Beyond moral dilemmas: The role of reasoning in five categories of utilitarian judgment

François Jaquet a, Florian Cova b,∗

a Centre de Recherche en Éthique, University of Montreal, Boul. Édouard-Monpetit 2910, CA - H3T 1J7 Montreal, Canada
b Swiss Center for Affective Sciences, University of Geneva, Campus Biotech, CISA - Université de Genève, Chemin des Mines 9, CH - 1202 Genève, Switzerland

A R T I C L E   I N F O

Keywords:
Joshua Greene
Moral dilemmas
Moral judgments
Moral psychology
Moral reasoning
Utilitarianism

A B S T R A C T

Over the past two decades, the study of moral reasoning has been heavily influenced by Joshua Greene’s dual-process model of moral judgment, according to which deontological judgments are typically supported by intuitive, automatic processes while utilitarian judgments are typically supported by reflective, conscious processes. However, most of the evidence gathered in support of this model comes from the study of people’s judgments about sacrificial dilemmas, such as Trolley Problems. To which extent does this model generalize to other debates in which deontological and utilitarian judgments conflict, such as the existence of harmless moral violations, the difference between actions and omissions, the extent of our duties of assistance, and the appropriate justification for punishment? To find out, we conducted a series of five studies on the role of reflection in these kinds of moral conundrums. In Study 1, participants were asked to answer under cognitive load. In Study 2, participants had to answer under a strict time constraint. In Studies 3 to 5, we sought to promote reflection through exposure to counter-intuitive reasoning problems or direct instruction. Overall, our results offer strong support to the extension of Greene’s dual-process model to moral debates on the existence of harmless violations and partial support to its extension to moral debates on the extent of our duties of assistance.

1. Introduction

In moral philosophy, utilitarianism is the theory according to which the right action is that which maximizes aggregate well-being.1 By opposition, deontological theories take actions to be right or wrong regardless of their consequences, such that some acts are wrong even though they maximize well-being and others right even though they fail to do so. Over the past two decades, Joshua Greene (2007, 2008) greatly contributed to importing this lexicon into moral psychology. In his terminology, a moral judgment is “characteristically utilitarian” if it is more easily justifiable with a utilitarian theory; “characteristically deontological” if it is more easily justifiable with a deontological theory (Greene, 2014). Of course, Greene’s influence goes way beyond this lexical innovation. The current psychological study of moral judgment is largely dominated by his dual-process model, which can be summarized with the following two claims (Greene, 2008, 2014):

(1) Characteristically deontological judgments are preferentially supported by automatic emotional responses.

(2) Characteristically utilitarian judgments are preferentially supported by conscious reasoning and allied processes of cognitive control.2

In this paper, we ignore the former claim and focus on the latter. Let’s just note that researchers have documented a wide array of evidence in favor of the former claim. First, it has been shown that deontological judgment is correlated positively with empathic concern and sensitivity to the suffering of others (Conway and Gawronski, 2013; Sinhababu, 2018).

1 All forms of utilitarianism assess acts in terms of their impact on aggregate well-being. As a reviewer notices, however, some do not demand that we maximize aggregate well-being. Thus, satisficing utilitarians hold that an act is right if and only if it produces a sufficient amount of well-being (Slote, 1984), and scalar utilitarians merely hold that an act is good to the extent that it produces well-being (Norcross, 2006; Sinhababu, 2018). Following the standard practice in moral psychology, this paper focuses on the classical, maximizing version of utilitarianism.

2 In the rest of the paper and for the sake of brevity, we omit the qualification “characteristically” before the expressions “deontological judgment” and “utilitarian judgment”.

https://doi.org/10.1016/j.cognition.2020.104572

Received 21 April 2020; Received in revised form 21 November 2020; Accepted 22 December 2020
Available online 2 January 2021

0010-0277/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
notorious trolley cases, in which the agent can divert a train about to run to react in a deontological way, elicit more activity in brain regions have shown that people to save social moral dilemmas processes that are allegedly conscious but also slow, and require (Paxton et al., 2012) while having participants answer under cognitive telers, 2008) or scores to the Cognitive Reflection Test (Paxton et al., 2012; reflective cognitive style, such as participants evidence: utilitarian judgment is positively correlated with measures of DLPFC activity within individuals (Greene et al., 2004). Moreover, deontological responses to these dilemmas are correlated with greater activity in the amygdala (Shenhar and Greene, 2014).

There is also much evidence in favor of the second claim, which can be sorted out in three main categories. First, there is correlational evidence: utilitarian judgment is positively correlated with measures of reflective cognitive style, such as participants’ need for cognition (Bartels, 2008) or scores to the Cognitive Reflection Test (Paxton et al., 2012; Byrd and Conway, 2019; see Hannikainen and Cova, 2020 for a meta-analysis). Second, there is experimental evidence: it has been shown that priming a counter-intuitive mindset increases utilitarian judgment (Paxton et al., 2012) while having participants answer under cognitive load or time pressure decreases utilitarian judgment (Conway and Gawronski, 2013; Suter and Hertwig, 2011) or at least makes it slower (Greene et al., 2008). Third, there is neuroscientific evidence, with studies showing that utilitarian judgments are associated with increased DLPFC activity within individuals (Greene et al., 2004).

In this paper, we focus on the second type of evidence, that is: experimental evidence for the claim that utilitarian judgments are preferentially supported by “conscious reasoning and allied processes of cognitive control” (Greene, 2014:699). As we will see, most of the studies from which this evidence is drawn do not directly investigate whether the psychological processes underlying utilitarian judgment are more conscious than those underlying deontological judgments. And only a handful can be thought to directly investigate whether utilitarian judgments are more likely to require cognitive control. Instead, some of these studies investigate features supposedly related to conscious reasoning and cognitive control, such as taking more time or requiring more cognitive resources (i.e., working memory). However, we leave these conceptual worries aside for the time being and will come back to them in our final general discussion. In the meantime, we will use the terms “reflection” and “reflective processes” to refer to the type of psychological processes supposed to underlie utilitarian judgments—processes that are allegedly conscious but also slow, and require cognitive control but also more cognitive resources.

Through five studies, we tried to generalize the results of previous studies that aimed to put Greene’s dual-process model to test by manipulating participants’ use of reflective reasoning (by hindering or promoting it) to new contexts in which deontological and utilitarian judgments tend to yield divergent conclusions.

2. Beyond sacrificial dilemmas: the generality problem

As we saw in the previous section, researchers have collected an impressive amount of evidence on the psychological underpinnings of deontological and utilitarian judgments, and in favor of the dual-process model of moral judgment. However, past studies suffer from a problem that we can call the “generality problem”: most of these studies did not investigate the psychological roots of deontological and utilitarian judgments beyond the context of sacrificial moral dilemmas. By “sacrificial moral dilemmas”, we mean cases in which an agent can sacrifice \( k \) people to save \( n \) people, where \( k < n \). Such dilemmas include the notorious trolley cases, in which the agent can divert a train about to run over five people towards a sixth person, or throw said sixth person from a bridge to stop the train (Foot, 1967; Thomson, 1976, 1985; see Cova, 2017 for a primer on the trolley problem).

One reason for this focus is that sacrificial moral dilemmas constitute a simple, convenient and easily-administered case of dissociation between deontological and utilitarian judgments: typically, utilitarians will find it permissible to sacrifice one person to save more, since doing so maximizes overall happiness, while deontologists will deem this sacrifice impermissible, as it violates moral rules such as “You shalt not kill.” However, conflicts between deontologists and utilitarians are not limited to sacrificial dilemmas. They involve other issues, such as the existence of harmless moral violations, the respective permissibility of active and passive euthanasia, the extent of our duty towards people in need, and the justification for punishment. One can thus wonder whether the conclusions drawn from sacrificial dilemmas can be extended to such debates. Will reflection favor the utilitarian conclusion in such matters as well?

Greene himself claims that his model is supposed to cover all cases of conflicts between deontological and utilitarian judgments, and he explicitly extends it to cases outside the context of sacrificial moral dilemmas:

1) Greene (2003) extends his model to cover debates about whether we must give large parts of our income to people in need.
2) Greene (2008) extends his model once again to debates about the extent of our duties of assistance, but also to debates about the justification of punishment and the existence of harmless moral violations.
3) Greene (2014) explicitly claims that his model applies to debates about the existence of harmless moral violations, the justification of punishment, and the extent of our duties of assistance.

Thus, in addition to sacrificial moral dilemmas, Greene claims that his model also applies to three other debates between deontology and utilitarianism: whether there can be harmless moral violations, the extent of our duties of assistance, and the appropriate justification for punishment. But do we have evidence that, in those debates, reflection will favor the utilitarian answer?

According to utilitarianism, an action can be wrong only in virtue of its consequences on overall welfare, such that a harmful action cannot be morally wrong (at worst, it will be morally neutral, and at best morally good if it pleases someone). On the other hand, deontologists are not bound by this restriction. Deontological theories can condemn harmful actions (such as violations of self-directed duties) for different motives. In the past years, harmless moral violations have attracted much attention from moral psychologists (see for example Haidt et al., 1993), notably because they constitute striking counter-examples to the traditional claim that morality is primarily about harm, fairness, and cooperation (Graham et al., 2011; Kelly et al., 2007). Though some researchers still doubt that one can condemn an action without perceiving

---

3 Though such dilemmas provide a fast and convenient way of dissociating deontological and utilitarian judgments, this distinction is nowhere perfect. For example, this method would flag as “utilitarian” a person who would find it acceptable to sacrifice someone even if there is no gain in return. To correct for these methodological shortcomings, Conway and Gawronski (2013) advocated the use of a “process-dissociation” approach that allows for a separate assessment of participants’ “deontological” and “utilitarian” tendencies. This approach can lead to observations that would have been impossible to make using only traditional sacrificial dilemmas (Conway et al., 2018). For example, Byrd and Conway (2019) found that logical reflection (as measured by the Belief Bias task) is positively correlated with both utilitarian and deontological inclinations (see Reynolds, Byrd & Conway, in prep. for a meta-analysis). However, since such a method did not seem directly applicable to some of our stimuli (such as harmless crimes or demanding ethics vignette), we chose a more traditional methodology.
it as harmful (e.g., Gray et al., 2014), this interest has spawned much research on the role of emotions (notably disgust) in the condemnation of harmless moral violations (e.g., Inbar et al., 2009, 2012). However, studies investigating the role of reflection in assessing such cases are scarce. As predicted by Greene, a handful of studies have found a positive correlation between reflection (as measured through CRT scores or need for cognition) and more permissive (and thus more utilitarian) assessment of harmless violations (Cova and Jaquet, 2017; Jaquet, 2015; Pennycook et al., 2014; Royzman et al., 2014a). Moreover, this correlation has been confirmed as robust by a recent meta-analysis (Hannikainen and Cova, 2020). On the manipulation side, Paxton et al. (2012) found that presenting participants with a strong argument in favor of tolerance and giving them time to deliberate leads them to be more tolerant of harmful violation, while Hannikainen and Rosas (2019) found that forcing participants to reflect on whether an action is harmful or not leads them to become more tolerant of harmful violation. Although these results show that reflection can sometimes lead to more utilitarian judgments when participants are nudged towards the utilitarian conclusion by considerations specifically relevant to utilitarian judgment, they do not show that reflection preferentially leads to utilitarian judgment in less specific contexts.

Data are also scarce concerning our duties of assistance, which constitute another point of disagreement between utilitarians and deontologists. Indeed, utilitarianism entails that we should sacrifice our resources to help others as soon as this sacrifice maximizes overall well-being. That is why utilitarianism has been criticized for being too demanding by its deontological opponents. Though the emotional determinants of helping behavior have been thoroughly investigated (e.g., Slovic, 2007; Small and Loewenstein, 2003), there has been very little research about moral intuitions concerning our duties of assistance (Nagel and Waldmann, 2013, 2016). However, as a solution to the generality problem, Kahane et al. (2018) distinguish between two components of utilitarian thought: permissive attitudes towards “instrumental harm” and “impartial beneficence” (see also Everett and Kahane, 2020). In their two-dimensional model of utilitarian psychology, impartial beneficence is measured by items such as “from a moral point of view, we should feel obliged to give one of our kidneys to a person with kidney failure since we do not need two kidneys to survive, but really only one to be healthy” or “it is morally wrong to keep money that one doesn’t really need if one can donate it to causes that provide effective help to those who will benefit a great deal.” The “impartial beneficence” subscale of their Oxford Utilitarianism Scale can thus be thought to probe intuitions about our duties of assistance. However, they found no relationship between impartial beneficence and participants’ need for cognition. Moreover, Capraro et al. (2019) found that priming reflection (vs. emotion) by asking participants to answer based on reason (vs. emotions) led participants to give more utilitarian answers for the “instrumental harm” subscale but not for “impartial beneficence.” There is therefore little support for the claim that reflection typically promotes a more utilitarian (and demanding) conception of our duties of assistance.

Finally, we have even fewer data about the role of reflection in moral judgments about punishment. Past research has established that people’s moral intuitions about punishment follow a deontological pattern, rather than a utilitarian one: the magnitude of punishment should depend on desert (i.e., the perpetrator’s intention and the magnitude of the harm inflicted) rather than efficiency (i.e., deterrence and perpetrator’s rehabilitation) (Carlsmith, 2006; Carlsmith et al., 2002; Sunstein et al., 2000). To our knowledge, however, no findings show that reflection leads participants to adopt a more utilitarian conception of punishment. The only suggestive evidence is that, while participants’ intuitions about particular cases are clearly deontological, probing their justification for punishment in a more abstract way leads them to be more utilitarian (Carlsmith, 2008).

In sum, there is a lack of evidence about the impact of reflection on utilitarian judgment in contexts outside sacrificial dilemmas. The present paper aims to fill this gap by systematically investigating the effect of reflection manipulation (hindering reflection in Studies 1 and 2, promoting reflection in Studies 3 to 5) on utilitarian judgment across different topics.

3. Material generation and validation

To investigate the impact of reflection on utilitarian judgment across various topics, we first needed to generate and validate different sets of scenarios that would pit utilitarian answers against deontological ones. Though the detailed description of this generation and validation process is the topic of another paper (Jaquet and Cova, 2020; see Cova and Jaquet, 2017 and Jaquet, 2015 for early steps), we provide a short summary in this section. However, a full description of the whole procedure and results can already be found online at osf.io/muj2m/.

3.1. Material generation

We first generated six sets of 10 vignettes (for a total of 60 vignettes), each set representing a different topic of disagreement between deontology and utilitarianism.

3.1.1. Utilitarian dilemmas (UD)

Utilitarian dilemmas are vignettes in which the agent decides to harm (or kill) k persons to save n persons, where k < n and the people sacrificed would have been left unharmed if the agent had not acted. This includes the traditional Footbridge case.

3.1.2. Pareto dilemmas (PD)

Pareto dilemmas are moral dilemmas in which the agent decides to harm (or kill) k persons to save n persons, where k < n and the people sacrificed would have been harmed (or killed) anyway if the agent had not acted. This included the traditional Crying Baby case.

3.1.3. Harmless crimes (HC)

Harmless crimes are vignettes featuring actions that do not harm anybody but are still disturbing enough for some participants to consider them morally wrong. This includes traditional scenarios, such as Jonathan Haidt’s Chicken, Incest, Flag, and Dog cases (Haidt et al., 1993). In addition to these, we tried to include cases that would not trigger disgust (see, for example, Royzman et al., 2014b).

3.1.4. Action/Omission (AO)

Action/Omission vignettes are actually pairs of vignettes in which the same consequences are brought about either by the agent’s action or by the agent’s omission. Here is an example (drawn from Rachels, 1975):

Action—Jane’s brother and his wife recently died, leaving behind a six-year-old orphan, Jane’s nephew. Because Jane is the orphan’s only remaining relative, it is now her duty to take care of him. Jane’s brother was very rich. Thus, if anything should happen to her nephew, Jane would gain a large inheritance. One evening, while the child is taking his bath, Jane sneaks into the bathroom, planning to drown him in it. She does so and arranges things so that it will look like an accident.
Omission—Paula’s brother and his wife recently died, leaving behind a six-year-old orphan, Paula’s nephew. Because Paula is the orphan’s only remaining relative, it is now her duty to take care of him. Paula’s brother was very rich. Thus, if anything should happen to her nephew, Paula would gain a large inheritance. One evening, while the child is taking his bath, Paula sneaks into the bathroom, planning to drown him in it. However, she then sees the child slip, hit his head, and fall face down in the water. The child drowns all by himself while Paula watches and does nothing.

**Question:** How much morally worse was it for Jane to drown her nephew than for Paula to let her nephew drown?

We decided to ask participants to compare both versions (action and omission) of each vignette rather than asking them to rate each version separately because there was no convenient way of implementing separate presentations in the design of Studies 1 and 2. One might worry that the joint presentation of the two cases (action and omission) might motivate participants to stay coherent and make the difference between action and omission disappear (see, for example, Hsee et al., 1999). But a pre-test allowed us to observe that using a single comparative question rather than two separate questions (∼= one for each case) actually made it easier to detect the difference between action and omission. Still, the worry remains that this joint presentation led participants to give less intuitive, more reflective answers.

Past psychological research suggests that participants make a difference between action and omission: even when outcomes are kept identical, people tend to find omissions less morally problematic than actions (Baron and Ritov, 2004). This is reflected by people’s systematic preference for passive euthanasia over active euthanasia (Hauser et al., 2009; Sugarman, 1986). From a utilitarian point of view, however, this difference makes no sense, as consequences on well-being are the same. This is why utilitarians typically deem the difference between action and omission morally irrelevant, while deontologists are more likely to find it relevant. But does reflection lead people to treat action and omission similarly? This has been suggested by Baron (1994), who argues that the difference between action and omission results from unreliable heuristics. There is some evidence pointing into this direction: Bartels (2008) found that more reflective participants (as measured by the Rational vs. Experiential Inventory) tend to make less of a difference between action and omission. And Cushman et al. (2011) observed that activation in the frontoparietal control network is associated with lower differences in judgments about action and omission. Moreover, using the CNI model, a method allowing to dissociate preference for inaction (over action) from sensitivity to norms and consequences in moral dilemmas, Gawronski et al. (2017) found that putting participants under cognitive load led them to give more weight to the action/inaction distinction. Nevertheless, using the same method, Körner & Gawronski (2020) found no significant relationship between need for cognition and sensitivity to the action/inaction distinction.

3.1.5. Demanding ethics (DE)

Demanding ethics vignettes feature agents who did not self-sacrifice to help others, even though utilitarianism typically entails that they should have. Here is an example (drawn from Unger, 1996):

One day, Samantha finds an envelope in her mailbox. The letter is from the UNICEF and asks her for a donation. The money would be used to vaccinate innocent children against malaria, in a distant country in which the disease regularly kills thousands of children. If Samantha transfers $100, thirty children will be vaccinated. If she doesn’t, they will contract malaria and die as a result. Samantha knows all that. As she prefers to keep her money, she tosses the envelope into the trash. Because they were not vaccinated, the thirty children die from malaria.

**Question:** How morally wrong was it for Samantha not to send money to the UNICEF?

3.1.6. Punishment (P)

Punishment vignettes feature an agent who decided to punish or not to punish a criminal based on utilitarian considerations, and in opposition to conflicting deontological considerations. Here are two examples (drawn from Kant, 1797/1996, and Carlsmit et al., 2002):

Scientists have warned the population that, due to changes in the activity in the sun, the earth will be entirely destroyed by a solar eruption within a month. As it turns out, they are right. In a prison is a man convicted for the murder of his lover, whom he was obsessed with, and very jealous of. The victim has no remaining relatives. The murderer was originally scheduled to be executed in two months. This situation raises a debate among judges. Some think that, given that the planet and all its inhabitants are doomed, there is no point in punishing the murderer: there is no future crime to deter, and rehabilitation is pointless. Thus, the murderer should be set free for the remaining days. Others think that the murderer should be executed before the planet disappears, because he really deserves to be punished. Finally, the judges decide that there is no point in punishing the murderer. They set him free for the remaining days before the destruction of the planet.

**Question:** How morally wrong is it for the judges to release the murderer?

Eva is a judge in a small country. One day, two hackers, James and Peter, are arrested for skimming money by abusing a security breach in a bank’s computer system. Both hackers acted independently and do not know each other. Both stole the same amount: $10,000. The difference between the two hackers is that Peter is much better than James. Indeed, the program James used is easily detectable. In fact, users of such programs are identified and arrested 9 times out of 10. On the contrary, the program Peter used is very subtle and hardly detectable. In fact, users of such programs are identified and arrested 1 time out of 50. It is only through an incredible series of coincidences that Peter was caught in this case. Because Eva wants to maximize deterrence while not overcrowding the few remaining prisons, and because deterrence is effective only when the punishment is proportional to the probability of the offender escaping justice, Eva decides to give James a 2-year jail sentence while giving Peter a 10-year jail sentence.

**Question:** How morally wrong is it for Eva to give a more severe punishment to Peter?

For all six categories of vignettes, we drew both on past philosophical and psychological literature to make sure that our stimuli would be relevant to philosophical debates and our results comparable to those of previous studies.

3.2. Material validation

To validate our material, we went through four steps. In the first step, we made sure that our stimuli constituted cases of conflict between deontological and utilitarian judgments by having them rated by 10 professional ethicists (i.e., 8 PhDs and 2 PhD students). For each scenario, participants were presented with a moral statement corresponding to the question that would be given to naïve participants and asked:

1) How easy is it to justify this statement in deontological terms? (−3 = “extremely difficult”, 3 = “extremely easy”)

2) How easy is it to justify this statement in utilitarian terms? (−3 = “extremely difficult”, 3 = “extremely easy”)

Participants’ answers to the second question were subtracted from
participants’ answers to the first question. We considered that the scenario was valid when the difference between the two scores was superior to 1 and went in the predicted direction. Average differences ranged from 1.69 for DE vignettes to 3.64 for PD vignettes.

In the second step, 100 naïve participants were presented with the full set of 60 vignettes. Based on their answers, we excluded scenarios for which there was too much agreement among participants (i.e., vignettes for which the average score was below one point away from one extremity of the scale). This was done to mimic the previous literature on sacrificial moral dilemmas, which focuses on “high-conflict” moral dilemmas that bring about disagreement among participants (see Greene et al., 2008; Koenigs et al., 2007).

Based on these first two steps, we excluded four scenarios. This led us, in a third step, to create four new alternate scenarios and go through the same validation procedure again (validation by professional ethicists then by naïve participants). All four new vignettes were successfully validated.

In a fourth and final step, we had 196 naïve participants rate all 60 vignettes to assess the internal consistency of each of our six categories of vignettes. The results can be found in Table 1. Overall, standard Cronbach’s alphas suggest that internal consistency was good for all six categories.

4. Overview of the studies

Our initial goal was to investigate the effect of hindering reflection in Studies 1 and 2 (Study 1: cognitive load, Study 2: time constraint) and of promoting reflection in Study 3. However, promoting reflection proved harder than expected, so that we had to run two additional studies. Studies 3 to 5 are thus dedicated to investigating the effect of promoting reflection on utilitarian judgment.

5. Study 1—the impact of cognitive load on utilitarian judgment

Prior studies suggest that putting participants under cognitive load makes them less utilitarian (Conway and Gawronski, 2013; see also Trémolière and Bonnefon, 2014), which suggests in turn that deontological judgments are driven by quick, automatic processes, while utilitarian judgments are the product of more careful deliberation. However, these studies investigate moral judgment only about sacrificial dilemmas (i.e., cases in which one life must be forfeited to save many lives). In this first study, our goal was to study this effect in a broader range of cases.

We preregistered our analysis plan on OSF. Preregistration can be found at: https://osf.io/5b86c/. Due to schedule constraints, however, data collection had already started when we made the preregistration.

5.1. Method

5.1.1. Rationale for sample size

To compute the sample size needed to have a reasonable chance to detect an effect, we took the effect size in Conway and Gawronski (2013)’s second study as a starting point (which we estimated to be $d = 0.61$). In order to find an optimal solution given the resources available to us, we computed what sample size would be required in a within-subject design while varying the target power (.80 and .95) and whether we should assume the original effect size or only half of it ($d = .31$). We found that, in a within-subject design, a total of 137 participants would be needed to reach .95 power for an effect size of $d = .31$. We thus decided to aim for a total of 137 participants.

5.1.2. Materials

We used a total of 48 scenarios, plus four additional scenarios (Training scenarios) used at the beginning of the experiment for a short training session. The 48 scenarios were divided into six categories within each category: Utilitarian dilemmas, Pareto dilemmas, Harmless crimes, Action/Omission, Demanding ethics, and Punishment.

5.1.3. Procedure

The study took place in our laboratory at the University of Geneva, with the experiment being programmed in Python 2.

Instructions. Participants were greeted by the experimenter and told that they would be presented with 48 short vignettes. They were also told that, for each vignette, they would be asked (1) to indicate whether the action performed by the main character was morally wrong (or morally worse than a comparison point, for the Action/Omission scenarios), and (2) to rate the wrongness of this action on a seven-point scale. For the first question, they would have to answer either “YES” or “NO” by clicking on the corresponding button on the mouse (participants were asked to choose which button was associated to which response, so that the answer scheme was the most intuitive to them). For the second question, they would have to select an answer (ranging from 1 to 7) on the screen.

Participants were then told that the experiment would take place in two stages. In one stage, they would simply have to read a sequence of eight characters before reading the vignette and answering the questions. In the other stage, they would have to memorize the sequence and type it down at the end of the trial (after reading the vignette and answering the two questions). The two conditions were presented in random order.

Training phase. Participants began with a training session allowing them to get familiar with the display and the control scheme. The training session involved the four Training scenarios. For the first two scenarios, participants were simply asked to read the eight-character sequence. For the last two scenarios, they were asked to memorize and type the sequence down. Thus, participants were acquainted with both conditions before starting the experiment itself.

Study’s overall structure. Participants were then presented with the 48 other scenarios, divided into two blocks (four scenarios of each type per block). One block constituted the No Load condition, in which participants only had to read the eight-character sequence without memorizing it. The other constituted the Load condition, in which participants were instructed to memorize the sequence at the start of the trial and type it down at the end. The two conditions were presented in random order.

Trial’s structure. For each trial, participants were first presented with a yellow screen instructing them to click any button on the mouse to begin the trial. The eight-character string (randomly generated but
containing at least one uppercase letter, one lowercase letter, one number, and one punctuation mark) was then displayed. After reading (and, possibly, memorizing) the character string, participants clicked on the mouse to move to the next step. If they did not click, the character string automatically disappeared after 30 seconds.

Participants were then presented with the vignette. Once they were done reading it, they clicked on the mouse to go to the next screen, which presented them with the first (YES/NO) question. Once the question was answered (by clicking on the corresponding button of the mouse), the second (scalar) question appeared. Participants answered it by clicking on the corresponding button on the screen.

In the Load condition, participants were then asked to input the character string they were instructed to memorize. After entering their answer, they were presented with a feedback screen telling them how well they performed. Participants with a score between 0 and 3 were presented with a RED feedback screen (“NOT GREAT”), participants with a score between 4 and 6 were presented with an ORANGE feedback screen (“NO GREAT”), and participants with a score between 7 and 8 were presented with a GREEN feedback screen (“VERY GREAT”).

5.2.1. Participants

Participants were 160 students at the University of Geneva who participated in exchange for credits or a CHF 15 retribution.4 There were 127 women, 25 men, and eight unidentified (Mage = 21.83, SDage = 3.88).

5.2.2. Correlations between scores across conditions

In this study, we decided to use a within-subject design to increase our statistical power. However, a within-subject design only increases power if participants’ scores in both conditions are positively correlated. We thus computed the Pearson’s product-moment correlation between scores in the No Load and Load conditions for each type of vignettes. For binary (YES/NO) answers, correlations were: UD = .53, PD = .60, HC = .37, AO = .59, DE = .48, P = .17. For scalar answers, correlations were: UD = .60, PD = .60, HC = .55, AO = .72, DE = .59, P = .34.

5.2.3. Computation and comparison of utilitarian scores (YES/NO)

Participants’ utilitarian scores were computed as the percentage of YES answers for DE vignettes, and as the percentage of NO answers for the UD, PD, HC, AO, and P vignettes. When analyzing the data, we discovered that one scenario in the P category had been mistranslated from English to French, which resulted in the impossibility to distinguish between deontological and utilitarian answers. This scenario was therefore excluded from analyses, meaning that utilitarian scores for the P category were sometimes computed on three scenarios rather than four. Mean and standard deviations for each score are presented in Table 2.

Table 2. Participants’ utilitarian scores (YES/NO) in function of condition and type of scenarios and results of the paired t-test comparing scores across conditions (Study 1).

|        | No load | Load | Comparison (Welch t-test) |
|--------|---------|------|--------------------------|
| UD     | 0.38 (0.33) | 0.33 (0.33) | t(159) = 1.87, p = .06, d = 0.15 |
| PD     | 0.43 (0.37) | 0.54 (0.35) | t(159) = 4.33, p < .001, d = -0.34 |
| HC     | 0.58 (0.21) | 0.59 (0.33) | t(159) = 7.44, p < .001, d = 0.59 |
| AO     | 0.35 (0.34) | 0.39 (0.30) | t(159) = 1.42, p = .16, d = -0.11 |
| DE     | 0.46 (0.37) | 0.39 (0.34) | t(159) = 2.81, p < .005, d = 0.22 |
| P      | 0.55 (0.33) | 0.42 (0.27) | t(159) = 4.34, p < .001, d = 0.34 |

5.2.4. Computation and comparison of utilitarian scores (Scales)

For participants’ answers on scales, utilitarian scores were computed by averaging answers within each category. Mean and standard deviations for each score are presented in Table 2, while the median and distribution are presented in Fig. 1. For each of the six types of vignettes, we used a paired Welch t-test to compare utilitarian scores between the No Load and Load conditions. The results for the t-tests are presented in Table 3.

5.3. Discussion

In this study, we investigated the effect of cognitive load on utilitarian judgment. We found an effect of our manipulation in the predicted direction (i.e., less utilitarian judgment under load) on both binary and scalar answers for scenarios about Harmless Crimes, Demanding Ethics, and Punishment. For Action/Omission, we also found a significant effect for scalar answers, but not for binary answers. Overall, this suggests that previous results about the effect of cognitive load on sacrificial dilemmas can be generalized to other kinds of conflicts between deontology and utilitarianism.

Nonetheless, we failed to replicate the effect of cognitive load on utilitarian judgment for sacrificial dilemmas. As a matter of fact, for Pareto Dilemmas, there was a significant effect in the opposite direction: answers were more utilitarian when given under load.

6. Study 2—the impact of time constraint on utilitarian judgment

Prior studies by Suter and Hertwig (2011) have shown that asking people to solve sacrificial dilemmas quickly (in less than eight seconds) rather than slowly (in no less than three minutes) tends to make their moral judgments less utilitarian. This suggests that careful deliberation (vs. quick, automatic answers) favors utilitarian judgment. Here, we decided to use a similar method to determine whether this conclusion could be generalized to utilitarian judgments outside the realm of sacrificial dilemmas. We presented participants with our 48 scenarios (eight for each type of vignettes) and asked them to give a fast answer (with a time limit of roughly 10 seconds) or slow answer (with participants being forced to wait at least one minute before answering).

We preregistered our analysis plan on OSF. Preregistration can be found at https://osf.io/29xcw. Due to schedule constraints, however, data collection had already started when we made the preregistration.

6.1. Method

6.1.1. Rationale for sample size

To compute the sample size needed to have a reasonable chance to detect an effect, we took the effect size in Suter and Hertwig (2011)’s first study as a starting point (r = -0.31, which we converted to d = 0.65). In order to find an optimal solution given our resources, we computed what sample size would be required in a within-subject design while varying the target power (.80 and .95) and whether we should assume the original effect size or only half of it (d = .33). We found that, in a within-subject design, a total of 121 participants would be needed to reach .95 power for an effect size of d = .33. We thus decided to aim for a total of 121 participants.

4 We overshoot our initial goal of 137 participants because psychology students needed to participate in experiments to validate credits.
6.1.2. Materials

The four training vignettes and the 48 experimental vignettes were the same as in Study 1.

6.1.3. Procedure

The study took place in our laboratory at the University of Geneva, with the experiment being programmed in Python 2.

Instructions. Participants were greeted by the experimenter and told that they would be presented with 48 short vignettes. They were also told that, for each vignette, they would be asked whether the action performed by the main character was morally wrong (or morally worse than a comparison point, for the Action/Omission scenarios), and would have to answer either “YES” or “NO” by pressing on the corresponding button of the mouse (participants were then asked to choose which button of the mouse they preferred to be associated with which response, in order for the answer scheme to be the most intuitive to them).

Participants were then told that the experiment would take place in two stages. In one stage, they would have to answer very quickly (the Fast condition; see below for the exact time). In the other, they would have to wait at least one minute before answering (the Slow condition). The two conditions were presented in random order.

Training phase. Participants then began with a training session to get familiar with the display and the control scheme. The training session involved the four training scenarios. For the first two scenarios, participants were asked to read the text, then wait for one minute before answering the moral question. For the last two scenarios, they were asked to read the text, then answer quickly, before a 10-second countdown displayed on the screen reached its end. Reading time was recorded and averaged over all four scenarios to estimate the participant’s reading speed. Reading time was then used throughout the experiment to estimate the exact time limit for the Fast condition.

Study’s overall structure. Participants were then presented with the 48 other vignettes, divided into two blocks (four vignettes of each type per block). One block constituted the Fast condition, in which participants had to answer quickly. The other constituted the Slow condition, in which participants had to wait one minute before answering. The two conditions were presented in random order.

Trials’ structure. For each trial, participants were first presented with a yellow screen telling them to press any button on the mouse to begin the trial (Fig. 2). The vignette was then displayed, and participants were instructed to press on the mouse as soon as they were done reading to trigger the question’s display. Nevertheless, if participants took too much time, the program automatically jumped to the question. The time limit for reading depended on the participants’ reading speed (time limit = character numbers * average reading speed on practice trials * 1.4).

In the Fast condition, the question was displayed in the middle of the screen (in black fonts). On the upper side of the screen, a countdown (in red fonts) indicated how much time participants had left to answer (“Only X seconds left to answer”). If participants failed to answer before the countdown was over, negative feedback appeared under the form of a red screen (“TOO LATE!”). The time limit depended on the vignette’s length and the participant’s reading speed (time limit = character numbers * average reading speed * 1.2 + 5 seconds).

In the Slow condition, the question was displayed in the middle of the screen (in black fonts). On the upper side of the screen, a countdown (in blue fonts) indicated how much time participants had left to wait before answering (“Wait X seconds before answering”). Participants had to wait 60 seconds before answering. Once the countdown was over, it was replaced with a message letting participants know that they could access
6.2. Results

6.2.1. Participants

Participants were 121 students at the University of Geneva who participated in exchange for credits or a CHF 15 retribution. There were 92 women and 29 men (M\text{age} = 23.57, SD\text{age} = 6.71). One participant was excluded from analysis for having too many missing answers in the Fast condition (this exclusion criterion was specified in the preregistration).

6.2.2. Average time constraint for the Fast condition

Since the time allocated to participants to answer in the Fast condition varied as a function of their reading time, we computed the average time limit in the Fast condition. The average time limit was 10235ms (SD = 2294.37).

6.2.3. Correlations between scores across conditions

In this study, we decided to use a within-subject design to increase our statistical power. However, a within-subject design only increases power if scores in both conditions are positively correlated. We thus computed the Pearson’s product-moment correlation between scores in the Slow and Fast conditions for each type of vignettes: UD = .65, PD = .47, HC = .41, AO = .74, DE = .44, P = .30.

6.2.4. Computation and comparison of utilitarian scores

Utilitarian scores were computed as the percentage of YES answers for DE vignettes and as the percentage of NO answers for the UD, PD, HC, AO, and P vignettes. In some cases, participants failed to answer one question in the Constraint condition. In these cases, the percentage was computed on a basis of three answers. Participants who failed to answer more than one question in the same category were excluded from analysis (see above). Mean and standard deviations for each score are presented in Table 4, while median and distribution are presented in Fig. 3.

For each of the six types of vignettes, we used a paired Welch t-test to compare utilitarian scores between the Fast and Slow conditions. The results for the t-tests are presented in Table 4.

6.3. Discussion

In this study, we investigated the effect of a time constraint (fast vs. slow answer) on utilitarian judgment. We found an effect of our manipulation for scenarios about Harmless Crimes and Demanding Ethics: in both cases, being forced to wait before answering led to more utilitarian judgments than being forced to answer quickly. These effects are in line with the results of Study 1. However, in contrast with Study 1, we did not find an effect of our manipulation for Action/Omission and Punishment. Moreover, we failed once again to replicate the effect of the manipulation on sacrificial dilemmas.

7. Study 3—investigating the effect of prior exposure to the cognitive reflection test on utilitarian judgment (first attempt)

In Studies 1 and 2, we examined the effect on moral judgment of interventions supposed to prevent reflection (through cognitive load and time constraint). In Study 3, our goal was to look in the other direction and study the effect on moral judgment of interventions purported to promote reflection.

In a famous experiment, Paxton et al. (2012) used the following method (inspired from Pinillos et al., 2011) in order to promote reflection in their participants: they had them answer all three questions of Fig. 2.
Frederick’s Cognitive Reflection Test (Frederick, 2005) before assessing a series of three sacrificial dilemmas. Participants in the control group had to answer the CRT after considering the three sacrificial dilemmas. After excluding participants who gave no correct answer to CRT (to ensure that the remaining participants actually reflected on the questions), Paxton, Ungar, and Greene observed that participants who received the CRT before considering the moral dilemmas gave more utilitarian answers than participants who received the CRT after.

Several studies have used the CRT to prime a counter-intuitive mindset—that is, a temporary disposition not to trust one’s gut feelings (see Paxton et al., 2013; Pinillos et al., 2011; Shenhav et al., 2012). However, one can criticize Paxton, Ungar, and Greene’s implementation of this idea on two grounds. First, their experimental condition did not exactly match their control condition: participants faced reasoning problems before answering moral dilemmas in the former but not in the latter condition. Thus, it is hard to conclude that the rise in utilitarian judgments is due to the induction of a counter-intuitive mindset rather than, say, a disposition to treat sacrificial dilemmas as arithmetic problems instead of moral ones. After all, Kleber et al. (2013) observed that highly numerated individuals were more likely to take into account utilitarian considerations when making hypothetical donations. And Byrd and Conway (2019) observed that, in the case of traditional sacrificial dilemmas, utilitarian judgments positively correlated with arithmetic measures of reflection (CRT) but not with non-arithmetic measures of reflection (Belief Bias). Second, their method was wasteful, as they were forced to exclude a lot of participants (around 40%) to make sure that the remaining ones actually reflected on the CRT problems.

We tried to address these shortcomings by using a related but different method: we compared two groups of participants who each received three reasoning problems before answering sacrificial dilemmas. Participants in the hard group received three counter-intuitive CRT problems, while participants in the easy group received three problems for which the intuitive answer was the right one. Moreover, participants who gave a wrong answer to one problem were notified and could not move to the next problem before finding the right answer. This was meant to ensure that participants actually took time to reflect on our problems.

We preregistered the whole study on OSF. Preregistration can be found at https://osf.io/4r9w6.

### 7.1. Method

#### 7.1.1. Rationale for sample size

To compute the sample size needed to have a reasonable chance to detect an effect, we took the effect size in Paxton et al. (2012)’s first study as a starting point (\(d = 0.43\)). We computed that reaching a .95 power would require 142 participants per group (total: 284) after exclusion. Based on the exclusion rate reported by Paxton et al. (2012), we estimated that a total of 463 participants would be needed. We rounded that up to 500.

#### 7.1.2. Procedure

Participants were asked to fill an online questionnaire, which was structured in the following order: (1) consent form, (2) reasoning problems, (3) vignettes, and (4) demographic questions (age, gender, education, religiosity, and political orientation).

**Reasoning problems.** After filling the consent form, participants were presented with three reasoning problems. The set of problems differed according to the condition (Easy or Hard). Participants in the Hard condition were presented with CRT-type problems in which participants tend to be drawn towards an intuitive but incorrect answer. However, we did not use Frederick (2005)’s traditional problems because we feared that participants recruited through Mechanical Turk would be too familiar with them (Stieger and Reips, 2016). Instead, we used a modified, less familiar version of the CRT, drawing on new CRT items designed by Finucane and Gullion (2010). Participants in the Easy condition received a modified version of the three CRT items. Items were modified so that the intuitive answer was the right one (see example in Table 5). We supposed that presenting our
lead them to stay in the intuitive mindset and would not elicit reflection. Participants with problems in which the intuitive answer is right would move to the next question. If participants selected the wrong answer, they were redirected to another page in which they received one of the following two messages:

- **Easy condition**: “This was not the right answer. Be careful!” Just trust your intuitions. The most obvious answer is sometimes the right one. Try again to find the right answer.
- **Hard condition**: “This was not the right answer. Be careful! Sometimes, our intuitions can deceive us, and the obvious answer is not necessarily the right one. Try again to find the right answer.”

After that, they were presented one more time with the problem they just failed and had to find the right answer to move to the next page. This procedure allowed us to ensure that participants who failed the CRT questions actually took the time to reflect upon them and were encouraged to adopt a more reflective mindset.

**Moral vignettes**. In this step (and the following), the UD and PD vignettes were merged to constitute a single category of scenarios (D = Dilemmas, leaving us with five types of vignettes: D = Dilemmas, HC = Harmless crimes, AO = Action/Omission, DE = Demanding ethics, P = Punishment). For the D category, we used the three vignettes originally used in Paxton et al. (2012) to maximize our chances of reproducing their results. For all four other categories, the vignettes were the same as in Studies 1 and 2. Two vignettes were selected at random from each category, meaning that each participant was presented with a total of 5×2 = 10 moral vignettes. After each vignette, participants had to answer the moral question on a seven-point scale (1 = “Not wrong at all”, 7 = “Extremely wrong”). We used fewer vignettes than in Studies 1 and 2 for fear that the effect of our priming manipulation would wane over time.

5 Nick Byrd rightfully pointed out that the formulation “Be careful!” might have in fact contributed to elicit a reflective mindset in the *Easy condition*, which clearly was not our goal. Fortunately, this message was only displayed to participants who failed a question. To make sure that exposure to this message did not interfere with our results, we re-ran the analyses presented in section 7.2.3 and 7.2.4 while only including participants who succeeded in all three questions. We did not find any effect of our condition (Easy vs. Hard) on utilitarian judgment.

### Table 5
Example of a pair of reasoning problems (*Easy* version on the left, *Hard* version on the right) (Study 3).

| Easy                                      | Hard                                      |
|-------------------------------------------|-------------------------------------------|
| If it takes one nurse 2 minutes to measure the blood pressure of 2 patients, how long would it take this one nurse to measure the blood pressure of 200 patients? ___ minutes | If it takes 2 nurses 2 minutes to measure the blood pressure of 2 patients, how long would it take 200 nurses to measure the blood pressure of 200 patients? ___ minutes |
| Choose the number corresponding to the right answer: ○ 2 ○ 20 ○ 200 ○ 2000 ○ 1 ○ 10 ○ 100 ○ 1000 | Choose the number corresponding to the right answer: ○ 2 ○ 20 ○ 200 ○ 2000 ○ 1 ○ 10 ○ 100 ○ 1000 |

participants with problems in which the intuitive answer is right would lead them to stay in the intuitive mindset and would not elicit reflection. For each problem, participants were presented with eight solutions and had to select one. If participants selected the right answer, they were redirected to another page in which they received one of the following two messages:

- **Easy condition**: “This was not the right answer. Be careful!” Just trust your intuitions. The most obvious answer is sometimes the right one. Try again to find the right answer.
- **Hard condition**: “This was not the right answer. Be careful! Sometimes, our intuitions can deceive us, and the obvious answer is not necessarily the right one. Try again to find the right answer.”

After that, they were presented one more time with the problem they just failed and had to find the right answer to move to the next page. This procedure allowed us to ensure that participants who failed the CRT questions actually took the time to reflect upon them and were encouraged to adopt a more reflective mindset.

**Moral vignettes**. In this step (and the following), the UD and PD vignettes were merged to constitute a single category of scenarios (D = Dilemmas, leaving us with five types of vignettes: D = Dilemmas, HC = Harmless crimes, AO = Action/Omission, DE = Demanding ethics, P = Punishment). For the D category, we used the three vignettes originally used in Paxton et al. (2012) to maximize our chances of reproducing their results. For all four other categories, the vignettes were the same as in Studies 1 and 2. Two vignettes were selected at random from each category, meaning that each participant was presented with a total of 5×2 = 10 moral vignettes. After each vignette, participants had to answer the moral question on a seven-point scale (1 = “Not wrong at all”, 7 = “Extremely wrong”). We used fewer vignettes than in Studies 1 and 2 for fear that the effect of our priming manipulation would wane over time.

### Table 6
Percentage of correct answers for each reasoning problem (Study 3).

| Easy (N = 253) | Hard (N = 244) |
|----------------|----------------|
| Q1             | 75.9%          | 70.9%          |
| Q2             | 94.9%          | 61.9%          |
| Q3             | 92.9%          | 59.4%          |

### 7.2. Results

#### 7.2.1. Participants
Participants were recruited through Amazon Mechanical Turk (United States residents only, number of HITs completed >50, success rate >95) and paid $1.00 for their participation. Four hundred ninety-nine participants completed the survey. Of those participants, two were excluded for refusing that their data be used for research and/or teaching purposes once the study’s aim was revealed at the end of the experiment. We were then left with 497 participants. Of those, 269 identified as men, 227 as women, and one as ‘other’. The mean age was 38.81 (SD = 11.52). There were 253 participants in the *Easy* condition and 244 in the *Hard* condition.

#### 7.2.2. Performance to reasoning task
The mean number of correct answers to the reasoning task was 2.64 in the *Easy* condition (SD = 0.64) and 1.92 in the *Hard* condition (SD = 1.09). There was a significant difference between the two conditions: t (388.77) = 8.86, p < .001, d = 0.8. This shows that the easy version of our reasoning problems was indeed easier than the hard version (see Table 6). In the *Easy* condition, one participant received a score of 0, 19 participants a score of 1, 51 participants a score of 2, and 182 participants a score of 3. In the *Hard* condition, 39 participants received a score of 0, 38 participants a score of 1, 70 participants a score of 2, and 97 participants a score of 3.

#### 7.2.3. Utilitarian answers (without exclusion)
Utilitarian scores were computed by averaging participants’ answers for each type of vignette. For D, HC, AO, and P vignettes, participants’ answers were reverse-scored so that higher scores indicate more utilitarian answers. The mean and standard deviations can be found in Table 7, while median and distribution can be found in Fig. 4.

We performed five Welch t-tests to compare participants’ moral judgments between the *Easy* and *Hard* condition for each of the five types of moral vignettes. The results for the five comparisons are presented in Table 7.

#### 7.2.4. Utilitarian answers (after exclusion)
As mentioned earlier, Paxton et al. (2012) excluded from their analyses participants who did not give at least one correct answer. To check whether any difference between our results and theirs could be due to this analysis choice, we reanalyzed our data after excluding participants who did not give at least one correct answer to the reasoning problems (this was planned in our registration). We performed five Welch t-tests to compare participants’ moral judgments between the *Easy* and *Hard* condition for each of the five types of moral vignettes. The results for the five comparisons are presented in Table 8.

#### 7.2.5. Correlations
Finally, we computed correlations between utilitarian scores and number of right answers to the easy and hard reasoning problems. Indeed, previous studies found a positive correlation between CRT scores and utilitarian judgment, and we wanted to know if the same effect could be found in our data. Though these analyses were not directly related to our main hypotheses, we still included them in our preregistration as exploratory analyses. The results are displayed in Table 9.
7.3. Discussion

Overall, we found no effect of our manipulation on utilitarian judgments. From these results, one might be tempted to conclude that priming a counter-intuitive mindset does not make people more utilitarian, no matter the type of moral vignette. However, such a conclusion would be premature. Indeed, it might simply be that we did not succeed in priming a more counter-intuitive mindset. We have no manipulation check that would help us decide between the two interpretations of our results.

This is further compounded by the fact that our experimental design deviates in many respects from Paxton, Ungar, and Greene’s original design. For example, it might be that solving our “easy” problems required enough effort from participants to elicit a counter-intuitive mindset, deleting all differences between the two conditions. Since, to our knowledge, no other study has actually used a similar design, we don’t know if ours worked.

To correct for these shortcomings, we decided to run another study with (i) a more traditional experimental design, and (ii) manipulation checks.

Table 7
Mean utilitarian scores (and standard deviations) in function of condition (Easy vs. Hard) and type of vignettes. The rightmost column indicates the results of the five Welch t-tests comparing utilitarian scores between the Easy and Hard conditions (Study 3).

|       | Easy (N = 253) | Hard (N = 244) | Welch t-test |
|-------|----------------|----------------|--------------|
| D     | 3.84 (1.58)    | 3.77 (1.67)    | t(491.1) = 0.503, p = 0.62, d = −0.05 |
| HC    | 3.72 (1.86)    | 3.53 (1.75)    | t(494.7) = 1.217, p = 0.22, d = −0.11 |
| AO    | 4.06 (1.70)    | 4.05 (1.64)    | t(494.6) = 0.966, p = 0.35, d = 0.09 |
| DE    | 3.33 (1.67)    | 3.10 (1.64)    | t(494.9) = 1.534, p = 0.13, d = −0.14 |
| P     | 3.88 (1.52)    | 3.75 (1.54)    | t(493.93) = 0.943, p = 0.35, d = 0.08 |

Table 8
Mean utilitarian scores (and standard deviations) in function of condition (Easy vs. Hard) and type of vignettes, after excluding participants who did not get at least one reasoning problem right. The rightmost column indicates the results of the five Welch t-tests comparing utilitarian scores between the Easy and Hard conditions (Study 5).

|       | Easy (N = 252) | Hard (N = 205) | Welch t-test |
|-------|----------------|----------------|--------------|
| D     | 3.85 (1.58)    | 3.81 (1.68)    | t(423.9) = 0.279, p = 0.78, d = −0.03 |
| HC    | 3.72 (1.86)    | 3.62 (1.77)    | t(443.4) = 0.539, p = 0.59, d = −0.05 |
| AO    | 4.06 (1.70)    | 4.11 (1.72)    | t(434.8) = 0.303, p = 0.76, d = 0.03 |
| DE    | 3.32 (1.67)    | 3.06 (1.61)    | t(442.1) = 1.701, p = 0.09, d = −0.16 |
| P     | 3.89 (1.51)    | 3.71 (1.55)    | t(431.75) = 1.236, p = 0.22, d = −0.12 |

Table 9
Correlations between utilitarian scores and number of correct answers to each set of reasoning problems (Easy and Hard).

|       | Easy (N = 253) | Hard (N = 244) |
|-------|----------------|----------------|
| D     | r = 0.05, p = 0.45 | r = 0.09, p = 0.16 |
| HC    | r = 0.05, p = .42  | r = 0.30, p < .001 |
| AO    | r = 0.16, p = .37  | r = 0.20, p = .002 |
| DE    | r = −0.25, p < .001 | r = −0.19, p = 0.002 |
| P     | r = 0.15, p = 0.02 | r = −0.01, p = 0.83 |

8. Study 4—investigating the effect of prior exposure to the cognitive reflection test on utilitarian judgment (second attempt)

In Study 3, we tried to prime a counter-intuitive mindset in our participants but failed to observe any effect on their moral judgments. We thus decided to conduct a second study, with two major modifications. First, we decided to use an experimental design closer to the one originally used by Paxton et al. (2012). Rather than presenting participants with the CRT either before or after answering the moral vignettes. Second, because we were worried that our manipulation might be ineffective, we introduced a potential manipulation check:

Fig. 4. Boxplot of utilitarian scores (scales) in function of condition (Easy vs. Hard) for each category of scenario (Study 3).
Belief Bias probes.

The whole study was preregistered on OSF. Preregistration can be found at https://osf.io/53rcd.

8.1. Method

8.1.1. Rationale for sample size

Following the same rationale as in Study 1, we aimed for a total of 500 participants, recruited through Amazon Mechanical Turk.

8.1.2. Procedure

Participants were asked to fill an online questionnaire. After filling a first consent form, they were randomly assigned to one of the following two conditions: Before or After. The two conditions differed with respect to the order in which participants were presented with the different parts of the questionnaire. Participants in the Before condition answered the CRT first, before answering the moral vignettes and then the Belief Bias probes. Participants in the After condition answered the moral vignettes first, then the Belief Bias probes, and finally the CRT. Following Paxton, Ungar, and Greene, we hypothesized that presenting the CRT first (vs. last) would put participants in a more counter-intuitive mindset, thus making their answers more utilitarian.

CRT. Contrary to what we did in Study 1, we used the three traditional CRT probes designed by Frederick (2005). This was done to stay as close as possible to Paxton et al. (2012)’s original design. Manipulation check—Belief Bias probes. One reason why we failed to observe any effect of our manipulation on participants’ moral judgments in Study 3 might simply be that our manipulation did not work and that we failed to induce a more counter-intuitive mindset in participants. This possibility makes it hard to interpret our results: either inducing a counter-intuitive mindset has no effect on moral judgment, or we failed to induce such a mindset. To tease apart these interpretations, we needed a way of checking that our manipulation does indeed work. Unfortunately, previous studies using the same experimental paradigm did not consider it necessary to include such a manipulation check (Paxton et al., 2012; Paxton et al., 2013; Pinillos et al., 2011; Shenhav et al., 2012). And it is not clear what would count as an appropriate one.

Still, we decided to include Belief Bias probes as a manipulation check. Belief bias probes are simple syllogisms (two premises, one conclusion) that are either valid but leading to a false conclusion, or invalid but leading to a true conclusion (Markovits and Nantel, 1989). Here is an example of the former case:

Suppose that:

(1) All mammals walk.

(2) Whales are mammals.

If these two statements are true, can we conclude from them that “whales walk”?

Belief Bias probes present a tension between an easy, intuitive and attractive answer (accepting or rejecting the syllogism based on the truth or falsity of its conclusion) and a more reflective one (assessing the syllogism based on its logical validity, while ignoring the truth or falsity of its conclusion). As a result, they presumably require a more counter-intuitive mindset to be solved. One would thus expect participants in a more counter-intuitive mindset to perform better at this task. Moreover, past studies have shown that good performance on Belief Bias probes is correlated with more utilitarian judgment (Baron et al., 2015). More recent studies even suggest that the link between CRT and utilitarian judgment is mediated by scores on Belief Bias probes (Byrd and Conway, 2019).

For these reasons, we decided to use participants’ performance to the Belief Bias probes as a manipulation check: participants in a counter-intuitive mindset should plausible score higher on this task. In total, we used four different probes: three from Baron et al. (2015) and one from De Neys and Franssens (2009).

Moral vignettes Moral vignettes were the same as in Study 3. Participants read a total of 10 vignettes (two for each type of vignette).

8.2. Results

8.2.1. Participants

Participants were recruited through Amazon Mechanical Turk (United States residents only, number of HITs completed >50, success rate >95) and paid $1.00 for their participation. Five hundred participants completed the survey. Of those, 12 participants were excluded for refusing that their data be used for research and/or teaching purposes once the aim of the study was revealed at the end of the experiment. We were then left with 488 participants. Of those, 269 identified as men, 218 as women, and one as ‘other’. The mean age was 36.85 (SD = 11.26). There were 248 participants in the Before condition and 240 in the After condition.

8.2.2. CRT scores

The mean number of correct answers to the CRT was 1.86 in the Before condition (SD = 1.20) and 1.84 in the After condition (SD = 1.23). There was no significant difference between the two conditions: t(484.1) = 0.19, p = .85, d = 0.01. Overall, 111 participants received a score of 0, 67 a score of 1, 93 a score of 2, and 217 a score of 3. Following Paxton et al. (2012), we excluded participants with a score of 0 from subsequent analyses, which left us with 193 participants in the Before condition and 184 in the After condition.

8.2.3. Manipulation check (Belief Bias probes)

Before exclusion, the correlation between CRT scores and numbers of correct answers to the Belief Bias probes was r = .44, showing a medium relationship between performance at both tasks. After exclusion, the mean number of correct answers to the Belief Bias probes was 2.42 in the Before condition (SD = 1.35) and 2.62 in the After condition (SD = 1.37). That the number of correct answers was higher in the After condition suggests that presenting the CRT before the Belief Bias probes did not lead participants to adopt a more counter-intuitive mindset.

8.2.4. Utilitarian answers

Mean and standard deviations for participants’ scores are presented in Table 10, while median and distribution are presented in Fig. 5. We performed five Welch t-tests to compare participants’ moral judgments between the Before and After conditions for each of the five types of moral vignettes. For D, HC, AO, and P vignettes, participants’ answers were reverse-scored so that higher scores indicate more utilitarian answers. The results for the five comparisons are presented in Table 10.

8.2.5. Correlations

Finally, we computed correlations between utilitarian scores for each type of vignettes and scores to the CRT and the Belief Bias probes. Participants with a CRT score of 0 were included in this analysis, the results of which are presented in Table 11. Though these analyses were not directly related to our main hypothesis, we included them in our preregistration as exploratory analyses.

Table 10

|                   | Before (N = 193) | After (N = 184) | Welch t-test |
|-------------------|-----------------|----------------|-------------|
| D                 | 4.05 (1.65)     | 4.10 (1.85)    | t(365.1) = 0.298, p = 0.77, d = 0.03 |
| HC                | 3.91 (1.95)     | 3.87 (1.83)    | t(374.9) = 0.191, p = 0.85, d = 0.02 |
| AO                | 4.13 (1.61)     | 4.10 (1.74)    | t(369.2) = 0.199, p = 0.84, d = 0.02 |
| DE                | 3.10 (1.61)     | 3.14 (1.62)    | t(373.9) = 0.289, p = 0.77, d = 0.03 |
| P                 | 3.90 (1.45)     | 3.74 (1.49)    | t(372.8) = 1.022, p = 0.31, d = 0.11 |
8.3. Discussion

Once again, we failed to observe any significant effect of our experimental manipulation on moral judgment. However, this might simply be because we failed to induce a more counter-intuitive mindset in our participants. Indeed, scores to the Belief Bias probes were unaffected by the manipulation, and presenting the CRT first did not improve performance. But, if our manipulation was indeed ineffective, we cannot draw solid conclusions about the impact of inducing a counter-intuitive mindset on utilitarian judgment.

Of course, it might be that Belief Bias probes are not the best manipulation check for this type of manipulation. Still, they are typically considered as “logical measures of reflection” in which obviously true (or false) conclusions tend to lure participants away from considering the actual validity of an argument (Byrd and Conway, 2019). Moreover, scores on the Belief Bias task typically correlate with scores on the CRT (as was the case in our study, but also in other studies such as Byrd and Conway, 2019; Fuhrer and Cova, 2020). It is thus reasonable to expect that participants in a more counter-intuitive mindset, and thus more suspicious of their intuitive answers, should perform better on the Belief Bias task.

One reason our manipulation was ineffective might have been that exposure to CRT problems is more likely to prime arithmetic thinking rather than a more counter-intuitive mindset. Indeed, there is growing evidence that mathematical reflection test performance is indistinguishable from general math test performance (Attali & Bar-hillel, 2020; Erceg et al., 2020).

These considerations motivated us to find another way to prime a counter-intuitive mindset in our participants. In Study 5, we directly instructed participants to rely on reason rather than intuition.

9. Study 5—investigating the effect of instructing participants to rely on reason vs. intuition on utilitarian judgment

In Studies 3 and 4, we tried to prime a more counter-intuitive mindset in participants by having them solve counter-intuitive reasoning problems, but we observed no effect of this manipulation on utilitarian judgments, and it was unclear that the manipulation succeeded. In this fifth study, we therefore decided to use a completely different way of inducing a counter-intuitive mindset, by directly asking participants to answer moral questions based on reason (rather than based on intuition). Similar methods were previously used in the literature and were reported to have a significant impact on moral judgments in the case of sacrificial dilemmas (Capraro et al., 2019; Suter and Hertwig, 2011).

The whole study was preregistered on OSF. Preregistration can be found at osf.io/2gdyr.

9.1. Method

9.1.1. Rationale for sample size

Our methods were inspired by Study 3 in Capraro et al. (2019). Consequently, we decided to use a comparable sample size. Earp and colleagues used an average of 230 participants per condition. Hence, given that we had three different conditions, we decided to recruit a total of 690 participants.

9.1.2. Procedure

Participants were asked to fill an online questionnaire. After filling a consent form, they were randomly assigned to one of the following three conditions: Intuition, Control, and Reason. The conditions differed with respect to the instruction given to participants. Participants in the Control condition received no specific instruction, while participants in
the Intuition condition were instructed to answer the questionnaire based on their intuitions, and participants in the Reason condition were instructed to answer based on their reason (see Table 12). Instructions are drawn from Capraro et al. (2019), except that we replaced the original instruction to “answer based on emotion” by a more general instruction to answer “based on intuition.”

After reading the instruction, participants in the Intuition and Reason conditions were asked a control question about the instruction they just received. They could not continue until they had correctly answered the question. Finally, the instruction to answer the questionnaire based on intuition or reason was repeated on each page, at the end of each question (see Table 12 for an example).

Moral vignettes. Then, participants answered a total of 10 moral vignettes (two for each type of vignettes). The vignettes were the same as in Capraro et al. (2019). Instructions within the questionnaire (example)

| Instructions at the beginning of the questionnaire | Intuition | Reason |
|-----------------------------------------------------|-----------|--------|
| Sometimes people make decisions by using feelings and relying on their intuition. Other times, people make decisions by using logic and relying on their reason. Many people believe that intuition leads to good decision-making. When we use feelings rather than logic, we make better decisions. Please answer the 14 following questions by relying on intuition rather than reason. How morally wrong was it for Denise to push the hiker in order to save the five workmen? (Rely on intuition) | Sometimes people make decisions by using logic and relying on their reason. Other times, people make decisions by using feelings and relying on their intuition. Many people believe that reason leads to good decision-making. When we use logic rather than feelings, we make better decisions. Please answer the 14 following questions by relying on reason rather than intuition. How morally wrong was it for Denise to push the hiker in order to save the five workmen? (Rely on reason) |

Table 12
Instructions at the beginning of the questionnaire and within the questionnaire for the Intuition and Reason conditions (Study 4).

9.2. Results

9.2.1. Participants

Participants were recruited through Amazon Mechanical Turk (United States residents only, number of HITs completed >50, success rate >95) and paid $1.00 for their participation. Six hundred eighty-nine participants completed the survey. Of those, 358 identified as men, 327 as women, and four as ‘other’. The mean age was 38.58 (SD = 11.15). There were 233 participants in the Intuition condition, 228 in the Control condition, and 228 in the Reason condition.

9.2.2. Belief Bias probes (Manipulation Check)

First, we examined the effect of our manipulation on scores to the Belief Bias probes. An ANOVA revealed a significant effect of our manipulation (see Table 13). Post-hoc Welch t-tests revealed a significant difference between the Intuition and Reason conditions: t(458.99) = 2.45, p = 0.01, d = 0.23, a significant difference between the Intuition and Control conditions: t(457.08) = 2.50, p = 0.01, d = 0.23, but no significant difference between the Control and Reason conditions: t(451.77) = 0.14, p = 0.89, d = 0.01. This means that asking participants to answer intuitively tended to decrease their performance at the Belief Bias task, whereas asking them to answer based on reason did not increase it. This suggests that we were successful in priming a more intuitive mindset (in the Intuition condition) but not in priming a more counter-intuitive mindset (in the Reason condition). Nevertheless, excluding participants who answered ‘NO’ to the second sincerity check question turned this effect non-significant (see the bottom row in Table 13).6

9.2.3. Utilitarian answers

Utilitarian scores were computed by averaging their answer within each category and reversing them for the D, HC, AO, and P categories. Mean and standard deviations are presented in Table 14, while median and distribution are presented in Fig. 6.

We performed five different ANOVAs with CONDITION as a between-subject factor and utilitarian scores as a dependent variable, one for each of the five types of moral vignettes. Results for the five comparisons are presented in Table 15. We found significant effects (all in the predicted direction) for the D, HC, and DE category. In the D category, post-hoc Welch t-tests revealed a significant difference between the Intuition and Reason conditions: t(458.95) = 5.24, p < .001, d = 0.49; a significant difference between the Control and Reason conditions: t(452.79) = 6.79, p < .001, d = 0.64; but no significant difference between the Intuition and Control conditions: t(458.2) = 1.41, p = 0.16, d = –0.13. In the HC category, we found a significant difference between the Intuition and Reason conditions: t(455.61) = 4.74, p < .001, d = 0.44; a significant difference between the Control and Reason conditions: t(454.61) = 3.83, p < .001, d = 0.35.

Table 13
Means (and standard deviations) for the number of right answers to the Belief Bias task in function of condition and whether participants who failed the second sincerity check were excluded (Study 5).

|                   | Intuition | Control | Reason | Anova |
|-------------------|-----------|---------|--------|-------|
| Before exclusion  | 2.26      | 2.26    | F(2,686) = 4.11, p = .02, |       |
|                   | (1.32)    | (1.32)  | \(\eta^2_p = 0.012\) |       |
|                   | N = 233   | N = 228 | N = 228 |       |
| After exclusion   | 2.22      | 2.22    | F(2,596) = 2.30, p = .10, |       |
|                   | (1.38)    | (1.38)  | \(\eta^2_p = 0.008\) |       |
|                   | N = 202   | N = 210 | N = 187 |       |

6 In our OSF preregistration, we made a mistake and announced that this second analysis would be performed after excluding participants who answered ‘NO’ to the first sincerity check.

---

The table and instructions are designed to reflect the natural flow of the content, ensuring clarity and coherence in the document's representation. The sections on participants, belief bias probes, and utilitarian answers are structured to facilitate understanding, with relevant tables and statistical summaries for easy reference.
Table 14
Mean utilitarian scores (and standard deviations) in function of condition (Intuition, Control, or Reason) and type of vignettes. The rightmost column indicates the results of the five ANOVAs examining the effect of our manipulation on utilitarian scores (Study 5).

| Type of vignettes | Intuition (N = 233) | Control (N = 228) | Reason (N = 228) | Anova |
|------------------|---------------------|-------------------|------------------|--------|
| D                | 4.00 (1.67)         | 3.79 (1.57)       | 4.81 (1.65)      | F(2, 686) = 25.08, p < .001, η² = 0.068 |
| HC               | 3.61 (1.73)         | 3.35 (1.81)       | 4.40 (1.85)      | F(2, 686) = 21.00, p < .001, η² = 0.058 |
| AO               | 3.80 (1.57)         | 4.05 (1.68)       | 4.69 (1.76)      | F(2, 686) = 1.52, p = .22, η² = 0.004 |
| DE               | 3.26 (1.59)         | 3.31 (1.70)       | 4.14 (1.64)      | F(2, 686) = 4.35, p = .01, η² = 0.013 |
| P                | 3.96 (1.63)         | 3.80 (1.52)       | 4.14 (1.64)      | F(2, 686) = 2.51, p = .08, η² = 0.007 |

Table 15
Percentage of YES answers to the two sincerity checks, depending on condition (Study 5).

| Condition | YES answers (Study 5) | χ² test |
|-----------|-----------------------|---------|
| (1) Intuition | 7.73%                 | χ² = 32.84, p < 0.001 |
| (2) Control | 5.26%                 |         |
| (3) Reason | 21.05%                |         |
| χ² test   | 13.30%                | χ² = 10.234, p = 0.006 |
| Type of scenarios | χ² test |         |
| D          | 7.73%                 |         |
| HC         | 5.26%                 |         |
| AO         | 17.98%                |         |
| DE         | 0.25%                 |         |
| P          | 0.32%                 |         |

Fig. 6. Boxplot of utilitarian scores (scales) in function of condition (Intuition, Control or Reason) for each category of scenario (Study 5).

In our OSF preregistration, we made a mistake and announced that this second analysis would be performed after excluding participants who answered “NO” to the second sincerity check.

9.2.4. Post-hoc analysis: participants’ completion time
In this study, we found an effect of our manipulation on utilitarian judgments, with judgments in the Reason condition being different from those in the Intuition and Control conditions. However, regarding scores to our manipulation check (the Belief Bias task), we found that scores in the Intuition condition were significantly different from scores in the Control and Reason conditions, but we found no significant difference between the Control and Reason conditions. Thus, it seems that the effect of our manipulation on utilitarian judgment cannot be explained by the fact that the Reason condition succeeded in inducing a more counter-
intuitive mindset.

An alternative explanation might be that participants in the Reason condition took more time to reflect on the moral vignettes before answering. To test this hypothesis, we analyzed the time people took to fill the survey. To test this hypothesis, we analyzed the time people took to answer. To test this hypothesis, we analyzed the time people took to fill the survey.

### Table 16

Mean utilitarian scores (and standard deviations) in function of condition (Intuition, Control, or Reason) and type of vignettes after exclusion of participants who answered YES to the first sincerity probe (Study 5).

|       | Intuition (N = 215) | Control (N = 216) | Reason (N = 180) | Anova |
|-------|---------------------|-------------------|------------------|-------|
| D     | 4.05 (1.66)         | 3.84 (1.56)       | 4.73 (1.62)      | F(2, 608) = 15.85, p < .001, η²_p = 0.050 |
| HC    | 3.64 (1.76)         | 3.40 (1.83)       | 4.31 (1.85)      | F(2, 608) = 13.12, p < .001, η²_p = 0.041 |
| AO    | 3.85 (1.59)         | 3.87 (1.66)       | 3.96 (1.66)      | F(2, 608) = 0.24, p = .79, η²_p = 0.0008 |
| DE    | 3.19 (1.56)         | 3.20 (1.64)       | 3.67 (1.76)      | F(2, 608) = 5.34, p = .005, η²_p = 0.017 |
| P     | 3.98 (1.62)         | 3.86 (1.51)       | 4.15 (1.62)      | F(2, 608) = 1.67, p = .19, η²_p = 0.005 |

Note that this analysis was not planned in the registration.

### 9.3. Discussion

In this study, we observed an effect of our manipulation on three categories of scenarios (D, HC, and DE) in the expected direction. Moreover, the effect was carried by the Reason condition (while there was no significant difference between the Intuition and Control conditions). However, scores to our manipulation check (the Belief Bias probe) did not differ between the Control and Reason conditions—rather, they differed between the Intuition and Control conditions. Post-hoc analyses suggested that differences in utilitarian judgments between the Control and Reason conditions could be due to the fact that participants took more time to reflect on their answers.

### 10. Meta-analysis of our results

Through five studies, we investigated the role of reasoning in utilitarian judgment by trying to hinder or promote reflection. To synthesize our results, we decided to conduct a series of mini meta-analyses on our results. This entailed a certain number of analytic decisions: (i) for Studies 1 and 2, we combined results for the UD and PD categories into a single D category, given that both categories were no longer separated in Studies 3 to 5; (ii) for Studies 1 and 2, we decided to focus on scalar scores rather than binary ones, because most of our other measures were already scalar; (iv) for Study 5, we decided to focus on the contrast between the Control and Reason conditions, leaving out the Intuition condition, because most of our effect was due to the contrast between these two conditions. Moreover, given that we used two different types of samples (student population in Studies 1 and 2, MTurk workers in Studies 3 to 5), we decided (v) to use a random-effects approach.

But the most crucial question was (vi) whether we should include Studies 3 and 4 in our meta-analysis. Indeed, as we saw, there are strong reasons to think that the manipulation we used in these studies simply does not work. By including them, we might thus have run the risk of underestimating the impact of inhibiting or promoting reasoning on utilitarian judgment. In the end, we decided to run our meta-analysis twice, with and without Studies 3 and 4, but to focus on the results obtained when excluding them to interpret our findings (see General Discussion).

The results of our meta-analyses can be found in Table 18a. Additionally, we also performed a second round of meta-analyses on the correlational results obtained in Studies 3 to 5. The results of this second round can be found in Table 18b.

### 11. General discussion

#### 11.1. Summary and limitations of our results

In this series of studies, our goal was to investigate the role of reflection in the formation of utilitarian judgments and determine whether Joshua Greene’s dual-process model of moral cognition could be generalized beyond thought experiments involving sacrificial dilemmas.

Putting aside Studies 3 and 4, the meta-analysis of our results identified three significant effects: a medium effect size for Harmless Crimes, a small effect size for Demanding Ethics, and a negligible effect size for Punishment. Results for Harmless Crimes were particularly impressive: in Studies 1 and 2, our manipulation led the rate of utilitarian answers to fall from 58% to 39% and from 56% to 38%. Moreover, participants who obtained higher scores to the CRT and Belief Bias tasks were more likely to give utilitarian answers for Harmless Crimes. Harmless Crimes therefore seem to constitute an ideal candidate to extend Greene’s dual-process model, according to which reflection promotes utilitarian judgment.

The meta-analytic estimate of the effect size was smaller for Demanding Ethics and Punishment. Still, in the case of Demanding Ethics, the effect of manipulation was significant in all three studies (1, 2, and 5), and thus we have reasons to think that this effect is robust and worth investigating further. In the case of Punishment, the effect turned out significant in only one study (Study 1), even though the final meta-analytic estimate was significantly different from zero. Given that the effect size is negligible, it is not clear to what extent this can be
considered strong support for a generalization of Greene’s dual-process model of moral cognition.

Moreover, our results raise two questions for further research. The first comes from the fact that we failed to replicate the effect of cognitive load and time constraint on sacrificial dilemmas. Though we obtained a significant effect of manipulation in Study 5, the total meta-analytic estimate did not significantly differ from zero. This is surprising, to say the least, but we are not the first to fail to replicate this kind of effect. For example, Abatista et al. (2018; see also Cova et al., 2020) failed to find no effect of their manipulation on scores to the “impartial beneficence” subscale. Since we and they used the same manipulation, and our Demanding Ethics vignettes and their “impartial beneficence” subscale are supposed to measure the same kind of moral intuitions, this is very surprising. Further studies should investigate to which extent our scenarios and the “impartial beneficence” subscale really tap into the same intuitions and which method is the most adequate to study utilitarian thinking about duties of assistance.

11.2. In which sense is utilitarian judgment reflective?

Another question is how exactly we should interpret our results. As mentioned above, our results seem to support one of Greene’s dual-process model’s two main tenets: utilitarian judgments are preferentially supported by reflection. But what exactly is “reflection” supposed to mean in this hypothesis?

Greene’s dual-process model is premised on the classical opposition between “System 1” and “System 2”, which pits two very broad types of psychological processes against each other. On the one side (System 1), we have processes that are supposed to be associative, fast, automatically processed, and not consciously represented. On the other side (System 2), we have processes that are supposed to be non-associative, slow, deliberately initiated, and consciously represented. However, it is unclear that these different features should always come together (Byrd, 2020). For example, it has recently been argued that paradigmatic reflective reasoning (such as giving a correct, non-intuitive assessment of valid syllogisms reaching obviously false conclusions) could actually occur very quickly (Bago and De Neys, 2017) and without much deliberation (Szaszi et al., 2017), and that propositional, non-associative processes can be automatic and unconscious (Byrd, 2020; Mandelbaum, 2016). Therefore, it is crucial to understand in which sense of “reflective” psychological processes underlying utilitarian judgments are supposed to be more reflective than psychological processes underlying deontological judgments.

In his most recent formulation of his dual-process model, Greene (2014) puts forward three characteristics of the cognitive processes underlying utilitarian judgments: (i) they are conscious (as opposed to non-conscious), (ii) voluntary (as opposed to automatically triggered by cues in our environment), and (iii) often experienced as effortful (as opposed to cost-free). However, these are not exactly the characteristics we investigated in our studies. Instead, we explored whether, compared to deontological judgments, utilitarian judgments instantiate the following properties:

1) Higher cognitive cost. In Study 1, we investigated whether making utilitarian judgments requires more cognitive resources than making deontological judgments (i.e., working memory).
2) Being slower. In Study 2, we investigated whether being forced to wait before giving their answer leads participants to make more utilitarian judgments.
3) Being less intuitive (i.e., intellectually attractive). In Studies 3 to 5, we investigated whether eliciting a counter-intuitive mindset led to less deontological judgments and more utilitarian judgments.
The results of Study 1 led us to conclude that certain forms of utilitarian judgments require more cognitive resources than their deontological counterparts. The results of Studies 2 and 5 suggest that making utilitarian judgments tends to take more time than making deontological judgments. Finally, because our manipulation was unsuccessful, Studies 3 and 4 do not allow us to conclude anything about the intuitiveness of utilitarian judgments. In Study 5, we were nonetheless successful in manipulating participants’ reliance on their intuitions (in the Intuition condition), but this did not seem to have any effect on the rate of utilitarian judgments. Overall, our results allow us to conclude that utilitarian judgments tend to rely on slower and more cognitively demanding processes than deontological judgments. But they do not allow us to conclude anything about characteristics such as automaticity or accessibility to consciousness. For example, while solving a very difficult math problem might take more time and require more resources than solving a mildly difficult math problem, there is no reason to think that the former process is more automatic and conscious than the latter—though it will clearly be experienced as more effortful. Moreover, although we could be tempted to conclude that our data show that, as predicted by Greene, utilitarian judgment is experienced as more effortful than deontological judgment, we collected no data about such experience—only data about the role of cognitive resources in utilitarian judgment. While it is reasonable to think that a process that requires more resources will be experienced as more effortful, we should keep in mind that this conclusion rests on an inferential leap.

Accordingly, our results are compatible with the idea that deontological and utilitarian judgments are both “intuitive” in a certain sense (the sense of being “intellectually attractive,” as are lure answers in the CRT and Belief Bias task). To explain answers to reasoning tasks including lure answers, Bago and De Neys (2017; see also De Neys, 2012) have put forward a hybrid dual-process model of reasoning. In this model, several intuitive answers are triggered in parallel, causing a conflict that requires further reasoning to be solved. A similar model could be applied to the way participants approached our scenarios. It might be that, when our participants read the vignette, both the deontological and utilitarian answers spontaneously presented themselves to their minds. After all, the utilitarian solution is often salient enough in moral vignettes (for example, vignettes depicting sacrificial dilemmas often spell out in an explicit manner that more lives can be saved if one person is sacrificed). And they seem easily accessible in our other vignettes as well. Most participants seem to understand that these situations are dilemmas that pit two plausible solutions against each other. There would thus be a conflict between two intuitive (i.e., intellectually salient and attractive) answers. Slow and costly reasoning would then come in a second stage to arbitrate this conflict, with a tendency to arbitrate it in favor of the utilitarian judgment (a similar proposal can be found in Landy and Royzman, 2018).

In such a model, utilitarian judgment is still intuitive in the sense that it is an option that automatically presents itself to our mind. But it is also reflective in the sense that it is favored over deontological judgment on the basis of slow, effortful reasoning. Thus, it will be important in further research to better distinguish between different interpretations of the claim that utilitarian judgment is “reflective” or preferentially supported by reflective processes.

11.3. A dissociation between correlational and experimental evidence

Finally, one puzzling (and, in our opinion, very interesting) observation was the discrepancy between the conclusions that could be drawn from manipulation effects and those that could be drawn from correlation effects (see Table 1B). For the Action/Omission category, we found a significant correlation effect but no significant manipulation effects. Conversely, for the Punishment category, we found a significant manipulation effect but no significant correlation effect. But the most striking case is the Demanding Ethics category, for which manipulation and correlation effects point in opposite directions: in such cases, while manipulation effects suggest that more reflection leads to more utilitarian judgments, correlation effects suggest that more reflective participants make less utilitarian judgments. Note that this is no isolated case. We described a similar dissociation for the CNI model in the introduction (see section 3.1.4), and such disconnections between correlational and experimental effects have been found for religious belief. For example, Yilmaz & Ilser (2019) found that, though CRT scores negatively correlate with belief in God, forcing non-believers to answer quickly leads them to report higher belief in God.

How are we to explain this disconnection? There are at least two separate possibilities. First, a modest proposal might appeal to the distinction we drew in the previous section and suggest that correlational evidence (CRT and Belief Bias scores) probes whether certain answers are intuitive or intellectually attractive while cognitive load and time constraint experiments probe the role of slow and effortful reasoning. In the case of Demanding Ethics, this would mean that the “utilitarian” answer is intuitive but still needs slow and effortful deliberation to be favored. Though the resulting picture of the situation might seem strange enough, it might be plausible if we suppose that there are several psychological paths towards the conclusion that we have the duty to sacrifice a lot of resources for people in need: an intuitive one and a reflective one. As mentioned earlier, past research has shown that intuitive cognitive style was linked to higher religiosity, which in turn was linked to higher moral concern and respect of social norms (Morgan et al., 2018). Moreover, promoting intuition leads to higher rates of unconditional cooperation (Rand, 2016). Given these findings, it is no wonder that reflective cognitive style correlates negatively with the seemingly “utilitarian” answer. But, at the same time, philosophical arguments can lead people to be more generous and ethical (Lindauer et al., 2020; Schwitzgebel et al., 2020). We can thus expect careful and deliberate reasoning to increase cooperation too.

A second option is more ambitious: it is to conclude that correlational evidence is not a reliable indicator of the underlying psychological mechanism, and that correlations between measures of reflective cognitive style and moral judgments are best explained by other factors. This conclusion should nonetheless be qualified by the fact that correlational evidence and the results of manipulation experiments seem to converge on the same conclusion for other cases (such as Harmless Crimes). Here again, future research should keep in mind that both methods of investigation will not necessarily yield the same results.

Despite all these reservations, we believe that our results warrant the conclusion that slow and effortful reasoning favors utilitarian judgments outside the context of sacrificial dilemmas.

Author statement

Florian Cova: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review and Editing, Visualization, Supervision, Project administration, Funding acquisition

François Jaquet: Conceptualization, Methodology, Validation, Investigation, Writing – Original Draft, Writing – Review and Editing, Supervision, Funding acquisition

Acknowledgements

The research was funded by an incentive funding of the Swiss Center for Affective Sciences. During work on the paper, F.C. was supported by an Eccellenza Professoral Fellowship of the Swiss National Science Foundation (“Eudaimonic emotions and the (meta)philosophy of well-being”), and F. J. was supported by a Postdoc.Mobility Fellowship from the Swiss National Science Foundation (“Debunking Arguments in Animal Ethics”). Given that the present work has its roots in F.C.’s Master thesis, we would like to thank Pierre Jacob, Emmanuel Dupoux and Dan Speerber for their supervision. We are also thankful to the Master students whose work on data collection made this research possible: Lisa...
Esposito, Claudia Ferreira, Daria Grigoryeva and Laura Leon Perez. Finally, we express our gratitude to the ten anonymous philosophers who assessed our vignettes.

References

Abatista, A. G. F., Cova, F., & Jaquet, F. (2018). Replication of Greene et al. (2008) “Cognitive load selectively interferes with utilitarian moral judgment.” Retrieved from osf.io/yh38c.

Bago, B., & De Neys, W. (2017). Fast logic: Examining the time course assumption of dual process theory. Cognition, 158, 90–109.

Baron, J. (1994). Nonconsequentialist decisions. Behavioral and Brain Sciences, 17(1), 1–10.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. Organizational Behavior and Human Decision Processes, 92(4), 34–45.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? Journal of Applied Research in Memory and Cognition, 4, 265–284.

Barret, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. Cognition, 108(2), 381–417.

Barret, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. Cognition, 121(1), 154–161.

Byrd, N. (2020). What we can (and can’t) infer about implicit bias from debiasing experiments. Synthese (in press).

Byrd, N., & Conway, P. (2019). Not all who ponder count costs: Arithmetic reflection predicts utilitarian moral judgment (and other things). Journal of Applied Research in Memory and Cognition, 4, 265–284.

Carlimuth, K. M. (2006). The roles of retribution and utility in determining punishment. Journal of Experimental Social Psychology, 42(4), 437–451.

Carlimuth, K. M. (2008). On justifying punishment: The discrepancy between words and actions. Social Justice Research, 21(2), 119–137.

Carlimuth, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. Journal of Personality and Social Psychology, 83(2), 84–92.

Capraro, V., Everett, J. A., & Eber, B. D. (2019). Priming intuition disfavors instrumental harm but not impartial beneficence. Journal of Experimental Social Psychology, 83, 142–149.

Carlimuth, K. M. (2006). The roles of retribution and utility in determining punishment. Journal of Experimental Social Psychology, 42(4), 437–451.

Carlimuth, K. M. (2008). On justifying punishment: The discrepancy between words and actions. Social Justice Research, 21(2), 119–137.

Carlimuth, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. Journal of Personality and Social Psychology, 83(2), 84–92.

Ciaralli, E., Muccioli, M., Ladavas, E., & di Pellegrino, G. (2007). Selective deficit in normal moral judgment taking following damage to ventromedial prefrontal cortex. Social Cognition and Affective Neuroscience, 2(2), 84–92.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. Journal of Personality and Social Psychology, 104(2), 216–236.

Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. Cognition, 179, 241–265.

Cota, A., Fousart, A., Hayakawa, S., Aparici, M., Apetegata, J., Heafner, J., & Keyser, B. (2014). Your morals depend on language. PLoS One, 9(4), e94842.

Cova, F. (2011). L’architectura de la cognición moral doctoral dissertation. Institut Jean Nicod.

Cova, F. (2017). What Happened to the Trolley Problem? Journal of Indian Council of Philosophical Research, 34(3), 543–564.

Cova, F., & Jaquet, F. (2017). Les intuitivités utilitaristes sont-elles vraiment utilitaristes ? Raisons Publiques, 22(2), 21–28.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., … Zhou, F. (2020). Estimating the reproducibility of experimental philosophy. Review of Philosophy and Psychology (in press).

Crockett, M. J., Clark, V. L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm avoidance. Proceedings of the National Academy of Sciences, 107(40), 17433–17438.

Cushman, F., Murray, D., Gordon-McKeon, S., Wharton, S., & Greene, J. D. (2011). Judgment before principle: engagement of the frontopolar control network in condemning harms of omission. Social Cognition and Affective Neuroscience, 7(8), 888–895.

De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: Not always present but at least convenient. Cognition, 112(1), 45–61.

Everett, J. A., & Kahane, G. (2020). Switching tracks? Towards a multidimensional model of utilitarian psychology. Trends in Cognitive Sciences, 24(2), 124–134.

Fiore, M., & Gullion, C. M. (2010). Developing a tool for measuring the decision-making competence of older adults. Psychology and Aging, 25(2), 271–288.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. Oxford Review, 5–15.

Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19, 5–25.

Fuhrer, J., & Cova, F. (2020). “Quick and dirty”: Intuitive cognitive style predicts trust in Didier Raoult and his hydroxychloroquine-based treatment against COVID-19. Judgment and Decision Making. 15(5), 25–42.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. Social and Personality Psychology Compass, 10(10), 535–549.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. Journal of Personality and Social Psychology, 101(2), 366.
Nagel, J., & Waldmann, M. R. (2013). Deconfounding distance effects in judgments of moral obligation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(1), 337.

Nagel, J., & Waldmann, M. R. (2016). On having very long arms: How the availability of technological means affects moral cognition. *Thinking & Reasoning, 22*(2), 184–208. 
Nnorcuass, A. (2006). The scalar approach to utilitarianism. In H. West (Ed.), *The Blackwell Guide to Mill’s Utilitarianism* (pp. 217–232). Hoboken: Wiley-Blackwell.

Patil, I. (2015). Trait psychopathy and utilitarian moral judgement: The mediating role of action aversion. *Journal of Cognitive Psychology, 27*(3), 349–366.

Patil, I., & Silani, G. (2014). Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in Psychology, 5*, 501.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*(1), 163–177.

Paxton, J. M., Brunni, T., & Greene, J. D. (2013). Are “counter-intuitive” deontological judgments really counter-intuitive? An empirical reply to Kahane. *Social Cognitive and Affective Neuroscience, 9*(9), 1368–1371.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2014). The role of analytic thinking in moral judgements and values. *Thinking & Reasoning, 20*(2), 188–214.

Pinillos, N.A., Smith, N., Nair, G. S., Marchetto, P., & Mun, C. (2011). Philosophy’s new challenge: Experiments and intentional action. *Mind & Language, 26*(1), 115–139.

Rachels, J. (1975). Active and passive euthanasia. *New England Journal of Medicine, 292*, 70–72.

Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science, 27*(9), 1192–1206.

Rosas, A., & Aguilar-Pardo, D. (2020). Extreme time-pressure reveals utilitarian intuitions in sacrificial dilemmas. *Thinking & Reasoning* (in press).

Royzman, E., Atanasov, P., Landy, J. F., Parks, A., & Gepty, A. (2014b). CAD or MAD? Anger (not disgust) as the predominant response to pathogen-free violations of the divinity code. *Emotion, 14*(5), 892–907.

Royzman, E. B., Landy, J. F., & Goodwin, G. P. (2014a). Are good reasoners more incest-friendly? Trait cognitive reflection predicts selective moralization in a sample of American adults. *Judgment and Decision making, 9*(3), 176–190.

Schwitzgebel, E., Cokelet, B., & Singer, P. (2020). Do ethics classes influence student behavior? Case study: Teaching the ethics of eating meat. *Cognition, 203*, Article 104397.

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. *Journal of Neuroscience, 34*(13), 4741–4749.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General, 141*(3), 425–428.

Sinhahabi, N. (2018). Scalar consequentialism the right way. *Philosophical Studies, 175*(12), 3131–3144.

Slote, M. (1984). Satisficing consequentialism. *Proceedings of the Aristotelian Society, Supplementary Volumes, 58*, 139–163.

Slovic, P. (2007). If I look at the mass I will never act: Psychic numbing and genocide. *Journal of Risk & Uncertainty, 26*(1), 5–16.

Stieger, S., & Reips, U. D. (2016). A limitation of the Cognitive Reflection Test: familiarity. *PeerJ, 4*, e2395.

Sugerman, D. B. (1986). Active versus passive Euthanasia: An attributional analysis. *Journal of Applied Social Psychology, 16*(11), 60–76.

Sunstein, C. R., Schkade, D., & Kahneman, D. (2000). Do people want optimal deterrence? *The Journal of Legal Studies, 29*(1), 237–253.

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

Szasz, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: exploring the ways individuals solve the test. *Thinking & Reasoning, 23*(3), 207–234.

Thomas, B. C., Croft, K. E., & Tranel, D. (2011). Harming kin to save strangers: further evidence for abnormally utilitarian moral judgments after ventromedial prefrontal damage. *Journal of Cognitive Neuroscience, 23*(9), 2186–2196.

Thomson, J. J. (1975). Killing, letting die, and the trolley problem. *The Monist, 59*(2), 204–217.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*(6), 1395–1415.

Tinghög, G., Andersson, D., Bonn, C., Johannesson, M., Kirchner, M., Koppell, L., & Vastfjäll, D. (2016). Intuition and moral decision-making—the effect of time pressure and cognitive load on moral judgment and altruistic behavior. *PLoS One, 11*(10), Article e0164012.

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.

Unger, P. K. (1996). *Living high and letting die: Our illusion of innocence*. USA: Oxford University Press.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17*(6), 476–477.