, while remaining unaware of the gap between their real motivation and their ex-post rationalisation (T. Wilson, 2002).

The immediate responses may be traced to moral emotions as opposed to moral reasoning. Through the analysis obtained with functional magnetic resonance imaging (fMRI) from subjects involved with both ethical problems of various kinds and non-ethical ones, it can be observed that the brain areas activated differ significantly (Greene et al., 2001, pp. 2106–2107; see also Greene and Haidt (2002) and Greene et al. (2004)). When facing moral-personal decisions – i.e. something that is 'up close and personal' in the fashion of the overweight stranger scenario, or the pushing of someone off a sinking lifeboat – those brain areas react (i.e. medial prefrontal cortex, posterior cingulate cortex/precuneus, and superior temporal sulcus/tempoparietal junction) that have been identified as correlates of emotional arousal. The same areas remain inactive in moral-impersonal decisions – i.e. something that is somehow distant, like the lever scenario or a case of keeping the money found in a lost wallet – and in non-moral ones – e.g. a choice between different means of transportation given some time constraint. The areas of the brain associated with working memory (i.e. middle frontal gyrus and parietal lobe) instead were less active in the moral-personal scenarios, but became activated in both the moral-impersonal and the non-moral decisions.

Also response times differ across scenarios. The fastest responses were judgements of inappropriateness to moral-personal decisions, signalling that they are virtually automatic because they elicit "prepotent, negative social-emotional responses" (Greene et al., 2004, p. 390). The slowest responses, on the other hand, were judgements of appropriateness to the same decisions, signalling that the subjects had to override the instinctive emotional response. Moral-impersonal and non-moral decisions' reaction times were in between (Greene et al., 2001, p. 2107). All these differences reflect on people's responses and can be responsible for seeming inconsistencies across scenarios.

Looking at the problem from a different angle may suggest that the questions and answers which puzzle philosophers may not amount to inconsistencies proper, but reflect real differences – perhaps differences which are not easily accessible to armchair philosophical speculation – in the perception and processing of information concerning scenarios, alternative conducts, and their moral significance. In other words, we express a warning that the association between the lever and the stranger variants of the trolley problem may be somewhat artificial: the two scenarios are cognitively and emotionally distinct and our brain treats them very differently although both scenarios reproduce a choice context in which a moral violation must be committed in order to ensure the maximisation of aggregate welfare. We may certainly imagine more cases, either abstract as here or actual as they would be in a real business decision, in which similar morally difficult choices ought to be made and we would like to bring our experimental results below to bear on the issue at large. Yet, we may not be fully entitled to directly comparing predominantly emotional and

predominantly reasoned decision-making simply because they could be constructed to apply to scenarios that resemble each other in any number of respects.

Aside from this warning, we subscribe to the notion that "moral thinking is driven largely by social-emotional dispositions built on those we inherited from our primate ancestors" (Greene and Haidt, 2002, p. 519). For instance, in an evolutionary perspective it makes sense to regard an emotional aversion to damaging other humans as a fitness-improving trait that confers upon its possessors some advantage in grouping successfully (D. Wilson, 2002). Such "adaptation would have arisen at a time when the scope of aggression was limited literally to a stone's throw" (Cohen, 2005, p. 12), while there was no need to avoid harming other humans at long distances, as this was not even, technologically speaking, a possibility. Beside emotions, we are also characterised by a capacity for elaborate abstract reasoning. Human moral judgement may thus be supposed to be "a complex interplay between (at least) two distinct types of processes: domain specific, social emotional responses and domain-neutral reasoning processes applied in moral contexts" (Greene and Haidt, 2002, p. 519).

Moral emotions seem to import eminently in the case of personal moral violations, while reasoning applies to impersonal violations. Joshua Greene and Jonathan Haidt (2002) characterise a personal violation as: physical harm to a specific person in such a way that is not determined by a pre-existing condition. Conversely, impersonal violations occur when there is a mediating object between the agent and his action, which on occasions seems to release the agent from responsibility. We should thus expect people to employ what Greene and Haidt call a ME HURT YOU heuristic.[4] Such heuristic may be employed to quickly discriminate among situations and viable conducts, and should result in refraining from personal violations. For the sake of the trolley problem, this amounts to pulling the switch and not pushing the overweight stranger.[5] This heuristic resembles the doctrine of double effect.

Since moral characters and opinions can be variegated, we also remind that the two leading moral schools of thought would each suggest a different conduct in the trolley problem, and that each may be translated into some heuristic as well (Chelini et al., 2007). One heuristic could be labelled SAVE THE MOST: people who follow this heuristic would push the stranger and pull the lever, in order to save the highest number of people. The prescriptions of this heuristic are in line with consequentialism. Another heuristic could instead be called DO NOT TOUCH. Participants who employ this heuristic would refrain from both pulling the lever and pushing the stranger. This heuristic is similar to the prescriptions of Kantian ethics. The latter two heuristics, therefore, may not so much discriminate among types of situations, to which one responds with certain behavioural scripts, but they promote a stable principle of conduct from which agents elaborate actions to be defined case by case. Since choice situations are complex along a variety of dimensions, heuristics may enable quick decision-making either by suggesting when to act in some way (ME HURT YOU) or to always act in certain ways (SAVE THE MOST and DO NOT TOUCH).

We now turn to the experimental results, which allow us to investigate which heuristics, if any, participants employ. We also examine whether question ordering has any effect on the heuristics employed. Finally, we elicit the general moral opinions of participants and compare these with previous responses.

The trolley experiment

On the 9th and the 13th November 2006 we conducted a series of experiments with a total of 62 undergraduate students of law at the University of Eastern Piedmont in Alessandria (Italy). The students were summoned for a class in the Seminar of Political Economy, which is not compulsory in their curriculum, and which grants them 3 credits (out of a total of 180 credits over 3 years required for graduation – i.e. 5% of the annual credit load), provided that they attend a total of three experiments and a concluding lecture. The credits constitute the compensation for participants.[6]

Participants were informed that they had to answer truthfully and fill in its entirety a questionnaire aimed at understanding their moral opinions, which had been prepared in a way that there existed no right or wrong answers. They knew the experiment would last about twenty minutes, and they

were guaranteed full anonymity. On both days, roughly half of the participants took part in either of two treatments: the one (Standard Treatment) featured the lever-pulling scenario first, followed by the overweight stranger scenario; the other featured the same scenarios in reversed order (Reversed Treatment). At the end of each scenario, the participants were required to answer several yes or no questions.

Scenario 1

(As in the previous scenario) A trolley without passengers and without conductor is travelling at full speed down a track. On the track there are five people, who will surely be killed if the trolley keeps riding on the actual path. There is also a side-track, on which there is one person.

A passer-by could pull a lever next to the track, and this way deviate the trolley onto the side-track. The passer-by realises that, if he does not pull the lever, the five people will be killed. If he pulls the lever instead, the five people will be saved. The passer-by is aware, however, that by pulling the lever the person on the side-track will be killed.

[*Question 1*] Under these circumstances, is it morally obligatory for the passer-by to pull the lever?
[*Question 2*] Under these circumstances, is it morally acceptable for the passer-by to pull the lever?
[*Question 3*] If the passer-by does not pull the lever, is he intentionally killing five people?
[*Question 4*] If the passer-by pulls the lever, is he intentionally killing one person?

Scenario 2

(As in the previous scenario) A trolley without passengers or conductor is travelling at full speed down a track. On the track there are five people, who will surely be killed if the trolley keeps riding on the actual path. A passer-by stands next to the track, and he could push a very fat stranger on the trolley's path, halting its ride.

The passer-by realises that, if he does not push the stranger, the five people will be killed. If he pushes the stranger instead, the five people will be saved. The passer-by is aware, however, that by pushing him, the stranger will be killed.

[*Question 5*] Under these circumstances, is it morally obligatory for the passer-by to push the stranger?
[*Question 6*] Under these circumstances, is it morally acceptable for the passer-by to push the stranger?
[*Question 7*] If the passer-by does not push the stranger, is he intentionally killing five people?
[*Question 8*] If the passer-by pushes the stranger, is he intentionally killing one person?

In the lights of the previous scenarios, our subjects were also asked to answer four additional questions, which we employed to check for consistency.

[*Question 9*] Is there a difference, in terms of moral responsibility, between intentionally killing someone and letting someone die?
[*Question 10*] If you answered affirmatively, is intentionally killing someone morally worse than letting someone die?
[*Question 11*] Is there a difference, in terms of legal responsibility, between intentionally killing someone and letting someone die?
[*Question 12*] If you answered affirmatively, is intentionally killing someone legally worse than letting someone die?

Results and discussion

The overall results (Figure 1) are not very much unanticipated. About 95% of the subjects say that pushing the stranger is not a moral obligation, and 53% say that it is morally unacceptable to do so. About 24%, instead, believe that pulling the lever is morally compelling, and 87% consider it at least morally acceptable. About 85% in the lever-scenario and 90% in the stranger-scenario believe that abstention from action does not amount to intentional murdering the five people on the track. By and large (90%), however, they consider pushing the stranger as a deliberate murder; a smaller proportion (42%) also say that pulling the lever amounts to an intentional killing of one.

When we observe the responses of two Treatments separately (Figures 2, 3) some interesting differences

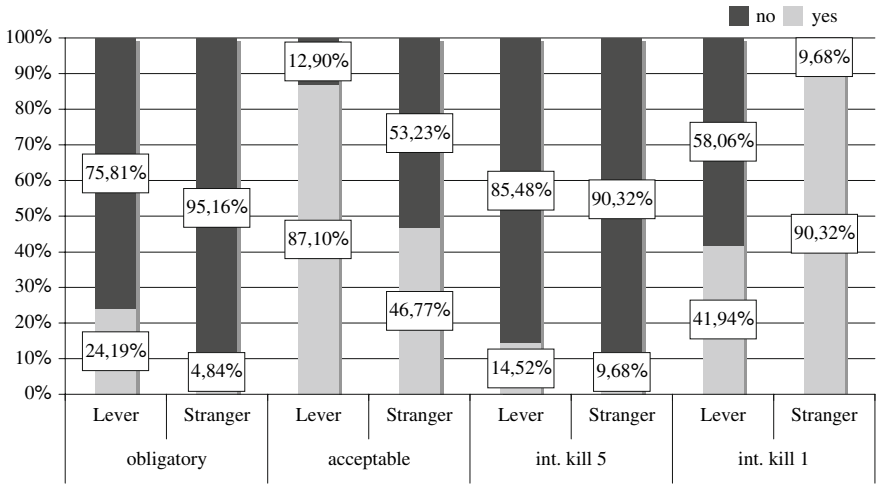

Figure 1. Responses, by scenario.

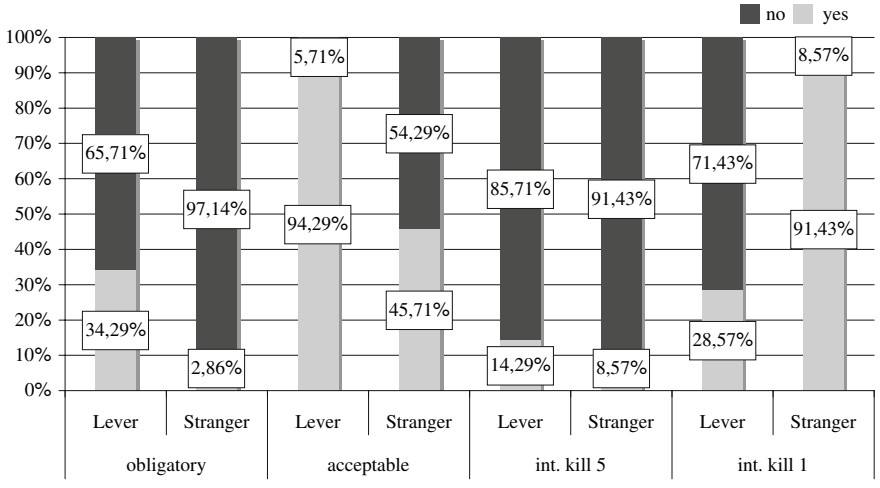

Figure 2. Responses, standard treatment.

can be pointed out. It is not easy to predict whether altering the sequence in which participants encounter the scenarios affects their responses and, if so, how.[7] Indeed, when the lever scenario is put second, fewer participants are willing to operate on the switch than when it is put first, but the responses to the stranger scenario remain unaffected.

If some moral emotions are evolutionarily sound and hard wired into our species, we should expect a very large majority of people to follow them so that we observe a pattern of dominant behaviour consistent with the emotion, while perhaps there will be a display of greater variety of moral outlooks as a result of moral reasoning. We also expect hard-wired moral-emotional behaviour to be more robust – that is less sensitive to contingent variations in experimental conditions – which can import on reasoning instead. Subjects in both treatments, indeed, respond similarly to the emotional stranger scenario.

Whereas in the Standard Treatment 34% of the subjects see room for moral compulsion and 94% for moral acceptability in pulling the lever, the figures scale down to 11% and 78%, respectively, for the Reversed Treatment. These differences are especially important because they show that there is no straightforward way to behave in the lever scenario and that the responses elicited among the participants

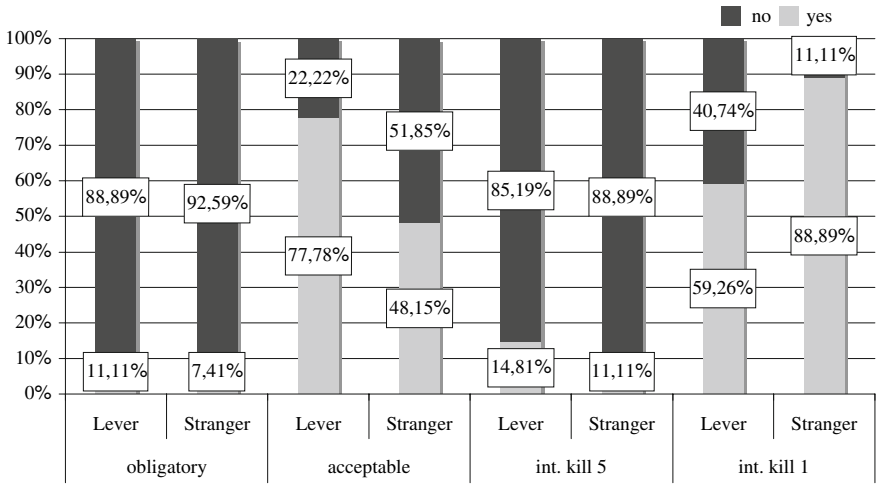

Figure 3. Responses, reversed treatment.

are crucially affected by the scenario sequence. On the other hand, the stranger scenario is overwhelmingly considered as a situation in which inaction is morally superior to action.

One major discrepancy of clear moral consequence can be identified between the two scenarios that accounts, at least in some measure, for these observations. The personal moral violation (i.e. pushing the stranger) is consistently considered intentional by a larger percentage of participants than the impersonal violation (i.e. pulling the lever). However, a discrepancy can be spotted here, too. In the Standard Treatment the percentage is lower than in the Reversed Treatment. Subjects in both treatments largely consider unintentional the death of five by means of not pulling the lever (86% in the Standard Treatment and 85% in the Reversed Treatment), but the murdering of the person on the side-track is considered intentional by 59% in the Reversed Treatment and by just 29% in the Standard Treatment. Our results, therefore, confirm that the stranger and the lever scenarios are perceived differently. They further show that the lever scenario may be perceived in more than one ways, depending for instance on whether the participants have already encountered the stranger scenario, while the stranger scenario is not thus affected by ordering.

The difference in the attributions of intentionality can be regarded as a turning point. Perhaps the emotional activation of the stranger scenario makes participants more alert to personal moral violations. It makes, as it were, the three features of personal moral violations, thereby including intentionality, more salient and thus more likely to be attributed. Even if we allow for the possibility that such difference only emerges as an ex-post justification, it is not immediately clear what else it would be an ex-post justification for.

We now turn to a more fine-grained analysis of heuristics at the individual level, by means of which we may point out additional and more specific differences.

Response heuristics

Since we don't ask our subjects to act directly, nor to state how they would act if they were involved with the decision first hand, and because experimental subjects usually suggest a third person to act more frequently than they would, we cannot easily say who is employing which heuristic – nor if anybody is employing any heuristic at all. We can, however, presume that some patterns of responses point decidedly towards the belief in one rule. We thus calculate the support for heuristics restrictively: namely, we consider the set of responses that can only result in an application of the heuristic at this stage (Table I). The SAVE THE MOST heuristic thus requires that a subject answers affirmatively to question 1 and to question 5 (i.e. she affirms it is morally obligatory to pull the lever and to push the stranger). For the DO NOT TOUCH heuristic, we require that a subject answers no to question 2 and to

TABLE I

Restrictive heuristics, by treatment

	St Treat	Rev Treat	Total
Save the Most	–	–	–
Do Not Touch	–	4 (15%)	4 (6%)
Me Hurt You	7 (20%)	1 (3%)	8 (13%)
N	35	27	62

TABLE II

Extended heuristics, by treatment

	St Treat	Rev Treat	Total
Save the Most	15 (43%)	11 (40%)	26 (42%)
Do Not Touch	23 (65%)	23 (85%)	46 (74%)
Me Hurt You	32 (91%)	20 (74%)	52 (84%)
N	35	27	62

question 6 (i.e. she states it is morally unacceptable both to pull the lever and to push the stranger). We consider as instances of the Me Hurt You heuristic those in which a subject answers yes to question 1 and no to question 6 (i.e. she declares that it is morally obligatory to pull the lever but morally unacceptable to push the stranger).

We also check for consistency, by means of the answers given to questions 3, 4, 7, 8, 9, and 10. For instance, when a subject holds the Save the Most heuristic, he should either believe that pushing the stranger and pulling the lever do not amount to an intentional murder (q3, q4, q7, q8); or, if he does believe that the killing is intentional, he should at least believe that the intentional killing of someone is not morally worse than letting someone die (q9, q10). Lacking these conditions amounts to some kind of inconsistency. A similar consistency-check was conducted on the other heuristics as well.

The first observation is that nobody restrictively supports the Save the Most rule and, though there are indications that some participants never approve of altering the path of the trolley (Do Not Touch), most of them intervene selectively (Me Hurt You). However, these are not evenly distributed, and in fact are strongly clustered between the two treatments. The Me Hurt You heuristic is more common in the Standard Treatment, with 7 instances.[8] In the Reversed Treatment, on the other hand, there is but one subject supporting the Me Hurt You heuristic, and four supporting the Do Not Touch heuristic.

We now allow for an extended definition of heuristics, as the set of responses that are merely compatible with an application of each heuristic (Table II). For the extensive Save the Most heuristic we require that a subject answers affirmatively to question 2 and to question 6 (i.e. she affirms it is morally acceptable to pull the lever and to push the stranger). The extensive Do Not Touch heuristic requires that a subject answers no to question 1 and question 5 (i.e. she states it is not morally compulsory to pull the lever nor to push the stranger). Finally, we consider extensive Me Hurt You heuristics those in which a subject answers yes to question 2 and no to question 5 (i.e. she declares that it is morally acceptable to pull the lever and not morally compulsory to push the stranger). This test is much less demanding, because some responses are obviously compatible with more than one heuristic, and in fact a rather common pattern of responses allows for both types of intervention – push the stranger and pull the lever – but does not consider either as obligatory, and is thus compatible with all of the three heuristics under consideration.

While roughly 40% of the participants give answers compatible with the Save the Most heuristic, we again observe a difference between the treatments for the other heuristics. The subjects in the Standard Treatment are more inclined to allow the Me Hurt You heuristic (91%) than the Do Not Touch one (65%), but the results are twisted in the Reversed Treatment, with 85% participants who seem acceptant of the Do Not Touch heuristic and only 74% who agree to the Me Hurt You one. Even the Me Hurt You heuristic, therefore, which has been proposed as a plausible account of how people face moral dilemmas such as the trolley problem may not be unavoidably the rule preferred by the most participants (see also Chelini et al., 2007).

Moral opinions and consistency

If scenario ordering affects the responses in terms of attributed intentionality and of required and accepted conducts, does it also influence more general moral and legal opinions?

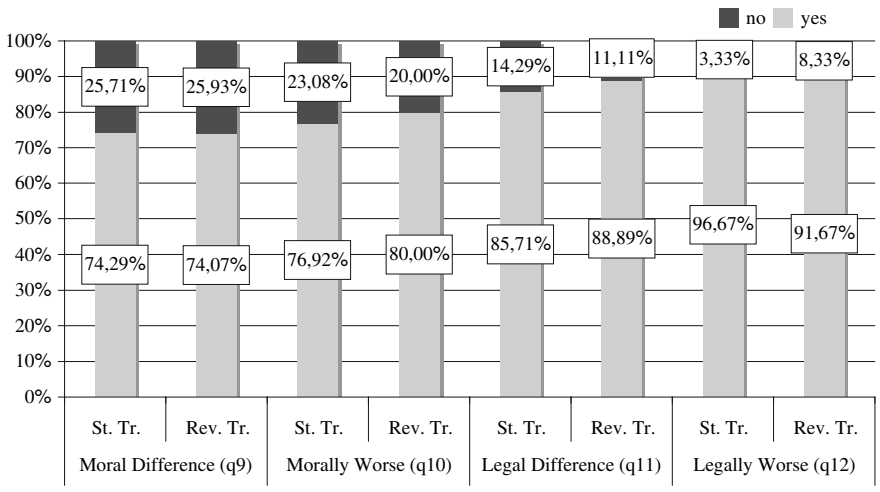

Figure 4. Opinions, by treatment.

No, this does not seem to be the case. There is a remarkable uniformity of responses to the questions (9–12) we employed to verify consistency across and within treatments (Figure 4).

Given the noticeable differences in the preceding responses, such homogeneity of moral opinions is quite unexpected. How does it come about?

Perhaps the earlier questions were formulated in such a way that there were several different ways to respond, even for participants who held a common moral outlook. This would be the case if those answers did not differ significantly – but they do – therefore we do not find this explanation persuasive. It is also possible that the ordering effect plays out in a morally meaningful way that is clear to our students but eludes us, and we can only suspend our judgement on this possibility. Finally, it has been suggested, both in connection with individual preferences in general (Licthenstein and Slovic, 2006) and specifically with moral judgements (Haidt, 2001), that preferences and opinions are constructed in the process of elicitation. This account seems to suggest that, perhaps participants are quite clear about their opinions, in an abstract sense, but when it comes to a specific application, their responses need not follow from those opinions and may instead be intuitively uttered or otherwise altered by emotional activation. Conversely, it may also mean that participants elaborate their moral opinions in a way that is at least in part disconnected from previous responses.

Analysing response patterns at the individual level (Table III), however, does not reveal major differences. In both treatments, about 55% affirm that there exist both moral and legal differences between killing someone and letting someone die, with the former being worse (YES – YES – YES – YES), about 20% declare that the difference is only legal (NO – YES – YES), and a further 15% suggest that the moral difference does not make the intentional kill worse than the alternative (YES – NO – ANY).

We do not know, however, whether the participants in our experiment had developed their moral and juridical opinions before answering questions 1–8. Perhaps the elicitation procedure is such that, when asked questions 9–12, almost everyone

TABLE III

Patterns of opinions, by treatment

Questions 9, 10, 11, 12	St Treat (%)	Rev Treat (%)
YES – YES – YES – YES	54.29	55.56
YES – YES – YES – NO	–	–
YES – YES – NO	2.86	3.70
YES – NO – YES – YES	11.43	3.70
YES – NO – YES – NO	2.86	7.41
YES – NO – NO	2.86	3.70
NO – YES – YES	17.14	22.22
NO – YES – NO	–	–
NO – NO	8.57	3.70
TOTAL	100.00	100.00

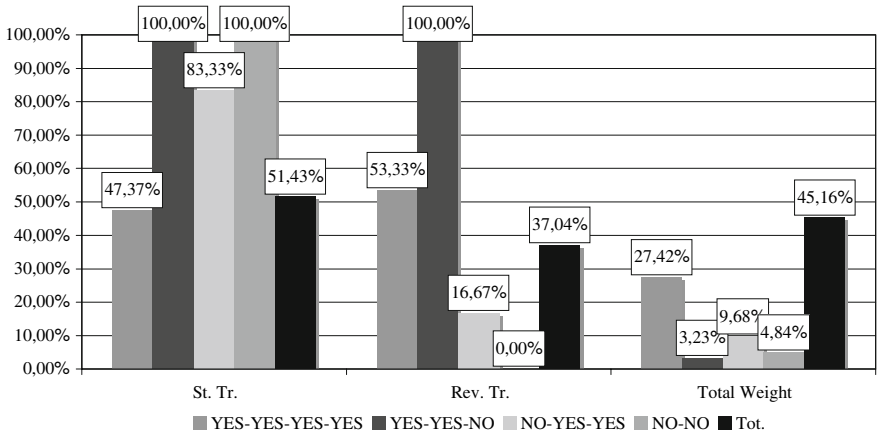

Figure 5. Inconsistent opinions, by treatment.

constructs the same preferences. If such is the case, however, the participants may now give answers that contradict answers previously given to the two scenarios.

And they do. Figure 5 reports the relative frequencies of four response patterns in which inconsistencies were observed, as well as their weight on the total number of observations.

How can one be inconsistent? First, we consider inconsistent suggestions that not pulling the lever and not pushing the stranger differ in terms of intentionality. Either one is intentionally killing five people when one does nothing or not. However, it cannot be plausibly argued that the intentionality of some deaths resulting from inaction depends on the alternatives to inaction.[9] We also count as inconsistency any combination of responses implying that it is ok to do something bad or that it is obligatory to choose an option not different from the alternative. For instance, for the YES – YES – ANY patterns, in which intentionality makes a death morally worse, we consider inconsistent the responses suggesting that pulling the lever amounts to an intentional murder and yet it is admissible, unless not pulling the lever is also considered an intentional kill. For the NO – ANY patterns, conversely, we define inconsistent any suggestion that an act is morally obligatory. How could this be the case, if it does not differ from abstention in a moral sense?

Opinions that there are moral and legal differences between intentionally killing someone and unintentionally letting someone die, but which are inconsistent with previous responses about what one is allowed or obliged to do, account for almost 30% of all the answers. About 10% of all answers are inconsistent with the opinions that the difference is only juridical and not moral. More generally, it is remarkable that 45% of all subjects are inconsistent in one way or another.

Concluding remarks

Should a carmaker introduce a safety system to guarantee additional protection to pregnant ladies in the passenger seat in the event of a side crash or should it modify the windshield so that it better protects both the driver and the passenger involved in any accidents? In addition, should it recruit low-income people to drive the cars into potentially harmful, voluntary accidents in order to test its innovations? Should the R&D department of a large pharmaceutical company develop a new drug to cure a relatively minor disease killing a few patients every year and for which there exists no therapy, or should it invest in a (or yet another) new drug to reduce the risk of heart failure in a much larger number of patients? Also, should it be entitled to endangering some people's or animals' lives when testing the new drug? In the presence of limited resources and uncertain returns to investments these are clearly financial/economic decisions. However, they are quite obviously also moral ones. On the one hand, they concern the allocation of scarce resources towards competing ends with different social costs and benefits, which may not be easily traded-off

against each other and which may have differing degrees of risk and uncertainty. On the other hand, they concern the commission of a moral violation in order to maximise aggregate welfare, and bear consequences on questions such as whether knowingly letting the few die in the attempt to save the many is morally admissible or whether deliberately killing the few in order to save the many is acceptable. In this article, we are concerned with the latter kind only.

We elaborate on the notion that humans employ different modalities of decision-making both along the cognitive/affective and along the controlled/automatic axes (Camerer et al., 2005). We also believe that these are often influenced by moral considerations, either in their emotional or reasoned versions. Our experimental study of the trolley problem allows us to explore some of the effects of moral cognitions and emotions on decision making, as well as some effects of their interplay. Specifically, our results substantiate the main finding that previously aroused emotions affect moral reasoning, but the vice versa does not hold. As in the real world people cannot choose in what order they face (and thus learn how to react to) ethically delicate situations, we believe it important to explore the scope and relevance of moral heuristics in a broad sense.

What we consider as a common weakness in the discussions of the trolley problem we came across in the literature is their lacking of hard data. To be sure, it is possible to speculate on alternative setups of the dilemma in the fashion of a thought experiment, or to gather circumstantial evidence in class and at seminars. Yet, the systematic and rigorous testing on substantial numbers of subjects allows much sounder (and richer) analyses. By means of this article, we also wish to contribute a set of observations that may be taken as a reference for those interested in the attempt.

The results of the trolley experiment presented above are consistent with previous suggestions that, when damage follows, indirect agency is often considered admissible while direct agency is not (Greene and Haidt, 2002). The data also confirm the broad acceptance of the ME HURT YOU heuristic and allows for alternative heuristics, which we labelled SAVE THE MOST and DO NOT TOUCH. Our goal, however, is not to proclaim any of these as morally sound procedure, nor as desirable ones.

Indeed the experiment we discussed does not make a compelling case for normativity: while many subjects tolerate the pull the switch option, it is not evident that they consider it a moral dogma. Conversely, virtually all participants do not consider pushing the overweight stranger as a moral obligation, but quite a number consider it acceptable in spite of the acknowledgement that it amounts to an intentional murder (perhaps some of them believe it admissible, under some circumstances, to deliberately kill someone). Even if our results pointed more evidently towards normative conclusions, we would still feel entitled to questioning the ethical authority of first year Law students. With due caveats of which we are aware, we therefore propose this research as an essay in positive cognitive-moral theory: our results uncover some novel and interesting facts about human decision-making in morally loaded contexts.

Human capacity for moral conduct might stem not so much from some reasoned principle, but from our biological profile. Human criteria for moral assessment might thus derive precisely from that capacity, instead of from some higher value handed down to lay people by means of moral theories. As shown in many experimental researches (including ours above), we are capable of quickly answering moral questions although we might ignore exactly why. Afterwards we can reason about the situation and try to make up a story that justifies our intuitive answer. Such justification, if successful, is likely to become some sort of rule that we keep following, reinforcing through time our conviction that we are doing the right thing. The doctrine of double effect might be interpreted like such justification. On the other hand, the general moral opinions we entertain may be inconsistent with our moral judgements and the justifications we assemble for them, perhaps because they, too, are intuitively generated.

Humans probably have a set of hard-wired moral emotions immediately triggered by some features in a choice situation – for instance, among others, personal-moral features. These do not seem to be susceptible to ordering effects. The more reflective set of cognitive tools that we employ for impersonal-moral (and non-moral) choices, on the other hand, can be 'disturbed' by previously activated emotions. Though this is the most plausible explanation we can conjure to account for our data, at the present stage

we can only advance it as a speculation and as a proposal for further testing for our colleagues in the neurosciences. We therefore wish to add to the difficulty of answering the questions such as those about the automobile and the pharmaceutical companies by emphasising that the way in which questions are framed is likely to bear direct consequences on the nature of answers. When arguing in favour or against either response and before dismissing the counterpart's outlook on the matter, therefore, one may want to stop and ponder about the modality through which she has come to an answer.

Notes

[1] The stranger ought to be described as significantly overweight both in order to ensure the plausibility of the scenario and to rule out (for most participants) the option of jumping in front of the trolley, thereby sacrificing oneself to save the five, but still sparing the stranger.

[2] The problem of whether there exist moral differences between killing and letting die has long troubled philosophers. We shall not enter the debate, but simply present the issue in the lights of the major moral theories before bringing psychological reflections that bear on the matter. For comprehensive reviews and discussions, see Norcross (2002) and Steinbock and Norcross (1994).

[3] The experimental data we shall shortly present, however, deny this remark in some ways.

[4] ME amounts to the requirement that the action must result from an agent's conscious will, and not by accident; HURT concerns physical harm; and YOU identifies a specific victim.

[5] We observe in passing that, while it is understandable that one should not commit personal moral violations and thus not push the overweight stranger, it is not very clear why one should commit impersonal moral violations and pull the lever.

[6] When they took part in the other experiments of the Seminar, the participants were paid according to their performance in specific tasks. An experiment based on the trolley problem, in which there is no 'correct' answer nor there is a way in which one can compute the collective outcome of individual decisions, cannot be rewarded based on individual performance. Such a reward might indeed backfire, because it would give the subjects an incentive to 'respond as they believe the experimenters want them to respond'.

[7] See, for instance, Unger (1996, p. 92). For more formal studies on the influence of questions ordering on participants' response see Alspach and Bishop (1991), Benton and Daly (1991), Crespi and Morris (1984), McFarland (1981), and Willits and Ke (1995).

[8] Two out of 7 ME HURT YOU restrictive-heuristics are inconsistent because the subjects later claim that there are no moral differences between intentionally murdering someone and letting someone die, so that it seems somewhat arbitrary to suggest it is morally compulsory to pull the lever, but morally unacceptable to push the stranger – the first not being an intentional murder and the second being intentional.

[9] Four participants in the Standard Treatment and 1 in the Reversed Treatment show this inconsistency. These inconsistencies are associated with at least another inconsistency so that whether we count them or not does not alter the data in Figure 5.

References

Alspach, S. and G. Bishop: 1991, 'Question Order Effect of Presidential Approval Ratings on Gubernatorial Approval Ratings: A Research Note', *Social Forces* **69**(4), 1241–1248.

Anscombe, G. E. M.: 1958, 'Modern Moral Philosophy', *Philosophy* **33**(124), 1–19.

Aquinas, T.: c. 1270, *Summa Theologica*, 2nd Rev. Edition, Fathers of the English Dominican Province (trans.) (online at: http://www.newadvent.org/summa).

Bargh, J. and T. Chartrand: 1999, 'The Unbearable Automaticity of Being', *American Psychologist* **54**(7), 462–479.

Baron, J.: 1994, '"Nonconsequentialist decisions" (with commentary and reply)', *Behavioural and Brain Sciences* **17**, 1–42.

Baron, J.: 1995, 'A Psychological View of Moral Intuition', *Harvard Review of Philosophy* **5**, 36–40.

Baron, J.: 1997, 'The Illusion of Morality as Self-Interest: A Reason to Cooperate in Social Dilemmas', *Psychological Science* **8**, 330–335.

Baron, J.: 1998, *Judgement Misguided: Intuition and Error in Public Decision Making* (Oxford University Press, Oxford).

Benton, E. and J. Daly: 1991, 'A Question Order Effect in a Local Government Survey', *The Public Opinion Quarterly* **55**(4), 640–642.

Camerer, C., G. Loewenstein and D. Prelec: 2005, 'Neuroeconomics: How Neuroscience Can Inform Economics', *Journal of Economic Literature* **43**(1), 9–64.

Chelini, C., A. Lanteri, and S. Rizzello: 2007, 'Moral Dilemmas and Decision-Making: An Experimental Investigation of the Trolley Problem', *Mimeo*.

Cohen, J.: 2005, 'The Vulcanization of Human Brain: A Neural Perspective on Interactions Between Cognition and Emotions', *Journal of Economic Perspectives* **19**(4), 3–24.

Crespi, I. and D. Morris: 1984, 'Question Order Effect and the Measurement of Candidate Preference in the 1982 Connecticut Elections', *The Public Opinion Quarterly* **48**(3), 578–591.

Damasio, A.: 1999, *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (Harcourt Brace, Orlando, FL).

Damasio, A.: 2000, *Descartes' Error. Emotion, Reason, and the Human Brain*, 2nd Edition (Quill, New York, NY).

Elster, J.: 1999, *Alchemies of the Mind: Rationality and the Emotions* (Cambridge University Press, Cambridge).

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

Gilovich, T., D. W. Griffin and D. Kahneman: 2002, *The Psychology of Intuitive Judgement: Heuristics and Biases* (Cambridge University Press, Cambridge).

Greene, J. and J. Haidt: 2002, 'How (and Where) Does Moral Judgement Work?', *Trends in Cognitive Sciences* **6**, 517–523.

Greene, J., R. Sommerville, L. Nyatrom, J. Darley and J. Cohen: 2001, 'An fMRI Investigation of Emotional Engagement in Moral Judgement', *Science* **293**, 2105–2108.

Greene, J., L. Nystrom, A. Engell, J. Darley and J. Cohen: 2004, 'The Neural Bases of Cognitive Conflict and Control in Moral Judgement', *Neuron* **44**, 389–400.

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

Haidt, J.: 2007, 'The New Synthesis in Moral Psychology', *Science* **316**, 998–1002.

Haidt, J. and C. Joseph: 2004, 'Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues', *Daedalus*, Special Issue on Human Nature, 55–66.

Hauser, M.: 2006, *Moral Minds: How Nature Designed our Universal Sense of Right and Wrong* (Harper Collins, New York, NY).

Kagan, S.: 1989, *The Limits of Morality* (Oxford University Press, Oxford).

Kahneman, D.: 2003, 'Maps of Bounded Rationality: Psychology for Behavioural Economics', *American Economic Review* **93**(5), 1449–1475.

Kahneman, D. and A. Tversky: 1979, 'Prospect Theory: An Analysis of Decisions Under Risk', *Econometrica* **47**(2), 263–291.

Kant, I.: 1784, *Groundwork of the Metaphysics of Morals*, M. Gregor (ed. & trans.) (Cambridge University Press, Cambridge).

Lichtenstein, S. and P. Slovic (eds.): 2006, *The Construction of Preferences* (Cambridge University Press, Cambridge).

McFarland, S.: 1981, 'Effects of Question Order on Survey Responses', *The Public Opinion Quarterly* **45**(2), 208–215.

Messick, D. and T. Schell: 1992, 'Evidence for an Equality Heuristic in Social Decision Making', *Acta Psychologica* **80**, 311–323.

Nagel, T.: 1993, ''Chapter 3'' from *Mortal Questions* (Cambridge University Press), reprinted in D. Statman (ed.), *Moral Luck* (State University of New York Press, New York, NY).

Norcross, A.: 2002, 'Killing and Letting Die', in R. G. Frey and C. H. Wellman (eds.), *The Blackwell Companion to Applied Ethics* (Blackwell, Oxford, UK).

Royzman, E. and J. Baron: 2002, 'The Preference for Indirect Harm', *Social Justice Research* **15**, 165–184.

Rubinstein, A.: 2007, 'Instinctive and Cognitive Reasoning: A Study of Response Times', *Economic Journal* **117**, 1243–1259.

Schelling, T.: 1984, *Choice and Consequence: Perspectives of an Errant Economist* (Harvard University Press, Cambridge, MA).

Simon, H. A.: 1955, 'A Behavioural Model of Rational Choice', *The Quarterly Journal of Economics* **69**(1), 99–118.

Simon, H. A.: 1961, *Administrative Behavior*, 2nd Edition (Mac Millan, New York, NY).

Simon, H. A.: 1978, 'On How to Decide What to Do', *Bell Journal of Economics* **9**(2), 494–507.

Singer, P.: 1999, ''On the Appeal to Intuitions in Ethics'' in D. Jamieson (ed.), *Singer and His Critics* (Blackwell).

Steinbock, B. and A. Norcross (eds.): 1994, *Killing and Letting Die*, 2nd Edition (Fordham University Press, New York, NY).

Sunstein, C.: 2004, 'Moral Heuristics', *Behavioural and Brain Sciences* **8**(4), 531–542.

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

Thomson, J.: 1985, 'The Trolley Problem', *Yale Law Journal* **94**(6), 1395–1415.

Thomson, J.: 1986, *Rights, Restitution, and Risk: Essays in Moral Theory* (Harvard University Press, Cambridge, MA).

Tversky, A. and D. Kahneman: 1981, 'The Framing of Decisions and the Psychology of Choice', *Science* **4481**, 453–458.

Unger, P.: 1996, *Living High and Letting Die* (Oxford University Press, Oxford, UK).

Willits, F. and B. Ke: 1995, 'Part-Whole Question Order Effects: Views of Rurality', *The Public Opinion Quarterly* **59**(3), 392–403.

Wilson, D.: 2002, *Darwin's Cathedral: Evolution, Religion, and the Nature of Society* (University of Chicago Press, Chicago, IL).

Wilson, T.: 2002, *Strangers to Ourselves: Discovering the Adaptive Unconscious* (Harvard University Press, Cambridge, MA).

Alessandro Lanteri
Department of Economic Science and Quantitative Methods (SEMeQ), Faculty of Economics, University of Eastern Piedmont at Novara, via Perrone 18, Novara 28100, Italy
E-mail: alessandro@lanteri.org

and

Erasmus Institute for Philosophy and Economics (EIPE), Faculty of Philosophy, Erasmus Universiteit Rotterdam, P. O. Box 1738, 3000 DR Rotterdam, The Netherlands

Chiara Chelini
Department of Economics, Turin University, via Po 53, Torino 10124, Italy
E-mail: chiara.chelini@unito.it

Salvatore Rizzello
Department of Legal and Economic Science, Faculty of Law, University of Eastern Piedmont at Alessandria, via Mondovi 6-10, Alessandria 15100, Italy
E-mail: salvatore.rizzello@unipmn.it