

# Predicting moral judgments from causal judgments\*

EMMANUEL CHEMLA  
*Laboratoire de Sciences Cognitives  
et Psycholinguistique\*\**

PAUL EGRÉ  
*Institut Jean-Nicod\*\**

PHILIPPE SCHLENKER  
*Institut Jean-Nicod\*\*  
& New York University*

July 18, 2012

## Abstract

Several factors have been put forward to explain the variability of moral judgments for superficially analogous moral dilemmas, in particular in the paradigm of trolley cases. In this paper we elaborate on Mikhail's view that (i) causal analysis is at the core of moral judgments and that (ii) causal judgments can be quantified by linguistic methods. According to this model, our moral judgments depend both on utilitarian considerations (whether positive effects outweigh negative effects) and on a representation of the causal structure of the action (whether the negative effects are essentially side-effects rather than main goals). However the exact contribution of each factor, as well as the precise way in which causal considerations interact with utilitarian considerations, has yet to be quantified and investigated. We present several variations on trolley dilemmas in which subjects had to assess the morality of the action and to evaluate their preference between two competing descriptions of the scene ('*a* caused the death of *m*, thereby saving *n*' vs. '*a* saved *n*, thereby causing the death of *m*'). Our main finding is that moral judgments are highly correlated with causal judgments in terms of such descriptions, which makes it possible to predict the former from the latter. Furthermore, we observe that the effect of causal judgments on the relative permissibility of actions is felt even in *anti-utilitarian* scenarios, namely scenarios for which the proposed action diminishes aggregate utility.

**Keywords:** moral sense, trolley dilemmas, linguistic judgments, causal structure of events, principle of double effect, utilitarianism

## 1 Introduction: Trolley dilemmas and the Principle of Double Effect

One of the most discussed principles in ethics and in recent work on moral psychology is the Principle of Double Effect, according to which "it may be permissible to harm an individual for the greater good if the harm is not the necessary means to the greater good but, rather, merely a foreseen side effect" (Hauser et al. 2007). The principle, whose first formulation is generally credited to Aquinas (McIntyre 2001, Mikhail 2011), is used to account for the morality of actions with mixed consequences, namely actions that have both good effects and bad effects. It features quite centrally in discussions concerning utilitarianism, and in particular in the explanation of judgmental contrasts found in moral dilemmas.

A paradigmatic illustration concerns trolley dilemmas (Foot 1967, Thomson 1985, Mikhail 2000, 2007, 2011, Greene et al. 2001, Hauser et al. 2007, Cushman and Greene 201x): in one scenario, an out of control train is threatening to kill five persons trapped on

---

\*We would like to thank Florian Cova, Emmanuel Dupoux, Xavier Gabaix, John Mikhail, Jérôme Sackur, Benjamin Spector and Leeat Yariv for comments and discussions. We would also like to thank Anne-Caroline Fievet and Inga Vendelin for their assistance with preliminary data collection. This work was supported in part by a 'Euryi' grant from the European Science Foundation ("Presupposition: A Formal Pragmatic Approach").

its main path; Denise, who has access to the control booth, can either deviate the train on a side track, where it will kill one bystander, or refrain from doing anything. In a second scenario, an out of control train is likewise threatening to kill five persons on its path; Frank, who is watching from a footbridge, can either shove a heavier person watching next to him onto the track, which will kill that person but stop the train and save the other five, or refrain from doing anything. As evidenced by the Moral Sense Test (see Hauser et al. 2007), an overwhelming majority of subjects judge that it is morally permissible to deviate the train in the former case, but that it is morally impermissible to shove the heavier man onto the track in the latter, despite the fact that the number of casualties and the number of people saved are held constant across the two scenarios. Several explanations for this contrast have been proposed. Mikhail (2007) and Hauser et al. (2007) consider that the best explanation is in terms of the principle of Double Effect: basically, in the first scenario, Denise intends to save five people, and only kills one person as a side-effect of deviating the train. By contrast, in the second scenario, Frank also intends to save five persons, but the shoving of the heavier man onto the track is a necessary means to that end.

There have been different proposals in recent years to assess the adequacy of this hypothesis, and in particular to investigate the precise components underlying the means/side effect distinction. Those include the question of how judgments about causation interact with judgments about intention proper (Greene et al. 2009, Cushman 2008) and the kind of causal intervention under consideration (whether direct or indirect, Royzman and Baron (2002); involving personal force or not, Greene et al. (2009); redirecting a threat or interposing a victim, Waldmann and Dieterich 2007, Waldmann and Wiegmann 2010). In this paper, we ask whether moral judgments in trolley dilemmas can be reliably predicted from judgments contrasting means and side-effects, that is with judgments concerning the causal structure of the action. To the best of our knowledge, although most accounts agree that moral judgments in trolley cases depend in part on the causal analysis of the scenarios, few systematic attempts have been made to correlate experimentally moral judgments with causal judgments contrasting means and side-effects. Two notable exceptions are Waldmann and Dieterich (2007), and Waldmann and Wiegmann (2010), who demonstrate the sensitivity of moral judgments to the locus of intervention (victim or threat). Like Waldmann and collaborators, we see the need for systematic variations on trolley dilemmas. Unlike them, here we restrict attention to only one kind of intervention, namely intervention on potential victims, but with the aim of showing that already for such cases systematic variations in the analysis of the causal structure of the action predict distinct moral evaluations.

We propose to do so within the framework proposed by Mikhail (2000, 2007, 2011). For our purposes, Mikhail makes two central proposals:

1. First, he offers an explicit connection between the causal structure of the situation and the moral judgments that are obtained:

“the key distinction that explains many of the standard cases in the literature is that the agent commits one or more distinct batteries prior to and as a means of achieving his good end in the impermissible conditions

(...) whereas these violations are subsequent side effects in the permissible conditions (...)." (Mikhail 2011:39)

2. Second, he provides two explicit linguistic tests (based on Goldman (1970)) to establish the causal relationship between the events that appear in moral dilemmas. These tests are designed to determine on independent (non-moral) grounds the precise causal structure of an action, a structure he describes in terms of *act trees* (Goldman 1970, Donagan 1977, Mikhail et al. 1998, Mikhail 2000) connecting causes and consequences:

"Descriptions using the word 'by' to connect individual nodes of these act trees in the downward direction (e.g., 'D turned the train by throwing the switch,' 'D killed the man by turning the train') will generally be deemed acceptable; by contrast, causal reversals using 'by' to connect nodes in the upward direction ('D threw the switch by turning the train,' 'D turned the train by killing the man') will generally be deemed unacceptable. Likewise, descriptions using connectors like 'in order to' or 'for the purpose of' to link nodes in the upward direction along the vertical chain of means and ends ('D threw the switch in order to turn the train') will generally be deemed acceptable. By contrast, descriptions of this type linking means with side effects ('D threw the switch in order to kill the man') will generally be deemed unacceptable." (Mikhail 2011:120)

## 2 Goals and hypotheses

### 2.1 Goals

Our goal in this piece is to submit a version of Mikhail's hypothesis to experimental scrutiny by investigating a possible correlation between moral and causal judgments. While intentions play an essential role in moral judgment, we try to isolate the specific contribution of the causal (rather than intentional) structure of the relevant situations. We take two steps to obtain the desired delineation of the causal factor. First, in all our scenarios, the agent knows the consequences of his or her actions, and also knows that the events reported do not happen by accident; in this way, the role of intention is held constant across our scenarios (but not suppressed). Second, the sentences we use to capture the means/side-effect distinction involve the verb "cause" and the construction "thereby" — rather the verb "intend" or the construction "in order to", which probe intention proper.

We will address three main questions.

1. *What is the right definition of the Causal Constraint that enters in the moral judgments obtained in trolley cases?* We show that an action is judged to be impermissible to the extent that the scenario can be described with a sentence of the form '[The agent] caused the death of  $m$  people, thereby saving  $n$  people'. We take this to suggest that, in the cases under study, an action is taken to be impermissible if in the causal order of things it first leads to some people's deaths and then saves people. If the

preferred causal order is the opposite, the action can be taken to be permissible (but only if its overall effect is to save more lives than it sacrifices).

2. *How does the Causal Constraint interact with utilitarian considerations?* On all standard accounts, action in trolley dilemmas is permissible only if it leads to a positive utilitarian outcome, in the sense that more lives are saved than sacrificed (the traditional focus is on the fact that this condition is not *sufficient*, but few would doubt that it is *necessary*). Still, this does not tell us how, in general, this 'Utilitarian Precondition' interacts with the Causal Constraint. We will test a gradient version of both, and show that two main analyses are compatible with our results. According to one, these constraints interact in an *additive* fashion; in particular, even when the is violated (so that more lives are sacrificed than saved), the effects of the Causal Constraint continue to make themselves felt. According to the other analysis, the two conditions interact in a *lexicographic* fashion, with the Utilitarian Precondition ranked above the Causal Constraint.
3. *How can theories of moral judgments be made predictive?* Our analysis also addresses a methodological problem. The nature of the difficulty is this: (i) any moral theory must take as an input the way subjects conceptualize a scenario, and yield as an output a moral judgment that these subjects are predicted to make; (ii) however, there is usually no *independent* way of assessing how a scenario is conceptualized; as a result, (iii) it is hard to derive bona fide *predictions* from the theories at hand. Usually researchers rely on their own semi-theoretical conceptualization of the scenarios to derive the desired predictions, but it is clear that if theories are to become formally precise such a measure won't suffice. We thus propose that linguistic judgments (specifically: semantic preferences between two descriptions) could help address this methodological problem, as they might provide an independent way of assessing the causal structure of a scenario.

While our analysis falls squarely within Mikhail's general framework, we depart from the letter of his analysis in three minor respects.

1. First, in order to establish the causal relations among the relevant events, we solely use a version of the *by* test. We set aside the *in order to* test because we believe that it assesses *intention* rather than causality *per se*.
2. Second, we use a test with two clauses connected by the anaphoric expression *thereby*, rather than Mikhail's monoclausal *by* test (e.g., 'D killed the man by turning the train'). Specifically, our hypothesis will take the following form:

(1) **Linguistic Test Hypothesis:**

Consider a trolley scenario in which an agent took an action *a* that caused an event E1 in which *n* people were saved and an event E2 in which *m* people died. A subject will conceptualize E1 as being causally prior to E2 to the extent that he prefers description (1a) to description (1b):

- a. [The agent] saved *n* people, thereby causing the death of *m* people.

- b. [The agent] caused the death of  $m$  people, thereby saving  $n$  people.

Originally, we worked with French and the ‘by’ test just wasn’t applicable. So the somewhat contingent reason for our choice was that we wanted a test that was more versatile and was easier to adapt to other languages.<sup>1</sup>

3. Third, Mikhail’s analysis allows for cases in which neither (1a) nor (1b) holds because the event of causing a death and the event of saving lives are on separate branches of a causal tree, e.g., with both being consequences of an earlier event (viz. Mikhail 2011:174). For simplicity, we only consider a choice between (1a) and (1b). We believe that this simplification won’t hurt our study because on Mikhail’s theory judgments concerning the truth of (1b), which are directly assessed by our experiments, should be correlated with impermissibility.

## 2.2 Hypotheses

In this section, we discuss the hypotheses we will be using: how moral judgments are impacted by utility considerations (section 2.2.1), by testable causal analyses (section 2.2.2), and by the interaction of the two (section 2.2.3). This discussion is designed specifically to account for the cases of trolley dilemmas we are concerned with. In the general case, the constraints we present may require a broader formulation and may interact with further constraints.

### 2.2.1 Statement of the Utilitarian Precondition

We propose to define the *utility*  $\mathcal{U}$  of a scenario or of an action in a scenario as the number of lives at the end of the scenario (this count will not explicitly include lives of people who are not explicitly threatened in the scenario, assuming that those stay constant across the scenarios we will compare). From this definition, we may formulate two rules that pertain to the utilitarian aspect of a scenario:

- (2) **Utilitarian Maximization** (standard rule):  
An action  $a$  is permissible to the extent that its utility  $\mathcal{U}(a)$  is maximal among the utilities of alternative possible actions.
- (3) **Utilitarian Precondition** (condition we will use):  
An action  $a$  is permissible to the extent that its utility  $\mathcal{U}(a)$  is not lower than the utility associated with inaction.

The rule in (2) is standard utilitarianism. The rule in (3) is a consequence of it: clearly, if utility is maximized, one will not engage in an action that yields lower utility than inaction. But the converse is not true: the special principle in (3) is compatible with views that are not at all maximization-based (in fact, standard utilitarianism is controversial, but

<sup>1</sup>A natural translation in French of *D killed the man by turning the train* is: *D a tué l’homme en réaiguillant le train*; but in general the *en* + *present participle* construction in French is ambiguous between a causal (*by*) reading and a simultaneous (*while*) reading. The French test we started from involved an appositive clause, roughly equivalent to: *[The agent] saved  $n$  people, which caused the death of  $m$  people*.

the principle in (3) has great initial plausibility). Now trolley scenarios were originally invented to refute maximization-based utilitarianism (see Thomson 1985, or judgments about the Footbridge case in Hauser et al. 2007). It is thus likely that any weight that (2) might have will be lower than that of other constraints, and specifically of the 'Causal Constraint' we discuss below (see also section 2.2.3 for further discussion). By contrast, (3) might well carry *more* weight than all other constraints. This will turn out to be crucial: both introspection and our experimental results suggest that in our scenarios an action that lowers aggregate utility while satisfying the Causal Constraint is *very* impermissible, more so than an action that increases aggregate utility but violates the Causal Constraint. This means that (3) is needed as a constraint that carries more weight than other constraints (a low-ranking (2) won't be able to 'override' the verdict of the Causal Constraint). Importantly, we only consider two versions of each scenario: one that lowers aggregate utility ('anti-utilitarian scenarios'), and one that increases it ('utilitarian scenarios'). As a result, once (3) has separated the two classes, there will be now work left to do for (2), since we do not have more fine-grained distinctions within each class. For this reason, the rest of this discussion is focused on (3) rather than on (2).

## 2.2.2 Statement of the Causal Constraint

We will state the Causal Constraint as follows:

### (4) Causal Constraint:

An action  $a$  is impermissible to the extent that it causes a death as a means to an end, rather than as a side-effect.

As announced, this constraint will be assessed by way of an auxiliary hypothesis described in (1), which states that subjects' analysis of the causal structure of a scenario is reflected in linguistic judgments involving apparently non-moral notions. To the extent that this Linguistic Test (1) is efficient, we can restate the Causal Constraint as follows:<sup>2</sup>

### (5) Causal Constraint (testable version):

Consider a trolley scenario in which an agent took an action  $a$  that caused an event E1 in which  $n$  people were saved and an event E2 in which  $m$  people died. The action  $a$  is permissible to the extent that it is preferably described as (1a) rather than as (1b), repeated below:

- a. [The agent] saved  $n$  people, thereby causing the death of  $m$  people.
- b. [The agent] caused the death of  $m$  people, thereby saving  $n$  people.

We will provide evidence in favor of this Causal Constraint. Notice however that an implicit aspect of the Linguistic Test (1) may be too strong: there may be a small 'retroaction' of moral judgments on causal judgments (see section 3.2.3).

<sup>2</sup>For the cases we are considering, in which a choice is given between (1a) and (1b), we can safely state the foregoing constraint in terms of permissibility, rather than impermissibility.

### 2.2.3 Interaction between the Causal Constraint and the Utilitarian Precondition

Our main goal is to evaluate the contribution of the Causal Constraint to moral judgments. How do the Causal Constraint and the Utilitarian Precondition interact though? We discuss three main possibilities. On the ‘conjunctive analysis’, an action is permissible just in case it satisfies the two (binary) constraints. On the ‘lexicographic analysis’, the constraints may be intrinsically gradient, and they are ranked: some take precedence over others. Finally, on the ‘additive analysis’, the constraints may also be gradient, but each makes a separate, additive contribution to the permissibility of an action.

#### A conjunctive analysis

On the conjunctive analysis, an action is permissible just in case it satisfies each of a number of constraints. Here we will restrict attention to the Utilitarian Precondition (3) and the Causal Constraint (5). We assume for the moment that these constraints are intrinsically binary: they are either satisfied or violated by a proposed course of action; and it is only in case both are satisfied that the action is permissible. Within this binary framework, we predict that actions should be uniformly impermissible in anti-utilitarian scenarios, since these systematically violate the Binary Utilitarian Precondition; by contrast, in utilitarian scenarios we predict that the proposed action should be permissible just in case the Binary Causal Constraint is satisfied. This prediction is illustrated in Figure 1.

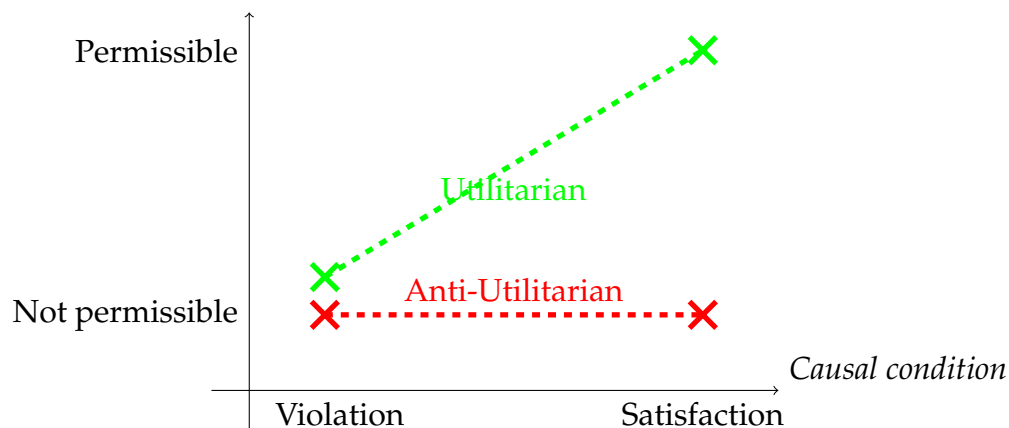


Figure 1: This figure illustrates the predictions of the conjunctive analysis. If the Causal Constraint is binary, the result is also binary, and we only obtain the four possible cases given by the crosses. If the Causal Constraint is gradient (or probabilistic), we also obtain gradient permissibility, as illustrated by the dashed lines.

If we were to stop here, we would obtain a simple bipartition of actions into permissible and impermissible ones. But since we are interested in *gradient acceptability judgments*, we must find a way to re-interpret this analysis within a continuous setting. We will assume that someone’s gradient judgment of permissibility reflects the *subjective probability* that this person assigns to the case in which the proposed action is permissible. We further assume that subjects have a perfect assessment of the (binary) Utilitarian Precondition: if the count of deaths is negative, the constraint is violated with certainty; otherwise, it is

satisfied with certainty. Therefore the only source of uncertainty lies in the Causal Constraint. We assume that subjects are typically unsure as to whether it is satisfied, and that they have gradient judgments which track the probability that they assign to the constraint's being satisfied. The *source* of this uncertainty is immaterial for present purposes: subjects might be unsure about the precise statement of the Causal Constraint; or the assessment of their own causal judgments by way of introspection might itself give rise to some uncertainty).

In sum, we assume that the permissibility of an action is based on the requirement that both the Utilitarian Precondition and the Causal Constraint be satisfied, but the former is evaluated with certainty while the latter is evaluated probabilistically. As a result, we obtain predictions that are a gradient version of Figure 1 (dashed lines). In anti-utilitarian scenarios, the Utilitarian Precondition is violated with certainty, and hence we obtain a constant function: the proposed action is impermissible; in utilitarian scenarios, the proposed action is permissible just in case the Causal Constraint is satisfied, and hence the probability that the proposed action is permissible tracks the probability that the Causal Constraint is satisfied. As we will see, this prediction is refuted by our data: in anti-utilitarian scenarios, the permissibility of the proposed action is not constant.

## A lexicographic analysis

The framework discussed in the preceding section offered a bipartition of actions into 'permissible' and 'impermissible' ones; gradient predictions were added after the fact, so to speak, by plugging a probabilistic component into the analysis. An alternative is to work from the start with a gradient version of the Causal Constraint and Utilitarian Precondition and of the permissibility of an action. In fact, we have formulated the constraints so that they can immediately be understood as gradient: actions may qualify as permissible *to the extent that* they satisfy some requirement, which requirement may itself be satisfied to different degrees.

For simplicity, we continue to take the Utilitarian Precondition to be binary: it is either satisfied or violated.<sup>3</sup> Importantly, however, the permissibility of an action is gradient, and the Causal Constraint can be satisfied to different degrees as well. How should the Causal Constraint (gradient) and the Utilitarian Precondition (binary) interact in this framework to produce gradable moral judgments? One possible view is that the constraints are *lexicographically ordered*: one constraint takes precedence over the other, and it is only in case of a tie that the lower-ranked constraint kicks in. Lexicographic orderings have been used in moral and political theorizing (notably Rawls 1971), and it would be

<sup>3</sup>Let us mention two sources of gradability for this constraint, even though they are not directly relevant for our dilemmas. First, the actual value of the utility  $\mathcal{U}$  of an action may be taken into account, as opposed to a mere comparison with competitors. For instance, killing one person for the sake of saving 100 lives may be different from killing one person for the sake of saving 2 lives. In our scenarios, this difference is immaterial since the utilitarian/anti-utilitarian outcome may always lead to one of only two cases: either 5 people are saved and 1 is killed, or 5 people are killed and 1 is saved. Second, we may replace utility with expected utility, as is standard. In other words, what matters to assess the permissibility of an action should not be its actual outcome, but rather the outcome that the agent was entitled to expect from it. Again, this will play no role in our scenarios which left no doubt about the consequences of the agent's actions.



rather natural to use them in the present context.<sup>4</sup>

To be concrete, let us give the following form to our assumptions:

- (6) a. Evaluation of the Permissibility (gradient):  
A function  $P$  maps scenarios into  $[0, 1]$  depending on how permissible the proposed action is.
- b. Evaluation of the Causal Constraint (gradient):  
A function  $\mathfrak{C}_C$  maps scenarios into  $[0, 1]$ , depending on how well the proposed action satisfies the Causal Constraint.
- c. Evaluation of the Utilitarian Precondition (binary):  
A function  $\mathfrak{C}_U$  maps scenarios into  $\{0, 1\}$ , depending on whether they satisfy the Utilitarian Precondition.

Within this framework, we can say precisely what it means for a function  $P$  to be lexicographic with the ordering Utilitarian Precondition  $\gg$  Causal Constraint:

- (7)  $P$  is lexicographic with the ordering Utilitarian Precondition  $\gg$  Causal Constraint just in case for all scenarios  $s_1$  and  $s_2$ , if  $\mathfrak{C}_U(s_1) < \mathfrak{C}_U(s_2)$ , then  $P(s_1) \leq P(s_2)$ .

Within this lexicographic framework, ranking the Utilitarian Precondition above the Causal Constraint, we predict that the permissibility curve for utilitarian scenarios should be entirely above the permissibility curve for anti-utilitarian scenarios: the fact that the Utilitarian Precondition is satisfied in utilitarian scenarios but not in anti-utilitarian ones should suffice to make the proposed action in *all* of the latter strictly less acceptable than in *any* of the former. This is illustrated in Figure 2.

A definition similar to (7) could be given with the opposite ordering Causal Constraint  $\gg$  Utilitarian Precondition. Although this is a theoretical option, this ordering can be disregarded because it would have the following undesirable consequence: for two scenarios  $s_1$  and  $s_2$  such that the first satisfies better the Causal Constraint ( $\mathfrak{C}_C(s_1) > \mathfrak{C}_C(s_2)$ ), the first scenario should be more permissible than the other, *even if*  $s_1$  is anti-utilitarian and  $s_2$  is not. This strong prediction is counter-intuitive and explicitly refuted by our data.

### An additive analysis

Finally, we may consider an analysis in which each constraint contributes separately, and additively, to the permissibility of an action, as is represented in the following formula:

$$(8) \quad P(s) = \alpha \times \mathfrak{C}_U(s) + \beta \times \mathfrak{C}_C(s) + \gamma$$

<sup>4</sup>Rawls's basic theory was based on two principles: a principle of equal liberty, and a 'maximin' principle that specified that the welfare of the most disadvantaged members of society should be maximized. For him, the latter principle was subordinate to the former: "...I shall, in fact, propose an ordering of this kind by ranking the principle of equal liberty prior to the principle regulating economic and social inequalities" (Rawls 1971:38). Lexicographic orderings more generally occupy central stage in Optimality Theory, which was developed in phonology to predict grammatical acceptability on the basis of *ranked constraints* (see, e.g., Kager 1999).

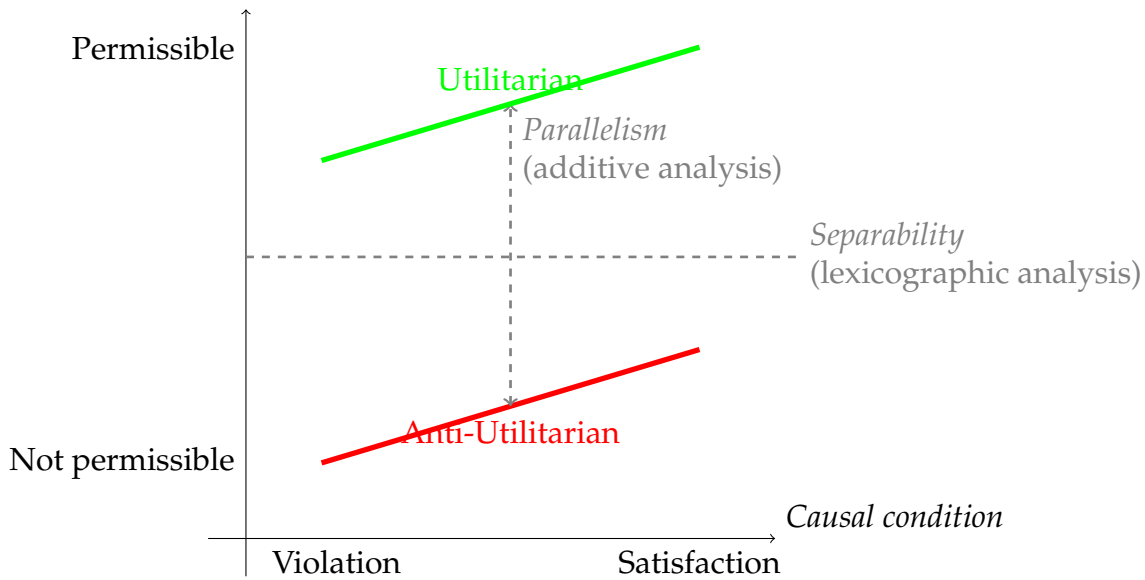


Figure 2: This schema represents the predictions of the lexicographic analysis (with the ordering Utilitarian Precondition  $\gg$  Causal Constraint) and of the additive analysis. The two views are committed to different aspects of this diagram. Under the lexicographic analysis, the shapes of the two curves may differ from each other (e.g., one may be flat and the other not, as in Figure 1), but they should be *separated*. Under the additive analysis, the two curves are predicted to be *parallel* but not separated (the highest point from the anti-utilitarian curve may be above the lowest point from the utilitarian curve).

It is natural to assume that both  $\alpha$  and  $\beta$  are positive; since by construction  $\mathfrak{C}_U$  yields a higher value when the scenario  $s$  is utilitarian than when  $s$  is anti-utilitarian, we predict that subjects' judgments should give rise to two parallel curves, with the 'utilitarian' curve above the 'anti-utilitarian' curve. The predictions of the additive analysis are thus well represented with Figure 2.

What is the empirical difference between the lexicographic analysis and the additive analysis then? Both are compatible with the outcome as it is presented in Figure 2. However, the two analyses are committed to different aspects of this schema. Two properties of this schema are relevant. The first is *separability*: the highest point from the anti-utilitarian curve is below the lowest point from the utilitarian curve. The second property is *parallelism*: the two curves are parallel (the fact that they are lines is irrelevant however). What is important is that the lexicographic analysis predicts and is committed to separability and not to parallelism, while the opposite is true for the additive analysis, which predicts and is committed to parallelism and not to separability.

## 2.2.4 Summary

In this section, we first introduced the Utilitarian Precondition. We discussed two possible versions of this constraint, presented in (2) and (3). Importantly, we are using the less classical version of the two, because it is the only one that is compatible with a lexicographic analysis (although both are compatible with an additive analysis). Let us repeat

why the classic Utilitarian Maximization constraint would not work in a lexicographic framework. First, if it was ranked lower than the Causal Constraint, it would reproduce the same problems produced by the Utilitarian Precondition we are currently using (see §2.2.3). Second, if it were ranked higher than the Causal Constraint, it would follow that inaction would be impermissible when it leads to an anti-utilitarian outcome, contrary to the facts (see §2.2.1).

We then discussed the Causal Constraint. Our main goal in this paper is to evaluate the role of this constraint, by which causal judgments serve as an input for moral judgment. To do so on safer grounds, we discussed three possible ways in which the two constraints may interact: the conjunctive analysis, the lexicographic analysis and the additive analysis. We showed that the distinction between the later two types of analyses was subtle but real.

In the experiment we present next, we provide evidence in favor of the existence of the Causal Constraint (our main target). We also show that the lexicographic and the additive analyses fare better than the conjunctive analysis. We would also like to argue that our results provides a subjective advantage to the additive analysis, but this will mostly remain a topic for future research.

### 3 Experiment

#### 3.1 Design

Participants were first introduced with a general context, described as follows:

A red locomotive in flames is moving at full speed on the tracks of an amusement park. It is out of control and threatens the lives of some people that are stuck in white wagons at a standstill on the tracks. From the control booth, the technician, John, can activate a device which will have different effects in the different scenarios.

Participants were then presented with a sequence of scenarios. Their task for each scenario was to provide a judgment within a continuous range of possible judgments between two anchors (see Figure 3). We coded each answer as the percentage of the line filled in red between the two anchors. All the scenarios were presented in random order in two different blocks, the ‘moral’ and the ‘causal’ blocks (which were themselves administered in random order to different participants). Participants saw only one scenario at a time, replaced by a new scenario after they had given their response, with no possibility to go back to compare stories or change their responses.

In the ‘moral’ block, participants were asked to provide a moral judgment, the two extreme anchors for the judgments being *Not moral at all* and *Perfectly moral*. Hence, answers ranged from 0% (*Not moral at all*) to 100% (*Perfectly moral*). In the ‘causal’ block, participants were asked to assess their preference between two descriptions. The descriptions that were indicated as the extreme anchors were one of the following pairs, depending on the utilitarian (9) or anti-utilitarian (10) aspect of the scenario:

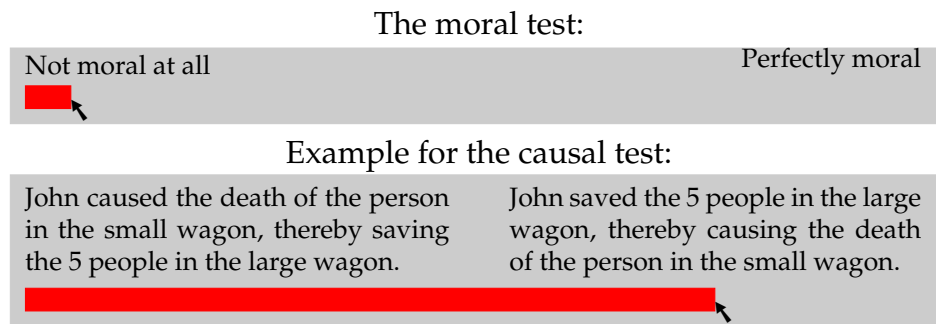


Figure 3: Response Scale. Participants were offered the possibility to situate their responses within a range of possibilities between two anchors, as above. Responses were coded as the percentage of the red line filled in red, 100% corresponding to an unambiguous ‘right’ response, and 0% to an unambiguous ‘left’ response. In the first, ‘moral’ example above, the answer would be around 5%, in the second, ‘causal’ example around 75%.

- (9)
  - a. John saved the 5 people in the large wagon, thereby causing the death of the person in the small wagon/of the pedestrian.
  - b. John caused the death of the person in the small wagon/of the pedestrian, thereby saving the 5 people in the large wagon.
- (10)
  - a. John saved the person in the small wagon, thereby causing the death of the 5 people in the large wagon/of the 5 pedestrians.
  - b. John caused the death of the 5 people in the large wagon/of the 5 pedestrians, thereby saving the person in the small wagon.

Our hypothesis was that preference for the (a) version of these descriptions would be found for scenarios leading to higher moral judgments. What distinguishes the (a) version from the (b) version is that ‘saving’ is presented as the causally prior action, and ‘thereby causing the death’ as causally secondary. In the (b) version, ‘causing the death’ is reported as the causally prior action, and ‘thereby saving’ as causally secondary. Our aim in testing this contrast was to get a handle on two ways of conceptualizing one and the same event, see Linguistic Test Hypothesis in (1).

Each scenario was presented twice in the ‘causal’ block: in one version, the (a) description would appear to the right, in the other version the (a) description would appear to the left. The goal of this manipulation was to avoid the risk of correlations emerging from a strategic and uninteresting tendency for a scenario to prompt a response to the right of the scale, rather than a tendency for the *Moral* answer to correlate with a preference with the (a) description, no matter whether this description would appear on the same side as the *Perfectly moral* anchor.

### 3.1.1 Material: Scenarios

Each scenario was presented by means of three vignettes, containing both text and a graphical illustration. The first vignette always contained a description of the situation, the second vignette a description of the two actions available to John and of his choice in

the situation, and the third a description of the predicted consequences of his action (see Figure 4).

We tested 28 different scenarios, which varied systematically according to the following factors.

**Utilitarian factor:** The scenarios were based on 14 basic stories, with two versions of each, a 'utilitarian' version and an 'anti-utilitarian' version. In the anti-utilitarian version, John chooses to save one life and to sacrifice five; in the utilitarian version, he chooses to save five lives and to sacrifice one (see Figure 5).

**Mode of action:** The 14 basic stories could differ depending on the mode of action available to John (interposition vs. extraction of a wagon, shoving of person(s) or wagon from a bridge), the type of connection between the small and large wagon (attached or not), but also the structure of the railway (plain intersection or loop). See Figure 6 for the utilitarian version of each story.

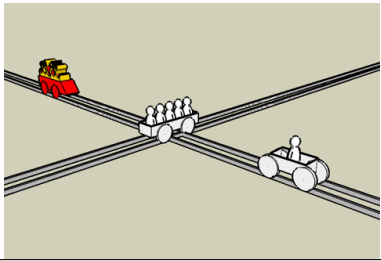
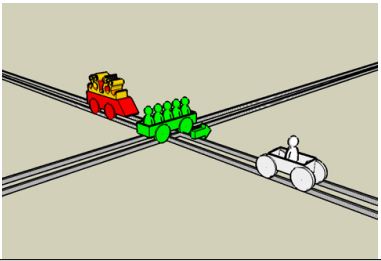
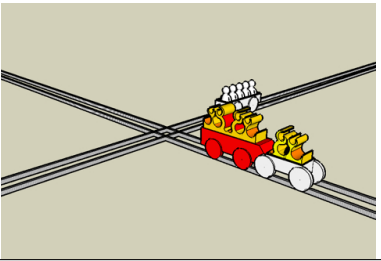
		
<p>The red locomotive is about to crash into the small wagon, which will kill all the people on board.</p>	<p>John only has two options, and he knows their consequences.</p> <ul style="list-style-type: none"> <li>•He may activate a device that will put the large wagon out of the trajectory of the locomotive. <b>In that case, the person in the small wagon will die.</b></li> <li>•Or he may do nothing. <b>In that case, the 5 people in the large wagon will die.</b></li> </ul> <p>He decides to activate the device.</p>	<p>The locomotive thus crashes into the small wagon, and the person on board dies.</p>

Figure 4: Example of a scenario as it was presented to participants with three vignettes.

The choice to have 14 stories (rather than 16 or 20 or more) is not essential to our design. The reason is that we are interested in the correlations between different types of judgments (causal vs. moral), and not so much in the judgments specific to particular scenarios. Our experimental constraint was just to include sufficiently many scenarios to be in a position to assess the correlations between the two types of judgments under study, namely moral judgments and causal judgments concerning the structure of the action. This is why, as far as possible, our scenarios varied along dimensions that could

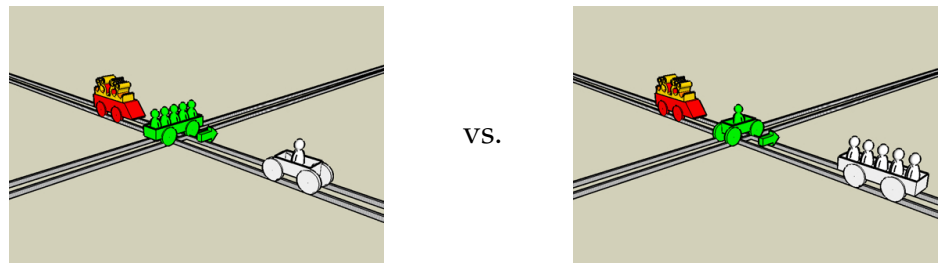


Figure 5: This figure represents the utilitarian vs. anti-utilitarian versions of a given story (The mode of action being: extraction, with non-attached wagons). The scenarios are reduced to the image from the second vignette, from which all the information can be recovered.

impact to various degrees causal and/or moral judgments (loop vs. no loop, link between wagons vs. no link between wagons, interposition vs. extraction, shoving a wagon with passengers vs. shoving pedestrians directly, etc., see Greene et al. 2001, Hauser et al. 2007, Mikhail 2011 for some of the variations we consider). One important feature common to all of our scenarios, finally, is that the threatening train is always an empty wagon in flames. In particular, in contrast to some of the scenarios proposed by Waldmann and Wiegmann (2010), we do not include cases in which the threatening train contains a passenger. Intervention, in our scenarios, is always on the potential victims rather than on the threat.

### 3.1.2 Participants

We recruited 52 participants on Amazon Mechanical Turk.<sup>5</sup> They all reported in a questionnaire after the experiment that they were native speakers of English.

We also ran three smaller experiments (around 20 participants each, with five participants overall being excluded for not reporting that they were native speakers of English). These were variants of the present experiment in which we either replaced the continuous scale of answers with an 8-point Likert scale type of answer, or we replaced the human lives at risk with rabbits in a similar situation,<sup>6</sup> or we made both changes (type of answers and rabbits/humans). We confine our analysis below to the original continuous task + human victims version; but in any event the other 3 smaller experiments provided faithful replications of the main results we present.

<sup>5</sup>See Sprouse (2011) for a quantitative study of the reliability of such participants in linguistic studies, albeit for judgments of a different type.

<sup>6</sup>Initially, we tested our rabbit scenarios because we thought they could show purely utilitarian results, thus providing causal judgments which would not be influenced by moral differences (see further below the discussion of possible ‘Knobe effects’). In fact, we found that, if anything, the moral judgments were more fine-grained with rabbits. This may be because the human cases, and their repetition, were judged somewhat implausible and thus yielded less careful and spontaneous responses.

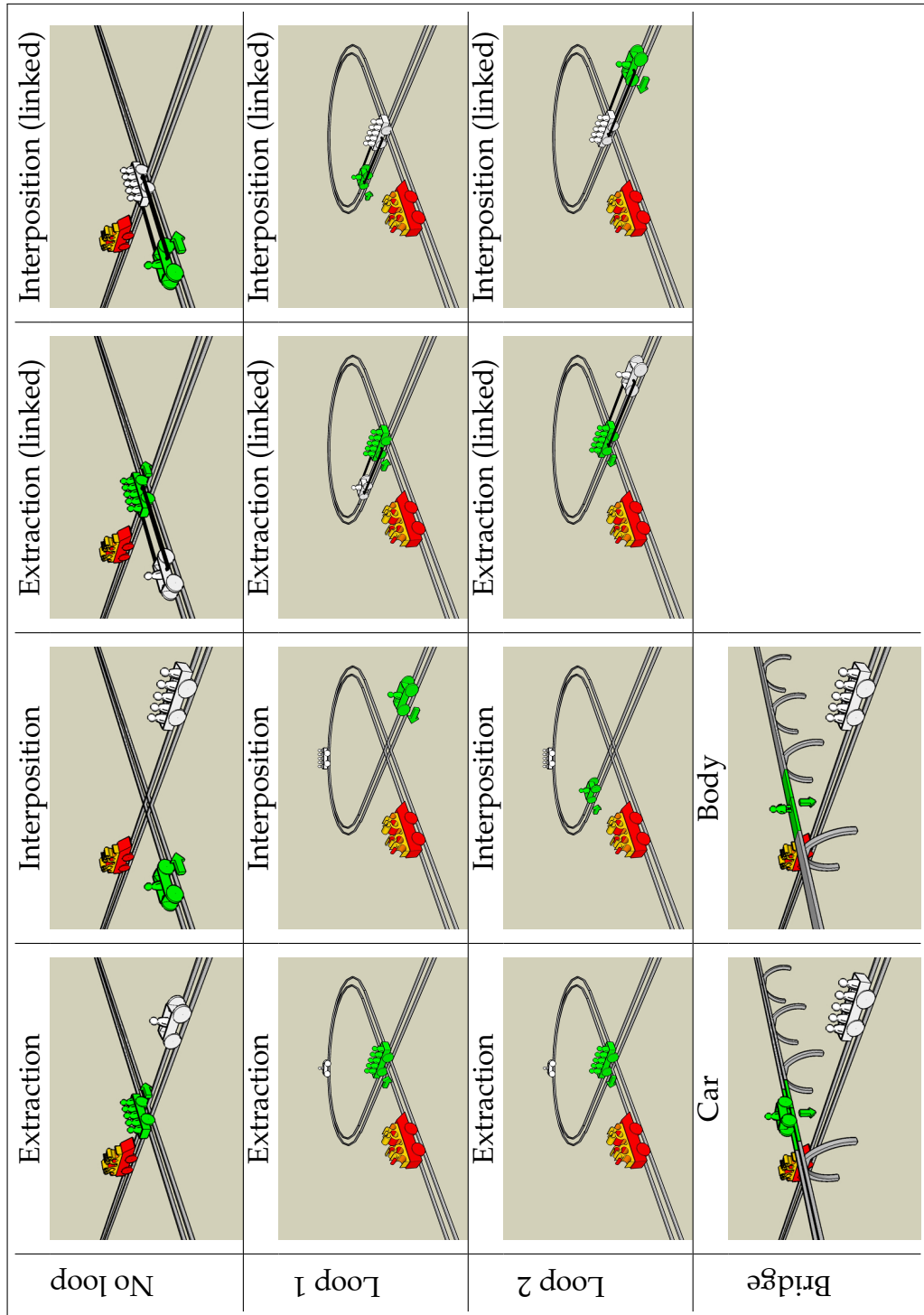


Figure 6: The 14 basic stories (utilitarian version, middle vignette with the action)

### 3.2 Results

We will first present the average results obtained for the moral and for the causal judgments (section 3.2.1). We present these results for the sake of completeness. We are mostly

interested in the *correlations* between the two types of judgments. Bare moral judgments are not useful without an *a priori* explanation of how they could be derived. Our specific position is that they can be (at least partially) derived from causal judgments. In section 3.2.2, we present correlation analyses that prove the relevance of the Causal Constraint we introduced (see §2.2.2). We discuss in more detail how our results constrain the type of the interaction that exists between the Utilitarian Precondition and the Causal Constraint. In section 3.2.3 (and in the appendix), we discuss the possibility that causal judgments may be polluted by moral judgments. Section 3.2.4 offers preliminary discussions of variations at individual levels.

### 3.2.1 Average results

#### Moral judgments

Figure 7a presents the average moral judgments obtained for the different scenarios. Because we are mostly interested in the correlations of these judgments with causal judgments, we will simply make two comments. First, utilitarian scenarios are judged more ‘moral’ than their anti-utilitarian counterparts. In fact, *all* utilitarian scenarios are judged more moral than *all* anti-utilitarian scenarios. This property is predicted by lexicographic analyses as presented in section 2.2.3. We will come back to this point in the next section. Second, the finer differences that may appear between pairs of scenarios seem coherent with our intuitions as well as with results obtained with similar scenarios in the literature. For instance, the scenario in which a person is shoved from a bridge (Body) receives a lower moral value than the scenario in which a wagon is extracted from tracks on which it is in danger (Extract). Again, our goal is not a study of such pairwise comparisons, but rather a global comparison of these results with causal judgments, to which we now turn.

#### Causal judgments

Figure 7b shows a composite measure of the causal judgments, obtained as follows. Let us call ‘moral description’ the sentence of the form ‘*x* saved..., thereby causing the death of...’. As described at the end of section 3.1, for each scenario, the causal question was asked twice: (A) once with the ‘moral description’ to the right as *John caused the death of 1, thereby saving 5* vs. *John saved 5, thereby causing the death of 1*, and (B) once in the reversed order with the ‘moral description’ to the left as *John saved 5, thereby causing the death of 1* vs. *John caused the death of 1, thereby saving 5*. According to our hypothesis, the scenario is morally permissible to the extent that participants prefer the moral description ‘John saved 5, thereby causing the death of 1’. Such a preference corresponds to a *high* response to (A) cases, and to a *low* response to (B) cases. Consequently, Figure 7b is obtained by subtracting (A)-responses to (B)-responses. Thus, a positive difference indicates preference for the ‘moral description’ (see (1a)), and a negative difference indicates preference for the opposite description (see (1b)).

All measures (the two ‘directional’ measures or their difference) show the same profile of results (formally, the two directional measures with the ‘moral anchor’ to the left or to the right are negatively correlated  $r^2 = .91, t(12) = -11, p < .001$ ). The difference



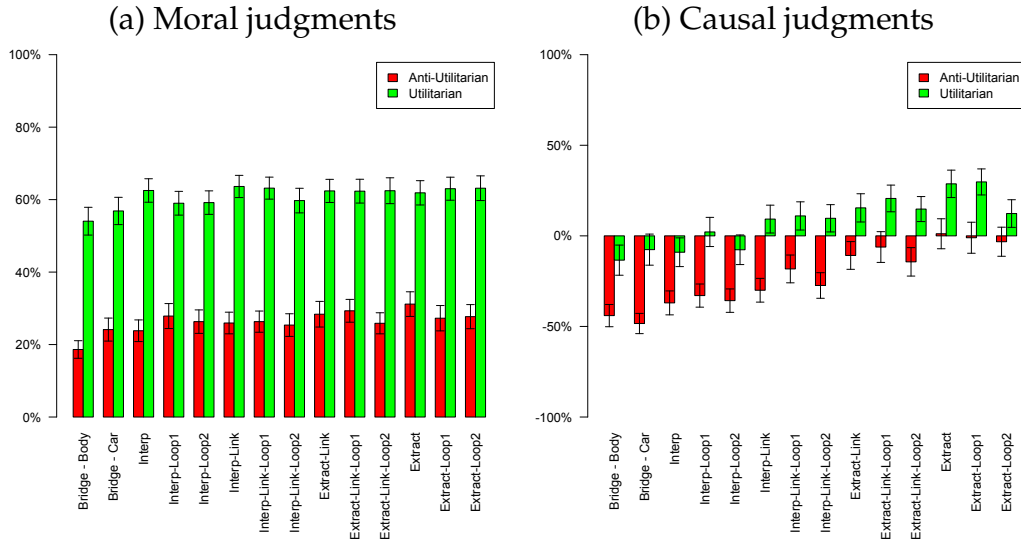


Figure 7: Moral judgments for each story in its utilitarian and anti-utilitarian versions are represented on Figure 7a. Figure 7b represents a composite measure of the causal judgments, such that answers given when the ‘moral’ description appeared to the right (on the same side as the *Perfectly moral* anchor) are subtracted from answers given when the ‘moral’ description appeared to the left (on the same side as the *Not moral at all* anchor.)

measure however protects us from a number of problematic response strategies and biases that may artificially pollute our results or inflate artificially the correlations we are interested in. For instance, participants who would have been through the moral block first may ignore the precise formulation of the question in the causal block. They may continue to answer as if the question was still the *not moral/moral* question. If that were the case, the answers to the (A) and (B) versions of the causal question would cancel out. This situation would thus not lead to an artificial correlation between moral and causal judgments, measured as the (A) minus (B) difference. Instead, our measure would simply be flat for causal judgments. Hence this worry does not block fruitful interpretations of correlations we will present.<sup>7</sup>

Now that we have explained the measure we will be using and why we will use it, let

<sup>7</sup>More precisely, since participants will have seen the same scenarios twice, there might be a bias  $\varepsilon$  for clicking on the same side for the causal question as for the moral question, or there may be some irrelevant, superficial property of each scenario that biases towards the left or the right. In both cases, one may worry that an irrelevant factor could create correlations between the two types of judgments as we gathered them. Let us make this more concrete. In all instances of the moral test, the positive answer to the moral question was on the right, while the negative answer was on the left. In the causal test, when the ‘moral description’ is on the right, the observed endorsement on a  $[0, 100]$  (left-right) scale should be the subjects’ ‘real’ judgment  $\alpha$ , augmented by the bias  $\varepsilon$  — hence  $(\alpha + \varepsilon)$ . When the ‘moral description’ is on the left, the observed endorsement on a  $[0, 100]$  scale (with 100 representing maximal endorsement of the *opposite* description) should be  $(100 - \alpha)$ , augmented as well with the bias  $\varepsilon$  — hence  $(100 - \alpha) + \varepsilon$ . By subtracting the endorsements obtained when the moral description was on the left from those obtained when the moral description was on the right, we obtain:  $(\alpha + \varepsilon) - ((100 - \alpha) + \varepsilon)$ , hence  $(2\alpha - 100)$ . This is just as good as  $\alpha$  to compute the correlations we are interested in — but the advantage is that we have eliminated the bias  $\varepsilon$ . Note that  $(2\alpha - 100)$  will take values between  $+100$  and  $-100$ , as is shown on the graph.

us make one more substantial comment on the actual results. Figure 7b shows that there is a systematic bias towards judging utilitarian scenarios ‘higher’ than their anti-utilitarian version. Such an effect is unsurprising when it concerns moral judgment (Figure 7a), but it is puzzling here, since we attempt to assess the purely causal analysis of a scene. Specifically, it goes against an implicit aspect of the test (1) we are using, namely that this test should reveal causal judgments without being polluted by moral judgments. We will comment on this retroaction effect in later sections (§3.2.3). In subsequent correlation analyses, we will abstract away from this effect by collapsing judgments obtained for utilitarian and anti-utilitarian scenarios, hoping that this will lead to a less biased measure of causal judgments. We could also decide to compute these correlations based on the causal judgments obtained from utilitarian scenarios only, or from anti-utilitarian scenarios only: the results would not be different.

### 3.2.2 The role of causal judgments in moral judgments, and the form of the interaction between the Utilitarian Precondition and the Causal Constraint

Figure 8a presents the correlations between moral judgments (both for utilitarian and anti-utilitarian scenarios) and corresponding causal judgments (aggregating utilitarian and anti-utilitarian scenarios, as explained above). The core of our proposal is that causal judgments influence moral judgments. We thus predict that the two types of judgments should be correlated.

We found that moral judgments to utilitarian scenarios are significantly correlated with causal judgments ( $r^2 = .69, t(12) = 5.1, p < .001$ ). Interestingly, we found that moral judgments to anti-utilitarian scenarios are also significantly correlated with causal judgments ( $r^2 = .76, t(12) = 6.2, p < .001$ ).<sup>8</sup> We will now examine the consequence of these two correlations for the three possible types of interactions between the Utilitarian Precondition and the Causal Constraint, as we introduced them in section 2.2.3.

#### Against the conjunctive analysis: no flat curve for the anti-utilitarian scenarios

The robustness of *both* correlations, and in particular the correlation that concerns anti-utilitarian scenarios, goes against a conjunctive analysis, which predicts that the anti-utilitarian curve should be flat (see Figure 1, §2.2.3).

#### In favor of the lexicographic analysis: separability found

The intercepts of the two correlations are different, i.e. one correlation line is above the other (confidence intervals are [64, 67] and [24, 26], with no overlap). This intercept difference corresponds to the effect of the utilitarian factor: utilitarian scenarios are judged more moral than their anti-utilitarian counterparts, a reassuring fact we already noticed

<sup>8</sup>We also computed correlations between moral judgments given in response to the utilitarian version of a scenario with causal judgments given to this scenario or to its anti-utilitarian counterpart (instead of the correlation with the composite causal measure). Correlations were similarly robust, which is expected given that utilitarian scenarios and their corresponding anti-utilitarian versions give rise to correlated causal judgments, as discussed in section 3.2.3.

from the average judgments in Figure 7a. More importantly, the two curves are *separable*: the highest point from the anti-utilitarian curve is below the lowest point from the utilitarian curve. This result is the property *separability* specifically predicted by the lexicographic analysis (see Figure 2, §2.2.3).

### In favor of the additive analysis: parallelism found

The slopes of the two correlations are *not* significantly different (the two-tailed confidence intervals at the .05 level for them overlap: [.082, .20] and [.097, .20]). The similarity of the slopes shows that if the causal analysis of a scene is a source of the observed moral judgments, this source is similarly active in both the utilitarian and anti-utilitarian scenarios. This result is the *parallelism* property specifically predicted by the additive analysis of the interaction between the Utilitarian Precondition and the Causal Constraint (see Figure 2, §2.2.3).

Let us confirm the similarity between the two curves. Figure 9a shows that utilitarian and anti-utilitarian versions of a given scenario yield significantly correlated moral judgments ( $r^2 = .43, t(12) = 3.0, p = .011$ ). Specifically, the moral judgments for utilitarian scenarios ( $\mathcal{M}_U$ ) are best predicted as the following linear function of the moral judgments in the anti-utilitarian scenarios ( $\mathcal{M}_A$ ):  $\mathcal{M}_U = .63 \times \mathcal{M}_A + 44$ . The 44 constant that appears at the end of this formula reveals that, unsurprisingly, utilitarian scenarios are judged as more moral than their anti-utilitarian counterpart. More importantly, the robustness of this correlation provides another piece of evidence in favor of the parallelism between the two curves, as predicted by an additive analysis of the interaction between the Utilitarian Precondition and the Causal Constraint.<sup>9</sup>

### Analysis of the first block of responses: against a strategic effect

The position of the descriptions in the causal task was counterbalanced as a protection against superficial improvements of the target correlations we discussed above (e.g., such undesirable improvements could arise if participants try to ‘reproduce’ previous answers, or if there exists a tendency for a scenario to prompt a response ‘to the right’). But one may worry that the presence of both types of questions in the same setting may play some role at a more abstract level and explain part of the robustness of the target correlations. For instance, participants who answered the moral questions first may be artificially inclined to include some moral evaluation in their assessment of the scenarios, which could have an impact on their answers to whatever non-moral questions they were asked later. To control for this artificial influence of one type of judgments on the other, we extracted the answers for the first block that any participants would have been administered. This restricted set of responses concerns either the moral or the causal questions for each participant, at the point of the experiment when he or she would not

<sup>9</sup>Note that the confidence interval for the slope in the formula above ([.17, 1.1]) includes the value 1, which leaves us with no reason to believe that the causal effect is different in any way for utilitarian and anti-utilitarian scenarios.

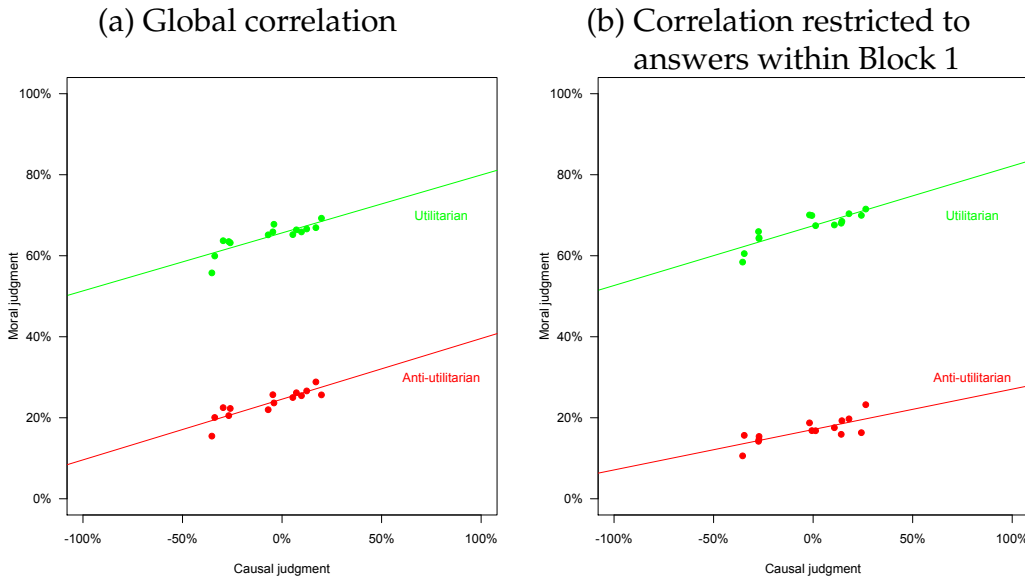


Figure 8: Correlations between moral judgments (y-axis, utilitarian scenarios in green, anti-utilitarian scenarios below in red) and causal judgments (x-axis, composite measure) to corresponding stories. Each dot thus corresponds to a ‘mode of action’ (see §3.1.1). Figure 8a represents global correlations, while Figure 8b is restricted to answers given within the first block of items, when participants could not be aware of the second kind of question (whether causal or moral).

be aware of the other type of question.<sup>10</sup> As illustrated in Figure 8b, the previous result is maintained: causal judgments correlate significantly with moral judgments both in the utilitarian case ( $r^2 = .50, t(12) = 3.5, p < .005$ ) and in the anti-utilitarian case ( $r^2 = .72, t(12) = 5.6, p < .001$ ).

### 3.2.3 A possible retroaction of moral judgments on causal judgments and its scope

Let us examine more closely the relation between causal judgments in utilitarian and anti-utilitarian versions of each of our scenarios. As illustrated in Figure 9b, causal judgments in utilitarian and corresponding anti-utilitarian scenarios are significantly correlated ( $r^2 = .88, t(12) = 9.4, p < .001$ ). However, the two are not *identical* either, despite the fact that a difference in terms of the final count of deaths is expected not to affect *at all* the causal analysis of a scene. Formally, causal judgments in utilitarian scenarios ( $C_U$ ) are best described as the following linear function of causal judgments ( $C_A$ ) in corresponding anti-utilitarian scenarios:  $C_U = 1. \times C_A + 26$ . This formula shows that causal judgments are identical for utilitarian and corresponding anti-utilitarian scenarios (cf. the striking ‘1.’ for the slope in the formula above), except that the utilitarian scenarios generate more ‘morally’ biased causal analyses (cf. the positive ‘26’ intercept, significantly different from 0, confidence interval at level .05: [21, 31]).

<sup>10</sup>Note that this new correlation is a pure between subject correlation, in the sense that we are now correlating the moral judgments of one set of participants with the causal judgments of a different set of participants.

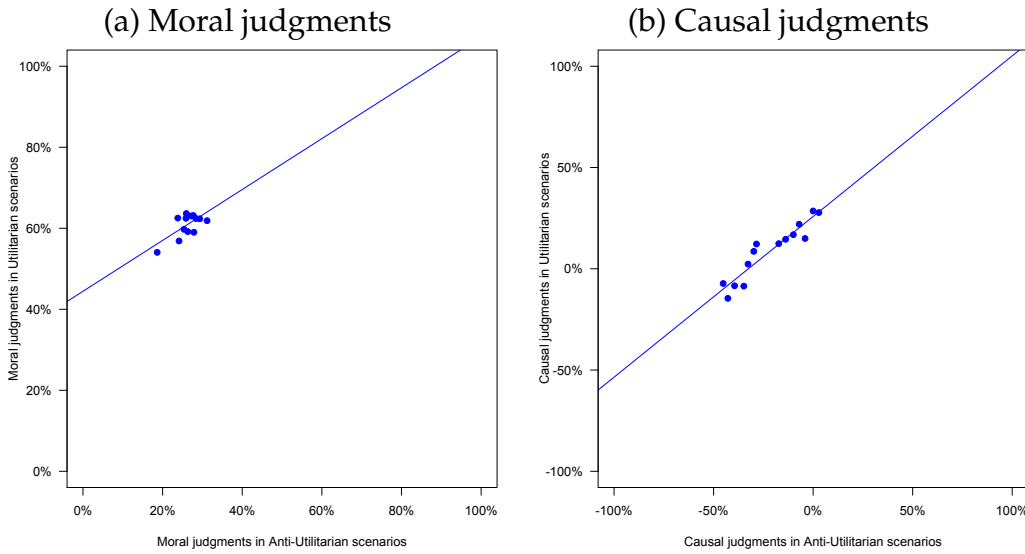


Figure 9: Correlations between utilitarian and corresponding anti-utilitarian scenarios, for (a) moral judgments and (b) causal judgments.

This finding creates the worry that our results might all be explained by a retroaction of moral judgments on causal judgments (see Alicke 1992, Knobe 201x): since the utilitarian factor impacts causal judgments as we measured them, this might be the source of the correlations we found so far. To rule out this possibility, let us measure the retroaction more precisely. The impact of the utilitarian factor is 44 for moral judgments (recall:  $\mathcal{M}_U = \alpha \times \mathcal{M}_A + 44$ ) and 26 for causal judgments. In other words, the utilitarian factor yields a change of 44 ‘moral points’, which translates into a change of 26 ‘causal points’. If the correlation between moral and causal judgments were solely due to the retroaction, we would thus expect the following relation between moral ( $\mathcal{M}$ ) and causal ( $\mathcal{C}$ ) judgments:  $\mathcal{M} = \frac{44}{26} \times \mathcal{C} \approx 1.7 \times \mathcal{C}$ . The facts are quite different: the slopes are significantly different from 1.7 since they lie in the interval between .08 and .20 (at confidence level .05). This would correspond to a much bigger intercept (at least 80, rather than 26). Hence, there is more to the correlation between moral and causal judgments than a retroaction.<sup>11</sup>

We conclude that (i) there is some retroaction of moral judgments on causal judgments, but (ii) the correlations between moral and causal judgment that we used to defend the role of the Causal Constraint in moral judgments is not reducible to this retroaction. How should we explain this retroaction, however? Two explanations suggest themselves. The first is that indeed moral judgments intrude in our test for causal judgments (in agreement with Alicke 1992, Knobe 201x). Another more superficial explanation is that we are seeing a superficial effect due to the form of our test. Specifically, if participants focus their attention on the event involving *more people*, they might prefer descriptions in which the most populated event is mentioned first. This attentional preference would explain

<sup>11</sup>There is a simplification here: these are confidence intervals for the slopes of the correlation between moral judgments and the aggregated measure of the causal judgment, which may vary as twice the bare causal judgments (see footnote 7). Arguably, the retroaction effect is too small to account for our data even if we do not make this simplification:  $1.7 \notin [.08 \times 2, .20 \times 2]$  and  $80/2 > 26$ .

that in anti-utilitarian scenarios, participants may be biased towards what corresponds to the ‘non-moral’ description (‘John caused the death **of the 5**, thereby...’), whereas they would be biased towards the ‘moral’ description in utilitarian scenarios (‘John saved **the 5**, thereby ...’). This alternative explanation is based on a simple strategic effect, and it does not make reference to moral judgments at all. We set aside a more detailed discussion of the choice between these two explanations, since the retroaction effect we observe is in any case not essential to the correlation we found.

### 3.2.4 Individual variations

As discussed in the introduction, our approach does not predict that moral judgments should be uniform in the population. Rather, moral judgments may vary from one individual to the next, provided that their causal analyses vary accordingly. We would like to show how the kind of results we gathered may on a larger scale help set out to investigate this issue. Let us first present two pieces of analyses.

1. First, the coherence of the judgments across participants is weak. Coherence (computed as the average of the Kendall  $\tau$  correlation coefficients for all pairs of participants)<sup>12</sup> is .029 for moral judgments and .042 for causal judgments.<sup>13</sup>
2. Second, rather than computing the correlation between average moral judgments in the population and average causal judgments in the population, we computed the moral/causal correlation *for each participant*. On average, the resulting correlation is significantly weaker: adjusted- $r^2$ s are .023 (average of the correlations obtained for each participant) vs. .49 (global correlation in the population),  $t(48) = -19, p < .001$ .<sup>14</sup>

While the first result above suggests that there is variation in the population, both for moral and causal judgments, the second result suggests that if we track these variations down to the individual levels, these variations are not ‘parallel’. This non-parallelism between moral and causal judgments at the individual level seems to go against our approach. However, individual judgments are noisy, because they rely on one data point per scenario and per judgment, while average judgments rely on as many such data points as there are participants. Individual results will also be influenced by order effects and, overall, they are expected to show more irrelevant variability. Hence, we would need more robust individual data to reach a firm conclusion, but we hope that we have illustrated clearly the kind of predictions and analyses that could be investigated at the individual level.

<sup>12</sup>Other correlation measures yield similar results.

<sup>13</sup>These measures are given based on the utilitarian cases only. Thus, the coherence measure is not artificially boosted by the utilitarian/anti-utilitarian split on which all participants should agree.

<sup>14</sup>Other correlation measures yield similar results.

## 4 Concluding remarks

The results of our study suggest three main conclusions. The first is that the correlation we found between moral judgments and the preference for one causal description of the scenarios confirms the analysis of moral judgments proposed by Mikhail for trolley problems. Mikhail, in particular, argues that the distinction between goal and side-effect is essential to the conceptual repertoire by which we compute mental representations of human acts and decide of their morality. A proper semantic analysis of the meaning of bi-clausal constructions involving ‘thereby’ remains to be undertaken, but we see that such constructions can be used to encode a distinction between a causally prior and a causally secondary action, and that the decision of which action to mark as prior or secondary reflects which action is given the higher moral value.

Secondly, our study casts further light on the Principle of Double Effect and its scope. In Mikhail’s words (Mikhail 2011, p. 149), the principle states that:

“an otherwise prohibited action (...) which has both good and bad effects may be permissible if

- [1] the prohibited act itself is not directly intended,
- [2] the good but not the bad effects are directly intended,
- [3] the good effects outweigh the bad effects, and
- [4] no morally preferable alternative is available.”

The most important lesson from our results on this issue concerns the interpretation of the Utilitarian Precondition (as expressed in clauses [3] and [4] of Mikhail’s formulation) and the Causal Constraint (as expressed in clauses [1] and [2]), and the interaction between them. What our results show is that even in cases in which the Utilitarian Precondition is violated, namely in anti-utilitarian scenarios where neither clause [3] nor clause [4] of Mikhail’s formulation is satisfied, an action may still be judged moral *to the extent* that one causal analysis is preferred over the other, that is to the extent that the act of saving is seen as causally prior to the act of killing.<sup>15</sup> As we have argued, this implies that a conjunctive analysis of the interaction between the Utilitarian Precondition and the Causal Constraint would be an inadequate articulation of the Principle of Double Effect. Instead, we have seen that our results are compatible with two alternative models for this interaction: one in which the Utilitarian Precondition is ranked lexicographically above the Causal Constraint; another in which the two constraints are not necessarily ranked one above the other, but in which they interact additively. So far our results do not allow us to choose between the latter two models, because each of them has some advantage: the lexicographic model accurately predicts what we have called *separability* between the utilitarian and the anti-utilitarian scenarios, while the additive model accurately predicts a striking *parallelism* between them.

<sup>15</sup>Granted, the Principle of Double Effect concerns the moral *permissibility* of an action; in our tests, we asked subjects to evaluate whether an action was *moral* or not, rather than morally *permissible* or not. This is an important difference, but we believe the judgments of morality should bear on judgments of moral permissibility.

Our third conclusion is more methodological and concerns the predictive character of our theory. Most theories of trolley dilemmas consist in *post hoc* analyses or rationalizations of the contrasts obtained in various conditions. Our analysis shows how to make such non-predictive analyses predictive. That is, we have proposed a way to predict moral judgments based on a correlation with an independent linguistic conceptualization of the scenarios at hand. Further work remains to be done, however, in particular to establish the robustness of the correlations we found at the individual level, and also to probe cases of causal intervention on threats as opposed to victims, as in Waldmann and Wiegmann (2010). For such cases, the linguistic judgments we used may no longer adequately predict moral acceptability, in particular for Waldmann and Wiegmann's most complex scenarios, in which an intervention on a threat is also an intervention on a potential victim.

## References

- Alicke, M. (1992). Culpable causation. *Journal of Personality and Social Psychology* 63(3), 368.
- Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108(2), 353–380.
- Cushman, F. and J. D. Greene (201x). Finding faults: How moral dilemmas illuminate cognitive structure. In J. Decety and J. Cacioppo (Eds.), *The Handbook of Social Neuroscience*. Oxford University Press. In press.
- Donagan, A. (1977). *The Theory of Morality*. Chicago University Press.
- Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review* 5, 5–15.
- Goldman, A. (1970). *A Theory of Human Action*. Princeton University Press.
- Greene, J., F. Cushman, L. Stewart, K. Lowenberg, L. Nystrom, and J. Cohen (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition* 111(3), 364–371.
- Greene, J., R. Sommerville, L. Nystrom, J. Darley, and J. Cohen (2001). An fMRI investigation of emotional engagement in moral judgment. *Science* 293(5537), 2105–2108.
- Hauser, M., F. Cushman, L. Young, R. Kang-Xin J., and J. Mikhail (2007). A dissociation between moral judgments and justifications. *Mind & Language* 22(1), 1–21.
- Kager, R. (1999). *Optimality Theory*. Cambridge: Cambridge University Press.
- Knobe, J. (201x). Actions trees and moral judgment. *Topics in Cognitive science*. Forthcoming.
- McIntyre, A. (2001). Doing away with double effect. *Ethics* 111(2), 219–255.
- Mikhail, J. (2000). *Rawls' Linguistic Analogy: A study of the 'generative grammar' model of moral theory described by J. Rawls in A Theory of Justice*. PhD Dissertation, MIT.
- Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences* 11(4), 143–152.
- Mikhail, J. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge University Press.
- Mikhail, J., C. Sorrentino, and E. Spelke (1998). Toward a universal moral grammar. In



- M. Gerbascher and S. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum Associates.
- Rawls, J. B. (1971). *A theory of justice*. The Belknap Press of Harvard University Press. Revised edition 2003.
- Royzman, E. and J. Baron (2002). The preference for indirect harm. *Social Justice Research* 15(2), 165–184.
- Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, 1–13.
- Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal* 94(6), 1395–1415.
- Waldmann, M. and J. Dieterich (2007). Throwing a bomb on a person versus throwing a person on a bomb intervention myopia in moral intuitions. *Psychological science* 18(3), 247–253.
- Waldmann, M. and A. Wiegmann (2010). A double causal contrast theory of moral intuitions in trolley dilemmas. In *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society*, pp. 2589–2594.