Filosofia Unisinos
**14**(2):139-151, may/aug 2013


# Our deontological-utilitarian (deontoutilitarian) minds

## As nossas mentes deontoutilitaristas

Cinara Nahra[1]
Universidade Federal do Rio Grande do Norte

**Abstract**
In this article I discuss the standard responses given by people for moral dilemmas related to life and death, proposing a philosophical model that I call "the utilitarian-deontological (deontoutilitarian) model " to explain how the majority of people respond to these dilemmas. I suggest that when people make moral judgements they use a system of judgments that combines utilitarian and deontological considerations, a system that is primarily deontological, and accept that to kill innocent people is wrong. Faced, however, with the necessity of having to kill someone to save more people, they will typically say that it is right to do this, unless they have to use their own personal force or have to have some kind of personal contact with the person(s) to be sacrificed. In these cases they will refrain from carrying out the act. However, typically they will agree that it is right to kill the person even using their personal force or having some kind of contact with the person if death is inevitable or in catastrophic situations, in which unless you kill one person hundreds of people will die as a consequence. The majority of us, however, will come back to their deontological judgment again if we are faced for example with a "blackmail" situation, in which someone asks us to carry out a killing and if we refuse they threaten to kill everybody. In this situation people are outraged with the offer and will typically judge the situation in a deontological way again, saying that it is wrong to carry out the killing.

**Key words:** trolley problems, moral dilemmas, deontoutilitarianism, utilitarian-deontological model.

**Resumo**
Neste artigo são discutidas as respostas padrão dadas pelas pessoas para dilemas morais relacionados a vida e morte, propondo um modelo filosófico que chamaremos de "modelo deontoutilitarista". Sugere-se que, quando as pessoas julgam moralmente, elas usam um padrão que combina considerações deontológicas e utilitaristas, um sistema que é primeiramente deontológico e aceita que é errado matar pessoas inocentes. Confrontados, porém, com a necessidade de matar alguém para salvar mais pessoas, elas tipicamente dirão que é correto fazê-lo, a menos que, para isso, elas tenham de usar sua força física ou tenham de ter algum tipo de contato pessoal com a(s) pessoa(s) que será(ão) sacrificada(s). Neste caso, elas não praticarão a ação. Entretanto, elas tipicamente responderão que é correto usar a força física ou ter algum contato com a pessoa

[1] Universidade Federal do Rio Grande do Norte. Campus Universitário, s/n, Lagoa Nova, 59078-970, Natal, RN, Brasil. E-mail: cinaranahra@hotmail.com

a ser morta se a morte for inevitável ou em "situações de catástrofe" em que, a menos que uma pessoa seja morta, milhares de outras pessoas irão morrer. A maioria de nós, entretanto, voltará a julgar deontologicamente se estiver diante de uma "situação de chantagem" em que alguém ordena a outrem que mate um terceiro, ameaçando, no caso de que a pessoa se negue a fazê-lo, matar um número maior de pessoas. Nestas situações, a maioria das pessoas se sente ultrajada com a oferta e, tipicamente, voltará a julgar de modo deontológico, afirmando que seria errado praticar a ação.

**Palavras-chave:** dilemas do vagão, dilemas morais, deontoutilitarismo, modelo deontológico-utilitarista.

Marc Hauser provides an interesting example of how little people have access to the principles underlying their moral judgements, even when they think they do (Hauser, 2009). He asks his father, a physicist, to give his answer to some of the trolley dilemmas. His father says he judged that it is permissible for someone to flip a switch diverting a trolley and save 5 people at the cost of saving 1. He also judged that it was permissible to push a large person onto the tracks of a train with the same purpose and he justified this by saying that the cases were both the same, as they reduce the number of people being killed. Hauser, then, asked his father if it would be permissible for a doctor to take the life of an innocent person who walked into a hospital, using 5 organs from this person in order to save the lives of 5 different people in the hospital who would die unless they had their organs transplanted. Hauser's father judged this act as being impermissible (Hauser, 2009)[2]. Then, realising that his justification for the earlier cases (it saves more lives) did not hold up he said that the previous cases were all artificial.

This is a good example, showing that not only people do not have access to the moral principles that they use when making moral judgements, but also, it illustrates my point in this article, that when people make moral judgements they actually use (without having access to them) a system of judgments that combines utilitarian and deontological considerations, a system that is primarily deontological, but allows people to breach the deontological rules for utilitarian considerations. There is, however, a limit for this, and this limit is set, again, by deontological considerations that can be overridden, over again, by utilitarian considerations only in very special cases. However, even in these special cases these utilitarian considerations can also be overridden once more by deontological considerations under certain circumstances.

What happens when people make judgements about the permissibility of killing other human beings in moral dilemmas seems to follow the model below:

(i) People in general judge that killing innocent people is wrong (first deontological constraint).

(ii) However, they are willing to breach this rule in order to maximise the number of people saved (first utilitarian consideration) (Greene *et al.*, 2001; Greene *et al.*, 2004; Koenigs *et al.*, 2007; Nichols and Mallon, 2006).[3]

[2] See also Harris (1975).

[3] All these articles present findings on people's judgements when they are presented with the choice of whether or not to sacrifice one person's life to save the lives of others as in the trolley dilemma where a train is heading directly to kill 5 people on the track, and they will be killed unless you press a switch that diverts the trolley onto an alternate set of tracks killing only one person. In this case, people typically say that they would press the switch. But in the footbridge dilemma where there is (as in the other dilemma) a trolley heading directly towards 5 people but now the only way to save these people is by pushing a stranger off the bridge onto the tracks, killing the stranger to save the 5 lives, people typically answer that they would not push the stranger (although in the study of Koenigs and others it is shown that in scenarios like the footbridge dilemma people with damage in the ventromedial prefrontal cortex VPMC are more likely to endorse the proposed action than other groups).

(iii) Not everything, however, is morally allowed to be done in order to save more lives (a deontological constraint again). What we are actually morally allowed to do in order to save more lives will depend on personal considerations and personal variations. In general when the killing involves some kind of physical contact or proximity with the person who will be killed, people tend to judge deontologicaly again[4] .

(iv) In a few special cases, for utilitarian reasons, we are allowed to violate these deontological constraints. These reasons could be (a) the inevitability of deaths, i.e., when the person will die anyway, in these situation people tend to make utilitarian judgements again, and/or (b) when the cost/benefit of overcoming the deontological constraint is very high, with many lives being saved.[5]

(v) We become deontological again if the killing has to be done to satisfy, say for example, the outrageous requirements of a perceived evil person who blackmails you, threatening to kill more people if you refuse to comply and demands that you actively carry out the killing. Here we have a conjunction of two factors: (a) blackmail from a perceived evil person and (b) the killing involves an act (you have to carry out the killing), not an omission (you are not required to leave the person to be killed, you are required to kill the person, shooting her/him or even pressing a bottom)[6].

Through the supplementary materials for Greene's "Cognitive Load Selectively Interferes with Utilitarian Moral Judgment" (Greene *et al.*, 2008b) we will have some important clues on how people make their moral judgements. Here when interviewee's are asked to answer the question if it is appropriated for people to kill their despicable boss who makes everyone lives a misery (*the architect* example) only 1% of the people interviewed said that yes, it was appropriated. It was also only 5% of people who answered that it was appropriated for a pregnant 15 years girl to kill her newborn child in order to move on with her life. The very low percentage figure of people who answered that it is wrong to kill these people (the boss and the baby) suggests that the greater majority of us really abide by the rule that to kill innocent people is generally wrong (let us call this rule 1).

However, people are willing to make exceptions to this rule under certain circumstances, for example, when in order to avoid the death of a larger number of people you do something that will cause the death of a smaller number. The typical example of this is the famous *trolley* case (already quoted) in which 85% of people said it was permissible to flip the switch diverting the trolley saving 5 people at the

[4] For an account on the effect of personal force in people's judgement on the morality of sacrificing one person's life in order to save other lives, see Greene *et al.* (2009); Cushman *et al.* (2006).

[5] For the influence of the inevitability of death on people's moral judgement see Moore *et al.* (2008) and for an account of the catastrophe effect where a huge number of people will be lost unless someone or a smaller group of people is killed see Nichols and Mallon (2006).

[6] Foot (2002) writes: "Suppose, for example that some tyrant should threaten to torture five men if we ourselves would not torture one. Would it be our duty to do so, supposing we believed him, because this would be no different from choosing to rescue five men from his torturers rather than one? If so, anyone who wants us to do something we think wrong has only to threaten that otherwise he himself will do something we think worse". Foot (2002, p. 28) continues stating that "In the examples involving the torturing of one man or five men the principle seems to be the same as for the last pair. If we are bringing aid we must obviously rescue the larger than the smaller group. It does not follow however that we would be justified in inflicting the injury or getting a third person to do so, in order to save the five. We may therefore refuse to be forced into action by the threats of bad men". Foot's conclusion is that the distinction between direct and oblique intention plays only a quite subsidiary role in determining what we say in these cases, while the distinction between avoiding injury and bringing aid is very important. See also Jim dilemma (Williams, 1973) and the *modified safari dilemma* (Greene *et al*., 2008b).

cost of saving 1.[7] The percentage is also high (76%), of people who gave the utilitarian answer in the *standard fumes* dilemma in which you hit a switch in order to divert fumes killing one patient in a hospital instead of three (Greene *et al.*, 2008b)[8].

But not everybody judges that even these kinds of exceptions should be allowed (think back to the 15% of people who answered that it is wrong to divert the train or the 24% of people who think that it is wrong to divert the fumes). The conditions under which people make exceptions or not for this deontological constraint (we shall not kill!) will vary from person to person. However, it does seem that there is a pattern for the way that the majority of people will judge. It appears that the majority of us will make exceptions in order to save the most possible number of lives (we can verify this by the *trolley dilemma* and also by the *standard fumes*) but stick with deontology again when killing implies that they may have to have some kind of physical contact or proximity with the person to be killed as we can see in the *footbridge dilemma* in which people are asked to answer if it is appropriate to push a man off a bridge in order to save 5 people, and only 12% of people give the utilitarian answer in the figures provided by Hauser *et al.* (2007). In Greene's figures (Greene *et al.*, 2008b) it rises to 21%. Nevertheless, it seems that when deaths are inevitable (i.e., when some people in the group or everyone will die anyway) and in order to maximise the number of people saved someone has to be sacrificed, people tend to make utilitarian judgements again. Here the percentage is very high (91%) of people who answered that it is appropriated for a captain to kill the injured people in order to provide enough oxygen for the majority to survive, in the *submarine* dilemma (Greene *et al.*, 2008b). We can also see here that there is a high percentage of utilitarian answers (71%) in the modified *lifeboat* dilemma (Greene *et al.*, 2008b), in which you throw into the water someone who will not survive anyway in order to save everyone's lives, 60% in the *crying baby* dilemma (Greene *et al.*, 2008b) in which you would have to smother your children in order to avoid the enemies soldiers killing the whole group of people including your children and even 62% of utilitarian judgements in *Sophie's choice* (Greene *et al.*, 2008b).

It seems that people tend to make utilitarian judgements not only when the death is inevitable but also when the cost/benefits in relation to saving lives is high, as in the *catastrophe case* dilemma proposed by Nichols and Mallon (2006). In this dilemma a train is transporting an extremely dangerous artificially produced virus to a safe disposal site. The virus is profoundly contagious and nearly always leads to the death of the victim within a matter of weeks. If the virus were to be released into the atmosphere, billions of people would die from it, and there is even a chance that it would kill more than half of the human population. In Nichols dilemma, Jonas sees that there is a bomb planted on the tracks and the only way to prevent it from exploding is to stop the train, pushing a stranger onto the rails. Nichols then found that 68% of the people who were asked to respond to this dilemma said that Jonas broke a moral rule, but only 24% said that the action was, after all things considered, the wrong thing to do.

[7] According to Hauser *et al.* (2007), voluntary visitors to the Moral Sense Test website (http:// www.moral.wjh.harvard.edu) from September 2003 to January 2004 answer that it is morally permissible for someone to divert the train (85%) and that it is not permissible to push the man from the bridge (12%). In Greene *et al.* (2008b) in the *standard trolley* dilemma the reported percentage of those who gave the utilitarian answer decreases to 82%.

[8] The *standard fumes* dilemma is the following: you are a late-night watchman in a hospital and due to an accident in the building next door, there are deadly fumes rising up through the hospital's ventilation system. In a certain room of the hospital there are three patients. In another room there is a single patient. If you do nothing the fumes will rise up into the room containing the three patients and cause their deaths. The only way to avoid the deaths of these patients is to press a switch, which will cause the fumes to bypass the room containing the three patients. As a result of doing this the fumes will enter the room containing the single patient, causing his death. Is it appropriate for you to hit the switch in order to avoid the deaths of the three patients?

Nichols explains what is going on. His hypothesis is that even if an action is thought to violate a rule, it might also be regarded as acceptable, all things considered. To judge that an action has violated a rule will be called judgments of "weak impermissibility". To judge that an action was wrong, all things considered, will be called judgments of "all-in impermissibility". According to Nichols and Mallon (2006), the findings reinforce the familiar problem posed by catastrophe cases: they indicate that most people are not absolutist deontologists. People think that sometimes it is all-in permissible to do something that violates a moral rule, including the rule that forbids killing innocent people. Nichols also states:

> The results also support the idea that there are two partly independent mechanisms underlying moral judgment. On the one hand, people have a general capacity to reason about how to minimize bad outcomes. On the other hand, people have a body of rules proscribing certain actions. This body of rules cannot be subsumed under the capacity to reason about how to minimize bad outcomes (Nichols and Mallon, 2006, p. 539).

Nichols proposes that the assessment of all-in impermissibility implicates three factors: cost/benefit analysis, checking for rule violations and emotional activations. For him in the cases of personal and impersonal trolleys the judgement of all-in impermissibility depends on both the presence of an emotion and the judgement that a rule has been violated, but in the absence of emotion the cost-benefit analysis typically wins. Emotional activation and thinking that a rule has been violated does not, however, necessitate a judgement that an action is all-in impermissible, since when the cost-benefit ratio is sufficiently high, people tend not to judge the action as all-in impermissible, as it was shown in the *catastrophe dilemma*.

Nichols gives an excellent account on how people make moral judgements, but the whole story has still to be completed. I suggest that not only is it the high number of people to be saved (the cost-benefit ratio), but also the inevitability of death which are the two main reasons why people alternate between the deontological reasoning that it is wrong to kill innocent's lives in a personal way (requiring some kind of proximity or body contact), to a utilitarian one that admits the killing even in that more personal way. However, we become deontological once again when asked to personally kill a person in order to satisfy a requirement of someone that you perceive as being evil. We do not have access to an experiment designed to test a proper *catastrophic* dilemma versus what I will call here the *blackmail dilemma.* However, the *modified safari dilemma* (Greene *et al.*, 2008b)[9] in which a group of terrorists promises to save your life and the lives of the children, if you personally kill one of the hostages who is being held with you, suggests that the cost-benefit utilitarian reasoning can still be overcome, at least in *semi-catastrophic* cases, by deontological considerations. It means that stage 4 (that we have mentioned in our model above) can still be overcome by deontological considerations as hypothesised in 5, i.e., if this killing has to be done to satisfy the outrageous requirement of a perceived evil person who blackmails you, threatening to kill a larger number of people if you refuse to carry out the killing, and not only this, requires your action (not merely omission). Nevertheless, why would some people still be willing to overcome the utilitarian cost/benefit analyses favouring a deontological norm, overcoming stage 4? A possible answer can be given if we admit that we have a

[9] In this dilemma a group of terrorists promises to save your life and the lives of children, if you kill one of the hostages who are being kept with you. The percentage of utilitarian answers in the *modified safari* is only 22%, according to Greene figures.

sense of dignity that makes us react emotionally to unfairness, rejecting it[10]. If this is the case, offers to save the lives of a group of people made by someone evil at the cost of a third innocent person being forced to carry out the killing to avoid the worse outcome, would trigger powerful emotional responses that would make people reject the offer. In this case, the response would not be as strong if they were not asked to personally kill the innocent person (which makes the majority of people still give the utilitarian answer in the *Sophie's choice* dilemma [62%]) (Greene *et al.*, 2008b) but in combination with the demand that the killing has to be carried out by the person who perceives the offer as outrageous, it would make people give the deontological response to the dilemma, saying that it is wrong to carry out the killing, exactly as it happens in the *modified safari* dilemma in which the percentage of utilitarian judgments is only 22% (Greene *et al.*, 2008b).

But not everybody judges as the majority of people do, so it will be important to understand and classify the various psychological/moral types according to how far they are in this chain, namely, how much they are willing to accept utilitarian reasons to overcome deontological constraints[11]. The majority of us are probably situated somewhere in the middle. The pure deontological type, are those who, for example, think that we are not allowed even to flip the switch diverting the trolley. The other extreme is the pure utilitarian types, probably be the 2% quoted by Hauser who answered that in a plane crash - the plane crash dilemma - it is appropriated for you and another man to sacrifice the life of a wounded boy that you conclude has no chance of survival in order to eat him and survive (Greene *et al.*, 2008b). In between these two pure deontological and utilitarian types we have all the other psychological/moral types. I have to stress that this is at the moment an entirely hypothetical philosophical model, and further work must be carried out by moral psychologists and neuroscientists in order to test and refine this model, as well as to establish exactly the different moral types.

## Greene's dual process theory of moral judgement

In order to try to understand the mechanisms underlying moral judgements in moral dilemmas involving life and death we could benefit from the Studies of Greene. Greene and his collaborators developed a dual-process theory of moral judgement (Greene *et al.*, 2001; Greene *et al.*, 2004; Greene, 2007; Greene *et al.* 2008a; Greene, 2009). In this model deontological judgements are driven by automatic emotional responses, while characteristically utilitarian judgments are driven by controlled cognitive processes.

[10] There is no experiment to test catastrophic versus blackmail dilemmas, but fiction could give us some important clues about how people would react in these cases. In the film "Batman: the dark knight" the Joker put people in two different boats with two detonators , and asked each group to detonate the bomb in the other boat, saying the first group to do this would have their lives saved. The main criminals in Gordon City were in one of the groups, but even this group refused the Joker's offer.

[11] I propose to set up the bases for a scale in terms of judging in a deontological or utilitarian way, a scale that ranges from +N to -N, being that the psychological/moral type scoring highest in the positive side is entirely deontological (not accepting any exceptions for the rule one should not kill) and the psychological type scoring lowest in the negative side is entirely utilitarian, accepting that when lives are at stake we should always and in any circumstances save the highest possible number of people, even if you have to kill some people to reach this result. The more a person is willing to abide by the rule that we should not kill and the more a person is not willing to accept any special cases where it is permissible to kill someone to save more people, the higher he/she will score as the deontological type, and the more she/he is willing to consider utilitarian reasons to break this rule, the higher he/she scores as the utilitarian type.

According to Greene (2005, p. 350):

> Multiple sources of evidence point toward the existence of at least two relatively independent systems that contribute to moral judgment: (i) an affective system that (a) has its roots in primate social emotion and behaviour; (b) is selectively damaged in psychopaths and certain patients with frontal brain lesions; and (c) is selectively triggered by personal moral violations, perceived unfairness, and, more generally, socially significant behaviours that existed in our ancestral environment. (ii) a "cognitive" system that (a) is far more developed in humans than in other animals; (b) is selectively preserved in the aforementioned lesion patients and psychopaths; and (c) is not triggered in a stereotyped way by social stimuli.

For Greene (2008), "cognitive" representations are inherently neutral representations, ones that do not automatically trigger particular behavioural responses or dispositions, whilst "emotional" representations do have such automatic effects. Emotion tends to be associated with parts of the brain, such as the amygdale and the medial surfaces of the frontal and parietal lobes. On the other hand, "cognitive" processes are especially important for reasoning, planning, manipulating information in the working memory, controlling impulses, and "higher executive functions" more generally. These cognitive functions tend to be associated with certain parts of the brain, primarily the dorsolateral surfaces of the prefrontal cortex and parietal lobes (Greene, 2008). Borg defines emotion and reason in a similar way (Borg *et al.*, 2006). For him ''emotions'' are immediate valenced reactions that may or may not be conscious. In contrast, ''reason'' is neither valenced nor immediate insofar as reasoning need not incline us toward any specific feeling and combines prior information with new beliefs or conclusions and usually comes in the form of cognitive manipulations (such as evaluating alternatives) that require working memory. He points out that emotion might still affect, or even be necessary for, reasoning but emotion and reasoning remain distinct components in an overall process of decision making.

Greene puts forward the personal/impersonal distinction (Greene *et al.*, 2001; Greene *et al.*, 2004). A personal moral violation is one in which (a) the violation must be likely to cause serious bodily harm, (b) this harm must befall a particular person or a set of persons and (c) the harm must not result from the deflection of an existing threat onto a different party. Dilemmas that fail to meet these three criteria are classified as impersonal. For Greene, dilemmas such as the *standard trolley* dilemma are impersonal, whilst the *footbridge dilemma* is personal. Even if we do not accept the personal/impersonal distinction[12] suggested by Greene, the fMRI data in his research shows, at least, that there is a crucial difference between the trolley dilemma and the footbridge dilemma, the main difference being that the footbridge dilemma engages people's emotions in a way that the trolley does not. Greene proposed that the thought of pushing someone to his death is more emotionally salient than the thought of hitting a switch that will produce similar consequences. As it was observed through brain images (Greene *et al.*, 2001) the contemplation of personal moral dilemmas like the *footbridge* case produces increased neural activity in brain regions associated with emotional response and social cognition (typically the posterior cingulated cortex, the medial prefrontal cortex and

[12] McGuire reanalysed the RT data from the Greene research and claimed (a) that there is no reason to assume that emotionally salient moral decisions are processed in a qualitatively different way to those dilemmas that are not emotionally salient and (b) that there is no evidence here to support the theory that there are two competing moral systems at work (McGuire *et al.*, 2009). Greene (2009) replies to the objection emphasizing that the dual-process theory is independent of the personal/impersonal distinction. The basic idea of the reply is that even if the distinction personal/impersonal does not hold up, the dual-process theory does.

the amygdale, as well as the superior temporal sulcus), whilst the contemplation of impersonal moral dilemmas such as the *trolley* case produces relatively greater activity in brain regions associated with "higher cognition" (as the dorsolateral prefrontal cortex and the inferior parietal lobe) (Greene, 2005).

Greene proposes then that the tension between the deontological and utilitarian perspectives in moral philosophy reflects a more fundamental tension arising from the structure of the human brain (Greene *et al.*, 2004). For him the social-emotional responses that we have inherited from our primate ancestors, shaped and refined by culture underpin the deontological absolute prohibitions whilst the "moral calculus" that defines utilitarianism is made possible by more recently evolved structures in the frontal lobes that support abstract thinking and high-level cognitive control. He supports his claim by showing that there is evidence of increasing emotional-social processing in cases in which deontological intuitions are prominent and greater activity in brain regions associated with cognitive control where utilitarian judgements prevail. There was further support for this claim when it was found (Greene, 2008) that cognitive load selectively increased RT (response time) for utilitarian judgment, yielding the predicted interaction between load and judgment type. According to him, in the full sample, load increased the average RT for utilitarian judgments by three quarters of a second, but did not increase average RT for non-utilitarian judgments at all and the predicted RT effects were observed in participants who tend to lean toward utilitarian judgment as well as those who do not. These results, concluded Greene, provided direct evidence for the hypothesised asymmetry between utilitarian and non-utilitarian judgments, with the former driven by controlled cognitive processes and the latter driven by more automatic processes.

On the other hand, the fMRI data obtained by Borg *et al.* (2006) suggests that some deontological responses can be mediated by reason, where other deontological responses can be mediated by emotion. They point out that the individual will use varying combinations of cognitive and emotive facilities to address moral challenges, but, overall, certain types of moral scenarios are likely to be processed in characteristic ways.

Only further research will give the definitive answer on up to which point deontological judgements are essentially emotionally driven and utilitarian judgements are essentially cognitively driven. Putting aside for further discussion how these interactions emotion/cognition operate in the brain, it is reasonable to suppose that there is a typical way of processing and solving moral dilemmas involving killing which involves a combination of deontological prohibitions and utilitarian calculus, although people are not conscious of how they operate (Hauser *et al.*, 2007). It is not clear whether or not people can switch on and off their deontological and utilitarian ways of thinking, switching also on and off their emotional and cognitive systems and whether or not their deontological and utilitarian responses always match the deontological/emotional and utilitarian/cognitive system. There are some indications, however, that people reason in a deontological/utilitarian way when responding to these dilemmas, and some typical conditions (as the possibility of saving more lives. personal force/proximity, cost/benefit, inevitability of death, or even outrage and integrity as we will see now) are able to trigger the final deontological or utilitarian response.

## Integrity, fairness and the ultimatum-game



The ultimatum game is a designed experiment to test, among other things, how people react to unfairness. In this game the proponent (first player) is given

a sum of money and he is asked to share the sum with the second player. He can choose the amount of money that he will offer to the second player, but if the second player refuses the offer, none of them will earn anything. It seems correct to deduce from this that if people are interested only in obtaining economical benefits the second player would accept all the offers, as something is always better than nothing. However, what generally happens is that offers of less than 20% of the total amount are frequently rejected (Heinrich, 2000)[13]. So what can we infer from this result? Basically, we can deduce that human beings have a sense of fairness so strong that even in situations where we know that we have nothing to loose we are not willing to put up with unfairness. The ultimatum game seems to show that to a higher or lesser degree human beings are not absolutely determined by the desire of obtaining advantages at any cost.

This behaviour suggesting inequality aversion, in some rudimental forms, seems to be shared with other primates as shown by the behaviour of female capuchins (De Waal and Brosnan, 2003). In an experiment De Waal showed that when these females capuchins received cucumber whilst the other female participants received grapes (grapes being a much more favoured food than cucumber for capuchins), they refused to cooperate or even to eat the food. De Waal found that the presence of high-value rewards (grapes) reduced the tendency to exchange for low-value rewards (cucumber) being the strongest increase of refusal to occur if another capuchine received better rewards without any effort (De Waal and Brosnan, 2003). De Waal's hypothesis is that even non human primates are guided by species-typical expectations about the way in which one (or others) should be treated and how resources should be divided.

The results of the ultimatum game together with Waals experiments might be a strong indication that human beings have evolved to have dignity and this would account for some deontological lines that we are not willing to cross, despite the price we have to pay for it. This could explain the ultimate deontological barrier, specified in step 5 of our model where people refuse to accept being blackmailed, and could also explain why we have such a low percentage (22%) of utilitarian answers to the *modified safari* dilemma (Greene *et al.,* 2008b). Here, personally killing an innocent person in order to satisfy the outrageous demands of an evil being is perceived as absolutely prohibited, triggering our deontological buttons despite of the minimisation of negative consequences that it would bring about (less deaths). Accepting the offer would violate our sense of dignity to such an extent that the majority of people prefer to decline the offer. There is a clear similarity here with the rejection of unfair offers in the ultimatum game. The offer is so clearly perceived as outrageous that despite all utilitarian considerations (save as many lives as you can) people prefer to say no.

Bernard Williams has already proposed a similar dilemma (the Jim Dilemma) in order to object to utilitarianism (Williams, 1973). In Williams example Jim got lost whilst on a botanical expedition and found himself in a small town where a row of twenty Indians were tied up against the wall and in front of them there were several armed men in uniform. The captain in charge (Pedro) explained to Jim that the Indians were a random group of inhabitants who, after a recent protest against the government, were just about to be killed as a reminder to other protesters of why they should not protest. The captain then offers to Jim the privilege of killing one of the Indians himself. If Jim accepts the captain will release all the others, but if Jim refuses Pedro (the captain) will carry out what he was about to do before Jim

[13] According to Heinrich (2000), the Machiguenga of the Peruvian Amazon are an exception to the rule of typically rejecting offers lower than 20%, as they typically accept even very low offers.

arrived, and will kill all of the Indians. Williams uses this example to criticise the strong notion of negative responsibility that he thinks is attached to consequentialism: if I know that if I do X, O1 will eventuate, and if I refrain from doing X, O2 will, and that O2 is worse than O1, then I am responsible for O2 (Williams, 1973). In the case of Jim, if he refuses to accept the offer and kill only one Indian it will make him responsible for all the other deaths. For Williams this is not a good interpretation of what is happening. He also uses this example to claim that utilitarianism does not leave room for personal integrity.

Williams argues that it is misleading to focus on Jim. The person undoubtedly responsible for what happens is Pedro and so we should be thinking about the effect of Pedro's project on Jim's decision. For Williams the utilitarian approach to the question "makes Jim a channel between the input of everyone's projects, including his own, and an output of optimistic decision; but this is to neglect the extent to which *his* actions and *his* decisions have to be seen as the actions and decisions which flow from the projects and attitudes with which he is most closely identified. It is thus, in the most literal sense, an attack on his integrity" (Williams, 1973, p. 116).

The view of Williams that somehow Jim's integrity is being attacked in this dilemma echoes a common feeling that people have, and it can be confirmed in the modified safari dilemma (Greene *et al.,* 2008b). It seems that, somehow, common sense captures the idea of an attack on our integrity in people being used in someone's malign projects. The majority of us refuses to be used by others to carry out *their* evil projects, actively killing innocent people, even if this refusal does not fit into a cost/benefit model. There are some acts that attack our dignity so strongly that people still refuse to carry them out even knowing that the act will cause the least damaging outcome in certain circumstances. The only way to explain why we refuse to carry out these acts is appealing for deontological notions of integrity and dignity. Here, going beyond Williams, Kant's considerations on morality, dignity and integrity seem to be able to teach us something.

## Vindicating Kant

Hauser (2009) sustains that much of our knowledge of morality is intuitive and based on inaccessible principles that guide our judgements, and not based on a conscious reflection on these principles. Given the apparent incapacity of the average person to supply the reasons for their moral judgements, these would go totally against what Kant theorises. But could it really be that, in fact, the studies of Hauser, Haidt (2001) and others really contradict Kant's moral theory? If really these studies challenge it, where is the point of collision?

The studies showing that people make moral decisions without consciously reasoning via moral principles does not refute what Kant affirms about the way people make their moral judgements. In fact Kant explicitly tells us that to act morally, in other words, to act "from the motive of duty" requires that people act under reflection, applying the Categorical Imperative: "Hence nothing other than the representation of the law in itself, which can of course occur only in a rational being, insofar as it and not the hoped-for effect is the determining ground of the will, can constitute the preeminent good we call moral, which is already present in the person himself who acts in accordance with this representation and need not wait upon the effect of his action" (Kant, 1997, p. 14). However, Kant admits that many of the human actions are made "in conformity with duty", i.e., they coincide with the duty but they are not carried out from the representation and application of the categorical imperative, in other words, they are not "from duty". Kant tells us

that moral actions must be those of the second type, but he has always appeared sceptical about the effective existence of these actions, having demonstrated their possibility, but never their existence. According to Kant "there is, however, something so strange in this idea of the absolute worth of a mere will, in the estimation of which no allowance is made for any usefulness, that, despite all the agreement even of common understanding with this idea, a suspicion must yet arise that its covert basis is perhaps mere high-flown fantasy and that we may have misunderstood the purpose of nature in assigning reason to our will as its governor. Therefore, we shall put this idea to the test from this point of view" (Kant, 1997, p. 8). Kant puts the idea of good will at stake and at no point demonstrates that the pure reason determines the will, which means he does not prove that moral actions exist. Kant believes that it is absolutely impossible by means of experience to make out with complete certainty a single case in which the maxim of an action otherwise in conformity with duty rested simply on moral grounds and on the representation of one's duty. What Kant actually shows us is that pure reason *can* determine our actions, in other words, that we *can* act morally. He shows that we can act based only on the representation of the Categorical Imperative, but, at no time does he prove that we actually act based on the representation of the Categorical Imperative, that we act "from duty" (Kant, 1997), that we act morally.

So, the conclusion seems clear: if Kant does not affirm that people in fact act from duty, namely, for pure and simple representation of the categorical imperative, then studies which conclude that people do not judge morally from a conscious reflection on principles would not affect Kant's moral view. These studies would simply point to the fact that most people do not act "from duty", something that Kant had already suspected. On the other side, there is another sense, a non-trivial sense in which Kant's moral philosophy might be questioned. In this non-trivial sense, the moral philosophy of Kant might be questioned from the evidences that the moral judgements made by the majority of people would not correspond even to what Kant calls "in conformity with duty". Bearing this in mind and returning to the *trolley problem*, what would really threaten Kant's theory is the observation that 85% of people think that it is right to divert the train killing one person instead of five (Hauser *et al.*, 2007). If that is the case, the basic intuitions of people about what is right or wrong would not corroborate, at least not in this case, what the categorical imperative prescribes to us, i.e., that people must always be treated as an end and never as a means. Killing one person to save five, even being a consequence of a double effect, would be seen as immoral when we apply the categorical imperative. The fact that 85% of the people do not agree with this would suggest that the agreement between the common sense and the moral theory of Kant[14] might be questioned. So Kant's philosophy would be open to question not because people do not make moral judgements based on principles, but because people's ordinary moral judgements about what is right or wrong does not coincide with what the categorical imperative prescribes as being right or wrong.

So what are the elements (or element) in the common judgement that are not present in Kant´s theory? My hypothesis here is that it is the utilitarian element. In certain situations people think that they are allowed to violate deontological prohibitions usually using utilitarian criteria to do so. It seems that we humans are willing to maximise welfare and willing to save as many lives as possible. Nevertheless, there is a limit up to which we are prepared to go. If, as it seems to be , there is a

[14] In the first section of *GM* Kant (1997) establishes that the layman, without any philosophical education, knows already what is right or wrong. There is, then, an agreement between the categorical imperative and common reason, there is an agreement between the ordinary moral knowledge and philosophical knowledge.

deontological-utilitarian way in which our minds work when we have to judge and make our moral decisions, it would be worthwhile to follow this path to investigate it further. If people use in general deontological criteria to make moral judgements, but replace these deontological criteria for utilitarian ones in some circumstances and vice-versa, it would be very promising to establish a better dialogue between deontology and utilitarianism, between Kant and Mill, in order to decipher our deontological-utilitarian (deontoutilitarian) minds.

## References


BORG, J.S.; HYNES, C.; VAN HORN, J.; GRAFTON, S.; SINNOTT-ARMSTRONG, W. 2006. Consequences, Action, and Intention as Factors in Moral Judgments: An FMRI Investigation. *Journal of Cognitive Neuroscience,* **18**(5):803-817. http://dx.doi.org/10.1162/jocn.2006.18.5.803

CUSHMAN, F.; YOUNG, L.; HAUSER, M. 2006. The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm. *Psychological Science,* **17(**12):1082-1089. http://dx.doi.org/10.1111/j.1467-9280.2006.01834.x

DE WAAL, F.; BROSNAN, S. 2003. Monkeys reject unequal pay. *Nature,* **425**:297-299. http://dx.doi.org/10.1038/nature01963

FOOT, P. 2002. The problem of Abortion. *In:* P. FOOT, *Virtues and Vices.* Oxford, Clarendon Press, p. 19-32. http://dx.doi.org/10.1093/0199252866.003.0002

GREENE, J.; SOMMERVILLE, B.; NYSTROM, L.; DARLEY, J.; COHEN, J. 2001. An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science,* **293**:2105-2108. http://dx.doi.org/10.1126/science.1062872

GREENE, J.; NYSTROM, L.; ENGELL, A.; DARLEY, J.; COHEN, J. 2004. The Neural Bases of Cognitive Conflict and Control in Moral Judgement. *Neuron*, **44**:389-400. http://dx.doi.org/10.1016/j.neuron.2004.09.027

GREENE, J. 2005. Cognitive Neuroscience and the Structure of the Moral Mind. *In:* S. LAURENCE; P. CARRUTHERS; S. STICH (eds.), *The Innate Mind: Structure and Contents.* New York, Oxford University Press, p. 338-353. http://dx.doi.org/10.1093/acprof:oso/9780195179675.003.0019

GREENE, J. 2007. Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences,* **11**(8):322-323. http://dx.doi.org/10.1016/j.tics.2007.06.004

GREENE, J.; MORELLI, S.; LOWENBERG, K.; NYSTROM, L.; COHEN, J. 2008a. Cognitive Load Selectively Interferes with Utilitarian, Moral Judgment. *Cognition,* **107**:1144-1154. http://dx.doi.org/10.1016/j.cognition.2007.11.004

GREENE, J.; MORELLI, S.; LOWENBERG, K.; NYSTROM, L.; COHEN, J. 2008b. Supplementary materials for Cognitive Load Selectively Interferes with Utilitarian, Moral Judgment. *Cognition,* **107**:1144-1154. Available at: http://wjh.harvard.edu/~mcl/materials/Greene-CogLoadSupMats.pdf. Accessed on: 10/09/2013.

GREENE, J. 2008. The Secret Joke of Kant's Soul. Available at: http://www.wjh.harvard.edu/~jgreene/GreeneWJH/Greene-KantSoul.pdf. Accessed on: 10/09/2013.

GREENE, J.; CUSHMAN, F.; LOWENBERG, K.; NYSTROM, L.; COHEN, J. 2009. Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition,* **111**(3):364-371. http://dx.doi.org/10.1016/j.cognition.2009.02.001

GREENE, J. 2009. Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. *Journal of Experimental Social Psychology*, **45**(3):581-584. http://dx.doi.org/10.1016/j.jesp.2009.01.003

HAIDT, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review,* **108**:814-834. http://dx.doi.org/10.1037/0033-295X.108.4.814

HARRIS, J. 1975. The Survival Lottery. *Philosophy,* **50**(191):81-87. http://dx.doi.org/10.1017/S0031819100059118

HAUSER, M.; CUSHMAN, F.; YOUNG, L.; JIN, K.-X.; MICKAIL, J. 2007. A dissociation between Moral Judgements and justifications. *Mind & Language*, **22**(1):1-21. http://dx.doi.org/10.1111/j.1468-0017.2006.00297.x

HAUSER, M.D. 2009. *Moral Minds.* London, Abacus, 493 p.

KANT, I. 1997. *Groundwork of the Metaphysics of Morals.* Cambridge, Cambridge University Press, 86 p.

KOENIGS, M.; YOUNG, L.; ADOLPHS, R.; TRANELL, D.; CUSHMAN, F.; HAUSER, M.; DAMASIO, A. 2007. Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature,* **446**:908-911. http://dx.doi.org/10.1038/nature05631

HENRICH, J. 2000. Does Culture matter in economic behaviour? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *The American Economic Review,* **90**(4):973-979. http://dx.doi.org/10.1257/aer.90.4.973

MCGUIRE, J.; LANGDON, R.; COLTHEART, M.; MACKENZIE, C. 2009. A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology,* **45**:577-580. http://dx.doi.org/10.1016/j.jesp.2009.01.002

MOORE, A.; CLARK, B.; KANE, M. 2008. Who Shalt Not Kill? Individual Differences in Working Memory Capacity, Executive Control, and Moral Judgment. *Psychological Science,* **19**(6):549-557. http://dx.doi.org/10.1111/j.1467-9280.2008.02122.x

NICHOLS, S.; MALLON, R. 2006. Moral dilemmas and moral rules. *Cognition,* **100**:530-542. http://dx.doi.org/10.1016/j.cognition.2005.07.005

WILLIAMS, B. 1973. A Critique of Utilitarianism. *In:* B. WILLIAMS; J.J.C. SMART, *Utilitarianism: For and Against.* Cambridge, Cambridge University Press, p. 75-155.

*Submitted on July 22, 2012*
*Accepted on August 10, 2013*