Judgment and Decision Making, Vol. 14, No. 5, September 2019, pp. 555–564

Decision conflict drives reaction times and utilitarian responses in sacrificial dilemmas

Alejandro Rosas* Juan Pablo Bermúdez† David Aguilar-Pardo‡

Abstract

In the sacrificial moral dilemma task, participants have to morally judge an action that saves several lives at the cost of killing one person. According to the dual process corrective model of moral judgment suggested by Greene and collaborators (2001; 2004; 2008), cognitive control is necessary to override the intuitive, deontological force of the norm against killing and endorse the utilitarian perspective. However, a conflict model has been proposed more recently to account for part of the evidence in favor of dual process models in moral and social decision making. In this model, conflict, moral responses and reaction times arise from the interplay between individually variable motivational factors and objective parameters intrinsic to the choices offered. To further explore this model in the moral dilemma task, we confronted three different samples with a set of dilemmas representing an objective gradient of utilitarian pull, and collected data on moral judgment and on conflict in a 4-point scale. Collapsing all cases along the gradient, participants in each sample felt less conflicted on average when they gave extreme responses (1 or 4 in the UR scale). They felt less conflicted on average when responding to either the low- or the high-pull cases. The correlation between utilitarian responses and conflict was positive in the low-pull and negative in the high-pull cases. This pattern of data suggests that moral responses to sacrificial dilemmas are driven by decision conflict, which in turn depends on the interplay between an objective gradient of utilitarian pull and the moral motivations which regulate individual responsiveness to this gradient.

Keywords: conflict, decision conflict, deontology, dual-process, moral dilemmas, utilitarianism

1 Introduction

Many of our cognitive processes are automatically completed outside our reflective control. Other processes require effort and conscious attention, and consume scarce cognitive resources (Evans & Stanovich 2013, Kaheneman 2011). Research into moral judgment with sacrificial dilemmas incorporated this dual-process account. The pioneering model of moral cognition advanced by Greene and collaborators (2001; 2004; 2008; hereafter 'Greene's model') brings two types of cognitive process – intuitive and reflective – in tidy alignment with the two types of moral judgment – deontological (respecting individual rights in conflict with the greater good) and utilitarian (willingness to violate individual rights to promote the greater good). This alignment allows researchers interested in the mechanism of moral cognition to import from cognitive science a direct way of testing the model. Greene's model predicts that manipulating participants into responding intuitively in sacrificial dilemma tasks (by e.g., incorporating in the design a limited time budget or a parallel task to load working memory) should decrease the proportion of utilitarian responses (hereafter: URs). However, experiments designed to elicit intuitions have not clearly confirmed the prediction. Studies go both ways: some confirm the prediction and some do not (Greene et al. 2008; Suter & Hertwig 2011; Trémolière & Bonnefon 2014; Tinghög et al. 2016; Gürçay & Baron 2017; Bago & De Neys, 2018; Rosas & Aguilar-Pardo, in press). These results counsel an open mind regarding whether hypothesizing a competition between utilitarian reflection and deontological intuition can effectively account for the data.

Researchers have used reaction times (RTs) as evidence for dual process models both when investigating cooperation in the face of selfish opportunities (Rand, Greene & Nowak 2012), and when investigating utilitarian decisions that involve violating a deontological norm against killing (Greene et al. 2001; 2004). However, some researchers in the former field have recently suggested that their data fit better with a decision conflict model that falls outside the framework of dual process theories (Evans et al. 2015; Krajbich et al. 2015). Their new hypothesis is that the choice to cooperate

We gratefully acknowledge helpful comments from Bence Bago, Michał Białek and Jonathan Baron.

This project was funded by the Research Division of the Universidad Nacional de Colombia, project 37159, 2017–2018 and by the Universidad Externado de Colombia, project 370072017–2018. Study 2 was conducted in collaboration with the LINCIPH lab at the Universidad Externado de Colombia.

*Department of Philosophy, Universidad Nacional de Colombia. Email: arosasl@unal.edu.co.

†Philosophy Program, Faculty of Social and Human Sciences Universidad Externado de Colombia, Bogotá, Colombia.

‡Psychology Program, Faculty of Psychology, Universidad Católica de Colombia

https://doi.org/10.1017/S193029750000485X Published online by Cambridge University Press

or not depends both on individual preference and on the specific choice problem constructed by experimenters. When the choice problem is held constant, participants with strong preferences regarding cooperation in tension with selfish temptation will respond faster than participants which have no clear preference and are torn between the options. If RTs are plotted on the Y-axis as a function of responses on a scale from strongly selfish to strongly cooperative on the X-axis, the result is an inverted-U pattern with fast responses at both extremes of the decision scale and slow responses at the middle. Alternatively, if participants' preferences were to show no variability in a sample, experimenters could nonetheless elicit variation in their responses. It would require only that some participants receive a choice problem where cooperation involves sacrificing a large monetary gain, while others receive a cooperation option involving only a small monetary loss compared to defection. Results in tasks where cooperation partially conflicts with selfishness, termed social dilemmas, have shown that RTs and levels of conflict can be explained by the interplay of these objective and subjective parameters, namely the choice problem on the one hand, and the variation in individual responsiveness to either cooperation or selfish defection, on the other (Evans et al. 2015, Krajbich et al. 2015).

A similar model has been applied to moral dilemmas research by Baron et al. (2012) and Baron and Gürçay (2017). They proposed a conflict model where the RTs of moral response depend on the interaction between dilemma difficulty and individual ability. "Difficulty" refers to how much the scenario discourages an UR, by offering only a small utilitarian benefit or a large loss from action. This is the objective parameter, represented in the contents of the scenario, i.e., the choice problem. Individual ability refers to the individual's disposition towards violating a deontological norm in order to achieve a greater good. Presumably, it varies normally in the population. The prediction of this model is that participants will be torn between the options when difficulty matches ability, and in those cases RTs will be slower. In contrast, when either difficulty or ability exceeds the other, RTs will be faster and the probability of either a deontological or utilitarian response will be greater. Hence, fast RTs can be found in both deontological and utilitarian responses. But dual process models predict that utilitarian responses should generally take longer than deontological ones even when difficulty and ability are equal, because this follows from the different nature of the cognitive processes supporting them and not from the variability of objective and subjective parameters and their interplay. However, a meta-analysis of RT to moral dilemmas in 26 data-sets supports the conflict model against the dual process model (Baron & Gürçay 2017).

In this paper we explore the virtues of the conflict model in research with moral dilemmas. We implement an experimental design that takes a leaf from the book of the research with social dilemmas. In the latter field, the levels of cooperation in a sample can be manipulated by changing the quantitative difference between the rewards of cooperation and the rewards of defection. Although individuals vary in their disposition to cooperate in the face of selfish temptations – individuals switch from defection to cooperation or vice versa at different thresholds – all individuals are responsive to the quantitative parameters. If you decrease said difference, levels of cooperation in a sample will increase. According to the conflict model, this happens simply because the (in many cases intuitively salient) quantitative parameters of the choice problem match the cooperation thresholds of a larger proportion of individuals. It has little to do with whether participants had more or less time or opportunity to reflect.

A similar manipulation of objective parameters in tasks with moral dilemmas would require a set of cases capable of eliciting, in any given sample, a roughly linear increase from a low to a high mean of UR. If the different cases elicit at least three statistically different levels of mean UR, and the effect is sustained throughout different samples, we can be confident that the set of cases represents a gradient of "utilitarian pull", from low over medium to high. This gradient would represent the objective parameter of the conflict model, the "difficulty" of dilemmas (Baron et al. 2012; Baron & Gürçay 2017). Measuring individual ability is trickier, and we shall not attempt it here. Instead, we shall assume that samples are representative and that mean ability is held constant across samples. This is of course a simplification. However, it is unlikely that expected deviations from the population mean ability would bias the results in favor of either a dual process or a conflict model. Their predictions are different and only those of the conflict model would vary depending on the distribution of subjective moral preferences in a given sample. The conflict model predicts that URs will increase, within a given sample, at cases higher up the gradient of utilitarian pull. By using only high-conflict dilemmas (see methods), we maintain in all cases the qualitative conflict between utilitarian and deontological judgment. In all cases presented utilitarian approval needs to override "up close and personal" killing. This makes it difficult for the corrective dual process model to account for a significant increase in URs despite other changes in scenario contents. The conflict model, in contrast, conceives responses in the task as the result of an interplay between the motivational factors (moral proclivities) and objective quantitative factors present in the scenarios, something like an utilitarian "rate of return" that varies per case. Also, different individual proclivities create conflict with different cases. For example, participants disposed to choose a higher score in the UR scale at the low end of the gradient and participants disposed to choose a lower score at the high end, should both be expected to report more conflict.

https://doi.org/10.1017/S193029750000485X Published online by Cambridge University Press

Additionally, we should see the same inverted-U pattern observed by research on cooperation when plotting conflict ratings and RTs as function of choice (Evans et al. 2015). The manifestation of this pattern should be twofold: within any particular sample and given a well-balanced gradient of utilitarian pull across Case, conflict should be higher at the intermediate than at the extreme cases. But similarly, when pulling all cases together in a sample, conflict reported should be higher for the intermediate than for the extreme scores in the moral response scale. In contrast, the dual process model conceives of conflict as a function of competition between types of cognitive processes: the intuitive response should be equally strong along the gradient, for in high-conflict personal scenarios killing is always "up close and personal"; and the reflective response should be equally strong in all cases, for all cases offer utilitarian benefits. Furthermore, a dual process model could not explain why conflict should vary between the intermediate and the extremes scores in a moral response scale. And finally, a the dual process advocate would be rattled if conflict were reported higher by utilitarian respondents at one extreme of the utilitarian gradient and by deontological respondents at the other.

2 Studies (1 to 3)

2.1 Method

2.1.1 Participants

Studies 1 (N=100) and 3 (N=284) recruited participants online at http://prolific.ac. Study 2 (N=80) recruited advanced undergraduate and graduate participants by word of mouth from two Universities in Bogotá, Colombia. Informed consent was obtained for anonymous data collection. M_{age} = 29.2, 24.7 and 32.6 and proportion of females was 35%, 54% and 35% in Studies 1, 2 and 3 respectively. No subject participated in more than one of these studies. Studies 1 and 2 were conducted in Spanish; Study 3 was conducted in English. Participants were mainly from Colombia, Mexico, Spain, the United Kingdom and the U.S.A. Cross-cultural comparisons were neither planned nor conducted.

2.1.2 Design

We designed a sequence of three studies using only high-conflict personal scenarios, i.e., scenarios where the sacrifice of one person involves contact and/or personal force and is knowingly performed as a means, not merely as a side-effect, to save others. To create a gradient of utilitarian appeal across high-conflict-personal scenarios, we varied the number of utilitarian incentives built into them – at least one and at most three different incentives. The baseline scenario was Footbridge-like: it only had the rather pallid utilitarian incentive of a 1:5 kill-save ratio. Other scenarios added features that increase utilitarian attractiveness:

- Extreme kill-save ratio: Sacrificing one person will save 100.000 people.
- Selfish: The agent of the utilitarian sacrifice is among the people saved.
- Doomed: The person to be sacrificed will inevitably die independently of the sacrifice.
- Guilty: The person to be sacrificed threatens five others with imminent death.

We included a self-report measure of conflict in all three studies; measured the reaction time of moral judgment on the response screen in the lab in Study 2 and explored the effect of cognitive load with two conditions between-subjects (Load/No Load) in Study 3. Thus, we tested the role of conflict in three different ways: conflict self-report (Study 1), conflict self-report + response-time (Study 2), and conflict self-report + cognitive load (Study 3).

2.1.3 Procedure

In the three studies, participants read the high-conflict personal dilemmas in a within-subjects design, counterbalanced for order. After each stimulus, participants answered two questions:

1) "How right or wrong is it to cause the death of the person in order to save N others?" Responses were registered in a 4-point scale: "Totally wrong" "More wrong than right", More right than wrong", "Totally right".

2) "If you experienced conflict when morally judging the action in the scenario, how intense was the conflict?" Responses were registered in a 4-point scale: "No conflict", "Low intensity", "Intermediate intensity", "High intensity".

Study 2 was carried out in the LINCIPH Laboratory at the Universidad Externado de Colombia. SuperLab was used to present the tasks and measure the time taken. For each scenario, participants saw the task divided into three separate screens: (1) they saw the scenario description; then (2) the moral question and the 4-point scale where they registered their response; and then (3) the conflict question and the response scale. The reaction time was measured as the time spent at screen (2). Studies 1 and 3 were conducted online and time was measured only as duration for the whole task. Study 3 randomly assigned participants to two conditions, Load and No-Load (see below).

Stimuli We used high-conflict personal cases only. We used three, five and four cases in Studies 1, 2 and 3 respectively. All cases included utilitarian incentives, but differed in including either just one or two or three incentives (see the scenarios in the Appendix). Shark includes only the baseline incentive, save five people. The other scenarios added one

https://doi.org/10.1017/S193029750000485X Published online by Cambridge University Press

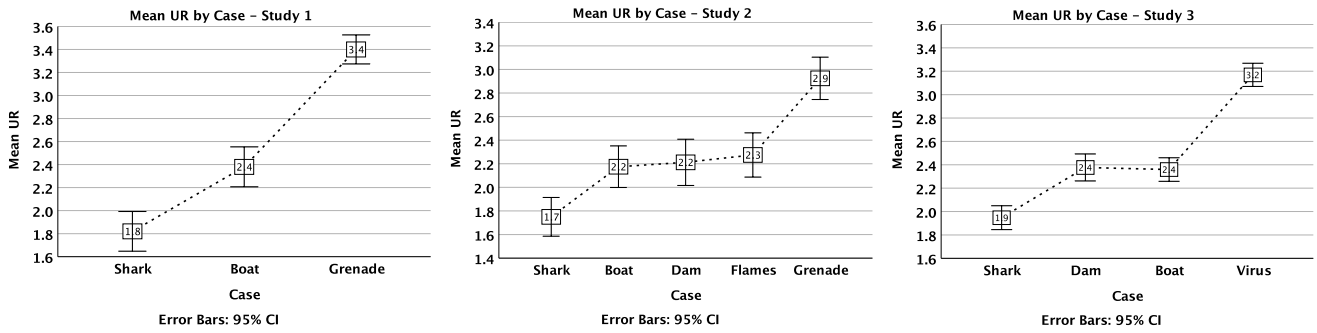

FIGURE 1: Mean of UR by Case in Studies 1, 2 and 3. Participants respond to the utilitarian incentives built into the scenarios by increasing URs. The scenarios create a gradient of utilitarian pull with three statistically distinct levels.

incentive to the baseline, unique to each scenario, excepting for Virus, which added two, thus including three incentives in total:

Shark: Includes only the baseline incentive: save five people.

Dam: Adds an extreme kill-save ratio: save 100.000 people.

Grenade: Adds a guilty victim, who credibly threatens the lives of five innocent others.

Flames: Adds a selfish incentive to the sacrifice: the agent saves herself and four others.

Boat: Adds a doomed victim, who will die anyway.

Virus: Adds both an extreme kill-save ratio and a doomed victim.

Some of the incentives used could be contested in their capacity to represent legitimate utilitarian reasons. However, besides the fact that they could arguably increase the utility – or decrease the disutility – of sacrificing one person, the plausibility of viewing them as legitimate components of an utilitarian theory is for us less important than their actual impact on participants responses. We call them utilitarian incentives if they move participants to shift their judgment in the direction of utilitarian approval. In this case, we side with folk wisdom rather than with explicit normative theories.

Previous research provides evidence of the influence of these incentives on URs (Moore, Clarke & Kane 2008; Huebner, Hauser & Pettit, 2011; Christensen et al. 2014; Tremolière & Bonnefon 2014; Rosas & Koenigs 2014; Bucciarelli, 2015; Gürçay & Baron 2017; Rosas et al. 2019). We included the selfish incentive only in Flames and eliminated it from our version of Boat (see Appendix). Study 3 introduces utilitarian incentives in a 2x2 design. Two levels of Kill-save ratio (5 or 100K) are combined with two levels of victim status (Doomed or Not-doomed): Shark (five saved, not-doomed), Dam (100k saved, not-doomed), Boat (five saved, doomed) and Virus (100K saved, doomed). We were not interested in mapping out the effect of different combinations of utilitarian incentives, but mainly in creating a positive gradient of utilitarian pull across scenarios, in each of the three studies.

Load in Study 3 In Study 3, we randomly assigned participants to a Load or no-Load condition. Load was implemented as a parallel dot memorization task (Bialek & de Neys 2017). Before reading each of the scenarios, participants saw for 2s a 4x4 dot matrix with five dots placed randomly. They were instructed to memorize it for future recognition. After answering the questions, participants were presented with four images of 4x4 matrices with five dots and were asked to identify the matrix they had seen a moment before. In the no-Load condition participants simply resolved the moral dilemma task in the absence of any parallel task.

2.2 Results

2.2.1 Utilitarian responses increase in response to utilitarian incentives

Visual inspection of the plot of mean of UR by Case (Figure 1) reveals that in all three studies the hypothesized utilitarian incentives significantly increase the mean of utilitarian response in the 4-point scale. It also reveals that Case creates a gradient of utilitarian pull with three statistically different levels. The gradient is well-balanced: the intermediate level is roughly midway between the low and the high levels. It is also consistent, for the cases included in the three studies consistently occupy the same level within the gradient.

2.2.2 Mean conflict is higher at the intermediate than at the extreme points of the utilitarian gradient

If conflict depends on a competition between deontological intuition and utilitarian reflection, it should not change significantly along the gradient. But if it depends on the interaction between individually variable moral preferences (how individuals differ in weighing deontological against utilitarian principles in high-conflict personal cases) and a quantitative utilitarian pull intrinsic to the choice and addressing the moral preferences in each individual, participants with clear preferences should weigh down the mean conflict ratings at the extremes of the gradient but keep it up at the intermediate level, where cases are balanced in their pull. In contrast,

https://doi.org/10.1017/S193029750000485X Published online by Cambridge University Press

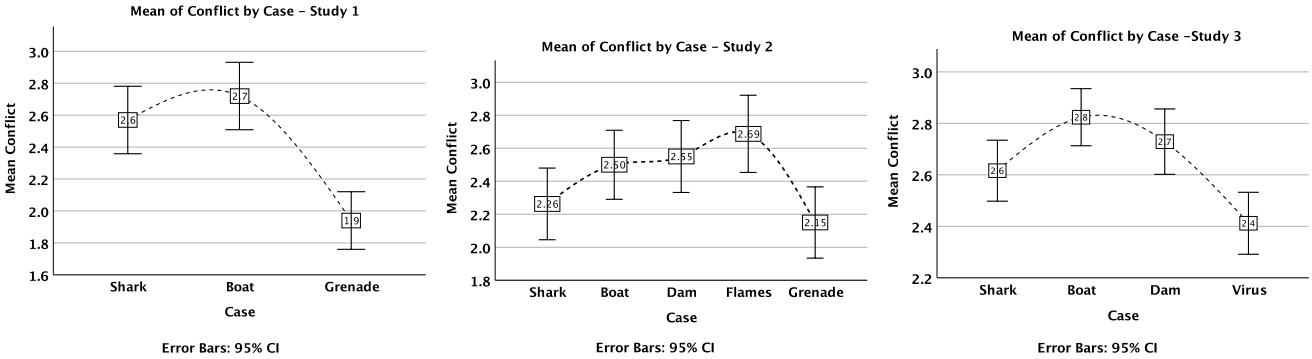

FIGURE 2: Mean of Conflict by Case in Studies 1, 2 and 3. The least conflictive cases are those that offer either the lowest or the highest utilitarian incentives.

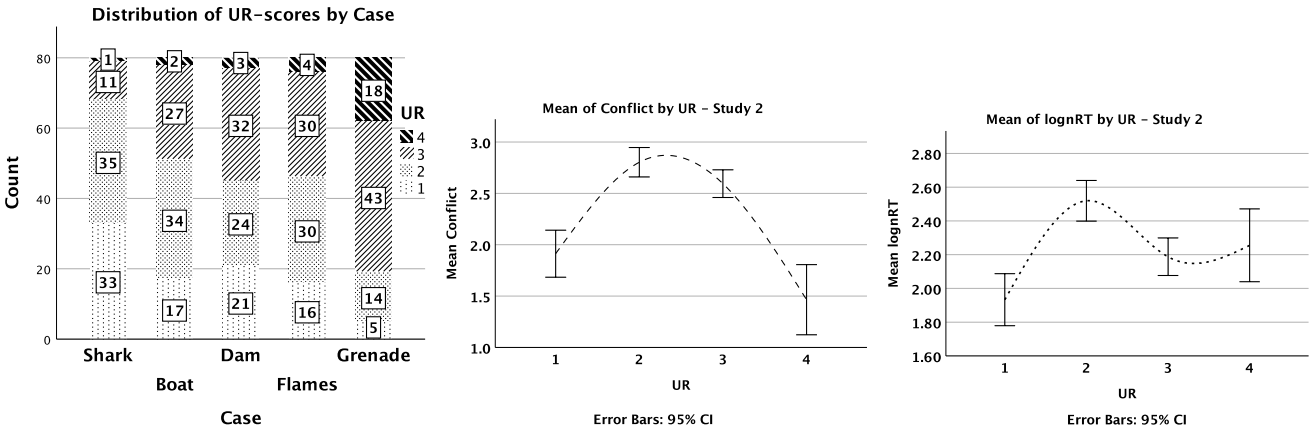

FIGURE 3: *Left panel*: Distribution of values in the 4-point scale of UR by Case in Study 2. *Middle panel*: Inverted-U pattern of mean Conflict by UR in Study 2: pooling cases together, mean conflict is lower when participants choose 1 or 4 from the 4-point scale of UR than when the choose 2 or 3. *Right panel*: deviation from the inverted-U pattern when mean lnRT is plotted as a function of the 4-point UR scale.

participants with no clear preferences should keep conflict high throughout. As a result, the intermediate levels of the gradient should exhibit higher mean conflict. This is what we observed (Figure 2).

To test the difference in conflict ratings between cases, we ran a Friedman analysis of variance by ranks with post hoc pairwise comparisons. In Study 1, Conflict was significantly lower in *Grenade* than in *Shark* and *Boat* (d =.6902 and .9687; $ps < 0.001$, Bonferroni corrected); in Study 2, Conflict was lower in *Grenade* than in *Dam* and *Flames* (d = .5496, p = 0.008 and d = .6161, p = 0.002, Bonferroni corrected); and in Study 3, Conflict was significantly lower in *Virus* than in *Shark*, *Dam* and *Boat* (d = .2593, .3755 and .4848; ps = 0.022, < 0.001, < 0.001 respectively, Bonferroni corrected). *Shark* consistently elicited the second *weakest* mean conflict after *Grenade* or *Virus*; its difference to the intermediate scenarios trended towards – but did not reach – significance (Figure 2, all three panels).

2.2.3 Mean conflict is lower at the extreme than at the intermediate scores of the UR scale

Our three studies presented a well-balanced gradient of utilitarian pull, i.e., the predominance of lower scores at the low end is compensated by a predominance of higher scores at the high end. Study 2 was the least well-balanced (Figure 3, left panel). We pooled cases together within each study and examined the levels of conflict for the 4 points of the UR scale. We observed an inverted-U pattern: responses of 1 and 4 along the 4-point scale exhibited significantly lower conflict than responses of 2 and 3. A Kruskal-Wallis test revealed that the distribution of Conflict was not the same across values of UR: conflict ratings of participants choosing 1 and 4 were significantly lower than ratings of participants choosing 2 and 3 in all three studies (all ps < .001, Bonferroni corrected). RT (ln transformed) was correlated with Conflict, although the correlation was far from large (r_s = .281, p < .001). When we plotted mean lnRT as a function of moral judgment in the 4-point scale, the line obtained deviated

https://doi.org/10.1017/S193029750000485X Published online by Cambridge University Press

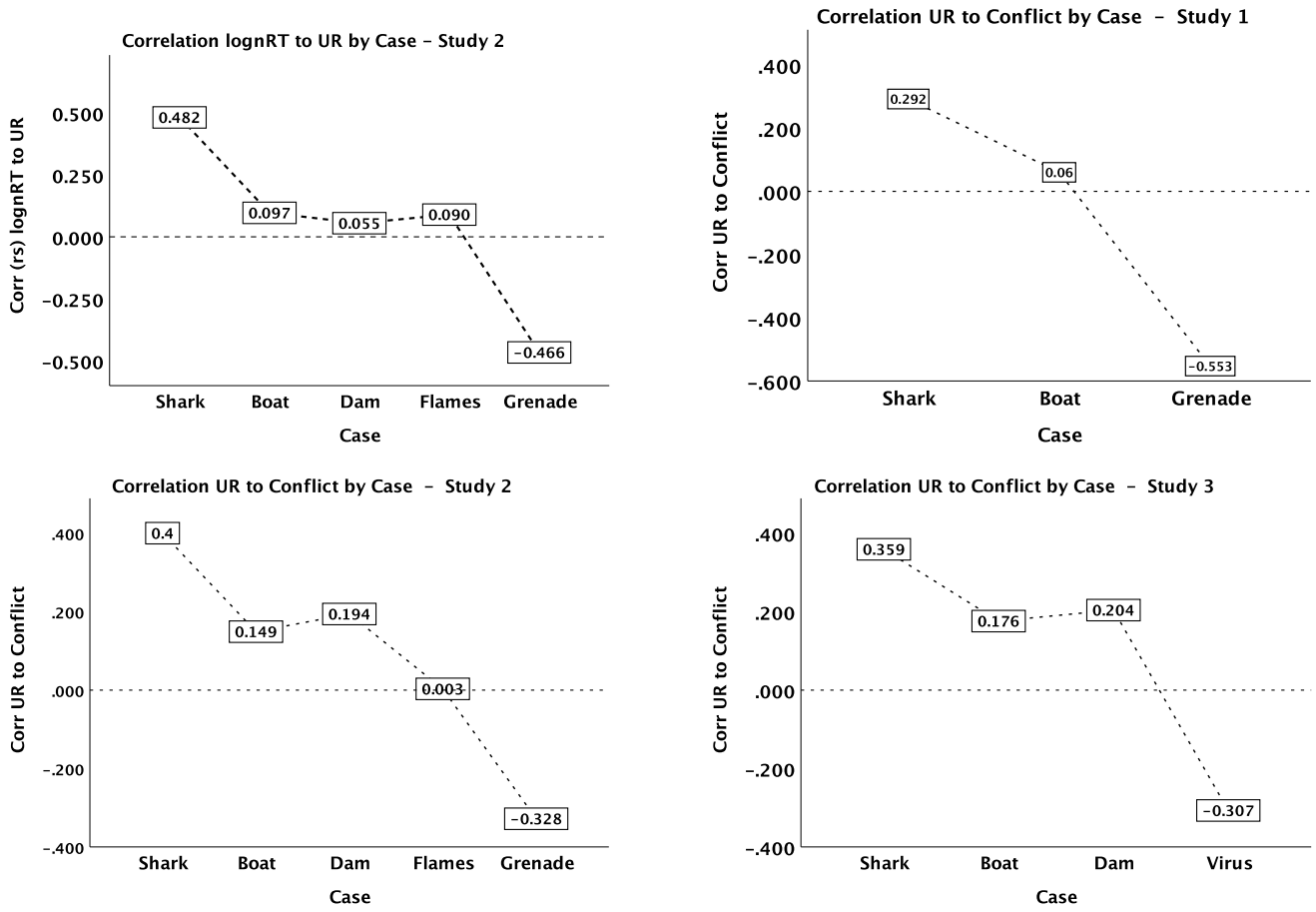

FIGURE 4: Spearman correlations between lnRT and UR in Study 2 (top left); and between Conflict and UR (top right to bottom right) by Case in Studies 1 to 3.

from the inverted-U pattern in the distribution of Conflict. A Kruskal-Wallis test revealed that the distribution of lnRTs was not the same across values of URs: RTs of participants choosing 1 and 3 were significantly faster than those of participants choosing 2 (*p*s < .002), but no other differences were significant. Checking our data, participants choosing 4 in the UR scale in *Shark*, *Boat* and *Flames* took longer, though they did not report higher conflict. We show both plots for Study 2, where we measured RT on the response screen (Figure 3, middle and right panel).

2.2.4 No correlation observed between UR and RT

The data supported no correlation between UR and lnRT in Study 2 ($r_s = 0.084$, $p = 0.110$), suggesting that, if RT data indicate reflection, they did not support a dual process model in our study. However, when the correlation was observed separately for each case, UR correlated positively with lnRT at the low end of the gradient (*Shark*) and negatively at the high end (*Grenade*). The same correlation pattern emerged in all three studies between Conflict and UR. Only the correlations at the low and high ends of the utilitarian gradient (*Shark*, *Grenade* and *Virus*) were consistently significant in all three studies (all ps ≤ .003). This indicates that participants within each sample feel conflict depending on the interplay between their particular moral preferences and the case judged. This pattern is preserved across samples, probably because the mean values in relative moral preference across samples are similar. We plot the correlations on Figure 4.

2.2.5 Effect of load in Study 3

One hundred and thirty two out of 144 participants (92%) in the Load condition correctly identified at least 3 of the 4 matrices accompanying the 4 cases. Also, total duration was positively correlated with Load ($r = .274$, $p < 0.001$), showing that participants under Load took significantly longer to complete the entire survey than participants in the no-Load condition. Load was coded "1" and no-Load "0". A linear regression of UR on Condition showed a very small effect of Load: Beta = −.065, $t = -2.188$, $p = .029$. Participants in the Load condition were slightly less utilitarian. This effect became marginally significant when controlling for Gender ($p = 0.054$). There were significantly more males than females

https://doi.org/10.1017/S193029750000485X Published online by Cambridge University Press

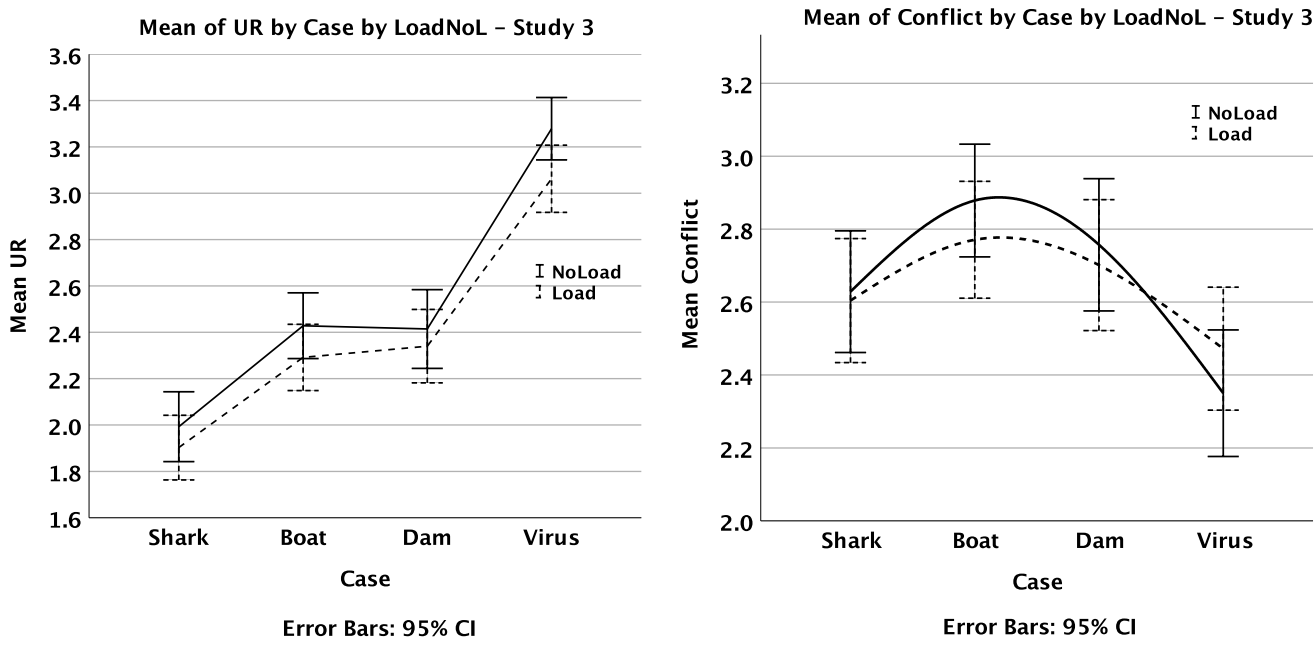

FIGURE 5: The effect of Case on UR was not importantly affected by Load (left); Load did not affect the inverted-U pattern in the distribution of Conflict across the values of the 4-point scale of UR (right).

in the no-Load condition, and males were significantly more utilitarian than females. Importantly, Load did not affect the patterns observed regarding the effect of Case on UR and the relationship between the scale of UR and Conflict (Figure 5). In relation to the later, a Kruskal-Wallis test revealed that the distribution of Conflict was not the same across the 4-point scale of UR: conflict ratings of participants choosing 1 and 4 were significantly lower than the ratings of participants choosing 2 and 3 for both the Load and no-Load conditions (all *ps* < .005, Bonferroni corrected).

3 General Discussion

In the three studies presented here, different samples were confronted with a set of moral dilemmas representing a gradient of utilitarian pull. We measured moral responses and conflict in a 4-point scale in all three studies, RTs on Study 2 and the effect of cognitive load in Study 3. The results fit better with a conflict model than with a dual process model. The conflict model was developed independently in research with moral (Baron et al., 2012; Baron & Gürçay 2017) and social dilemmas (Evans et al 2015; Krajbich et al. 2015). In research with social dilemmas, (prisoners' dilemmas and public good games) researchers found evidence for the hypothesis that cooperative and selfish decisions are differentially controlled by fast and slow cognitive processes (Rand, Greene and Nowak 2012). Newer research has found, however, that RTs depend more on the conflict between selfish and cooperative motivations in interplay with the quantitative parameters of the choice. In a representative sample, some individuals will be inclined to extreme selfishness, others to extreme altruism, and the majority to intermediate points. Individuals at the extremes have clear preferences one way or the other and will tend to make faster choices, while individuals at the middle of the motivational range will be torn between options and will make slower decisions. Also, the distribution and RTs of responses can be manipulated by changing the quantitative parameters of the choice problem. For example, by making the difference between the benefits of defection (always greater in one-shot social dilemmas) and the benefits of cooperation very small, participants who were conflicted confronting a larger difference and who cooperated or defected with equal probability will now cooperate with low or no conflict at all and with faster RTs. This illustrates how the interplay between subjective and objective parameters drives conflict, which in turn influences the probability of the choice made and the time taken to decide.

In our three studies we harnessed this insight and explored its applicability to research with moral dilemmas. Deontological and utilitarian judgments play here the part that selfish and cooperative decisions play in social dilemmas. They represent the motivational parameters that vary across individuals. Our strategy consisted in manipulating the parameters of the choice problem, and observe the changes in self-reported conflict, utilitarian responses and RTs. Guided by results from previous moral dilemma studies, we constructed a set of cases suited to create a gradient of utilitarian pull from low over medium to high. The changes we observed in

https://doi.org/10.1017/S193029750000485X Published online by Cambridge University Press

the data do not fit comfortably with a dual process corrective model. In this model, the important parameters consist in the "up close and personal" killing and its consequence in maximizing lives saved. The former triggers a prepotent deontological intuition and the latter triggers an utilitarian reflection. Given that all our cases were high-conflict personal dilemmas, both elements were a constant. The formal difference between deontological intuition and utilitarian reflection could not convincingly explain the increase in UR, unless one could show that participants are less aware of the utilitarian benefits at the low end of the gradient than at the high end. This seems unlikely, but we concede that it should be explicitly researched by including, e.g., comprehension questions, which we did not include in these studies. The increase in URs observed along the gradient have a good fit to a conflict model where the interplay between the moral preferences of subjects and the objective parameters of choice explain the difference in the moral salience of the options. The objective parameters in this case are the features of the scenarios we labeled "utilitarian incentives". They play the role that "rates of return" play in social dilemmas. They modulate the participants' responses in interplay with their moral preferences, as the conflict model predicts.

Three additional observations fit the conflict model better than the dual process model. First, since our gradient has three statistically different levels of utilitarian pull in all three studies (Figure 1) we can assume that participants who strongly disapprove of deontological violations feel little or no conflict at the low end of the gradient and high conflict at the high end, while the reverse applies to participants with a bent for utilitarian decisions. Participants more equally balanced in their inclinations will be torn by choices at the middle of the gradient, but more relieved at both the low and the high end. As a consequence, higher conflict should be reported on average at the mid-point of the gradient, and that is what we observed (Figure 2).

Second, this pattern has consequences for their choice of scores along the 4-point UR scale. Participants with no clear preferences in the task will be very conflicted at the mid-point of the gradient, but at the low and high ends they could yield into choosing scores 1 and 4 of the UR scale. In fact, we observe that choices of 1 increase at the low end, and of 4 at the high end of the gradient (Figure 3, left panel). This leads to choices of 1 and 4 being associated with significantly lower levels of conflict. When we pool all cases together and plot conflict as a function of the 4-point utilitarian scale, we observe an inverted-U pattern where the intermediate choices (2, 3) are associated to statistically higher conflict than are the extreme choices (1, 4) (Figure 3, middle panel). This observation is more readily explained by the conflict model and the interplay between the subjective and the objective factors of choice. It is more difficult for the corrective dual process model to explain why conflict is lower at choices 1 and 4, without involving both individual variations in moral preferences and their interplay with the objective gradient of utilitarian pull. Sure, there is still this difficulty: when we plot lnRTs as a function of the 4-point UR scale, we fail to get the inverted-U pattern. Choices of 4 in the scale are associated with lower conflict, but not with lower RTs, specially choices of 4 in *Flames*, *Shark* and *Boat*. This result, though, should be taken with caution, since we can only report it from one study (N=80) where we measured RTs at the response screen (Study 2).

And last, but not least, there is the fact that URs are positively correlated with conflict at the low end and negatively at the high end of the gradient. Again, this has a good fit with a conflict model. It is easy to understand why, when all the pull of the case is towards a deontological choice (1, 2 in the scale) as in *Shark* (save 5 people, victim innocent and not doomed) choosing 3 or 4 would be associated with higher conflict; and conversely, when the pull is towards an utilitarian choice as in *Grenade* (guilty victim) or *Virus* (100K saved, doomed victim), choosing 1 or 2 would provoke higher conflict. This follows not only from the difference in utilitarian pull from low to high, but from the fact that the reversal of the correlation is observed in one an the same sample, and we can assume that its mean stance regarding deontological and utilitarian proclivities remains constant as we measure its mean moral judgment across cases. The same reversal is also observed, in Study 2, in the correlation between UR and RTs. The corrective dual process model would have a hard time trying to explain this reversal of correlation, unless it would be willing to capitalize on the same resources of the conflict model. That said, we end by emphasizing that an attempt to combine the two models should not be excluded a priori. It must be noted however, that regardless of how they are combined, they will still compete for the explanation of the same observations, like RTs and URs, and they will need to sort out how much of the variance each can explain.

References

Bago, B., & De Neys, W. (2018). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General, 148*(10), 1782–1801. http://dx.doi.org/10.1037/xge0000533.

Baron, J., & Gürçay, B. (2017). A meta-analysis of response-time tests of the sequential two-systems model of moral judgment. *Memory & Cognition, 45*, 566–575. http://dx.doi.org/10.3758/s13421-016-0686-8.

Bialek, M., & De Neys, W. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment and Decision Making, 12*(2), 148–167.

Bucciarelli, M. (2015). Moral dilemmas in females: Children are more utilitarian than adults. *Frontiers in Psy-*

https://doi.org/10.1017/S193029750000485X Published online by Cambridge University Press

chology, 6, 1345. http://dx.doi.org/10.3389/fpsyg.2015.01345.

Christensen, J. F., Flexas, A., Calabrese, M., Gut, N., K. & Gomila, A. (2014). Moral Judgment Reloaded: A Moral Dilemma validation study". *Frontiers in Psychology 5*, 607. http://dx.doi.org/10.3389/fpsyg.2014.00607.

Evans, J. St. B. T., & Stanovich, K. E. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science 8*, 223–241. http://dx.doi.org/10.1177/1745691612460685.

Evans, A. M., Dillon, K. D., & Rand, D. G. (2015). Fast but not intuitive, slow but not reflective: Decision conflict drives reaction times in social dilemmas. *Journal of Experimental Psychology: General, 144*(5), 951–966. http://dx.doi.org/10.1037/xge0000107

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science 293*, 2105–2108.

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

Greene, J. D., Morelli, S. A., Lowenberg, K., & Nystrom, L. E. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*, 1144–1154.

Gürçay, B & Baron J. (2017). Challenges for the sequential two-system model of moral judgment, *Thinking & Reasoning, 23*(1), 49–80. http://dx.doi.org/10.1080/13546783.2016.1216011

Haidt, Jonathan. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Huebner, B., Hauser, M. D., & Pettit, P. (2011). How the source, inevitability and means of bringing about harm interact in folk-moral judgments. *Mind & Language, 26*, 210–233. http://dx.doi.org/10.1111/j.1468-0017.2011.01416.x.

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus, and Giroux.

Krajbich, I. & Bartling, B. & Hare, T. & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications. 6*, 7455. http://dx.doi.org/10.1038/ncomms8455.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science, 19*, 549–557. http://dx.doi.org/10.1111/j.1467-9280.2008.02122.x.

Rand, D. G., Greene, J. D. & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489*, 427–430.

Rosas, A. & Koenigs, M. (2014). Beyond 'utilitarianism': Maximizing the clinical impact of moral judgment research. *Social Neuroscience, 9*, 661–667.

Rosas, A. Viciana, H., Caviedes, E. & Arciniegas, A. (2019). Hot utilitarianism and cold deontology: Insights from a response patterns approach to sacrificial and real world dilemmas, *Social Neuroscience, 14*2, 125–135, http://dx.doi.org/10.1080/17470919.2018.1464945.

Rosas, A. & Aguilar Pardo, D. (in press). Extreme time pressure reveals utilitarian intuitions in sacrificial moral dilemmas. *Thinking and Reasoning*.

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition, 119*(3), 454–458.

Tinghög, G., Andersson, D., Bonn, C., Johannesson, M., Kirchler, M., Koppel, L., et al. (2016). Intuition and moral decision-making – The effect of time pressure and cognitive load on moral judgment and altruistic behavior. *PLoS ONE, 11*(10), e0164012. http://dx.doi.org/10.1371/journal.pone.0164012.

Trémolière, B., & Bonnefon, J.-F. (2014). Efficient kill–save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality and Social Psychology Bulletin, 40*(7), 923–930.

3.1 Appendix

English and Spanish versions of the scenarios used at discretion in the three studies. The utilitarian incentives built into each scenario are indicated in parentheses in the English version.

Shark (baseline incentive: save five)

You are watching an exhibition of sharks being fed in an aquarium pool. A metal fence suddenly collapses, and a group people fall into the pool. Their frantic movement in the water attracts the hungry sharks. A person next to you has a harmless episode of nose-bleeding. If you push this person into the water, their blood will draw the sharks away from the other five people. The bleeding person will die, but the five people will swim to safety. How right or wrong is it to cause the death of the person in order to save five people?

Dam (baseline and save100K)

You have a job in the dam management of a large river. Failure in the materials has caused a leak in the dam. The leak has opened a hole and the dam will soon burst apart. A huge volume of water will flood a nearby town. There is no time to warn the inhabitants.

On the dam's edge a worker looks concerned into the dam. If you push this person into the water, his body will stopple the hole, giving time to take the necessary actions. This person will drown, but you will save 100.000 people living in the town. How right or wrong is it to cause the death of the person in order to save 100.000 people?

Boat (baseline and doomed victim)

You escape on a lifeboat from a ship on fire. The lifeboat slowly sinks due to the weight of 2 passengers too many. You are a good swimmer, so you jump into the turbulent waters; but now with one passenger too many the boat continues

to slowly sink. None of the remaining passengers wear life jackets, and none of them can swim.

Nobody is willing to get off and the boat will soon sink. If you pull the person closest to you into the water and keep clear of his attempts to drag you down with him, this person will drown, but you will save five people in the boat. How right or wrong is it to cause the death of the person in order to save five others?

Flames (baseline and selfish incentive)

You and other five people are trapped in a burning building. There is only one emergency exit, but it is blocked by burning debris. A sixth person tries to crawl through a hole at the base of that exit. You and the other five lack the time to do the same thing.

If you push that person through the debris to unlock the exit, that person will die, but you and the other five people will exit to safety.

How right or wrong is it to cause the death of the person in order to save yourself and four others?

Grenade (baseline and guilty victim)

You stand on a rooftop and near you a man is about to throw a grenade onto a group of five people below. The group is unaware of the threat and there is no way to warn them. The five people will surely die in the explosion.

If you push the man with the grenade with a quick movement, he will fall and fail to activate the grenade. The fall will kill him, but the five people below will be saved.

How right or wrong is it to cause the death of the person in order to save five others?

Virus (baseline, save 100K and doomed victim)

Your co-worker, a lab technician accidentally infected with a lethal, incurable and highly contagious virus, heads towards a Super Bowl match completely ignorant of his lethal infection. You rush to prevent him from entering the stadium, but he has just entered as you arrive.

You carry a syringe with a drug that will stop his heart and cause a painless, immediate death, putting and end to the contagion as well. This one person will die, but 100.000 people will survive. How right or wrong is it to cause the death of the person in order to save 100.000 lives?

Spanish version

Tiburón

Visitas un acuario que exhibe tiburones mientras son alimentados. Muy cerca de ti, la reja donde se agolpan los visitantes cede y 5 personas caen al agua. Su caída atrae a los tiburones hambrientos.

Junto a ti, otro visitante sangra excesivamente por la nariz. Si lo empujas al agua, la sangre atraerá a los tiburones y los alejará de las 5 personas que están a punto de ser devoradas. El hombre que sangra morirá, pero las otras cinco personas se salvarán.

Barco

Viajas en un barco que se incendia y te pones a salvo en un bote. El bote aguanta a cinco personas, pero lleva siete y comienza a hundirse por sobrepeso. Tú nadas muy bien y saltas al agua, pero el bote con seis se sigue hundiendo. Ninguno de los seis lleva chaleco salvavidas y ninguno sabe nadar. Todos van a morir si nadie salta al agua; pero nadie lo hace.

Si desde el agua tú jalas al hombre que está más cerca y lo sacas del bote sabiendo que se ahogará por las condiciones difíciles del mar, el bote dejará de hundirse y los otros cinco se salvarán.

Represa

Trabajas en la represa de un río muy caudaloso. Por fallas en el material se produjo un escape en el dique. La concentración de la presión en ese punto romperá el dique. El río caerá con ferocidad sobre una ciudad aledaña de cien mil habitantes. Ya no hay tiempo de avisar.

En el borde de la represa hay un trabajador. Si lo empujas al agua, la corriente lo arrastrará y tapará el hueco con su cuerpo dando tiempo a tomar medidas permanentes. El trabajador se ahogará, pero se salvarán cien mil vidas.

Llamas

Tú y cuatro personas más están atrapadas en un edificio en llamas. Hay una única salida de emergencia, pero está bloqueada por escombros ardientes. Un hombre intenta arrastrarse por un hueco en la base de dicha salida, pero las otras cuatro personas y tú no tienen tiempo de hacer lo mismo.

Si empujas al hombre que está intentando salir contra los escombros ardientes para desbloquear el paso, él morirá, pero tú y las otras cuatro personas se salvarán.

Granada

Estás en una azotea y te das cuenta de que cerca de ti hay un hombre que amenaza con arrojar una granada sobre un grupo de 5 personas reunidas en un parque aledaño. El grupo desconoce la amenaza y no hay forma de avisarles. Las 5 personas morirán con seguridad en la explosión.

Si empujas al hombre con un movimiento rápido, caerá desde la azotea sin tener tiempo de activar la granada. Esto lo matará, pero así se salvarán las 5 personas inocentes.

Virus

Un técnico de laboratorio con quien trabajas se ha infectado accidentalmente con un virus letal, incurable y altamente contagioso. Sin saber de su infección, va en camino a un espectáculo deportivo a un estadio con cien mil personas. Tú te apresuras a interceptarlo e impedir que entre, pero al darle alcance tu colega acaba de ingresar al estadio.

Llevas un jeringa con una droga que le causará una muerta instantánea e indolora, poniendo también fin al contagio. Con su muerte lograrás salvar a cien mil personas.

https://doi.org/10.1017/S193029750000485X Published online by Cambridge University Press