THE INTUITIVE GREATER GOOD: TESTING THE CORRECTIVE DUAL PROCESS MODEL OF MORAL COGNITION

Bence Bago[1,2] & Wim De Neys[2,3]

1 – Toulouse School of Economics, University of Toulouse, France

2 - Paris Descartes University, Sorbonne Paris Cité, UMR 8240 LaPsyDÉ, France

3 - CNRS, UMR 8240, LaPsyDÉ, France

In press – Journal of Experimental Psychology: General

Corresponding author:

Bence BAGO, Paris Descartes University, 46 Rue Saint-Jacques, FR-75005, Paris; Phone: +33(0)664 581 065; Email: bencebagok@gmail.com

ABSTRACT

Building on the old adage that the deliberate mind corrects the emotional heart, the influential dual process model of moral cognition has posited that utilitarian responding to moral dilemmas (i.e., choosing the greater good) requires deliberate correction of an intuitive deontological response. In the present paper we present four studies that force us to revise this longstanding "corrective" dual process assumption. We used a two-response paradigm in which participants had to give their first, initial response to moral dilemmas under time-pressure and cognitive load. Next, participants could take all the time they wanted to reflect on the problem and give a final response. This allowed us to identify the intuitively generated response that preceded the final response given after deliberation. Results consistently show that in the vast majority of cases (+70%) in which people opt for a utilitarian response after deliberation, the utilitarian response is already given in the initial phase. Hence, utilitarian responders do not need to deliberate to correct an initial deontological response. Their intuitive response is already utilitarian in nature. We show how this leads to a revised model in which moral judgments depend on the absolute and relative strength differences between competing deontological and utilitarian intuitions.

INTRODUCTION

In the spring of 2013 the Belgian federal health minister, Laurette Onkelinckx, faced a tough decision. In a highly mediatized case, the seven year old Viktor Ameys who suffered from a very rare immune system disorder, begged her to approve reimbursement of the drug Soliris—a life-saving but extremely expensive treatment costing up to $400 000 a year (Dolgin, 2011). By not approving reimbursement the health minister was basically condemning an innocent seven year old to death. On the other hand, the federal health care budget is limited. Money that is spent on covering Viktor's drugs cannot be spent on the reimbursement of drugs for more common, less expensive disorders that threaten the lives of far more patients. Hence, saving Viktor implied not saving many others (London, 2012). Eventually, the health minister—herself a mother of three—felt she could not bring herself to let Viktor die and the Soliris reimbursement was approved (Schellens, 2015).

The Viktor case illustrates a classic moral dilemma in which utilitarian and deontological considerations are in conflict. The moral principle of utilitarianism (e.g., Mill & Bentham, 1987) implies that the morality of an action is determined by its consequences. Therefore, harming an individual can be judged acceptable, if it prevents comparable harm to a greater number of people. One performs a cost benefit analysis and chooses the greater good. Hence, from a utilitarian point of view it is morally acceptable to deny Viktor's request and let him die because more people will be saved by reimbursing other drugs. Alternatively, the moral perspective of deontology (e.g., Kant, 1785/2002) implies that the morality of an action depends on the intrinsic nature of the action. Here harming someone is considered wrong regardless of its potential benefits. Hence, from a deontological point of view, not saving Viktor would always be judged unacceptable.

In recent years, cognitive scientists in the fields of psychology, philosophy, and behavioral economics have started to focus on the cognitive mechanisms underlying utilitarian and deontological reasoning (e.g., Conway & Gawronski, 2013; Greene, 2015; Kahane, 2015; Moore, Stevens, & Conway, 2011; Nichols, 2004; Valdesolo & DeSteno, 2006). A lot of this work has been influenced by the popular dual-process model of thinking (Evans & Stanovich, 2013; Evans, 2008; Kahneman, 2011; Sloman, 1996), which often describes cognition as an interplay of fast, effortless, and intuitive (i.e., so-called "System 1") processing on one hand, and slow, cognitively demanding, deliberate (i.e., so-called "System 2") processing on the other. Inspired by this dichotomy the dual process model of moral reasoning (Greene, 2013; Greene & Haidt, 2002) has associated utilitarian judgments with deliberate System 2 processing and deontological judgments with intuitive System 1 processing. A core idea is that giving a utilitarian response to moral dilemmas requires that one engages in System 2 thinking and allocates

cognitive resources to override an intuitively cued intuitive System 1 response that primes us not to harm others (Greene, 2007; Paxton, Ungar, & Greene, 2012).

There is little doubt that the dual process model of moral cognition presents an appealing account and it has proved to be highly influential (Sloman, 2015). However, the framework is also criticized (e.g., Baron, 2017; Baron, Scott, Fincher, & Metz, 2015; Białek & De Neys, 2017; Kahane, 2015; Tinghög et al., 2016; Trémolière & Bonnefon, 2014). A key problem is that the processing specifications of the alleged System 1 and 2 operations are not clear. A critical issue concerns the time-course of utilitarian responding. In a typical moral dilemma, giving a utilitarian response is assumed to require the correction of the fast, initial System 1 response. The idea is that our immediate System 1 gut-response is deontological in nature but that after some further System 2 deliberation we can replace it with a utilitarian response. Hence, the final utilitarian response is believed to be preceded by an initial deontological response. From an introspective point of view, this core "corrective" dual process assumption (De Neys, 2017) seems reasonable. When faced with a dilemma such as the Viktor case, it surely feels as if the "don't kill Viktor" sentiment pops up instantly. We're readily repulsed by the very act of sacrificing a young boy and correcting that judgment by taking the greater good into account seems to require more time and effort. Unfortunately, the available empirical evidence is less conclusive than our introspective impressions seem to imply.

Consider, for example, evidence from latency studies and time pressure manipulations. Some earlier studies found that utilitarian responses take more time than deontological ones (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). Likewise, experimentally limiting the time allowed to make a decision was also shown to reduce the number of utilitarian responses (Suter & Hertwig, 2011). However, in recent years conflicting findings have been reported. Limiting response time does not always have significant effects and sometimes deontological responses are even found to take longer than utilitarian ones (e.g., Baron & Gürçay, 2017; Gürçay & Baron, 2017; Tinghög et al., 2016). More critically, even if we were to unequivocally establish that utilitarian responses take more time than deontological responses, this does not imply that utilitarian responders generated the deontological response before arriving at the utilitarian one. They might have needed more time to complete the System 2 deliberations without ever having considered the deontological response.

Neuroimaging studies have also explored the neural correlates of deontological and utilitarian reasoning (e.g., Greene, Nystrom, Engell, Darley, & Cohen, 2004; Shenhav & Greene, 2014). In a nutshell, results typically indicate that deontological judgments are associated with activation in brain areas that are known to be involved in emotional processing (e.g., amygdala) whereas utilitarian decisions seem to recruit brain areas associated with controlled processing (e.g., dorsolateral prefrontal cortex).

This imaging work suggests that deontological and utilitarian responses might rely on a different type of processing. However, it does not allow us to make strong inferences concerning their precise time course.

A more direct test of the corrective dual process assumption comes from mouse tracking studies (Gürçay & Baron, 2017; Koop, 2013). After having read a moral dilemma, participants in these studies are asked whether they favor a deontological or utilitarian decision. To indicate their answer, they have to move the mouse pointer from the center of the screen towards the utilitarian or deontological response options that are presented in the opposite corners of the screen. In the mouse-tracking paradigm researchers typically examine the curvature in the mouse movement to test whether participants show "preference reversals" (Spivey, Grosjean, & Knoblich, 2005). For example, if utilitarian responders initially generate a deontological response, they can be expected to move first towards the deontological response and afterwards to the utilitarian one. This will result in a more curved mouse trajectory. Deontological responders on the other hand are expected to go straight towards the deontological option from the start. However, contrary to the dual process assumption, the mouse trajectories have been found to be equally curved for both types of responses (Gürçay & Baron, 2016; Koop, 2013).

There is also some converging evidence for the mouse-tracking findings. Białek and De Neys (2016, 2017) studied deontological responders' conflict sensitivity. They presented participants with classic moral dilemmas and control versions in which deontological and utilitarian considerations cued the same non-conflicting response. For example, a no-conflict control version of the introductory drug example might be a scenario in which reimbursing the drug would save many more patients than not reimbursing it. Bialek and De Neys reasoned that if deontological responders were only considering the deontological response option, they should not be affected by the presence or absence of conflict. Results indicated that the intrinsic conflict in the classic dilemmas also affected deontological responders, as reflected in higher response doubt and longer decision times for the conflict vs no-conflict versions. Critically, this increased doubt was still observed when System 2 deliberation was experimentally minimized with a concurrent load task (Bialek & De Neys, 2017). This suggests that our intuitive System 1 is also cueing utilitarian considerations. However, it should be noted that although deontological responders showed conflict sensitivity, they still selected the deontological response option. Consequently, proponents of the corrective dual process view can still claim that people who actually make the utilitarian decision will only do so after deliberate correction of their initial deontological answer.

In the present studies we adopt a two-response paradigm (Thompson, Turner, & Pennycook, 2011) to obtain a more conclusive test of the corrective dual process assumption. The two-response paradigm has been developed in logical and probabilistic reasoning studies to gain direct behavioral insight into the time-course of intuitive and deliberate response generation (Bago & De Neys, 2017;

Newman, Gibb, & Thompson, 2017; Pennycook & Thompson, 2012; Thompson & Johnson, 2014). In the paradigm participants are presented with a reasoning problem and have to respond as quickly as possible with the first response that comes to mind. Immediately afterwards they are presented with the problem again and can take as much time as they want to reflect on it and give a final answer. To make maximally sure that the first response is truly intuitive in nature participants are forced to give their first response within a strict deadline while their cognitive resources are also burdened with a concurrent load task (Bago & De Neys, 2017). The rationale is that System 2 processing, in contrast to System 1, is often conceived as time and resource demanding. By depriving participants from these resources one aims to "knock" out System 2 during the initial response phase (Bago & De Neys, 2017). The prediction in the moral reasoning case is straightforward. If the corrective assumption holds, the initial response to moral dilemmas should typically be deontological in nature and utilitarian responses should usually only appear in the final response stage. Put differently, individuals who manage to give a utilitarian response after deliberation in the final response stage should initially give a deontological response when they are forced to rely on more intuitive processing in the first response stage.

We present four studies in which we tested the robustness of the two-response findings. To foreshadow our key result, across all our studies we consistently observe that in the majority of cases in which people select a utilitarian responses after deliberation, the utilitarian response is already given in the initial phase. Hence, utilitarian responders do not necessarily need to deliberate to correct an initial deontological response. Their intuitive response is typically already utilitarian in nature. We will present a revised dual process model to account for the findings.

STUDY 1

Method

Participants

In Study 1, 107 Hungarian students (77 female, Mean age = 21.6 years, SD = 1.4 years) from the Eotvos Lorand University of Budapest were tested. A total of 94% of the participants reported high school as highest completed educational level, while 6% reported having a post-secondary education degree. Participants received course credit for taking part. Participants in Study 1 (and all other reported studies) completed the study online.

Sample size decision was based on our previous two-response studies in the logical reasoning field (Bago & De Neys, 2017) in which we also always tested approximately 100 participants per condition.

Materials

Moral reasoning problems. In total, nine moral reasoning problems were presented. Problem content was based on popular scenarios from the literature (e.g., Cushman, Young, & Hauser, 2006; Foot, 1967; Royzman & Baron, 2002). All problems had the same underlying structure and required subjects to decide whether or not to sacrifice the lives of one of two groups of scenario characters. To minimize inter-item noise and possible content confounds (e.g., Trémolière & De Neys, 2013) we stuck to the following content rules for all problems: a) the difference between the possible number of characters in the two groups was kept constant at 8 lives, b) all characters were adults, c) the to-be made sacrifice concerned the death of the characters, d) there was no established hierarchy among the to-be sacrificed characters, and e) the scenario protagonist's own life was never at stake. All problems are presented in the Supplementary Material, section A. All problems were translated to Hungarian (i.e., participants' mother tongue) for the actual experiment.

The problems were presented in two parts. First, the general background information was presented (non-bold text in example bellow) and participants clicked on a confirmation button when they finished reading it. Subsequently, participants were shown the second part of the problem that contained the critical conflicting (or non-conflicting, see further) dilemma information and asked them about their personal willingness to act and make the described sacrifice themselves ("Would you do X?"). Participants entered their answer by clicking on a corresponding bullet point ("Yes" or "No"). The first part of the problem remained on the screen when the second part was presented. The following example illustrates the full problem format:

> Due to an accident there are 11 miners stuck in one of the shafts of a copper mine. They are almost out of oxygen and will die if nothing is done. You are the leader of the rescue team.
>
> The only way for you to save the miners is to activate an emergency circuit that will transfer oxygen from a nearby shaft into the shaft where the 11 miners are stuck.
>
> **However, your team notices that there are 3 other miners trapped in the nearby shaft. If you activate the emergency circuit to transfer the oxygen, these 3 miners will be killed, but the 11 miners will be saved.**
>
> **Would you activate the emergency circuit?**
> **0 Yes**
> **0 No**

Four of the presented problems were traditional "conflict" versions in which participants were asked whether they were willing to sacrifice a small number of people to save several more. Four other problems were control "no-conflict" versions in which participants were asked whether they were willing

to sacrifice more people to save less (e.g., Bialek & De Neys, 2017). The following is an example of a no-conflict problem:

> You are a radar operator overseeing vessel movement near Greenland. Due to sudden ice movement a boat carrying 3 passengers is about to crash into an iceberg. If nothing is done, all passengers will die.
>
> The only way to save the 3 passengers is for you to order the captain to execute an emergency manoeuvre that will sharply alter the course of the boat.
>
> **However, the manoeuvre will cause the boat to overrun a life raft carrying 11 people. The life raft is floating next to the iceberg and out of sight of the captain. The 11 people will be killed if you order to execute the manoeuvre, but the 3 people on the boat will be saved.**
>
> **Would you order to execute the manoeuvre?**

Hence, on the conflict version the utilitarian response is to answer "yes" and the deontological response is to answer "no". On the no-conflict problems both utilitarian and deontological considerations cue a "no" answer (for simplicity, we will refer to these non-sacrificial greater good answers as "utilitarian responses"). We included the no-conflict versions to make the problems less predictable and avoid that participants would start to reason about the possible dilemma choice before presentation of the second problem part. For the same reason we also included a filler item in the middle of the experiment (i.e., after 4 test problems). In this filler problem saving more people did not involve any sacrifice (i.e., doing the action implied saving 6 and killing 0 characters).

Two problem sets were used to counterbalance the scenario content; scenario content that was used for the conflict problems in one set was used for the no-conflict problems in the other set, and vice-versa. Participants were randomly assigned to one of the sets. The presentation order of the problems was randomized in both sets.

Load task. We wanted to make maximally sure that participants' initial response was intuitive (i.e., System 2 engagement is minimized). Therefore, we used a cognitive load task (i.e., the dot memorization task, see Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001) to burden participants' cognitive resources The rationale behind the load manipulation is simple; it is often assumed that System 2 processing requires executive cognitive resources, while System 1 processing does not (Evans & Stanovich, 2013). Consequently, if we burden someone's executive resources while they are asked to solve a moral reasoning problem, System 2 engagement is less likely. We opted for the dot memorization task because it is specifically assumed to burden participant's executive resources (De Neys & Schaeken, 2007; De Neys & Verschueren, 2006; Franssens & De Neys, 2009; Miyake et al., 2001). The dot matrix task is visuo-spatial in nature. Note that although it has been shown that some visual load tasks can interfere with generation of deontological responses (i.e., because they make it harder to visually imagine

the sacrifice, Amit & Greene, 2012), previous studies indicate that the dot matrix task rather interferes with the generation of utilitarian responses (Bialek & De Neys, 2017; Trémolière, De Neys, & Bonnefon, 2012).

Before each reasoning problem participants were presented with a 3 x 3 grid, in which 4 dots were placed. Participants were instructed that it was critical to memorize the location of the dots even though it might be hard while solving the reasoning problem. After answering the reasoning problem participants were shown four different matrixes and they had to choose the correct, to-be-memorized pattern. They received feedback as to whether they chose the correct or incorrect pattern. The load was only applied during the initial response stage and not during the subsequent final response stage in which participants were allowed to deliberate and recruit System 2 (see further).

Procedure

Reading pre-test. Before we ran the main study we recruited an independent sample of 33 participants for a reading pre-test (28 female, Mean age = 19.5 years, SD = 1.03 years). Participants were also recruited from the Eotvos Lorand University of Budapest and received course credit in exchange. All participants reported high school as the highest completed educational level. The basic goal of the reading pre-test was to determine the response deadline which could be applied in the main moral reasoning study. The idea was to base the response deadline on the average reading time in the reading test (e.g., Bago & De Neys, 2017). The rationale here was very simple; if people are allotted the time they need to simply read the problem, we can expect that System 2 reasoning engagement is minimized. Thus, in the reading pre-test, participants were presented with the same items as in the main moral reasoning study. They were instructed to read the problems and randomly click on one of the answer options. The general instructions were as follows:

> **Welcome to the experiment!**
> **Please read these instructions carefully!**
> This experiment is composed of 9 questions and 1 practice question. It will take 5 minutes to complete and it demands your full attention. You can only do this experiment once.
>
> In this task we'll present you with a set of problems we are planning to use in future studies. Your task in the current study is pretty simple: you just need to read these problems. We want to know how long people need on average to read the material. In each problem you will be presented with two answer alternatives. You don't need to try to solve the problems or start thinking about them. Just read the problem and the answer alternatives and when you are finished reading you randomly click on one of the answers to advance to the next problem.
>
> The only thing we ask of you is that you stay focused and read the problems in the way you typically

would. Since we want to get an accurate reading time estimate please avoid wiping your nose, taking a phone call, sipping from your coffee, etc. before you finished reading.

At the end of the study we will present you with some easy verification questions to check whether you actually read the problems. This is simply to make sure that participants are complying with the instructions and actually read the problems (instead of clicking through them without paying attention). No worries, when you simply read the problems, you will have no trouble at all to answer the verification questions.

Please confirm below that you read these instructions carefully and then press the "Next" button.

Problems were presented in two parts (background information and critical dilemma information) as in the main study. Our interest concerned the reading time for the critical second problem part. To make sure that participants would actually read the problems, we informed subjects that they would be asked to answer two simple verification questions at the end of the experiment to check whether they read the material. The verification questions could be easily answered even by a very rough reading. The following illustrates the verification question:

We asked you to read a number of problems.
Which one of the following situations was not part of the experiment?

- o You were a soccer player
- o You were the leader of a rescue team
- o You were a railway controller
- o You were a late-night watchman

The correct answers were clearly different from the situations which were presented during the task. Only one of the participants did not manage to solve both verification questions correctly (97% solved both correctly). This one participant was excluded from the reading-time analysis. The average reading time for the critical dilemma part in the resulting sample was M = 11.3 s (SD = 1.5 s). Note that raw reaction time data were first logarithmically transformed prior to analysis. Mean and standard deviation were calculated on the transformed data, and then they were back-transformed into seconds. We wanted to give the participants some minimal leeway, thus we rounded the average reading time to the closest higher natural number; the response deadline was therefore set to 12 seconds.

One-response pre-test. To make sure that our 12 s initial response deadline was sufficiently challenging we also ran a traditional “one-response” reasoning pre-test without deadline or load. The same material as in the reading pre-test and main reasoning study was used. As in traditional moral reasoning studies, participants were simply asked to give one single answer for which they could take all the time they wanted. We recruited an independent sample of 55 participants (34 female, Mean age = 23.4 years, SD = 2.4 years) from the Eotvos Lorand University of Budapest who received course credit in

exchange. A total of 76% of the participants reported high school as highest completed educational level, while 24% reported having a post-secondary education degree. Raw reaction time data were first logarithmically transformed prior to analysis. Mean and standard deviation were calculated on the transformed data, and then they were back-transformed into seconds as with the reading pre-test data. Results showed that from the moment that the critical second problem part was presented participants needed on average 14.4 s (SD = 1.9 s) to enter a response. For the conflict problems this average reached 14.7 s (SD = 1.9 s; utilitarian response, mean = 14.6 s, SD = 1.8s; deontological response, mean = 15.5 s, SD = 2.2 s). In sum, these results indicate that the 12 s response deadline will put participants under considerable time-pressure (i.e., less than average one-response response time minus 1 SD).

The one-response pre-test also allowed us to test for a potential consistency confound in the two-response paradigm. That is, when people are asked to give two consecutive responses, they might be influenced by a desire to look consistent. Hence, where people might implicitly change their initial deontological intuition after deliberation in a one-response paradigm, they might refrain from doing so when they are forced to explicitate their initial response. Thereby, the two-response paradigm might underestimate the rate of final utilitarian responses and the associated correction rate. However, in our one-response pre-test we observed 85.4% (SD = 35.3%) of utilitarian responses on the conflict versions. This is virtually identical to the final utilitarian response rate of 84.5% (SD = 36.2) in our main two-response study (see main results). This directly argues against the consistency confound and validates the two-response paradigm.

Two-response moral reasoning task. The experiment was run online. Participants were specifically instructed at the beginning that we were interested in their very first, initial answer that came to mind. They were also told that they would have additional time afterwards to reflect on the problem and could take as much time as they needed to provide a final answer. The literal instructions that were used stated the following (translated from Hungarian):

> **Welcome to the experiment!**
> **Please read these instructions carefully!**
>
> This experiment is composed of 9 questions and a couple of practice questions. It will take about 12 minutes to complete and it demands your full attention. You can only do this experiment once.
>
> In this task we'll present you with a set of moral reasoning problems. We would like you to read every problem carefully and enter your response by clicking on it. There are no correct or incorrect decisions, we are interested in the response you personally feel is correct. We want to know what your initial, intuitive response to these problems is and how you respond after you have thought about the problem for some more time. Hence, as soon as the problem is presented, we will ask you to enter your initial response. We want you to respond with the

> very first answer that comes to mind. You don't need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible. Next, the problem will be presented again and you can take all the time you want to actively reflect on it. Once you have made up your mind you can enter your final response. You will have as much time as you need to indicate your second response.
>
> After you have entered your first and final answer we will also ask you to indicate your confidence in your response.
>
> In sum, keep in mind that it is really crucial that you give your first, initial response as fast as possible. Afterwards, you can take as much time as you want to reflect on the problem and select your final response.
>
> Please confirm below that you read these instructions carefully and then press the "Next" button.

After this general introduction, participants were presented with a more specific instruction page which explained them the upcoming task and informed them about the response deadline. The literal instructions were as follows:

> We are going to start with a couple of practice problems. First, a fixation cross will appear. Then, the first part of the problem will appear. When you finished reading this click on the "Next" button and the rest of the problem will be presented to you.
>
> As we told you we are interested in your initial, intuitive response. First, we want you to respond with the very first answer that comes to mind. You don't need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible. To assure this, a time limit was set for the first response, which is going to be 12 seconds. When there are 2 seconds left, the background colour will turn to yellow to let you know that the deadline is approaching. Please make sure to answer before the deadline passes. Next, the problem will be presented again and you can take all the time you want to actively reflect on it. Once you have made up your mind you enter your final response.
>
> After you made your choice and clicked on it, you will be automatically taken to the next page. After you have entered your first and final answer we will also ask you to indicate your confidence in the correctness of your response.
>
> Press "Next" if you are ready to start the practice session!

After the specific instruction page participants solved two unrelated practice reasoning problems to familiarize them with the procedure. Next, they solved two practice dot matrix problems (without concurrent reasoning problem). Finally, at the end of the practice, they had to solve the two earlier practice reasoning problems under cognitive load.

Each problem started with the presentation of a fixation cross for 1000 ms. Then, the dot matrix appeared and stayed on the screen for 2000 ms. Next, the first part of the moral reasoning problem with the background information appeared. Participants could take all the time they wanted to read the background information and clicked on the next button when they were ready. Next, the remaining part of

the problem appeared (while the first part stayed on screen). From this point onwards participants had 12 s to give an answer; after 10 s the background of the screen turned yellow to warn participants about the upcoming deadline. If they did not provide an answer before the deadline, they were asked to pay attention to provide an answer within the deadline on subsequent trials

After the initial response, participants were asked to enter their confidence in the correctness of their answer on a scale from 0% to 100%, with the following question: "How confident are you in your answer? Please type a number from 0 (absolutely not confident) to 100 (absolutely confident)". After indicating their confidence, they were presented with four dot matrix options, from which they had to choose the correct, to-be-memorized pattern. Once they provided their memorization answer, they received feedback as to whether it was correct. If the answer was not correct, they were also asked to pay more attention to memorizing the correct dot pattern on subsequent trials.

Finally, the full problem was presented again, and participants were asked to provide a final response. Once they clicked on one of the answer options they were automatically advanced to the next page where they had to provide their confidence level again.

The colour of the answer options was green during the first response, and blue during the final response phase, to visually remind participants which question they were answering. Therefore, right under the question we also presented a reminder sentence: "Please indicate your very first, intuitive answer!" and "Please give your final answer", respectively, which was also coloured as the answer options.

At the end of the study participants completed a page with standard demographic questions.

Exclusion criteria. Participants failed to provide a first response before the deadline in 7% of the trials. In addition, in 8.3% of the trials participants responded incorrectly to the dot memorization load task. All these trials were removed from the analysis because it cannot be guaranteed that the initial response resulted from mere System 1 processing: If participants took longer than the deadline, they might have engaged in deliberation. If they fail the load task, we cannot be sure that they tried to memorize the dot pattern and System 2 was successfully burdened. In these cases we cannot claim that possible utilitarian responding at the initial response stage is intuitive in nature. Hence, removing trials that did not meet the inclusion criteria gives us the purest possible test of our hypothesis.

In total, 14.8% of trials were excluded and 821 trials (out of 963) were further analysed (initial and final response for the same item counted as 1 trial).

Statistical analysis. Throughout the article we used mixed-effect regression models to analyse our results (Baayen, Davidson, & Bates, 2008; Kuznetsova, Brockhoff, & Christensen, 2015), accounting for

the random intercept of participants and items. For the binary choice data we used logistic regression while for the continuous confidence and reaction time data we used linear regression.

Results

Table 1 gives a general overview of the results. We first focus on the response distributions for the final response. As one might expect, on the no-conflict problems in which utilitarian and deontological considerations cued the same response and choosing the greater good did not involve a sacrifice, the rate of utilitarian responses was near ceiling (95.4%). Not surprisingly, the utilitarian response rate was lower (84.5%) on the conflict problems in which choosing the greater good did require to sacrifice lives, χ^2 (1) = 11.1, p = 0.0009, b = -0.99. The key finding, however, was that the utilitarian response was also frequently given as the initial, intuitive response on the critical conflict problems (79.7% of initial responses). This suggests that participants can give intuitive utilitarian responses to classic moral dilemmas.

However, the raw percentage of intuitive utilitarian conflict problem responses is not fully informative. We can obtain a deeper insight into the results by performing a Direction of Change analysis on the conflict trials (Bago & De Neys, 2017). This means that we look at the way a given person in a specific trial changed (or didn't change) her initial answer after the deliberation phase. More specifically, people can give a utilitarian and a deontological response in each of the two response stages. Hence, in theory this can result in four different types of answer change patterns ("DD", deontological response in both stages; "UU", utilitarian response in both stages; "DU", initial deontological and final utilitarian response; "UD", initial utilitarian and final deontological response). Based on the corrective dual process assumption, one can expect that people will either give "DD" responses, meaning that they had the deontological intuition in the beginning and did not correct it in the final stage, or "DU" responses meaning that they initially generated a deontological response, but then, after deliberation, they changed it to a utilitarian response.

Table 2 shows the direction of change category frequencies for the conflict problems. First of all, we observed a non-negligible amount of DD (9.7%) and DU (10.6%) responses. In and by itself, these patterns are in accordance with the corrective predictions; reasoners generated the deontological response initially, and in the final response stage they either managed to override it (DU) or they did not (DD). However, what is surprising and problematic for the corrective perspective is the high percentage of UU responses (73.9% of the trials). Indeed, in the vast majority of the cases in which participants managed to give a utilitarian answer as final response, they already gave it as their first, intuitive response (i.e., 87.5% of cases). We refer to this critical number [(i.e., UU/(UU+DU) ratio] as the % of non-corrective utilitarian

responses or non-correction rate in short. Overall, this means that utilitarian reasoners did not necessarily need their deliberate System 2 to correct their initial deontological intuition; their intuitive System 1 response was typically already utilitarian in nature.

STUDY 2

Our Study 1 results are challenging for the corrective dual process assumption: utilitarian responses to moral dilemmas were typically generated intuitively. However, one potential issue is that although the rate of utilitarian responding on the critical conflict items was lower than on the no-conflict problems, it was still fairly high. A critic might utter that the dual process model does not necessarily entail that all moral decisions require a deliberate correction process. The prototypical case on which the model has primarily focused concerns "high-conflict"[1] situations in which a dilemma cues a strong conflicting emotional response which renders the utilitarian override particularly difficult (Greene, 2009; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Shenhav & Greene, 2014). The high rate of final utilitarian responding on our Study 1 conflict problems might be argued to indicate that the problem did not evoke a particularly strong emotional response. Hence, the corrective assumption might be maintained in cases in which utilitarian responding is rarer (i.e., more demanding). Study 2 was run to address this issue. We used a manipulation (i.e., one of the to-be sacrificed persons in the dilemma was a close family member, e.g., Hao, Liu, & Li, 2015; Tassy, Oullier, Mancini, & Wicker, 2013) that has been shown to increase the emotional averseness of the sacrificial option. We expected the manipulation to decrease the rate of utilitarian responding on the conflict problems. The critical question concerned the non-correction rate. If the Study 1 critique is right, final utilitarian decisions in Study 2 should be typically preceded by initial deontological responses leading to a floored non-correction rate.

Method

Participants

A total of 107 participants (68 female, Mean age = 20.6 years, SD = 1.9 years) from the Eotvos Lorand University of Budapest were tested. A total of 87.9% of the participants reported high school as

[1] The a priori operationalisation of what constitutes a "high-conflict" situation has been shown to be somewhat controversial (Gürcay & Baron, 2017; Greene, 2009). The simple point here is that we wanted to make sure that our non-correction rate results are robust and are not driven by specific idiosyncratic features of our dilemmas.

highest completed educational level, while 12.1% reported having a post-secondary education degree. Participants received course credit for taking part.

Materials and Procedure

Moral reasoning task. The same scenario topics as in Study 1 were used. The only modification was that one of the-to be sacrificed persons in the dilemma was always a close family member (father, mother, brother, or sister). This "family member" manipulation has been shown to increase the emotional averseness of the sacrificial option and decrease the rate of utilitarian conflict responses (Hao et al., 2015; Tassy et al., 2013). Here is an example of a conflict problem:

> Due to an accident there are 11 miners stuck in one of the shafts of a copper mine. They are almost out of oxygen and will die if nothing is done. You are the leader of the rescue team.
>
> The only way for you to save the miners is to activate an emergency circuit that will transfer oxygen from a nearby shaft into the shaft where the 11 miners are stuck.
>
> **However, your team notices that your own father and two other miners are trapped in the nearby shaft. If you activate the emergency circuit to transfer the oxygen, your father and the two other miners will be killed, but the 11 miners will be saved.**
>
> **Would you activate the emergency circuit?**
> **0 Yes**
> **0 No**

The following is an example of a no-conflict problem:

> You are a radar operator overseeing vessel movement near Greenland. Due to sudden ice movement a boat carrying 3 passengers is about to crash into an iceberg. If nothing is done, all passengers will die.
>
> The only way to save the 3 passengers is for you to order the captain to execute an emergency manoeuvre that will sharply alter the course of the boat.
>
> **However, the manoeuvre will cause the boat to overrun a life raft carrying your own father and 10 other people. The life raft is floating next to the iceberg and out of sight of the captain. Your father and the other 10 people will be killed if you order to execute the manoeuvre, but the 3 people on the boat will be saved.**
>
> **Would you order to execute the manoeuvre?**
> **0 Yes**
> **0 No**

As in Study 1, participants evaluated four conflict, one filler, and four no-conflict problems in a randomized order. Scenario content of the conflict and no-conflict problems was counterbalanced. We also adopted the exact same two-response procedure as in Study 1. Hence, except for the modified scenario content, the procedure was completely identical to Study 1. All Study 2 problems are presented in the Supplementary Material, section A

Exclusion criteria. The same exclusion criteria were applied as in Study 1. Participants failed to provide a first response before the deadline in 8.4% of the trials. In addition, in 7.1% of the trials participants responded incorrectly to the dot memorization load task. All these trials (15.1% of trials in total) were excluded and 818 trials (out of 963) were further analyzed (initial and final response for the same item counted as 1 trial).

Results

Table 1 gives an overview of the results. In line with expectations, we see that the percentage of utilitarian responses on the conflict items is much lower in Study 2 than in Study 1, both at the initial (17.5%) and final response (21.2%) stage. On the no-conflict items—in which choosing the greater good and saving the family member did not entail a sacrifice— the utilitarian response rate remained at ceiling. We tested this trend statistically by testing the interaction of conflict and family member condition (data from Study 1 as no-family condition, data from Study 2 as family condition). This interaction was indeed significant both at the initial, χ^2 (3) = 108.45, $p < 0.0001$, b = -3.81, and final, χ^2 (3) = 76.6, $p < 0.0001$, b = -4.37, response stage. This supports the claim that the family member manipulation increases the emotional averseness of a utilitarian sacrifice (Hao et al., 2015; Tassy et al., 2013).

Reflecting the lower overall rate of initial and final utilitarian responses, the direction of change results in Table 2 indicate that there were fewer "UU" and "DU" responses in Study 2. However, the critical finding is that despite the overall decease, the "UU" responses (12.7%) are still twice as frequent as the "DU" (4.8%) responses. Hence, far from being floored, the non-correction rate remained high at 72.6%. In sum, in those cases that utilitarian responses are generated, they are still predominantly generated at the initial, intuitive response stage. This confirms the Study 1 finding and further argues against the corrective dual process assumption: Even in "high conflict" situations, utilitarian responding does not necessarily require reasoners to deliberately correct their initial deontological response.

The low level of initial utilitarian conflict responses in Study 2 might give rise to the objection that these rare responses result from mere guessing. After all, our task is quite challenging; people had to respond within a strict response deadline and under secondary task load. In theory, it might be that participants found it too hard and just randomly clicked on one of the answer options. However, the ceiled performance on the no-conflict problems (94.5% utilitarian response) argues against such a guessing account. If participants were guessing, their performance on the conflict and no-conflict problems should be closer to 50% in both cases. We also conducted a so-called stability analysis (Bago & De Neys, 2017) to further test for a guessing account. We calculated for every participant on how many conflict problems

they displayed the same direction of change category. We refer to this measure as the stability index. For example, if an individual shows the same type of direction of change on all four conflict problems, the stability index would be 100%. If the same direction of change is only observed on two trials, the stability index would be 50%[2] etc. Results showed that the average stability index in Study 2 reached 83.8% (similar high stability rates were observed in all our studies, see Table S1 in the Supplementary Material). This indicates that that the direction of change pattern is highly stable on the individual level and argues against a guessing account; if people were guessing, they should not tend to show the same response pattern consistently.

STUDY 3

Study 3 was run to further test the robustness of our findings. One might argue that Study 1 and 2 focused on two more extreme cases: utilitarian responding was either very rare or very prevalent. In Study 3 we looked at a more "intermediate" case. We therefore combined the family member manipulation with a manipulation that has been shown to facilitate utilitarian responding. Trémolière and Bonnefon (2014) previously showed that increasing the kill-save ratio of a sacrifice (i.e., more people are saved), promoted utilitarian responding. Hence, by making the kill-save ratio more extreme we expected to increase the rate of utilitarian responding in comparison with Study 2. We again were interested in the non-correction rate. If the high non-correction rate is consistently observed with different scenario characteristics, this indicates that the findings are robust.

Method

Participants

In Study 3, 230 Hungarian students (171 female, Mean age = 22.6 years, SD = 21.7 years) from the Eotvos Lorand University of Budapest were tested. A total of 83% of the participants reported high school as highest completed educational level, while 17% reported having a post-secondary education degree. Participants received course credit for taking part.

Materials and procedure

[2] Note that due to methodological restrictions (we excluded items with incorrect load questions and items where response was not given within the deadline) some participants had less than four responses available. For these participants, stability was calculated based on the available items.

Moral reasoning task. The same scenario topics as in the previous studies were used. The only modification was the kill-save ratio. Therefore, we multiplied the number of lives at stake in the largest group by a factor 5. Hence, in Study 1 and 2 the ratio was moderate (e.g., kill 3 to save 11; all ratios between 20%-30%), in Study 3 the ratio was more extreme (e.g., kill 3 to save 55; all ratios between 4-8%). Note that we made the ratio as extreme as the scenario content would allow (e.g., a life raft/plane carrying 5000 passengers would not be realistic, e.g., Trémolière & Bonnefon, 2014). For half of the sample the extreme ratios were combined with the same "family member" scenario content that was used in Study 2. For completeness, for the other half of the sample we combined the extreme ratios with the original "no family" scenario content that was used in Study 1. Participants were randomly allocated to one of the two conditions. Hence, over the three studies the kill-save ratio and family member manipulations were fully crossed.

As in Study 1 and 2, participants evaluated four conflict, one filler, and four no-conflict problems in a randomized order. Scenario content of the conflict and no-conflict problems was counterbalanced. We also adopted the exact same two-response procedure as in Study 1 and 2. Hence, except for the modified kill-save ratio scenario content, the procedure was identical to Study 1 and 2.

Exclusion criteria. The same exclusion criteria were applied as in Study 1 and 2. Participants failed to provide a first response before the deadline in 8.1% of the trials. In addition, in 6.4% of the trials participants responded incorrectly to the dot memorization load task. All these trials (13.8% of trials in total) were excluded and 1784 trials (out of 2070) were further analysed (initial and final response for the same item counted as 1 trial).

Results and discussion

Table 1 gives an overview of the results. As before, the no-conflict items remained at ceiling throughout. As expected, the extremer kill-save ratios in Study 3 resulted in a slightly higher initial and final utilitarian conflict problem response rate in comparison with the moderate kill-save results in Study 1 and 2. This trend was most pronounced in the "family" condition in which the utilitarian response rate with moderate ratios was lowest. Statistical testing showed that the main effect of the extremity manipulation (after accounting for the effect of "Family" condition) was significant at the final response, χ^2 (2) = 11.97, p = 0.0005, b = -1.06, but not at the initial response stage, χ^2 (2) = 3.28, p = 0.07, b = -0.43. The interaction trend with the family member manipulation failed to reach significance both at the initial, χ^2 (3) = 0.56, p = 0.45, and final response stage, χ^2 (3) = 1.85, p = 0.17. Note that the more limited impact

of the extremity manipulation might be due to the fact that our ratios were less extreme than in previous work (e.g., Trémolière & Bonnefon, 2014).

Nevertheless, the key point is that we observed a higher absolute descriptive number of utilitarian responses in Study 3, especially in the family condition (31.4% final utilitarian vs. 17.5% in Study 2) which allows us to test the generalizability of our non-correction findings across various levels of ultimate utilitarian responding. Table 2 shows the direction of change results. The critical finding is that we again observe very high non-correction rates in Study 3, both when the life of a family member was a stake (72.6%) or not (87.8%). Hence, across our three studies with varying dilemma characteristics and absolute levels of utilitarian responding, we consistently observe that although correction is sometimes observed, it is far less likely than non-correction. In more than 70% of the cases, utilitarian responders do not need to correct an initial, deontological response, their initial intuitive response is already utilitarian in nature.

Additional analyses. After having established the robustness of the non-correction findings in our three studies, we can explore a number of additional two-response data questions. For example, one can contrast response latencies and confidence ratings for the different direction of change categories. Previous two-response studies on logical reasoning (e.g., Bago & De Neys, 2017; Thompson et al., 2011; Thompson & Johnson, 2014) established that the initial response confidence is typically lower for responses that get subsequently changed after deliberation (e.g., "DU" and "UD" in the present case) than for responses that are not changed (e.g., "UU" and "DD" in the present case). It has been suggested that this lower initial confidence (or "Feeling of Rightness" as Thompson et al. refer to it) would be one factor that determines whether reasoners will engage in System 2 deliberation (e.g., Thompson et al., 2011). Changed responses have also been shown to be associated with longer "re-thinking times" (i.e., response latencies) in the final response stage. To explore these trends in the moral reasoning case, Figures 1 and 2 plot the average confidence ratings and response latencies findings across our three studies. As the figures indicate our moral reasoning findings are consistent with the logical reasoning trends. Initial response confidence (Figure 1, top panel) for the "UD" and "DU" categories in which the initial response is changed after deliberation is lower than for "UU" and "DD" categories in which the initial responses is not changed. Final response times (Figure 2, bottom panel) are also longer for the change categories (i.e., "DU" and "UD") than for the no-change ones. To test these trends statistically we entered direction of change category (change vs no-change) as fixed factor to the models. All latency data were log-transformed prior to analysis. Both the confidence, χ^2 (1) = 104.7, $p < 0.0001$, b = -15.8, and latency, χ^2 (1) = 49.03, $p < 0.0001$, b = 0.17, trends were significant. One additional trend that visually pops-out is that for the "DU" category in which an initial deontological response is corrected, there is a sharp

confidence increase when contrasting initial and final confidence, χ^2 (1) = 49.1, p < 0.0001, b = 21.4. After deliberation, the response confidence attains the level of intuitively generated utilitarian responses in the "UU" case. Hence, perhaps not surprisingly, in those cases that deliberate correction occurs it seems to alleviate one's initial doubt. We also note that with respect to the initial response latencies (Figure 2, top panel), the rare UD trials seemed to be generated slightly faster than the others, χ^2 (1) = 7.3, p = 0.007, b = -0.05. For completeness, the interested reader can find a full overview of the confidence and latency data in the Supplementary material (Table S2 and S3).

A related issue we can explore with our confidence data is whether intuitive utilitarian responders are actually faced with two competing intuitions at the first response stage. That is, a possible reason for why people in the "UU" category manage to give a utilitarian initial response might be that the problem simply does not generate an intuitive deontological response for them. Hence, they would only generate a utilitarian response and would not be faced with an interfering deontological one. Alternatively, they might generate two competing intuitions, but the utilitarian intuition might be stronger and therefore dominate (Bago & De Neys, 2017).

We can address this question by looking at the confidence contrast between conflict and no-conflict control problems. If conflict problems cue two conflicting initial intuitive responses, people should process the problems differently than the no-conflict problems (in which such conflict is absent) in the initial response stage. Studies on conflict detection during moral reasoning that used a classic single response paradigm have shown that processing conflict problems typically results in lower response confidence (e.g., Bialek & De Neys, 2016, 2017). The question that we want to answer here is whether this is also the case at the initial response stage. Therefore, we contrasted the confidence ratings for the initial response on the conflict problems with those for the initial response on the no-conflict problems[3]. Our central interest here concerns the "UU" cases but a full analysis and discussion for each direction of change category is presented in the Supplementary Material (section C). In sum, results across our studies indeed indicate that "UU" responders showed a decreased confidence (average decrease = 6.1%, SE = 1.1, χ^2 (1) = 21.4, p < 0.0001, b = -6.76) on the conflict vs no-conflict problems. This supports the hypothesis that in addition to their dominant utilitarian intuition the alternative deontological response is also being cued.

STUDY 4

[3] In general, response latencies can also be used to study conflict detection (e.g., Botvinick, 2007; De Neys & Glumicic, 2008; Pennycook, Fugelsang, & Koehler, 2015). However, we refrained from focusing on response latencies in the current context given that they have been found to be a less reliable conflict indicator in the moral reasoning domain (Bialek & De Neys, 2017). An overview of the latency data can be found in the Supplementary Material C, Table S3.

The results of our first three studies question the corrective dual process assumption. In Study 4 we introduced additional design modifications to test the robustness of the findings. First, we changed the dilemma question. In Study 1-3 participants were asked about their willingness to act in the dilemma (e.g., "would you pull the switch?"). As one of our reviewers noted, although this question format is not uncommon, most moral reasoning studies have asked whether participants find the described action morally acceptable (e.g., "do you think it is morally acceptable to pull the switch?"). Literature suggests that the question format can affect moral decisions (e.g., Baron, 1992; Patil, Cogoni, Zangrando, Chittaro, & Silani, 2014; Tassy et al., 2013). Hence, in theory, it is possible that the Study 1-3 non-correction findings would be driven by the dilemma question.

Second, in Study 1-3 we used a variety of dilemmas with a range of absolute utilitarian response rates and emotional averseness of the scenarios. Nevertheless, one might note that all our scenarios concerned cases in which the sacrifice that the agent had to make did not involve physical contact or the use of personal force. Such scenarios have traditionally been labeled "impersonal harm" scenarios and can be contrasted with "personal harm" scenarios (e.g., pulling switch vs pushing a man on the track, e.g., Greene et al., 2001; but see also Greene, 2009). In Study 4 we included additional "personal harm" scenarios from the work of Greene and colleagues to test the generalizability of the non-correction findings.

Finally, in Study 4 we also used an even more demanding load task (Trémolière et al., 2012) and response deadline in the initial response stage to further minimize the possibility that the initial responses resulted from deliberate processing.

Method

Participants

In total, 101 people (59 female, Mean age = 31.5 years, SD = 10.8 years) were tested. Participants were recruited online on the Prolific platform. A total of 40.6% of the participants reported high school as highest completed educational level, while 57.4% reported having a post-secondary education degree (2% reported to have an education level less than high school). Participants received £1.25 for taking part.

Materials and procedure

Moral reasoning task. Participants were presented with a total of 12 dilemmas. These included the four conflict and four no-conflict items from Study 2 (family member/moderate ratio condition). We also presented four additional "high conflict personal dilemmas" (i.e., "Crying baby", "Lawrence of Arabia", "Submarine", & "Sacrifice") from the study of Greene et al. (2008). In these personal harm dilemmas the sacrifice involved a "personal force" of the actor (e.g., "smothering a child", "beheading someone", etc., see Supplementary Material A for an overview). In all dilemmas participants were asked about the moral acceptability of the sacrifice (e.g., "Is it morally acceptable for you to do X?"). Except for the specific deadline and load task (see below) the two-response procedure was similar to the Study 1-3 design.

The 12 dilemmas were presented in random order. Note that in Study 4 we did not present a filler item. Participants were also not asked to provide a confidence judgment after they entered their responses.

Response deadline. In Study 1-3 the response deadline for the initial response was set to 12 s based on our reading and one-response pretests. The initial response latencies in Study 1-3 (see Figure 2 top panel) indicated that on average participants managed to respond before the deadline (e.g., average initial conflict response latency = 7.8 s, SD = 1.58). On the basis of these data we decided to try to decrease the deadline further to 10 s. Note that as in Study 1-3, scenarios were presented in two parts (first background information followed by the second part with the critical conflicting or non-conflicting dilemma information and dilemma question). Hence, as in Study 1-3, the allotted response time comprised both the time needed to read the second part and enter a response. It should be clear that this entails a challenging deadline (i.e., less than one-response average minus 2 SDs, see Study 1 pretest). This should further minimize the possibility that participants engaged in deliberate reasoning during the initial response phase. Since the additional personal harm dilemmas had approximately the same length as our other materials we decided to stick to the same deadline. As in our previous studies, 2 s before the deadline the screen background turned yellow to urge participants to enter their response.

Load task. In Study 4 we also used a more demanding load task during the initial response stage. In Study 1-3 participants had to memorize a complex 4-dot pattern in a 3x3 grid during the initial response stage. In Study 4 we presented an even more complex 5-dot pattern in a 4x4 grid (e.g., Bialek & De Neys, 2017; Trémolière & Bonnefon, 2014). Trémolière et al. (2012) established that this 5-dot pattern task was more effective at hindering utilitarian responding during moral reasoning. Except for the precise load pattern the load procedure was similar to Study 1-3. The pattern was shown for 2 s before the dilemma was presented. After participants had entered their initial response, they were shown four different matrixes and they had to choose the correct, to-be-memorized pattern. As in Trémoliere et al.

(2012), all response options showed an interspersed pattern with 5 dots. There was always one incorrect matrix among the four options that shared 3 out of the 5 dots with the correct matrix. The two other incorrect matrices shared one of the dots with the correct matrix. Participants received feedback as to whether they chose the correct or incorrect pattern.

Exclusion criteria. The same exclusion criteria were applied as in Study 1-3. Participants failed to provide a first response before the deadline in 9.98% of the trials. In addition, in 17.7% of the trials participants responded incorrectly to the dot memorization load task. All these trials (25.1% of trials in total) were excluded and 908 trials (out of 1212) were further analyzed (initial and final response for the same item counted as 1 trial). Note that the proportion of excluded trials is slightly higher in Study 4 than in Study 1-3 (i.e., 25.1% vs approximately 15%) which presumably reflects the higher task demands.

Results and discussion

Results are presented in Table 1 and 2. Findings for the no-conflict problems—in which choosing the greater good did not entail a sacrifice— show that although the initial utilitarian response rate is slightly lower than what we observed with the similar scenario content in Study 2, χ^2 (1) = 7.55, p = 0.006, b = -1.74, it remains high at 84.8%. This indicates that participants managed to read the scenarios and did not simply enter random responses in the initial response stage[4]. Interestingly, Table 1 further shows that on our "family member/moderate ratio" conflict items, both the initial, χ^2 (1) = 34.39, $p <$ 0.0001, b = 2.29, and final, χ^2 (1) = 49.16, $p < 0.0001$, b = 3.65, utilitarian response rates are clearly higher than what we observed in Study 2. Consistent with Baron (1992) this might indicate that moral acceptability judgments (vs willingness to ask questions) result in increased utilitarian responding (but see Tassy et al., 2013). Alternatively, the difference might result from the different composition of the sample (i.e., university students vs online workers). Finally, one might note that the final utilitarian response rate on the "Greene personal harm" conflict scenarios is in line with what was reported by Greene et al. (2008).

However, the key question concerns the direction of change findings and the critical non-correction rate. Table 2 shows the results. As the table indicates, the overall higher utilitarian response rate on our Study 4 "family member/moderate ratio" dilemmas is reflected in a higher rate of both "UU" and "DU" responses. Indeed, the non-correction rate is similar to what we observed in Study 2. In the vast

[4] This is further confirmed by the stability index on the Study 4 conflict problems (see Table S2). As in all our studies, the direction of change pattern is stable with an average value of +71.1% (Greene personal harm items) and 78.9% (family-moderate ratio items).

majority of cases (76.3%), final utilitarian responses are preceded by initial utilitarian responses. Table 2 further indicates that the high non-correction rate (85.4%) is also observed on the personal harm problems. Hence, even on the traditional "high conflict personal harm" dilemmas that have been assumed to be a paradigmatic example of the utilitarian correction process, we observe that in the majority of cases utilitarian responders do not need to deliberate to correct an initial deontological intuitive response.

In sum, the Study 4 findings validate the Study 1-3 results. The alternative question framing, use of personal harm scenarios, and more demanding deadline and load did not affect the key non-correction findings. The fact that the high non-correction rate is consistently observed with different scenario and design characteristics indicates that the finding is robust.

GENERAL DISCUSSION

Our studies tested the claim that utilitarian responses to moral dilemmas require deliberate System 2 correction of an initial, intuitive deontological System 1 response. By adopting a two-response paradigm in which participants were required to give an initial response under time-pressure and cognitive load we aimed to empirically identify the intuitively generated response that preceded the final response given after deliberation. We ran four studies in which we tested a range of conflict dilemmas that gave rise to various absolute levels of utilitarian responding, including "high-conflict" cases in which there was a strong emotional averseness towards the sacrificial option. Our critical finding is that although there were some instances in which deliberate correction occurred, these were the exception rather than the rule. Across the studies, results consistently showed that in the vast majority of cases in which people opt for a utilitarian response after deliberation, the utilitarian response is already given in the initial phase. Hence, pace the corrective dual process assumption, utilitarian responders do not necessarily need to deliberate to correct an initial deontological response. Their intuitive response is typically already utilitarian in nature.

Our two-response findings point to the pervasiveness of an intuitive utilitarianism in which people intuitively prefer the greater good without any deliberation. One might note that the idea that utilitarian reasoning can be intuitive is not new. As Bialek and De Neys (2017) noted, at least since J. S. Mill various philosophers have characterized utilitarianism as a heuristic intuition or rule of thumb. At the empirical level, Kahane (2012, 2015; Wiech et al., 2013) demonstrated this by simply changing the severity of the deontological transgression. Kahane and colleagues showed that in cases where the deontological duty is trivial and the consequence is large (e.g., when one needs to decide whether it is acceptable to tell a lie in order to save someone's life) the utilitarian decision can be made intuitively. Likewise, Trémolière and Bonnefon (2014) showed that when the kill-save ratios (e.g., kill 1 to save

5000) were exceptionally inflated, people effortlessly made the utilitarian decision even when they were put under cognitive load. Hence, one could argue that these earlier empirical studies established that at least in some exceptional or extreme scenarios utilitarian responding can be intuitive. What the present findings indicate is that there is nothing exceptional about intuitive utilitarianism. The established high non-correction rate in the present studies implies that the intuitive generation of a utilitarian response is the rule rather than the exception. Moreover, the non-correction was observed in standard dilemmas with moderate, conventional kill-save ratios and severe deontological transgressions (i.e., killing) that were used to validate the standard dual process model of moral cognition. This indicates that utilitarian intuitions are not a curiosity that result from extreme or trivial scenario content but lie at the very core of the moral reasoning process (Bialek & De Neys, 2017).

Critiques

Critics of our work might argue that our conclusions only hold insofar as our methodology is effective at blocking deliberation during the initial response phase. A proponent of the corrective dual process model of moral reasoning can always try to argue that our methods were not demanding enough and reasoners still managed to successfully deliberate during the initial response stage. We anticipated this critique and therefore created one of the most challenging test conditions that have been used in the moral reasoning and dual process literature to date. To recap, previous work has used instruction (e.g., Thompson et al., 2011), time-pressure (e.g., Suter & Hertwig, 2011), or load manipulations (e.g., Amit & Greene, 2012; Trémolière et al., 2012) to isolate intuitive and deliberate processing. Each of these methods has been shown to interfere with deliberate thinking (Trémolière et al., 2018). In the present study we combined all three of them. We validated our specific deadline in two pretests (i.e., a reading and one-response pretest). In Study 4 we adopted an even more challenging load task and deadline to further minimize the possibility that reasoners would deliberate about their initial answer. The critical non-correction findings were consistent across our studies. These features make it highly unlikely that participants managed to successfully deliberate during the initial response phase.

Nevertheless, one might note that in all our studies participants tended to respond several seconds before the deadline. A critic could argue that this indicates that our participants still had ample time to deliberate. However, there are clear counterarguments against this specific suggestion. First, let us note that logically speaking the absolute response time of an answer generated under time pressure cannot be used to argue against its intuitive nature. People are instructed to respond as fast as possible with the first answer that comes to mind. In order to encourage this we set a deadline. Two seconds before the deadline people are alerted to it (i.e., screen is colored yellow). If the deadline is missed, the trial is excluded. By

definition, our response times will always be shorter than the deadline. This does not logically imply that participants were deliberating.

Second, and more critically, let us assume that that the critique is right. Consistent with the traditional corrective dual process view, our initial utilitarian responders would have first generated an intuitive deontological response but–despite the time pressure and load–would afterwards still have had time to engage in additional deliberation and replace it with a utilitarian response. However, in this case we should have observed that the initial utilitarian responses (which are assumed to result from additional time-demanding deliberation) take longer than the initial deontological responses (which are assumed to be truly intuitive). Our data show that this was not the case. Figure 2 already indicated that initial utilitarian and deontological response times do not differ. To test this directly, in an additional analysis we contrasted the response times for initial utilitarian and initial deontological responses across all our studies. Results clearly show that they do not differ ($n = 2061$, mean utilitarian response = 6.95s, mean deontological response = 7.04s; $\chi 2\ (1) = 0.1$, $p = 0.75$). This further argues against the claim that our initial utilitarian responses result from deliberation.

To avoid confusion, it is important to clarify that our conclusions hold both under a so-called serial and parallel interpretation of the dual process model. The serial and parallel processing view concern specific assumptions about the time-course of the System 1 and 2 interaction in dual process models (e.g., see Bialek & De Neys, 2017; De Neys, 2017, for an overview). The serial view model entails that at the start of the reasoning process only System 1 is activated by default. System 2 can be activated but this activation is optional and occurs later in the reasoning process (e.g., Evans & Stanovich, 2013; Kahneman, 2011). The parallel view (e.g., Sloman, 1996) entails that both System 1 and System 2 are activated simultaneously from the start. Hence, the serial model explains the fact that people did not give the alleged deliberate response (e.g., utilitarian response during moral reasoning) by assuming they did not engage System 2. The parallel model explains it by assuming that the System 2 computations were not completed by the time that the fast System 1 already finished computing a response.

However, what is critical is that although the serial and parallel view differ on when System 2 processing is assumed to start, both make the corrective assumption and hypothesize that computing the alleged (e.g., utilitarian) System 2 response will take time and effort. Consequently, even if both deliberative and intuitive processes start simultaneously, by definition, generating the deliberate response should still take longer. This is as fundamental an assumption of the parallel model as it is of the serial model. Hence, switching from a serial to a parallel model conceptualisation does not help to account for the present findings. Limiting response time and putting people under load in our initial response phase should make it even less likely that reasoners will manage to complete the deliberate process. In case they somehow managed to pull this off (pace previous evidence that validates the effectiveness of the load and

deadline procedure), it should come at the cost of additional processing time. Since deliberation is expected to run slower, response times for utilitarian responses should be relatively longer than deontological responses. As we clarified, there was not the slightest hint of such a trend in our data.

Taken together, there is no good evidence to claim that the utilitarian responses in our initial response stage result from deliberate processing.

Towards a new dual process model of moral cognition

The evidence in favor of intuitive utilitarianism and against the corrective assumption forces us to revise the dual process model of moral cognition. So what type of model or architecture do we need to account for the present findings? We already clarified that neither the serial, nor the parallel dual process variant is a viable option. However, an interesting recent alternative to the more traditional serial and parallel models is the so-called "hybrid[5]" model view (Bago & De Neys, 2017; Ball, Thompson, & Stupple, 2017; Banks, 2017; Białek & De Neys, 2017; De Neys, 2012, 2017; Handley & Trippas, 2015; Pennycook, Fugelsang, & Koehler, 2015; Thompson & Newman, 2017; Thompson, Pennycook, Trippas & Evans, 2018; Trippas & Handley, 2017, see also: Stanovich, 2018). Put bluntly, at the most general level this model simply entails that the response that is traditionally assumed to be cued by System 2 can also be cued by System 1. Hence, in the case of moral reasoning the idea is that System 1 is simultaneously generating both a deontological and utilitarian intuition (e.g., Bialek & De Neys, 2016, 2017; see also Gürçay & Baron, 2017; Rosas, 2017). This allows us to account for the fact that utilitarian responses can be intuitive and non-corrective in nature. However, this does not suffice. The key challenge for the dual process model is to account for the direction of change results. Indeed, although we observed that final utilitarian responders predominantly generate the utilitarian response intuitively, many reasoners did not generate utilitarian responses and stuck to a deontological response throughout. Likewise, there were also cases in which correction occurred and the utilitarian response was only generated after deliberate correction. How can we explain these different response patterns?

Here it is important to underline that the hybrid model—such as it has been presented in the logical/probabilistic reasoning field—posits that although System 1 will generate different types of intuitions, this does not entail that all these intuitions are equally strong (Bago & De Neys, 2017; Pennycook, 2017; Pennycook et al., 2015; Thompson et al. 2017). They can vary in their strength or activation level. More specifically, the model proposes that we need to consider both absolute (which one of the two intuitions is strongest?) and relative (how pronounced is the activation difference between both

[5] We use the "hybrid" model label to refer to core features that seem to be shared – under our interpretation – by the recent theoretical proposals of various authors. It should be clear that this does not imply that these proposals are completely similar. We are talking about a general family resemblance rather than full correspondence and focus on commonalities rather than the differences.

intuitions?) strength differences between competing intuitions (Bago & De Neys, 2017). The initial response will be determined by the absolute strength level. Whichever intuition is strongest will be selected as initial response. Whether or not the initial response gets subsequently deliberately changed will be determined by the relative strength difference between both intuitions. The smaller the difference, the less confident one will be, and the more likely that the initial response will be changed after deliberation. Bago and De Neys (2017) already showed that such a model accounted for the two-response findings in logical/probabilistic reasoning. Here we propose to apply the same principles to the moral reasoning case.

Figure 3 illustrates the idea. In the figure we have plotted the hypothetical strength of the utilitarian and deontological intuition for each of the four direction of change categories in imaginary activation strength "units". For example, in the UU case, the utilitarian intuition might be 4 units strong whereas the deontological intuition might be only 1 unit strong. In the DD case, we would have the opposite situation with a 4 unit strong deontological intuition and a much weaker, 1 unit utilitarian intuition. In the two change categories, one of the two intuitions will also dominate the other but the relative difference will be less pronounced. For example, in the DU case the deontological intuition might have strength level 3 whereas the utilitarian intuition has strength level 2. Because the relative difference is less pronounced, there will be more doubt and this will be associated with longer final rethinking and answer change. In other words, in each of the four direction of change categories there will be differences in which intuition is the dominant one and how dominant the intuition is. The more dominant an intuition is, the more likely that it will be selected as initial response, and the less likely that it will be corrected by deliberate System 2 processing.

It should be clear that Figure 3 presents a hypothetical model of the strength levels. The strength levels were set to illustrate the core hybrid model principles. However, the principles themselves are general and were independently established in the logical/probabilistic reasoning field. The key point is that by allowing utilitarian intuitions within System 1 and considering strength differences between competing System 1 intuitions we can readily explain why utilitarian responses can be generated intuitively and why sometimes people will correct their initial responses after deliberation.

The hybrid model illustrates how one can make theoretical sense of the observed findings. Interestingly, it also makes new predictions. That is, given the core principles one can expect that changes in the strength levels of competing intuitions should lead to predictable consequences. For example, our studies showed that the family member manipulation had a profound impact on the rate of utilitarian responding. It is not unreasonable to assume that putting the life of a close family member at stake will increase the strength of the deontological intuition. It follows from the hybrid model principles that the prospect of sacrificing a family member should not only decrease the utilitarian response rate (which we

observed but is fairly trivial) but also affect the associated response confidence. Consider two reasoners in the family and no family condition who both give an intuitive deontological response. The fact that they give an initial deontological response implies that the absolute strength of their deontological intuition dominates their utilitarian intuition. Putting the life of a family member at stake in the family condition will further increase the strength of the deontological intuition. Hence, for a deontological responder in the family condition the strength difference with the competing utilitarian intuition will increase. Therefore, for a deontological responder it should become less likely to experience conflict in the family vs no-family condition, and their response should be doubted less.

Furthermore, based on the same principles, one can expect that the strength manipulation should have the exact opposite impact for intuitive utilitarian responders. The fact that someone gives the utilitarian response implies that the absolute strength of their utilitarian intuition will dominate their deontological intuition. However, since putting the life of a family member at stake will increase the strength of the deontological intuition, the relative difference between the two intuitions will be smaller for the utilitarian responder in the family condition. That is, the utilitarian responder in the family condition will now face a deontological intuition which strength is closer to their utilitarian intuition strength. Consequently, given the smaller difference, they should experience more conflict and show more response doubt. Figure 4 plots the average confidence data for initial deontological and utilitarian responses in the family and no family conditions across our studies. The expected trend is indeed observed[6]. Making the deontological intuition stronger makes utilitarian responders less confident about their decision (i.e., 19.6% decrease) whereas deontological responders grow more confident (i.e., 15.6% increase). Statistical testing indicated that this interaction was significant, χ^2 (3) = 76.63, $p < 0.0001$, b = -25.7.

In sum, we hope to have demonstrated how an application of the hybrid dual process principles can account for the observed findings—and makes testable predictions. We believe this underlines the potential of the hybrid model view as an alternative to the traditional dual process model. At the very least the present studies should make it clear that the traditional corrective model is untenable. Although the hybrid model will need to be further validated and developed the present studies indicate that it core principle stands: Any viable dual process model of moral cognition will need to allow for the generation of both utilitarian and deontological intuitions within System 1 and consider competition between these intuitions.

Why do we need System 2 deliberation?

[6] A related, albeit less pronounced trend, can be observed for the kill-save manipulation (see Supplementary Material D).

The hybrid model and the evidence for intuitive utilitarianism imply that we need to upgrade the role of System 1: Utilitarian judgments do not necessarily require System 2 deliberation but can be generated by System 1. Here it is important to stress that upgrading the role of System 1 does not imply a downgrade of System 2. First, in all our studies we observed that correction does sometimes occur. Hence, although it is more exceptional, System 2 *can* be used to deliberately correct one's intuitive response. Second, and more critically, the fact that deliberation is not typically used for correction does not imply it cannot be important for other functions. For example, one of the features that is often associated with deliberation is its *cognitive transparency* (Bonnefon, 2016). Deliberate decisions can typically be justified; we can explain why we opt for a certain response after we reflected on it. Intuitive processes often lack this explanatory property: People have little insight into their intuitive processes and do typically not manage to justify their "gut-feelings" (Marewski & Hoffrage, 2015; Mega & Volz, 2014). Hence, one suggestion is that people might be using deliberation to look for an explicit justification or validation of their intuitive insight (Bago & De Neys, 2018). For example, Bago and De Neys (2018) observed that although reasoners could often intuitively generate the correct solution to logical reasoning problems, they struggled to properly explain why their answer was correct. Such justifications were more likely after people were given the time to deliberate. A similar process might be at play during moral reasoning. In the Supplementary Material (section E) we present the results of an exploratory pilot study in which people were given moral dilemmas and were asked to give a justification after both their initial and final response. We were specifically interested in proper utilitarian justifications that explicitly mentioned the greater good (e.g., "I opted for this decision because more people will be saved"). The study replicated the finding that final utilitarian decisions were typically preceded by initial utilitarian responses (i.e., high non-correction rate). Critically, however, proper utilitarian justifications for a utilitarian response were more likely in the final response stage (i.e., up to +20% increase when the life of a family member was at stake[7]). Hence, although utilitarian responses can be generated intuitively, additional deliberation might make it more likely that we will manage to properly justify it.

In general, being able to justify one's response and producing explicit arguments to support it might be more crucial for reasoning than it was often believed to be in the past (Mercier & Sperber, 2011, 2017). The work of Mercier and Sperber, for example, underscores that arguments are critical for communicative purposes. We will not be very successful in convincing others that our decision is acceptable, if we can only tell them that we "feel it is right". If we come up with a good explanation, people will be more likely to change their mind and accept our view (Trouche, Sander, & Mercier, 2014;

[7] One limitation of the study is that participants can use the justification to deliberate about their initial response. This might inflate proper utilitarian justifications at the initial response phase. However, the point is that despite this limitation we still observed an increase in utilitarian justifications in the final response phase.

but see also Stanley, Dougherty, Yang, Henne, & De Brigard, 2018). If System 2 deliberation plays a role in this process, it should obviously not be downplayed.

Interestingly, at least one tradition within moral reasoning research has characterized deliberate justifications as post hoc constructions or "rationalizations" (Haidt, 2001). This "social intuitionist" approach has stressed the primacy of intuitive processes for moral reasoning. By and large, moral reasoning would be driven by mere intuitive processes. Interestingly, the traditional dual process model of moral cognition reacted against this "intuitionist" view by arguing that corrective deliberate processes were also central to moral reasoning (Greene & Haidt, 2002). By presenting evidence against the corrective assumption the current paper might seem to support the social intutionist framework. We simply want to highlight here that although the hybrid model shares the upgraded view of intuitive processes, it does not conceive deliberation as epiphenomenal or extrinsic to the reasoning process. Whatever one's position in this debate might be, our point here is that that the case against the corrective dual process assumption should not be taken as an argument against the role or importance of deliberation in human cognition. Our goal is not to contest that deliberation might be important for human reasoning. The point is simply that this importance does not necessarily lie in a correction process.

In closing

Finally, we want to highlight the close link between the current work on moral reasoning and related dual process work in the logical reasoning field. As we noted, our two-response paradigm and the theoretical hybrid dual process model we proposed were inspired by recent dual process advances on logical reasoning. In the past, dual process research in the moral and logical reasoning fields has been occurring in somewhat isolation (Bonnefon & Trémolière, 2017; Gürçay & Baron, 2017) and we hope that the present study can stimulate a closer interaction (Białek & De Neys, 2017; Gürçay & Baron, 2017; Trémolière, De Neys, & Bonnefon, 2018). In our view, such interaction is the critical stepping stone to arrive at a unified domain-general model of the interplay between intuitive and deliberate processes in human cognition. Our evidence against the corrective dual process view suggests that such a model will need to be build on a hybrid processing architecture in which absolute and relative strength differences between competing intuitions determine our reasoning performance.

ACKNOWLEDGMENTS

Bence Bago was supported by a fellowship from the Ecole des Neurosciences de Paris Ile-de-France and the Scientific Research Fund Flanders (FWO-Vlaanderen). This research was also supported by a research grant (DIAGNOR, ANR-16-CE28-0010-01) from the Agence National de la Recherche. We like to thank

Balazs Aczel for granting us access to his university participant pool. All raw data can be retrieved from https://osf.io/6bw8n/?view_only=6ff5c8012c9c433b8d9999a64adf7851

REFERENCES

Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means: Visual imagery and moral judgment. *Psychological science*, *23*, 861-868.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

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

Bago, B., & De Neys, W. (2018). The smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning.* Advance online publication.

Ball, L., Thompson, V., & Stupple, E. (2017). Conflict and dual process theory: the case of belief bias. In W. De Neys (Ed.), *Dual Process Theory 2.0*. Oxon, UK: Routledge.

Baron, J. (1992). The effect of normative beliefs on anticipated emotions. *Journal of Personality and Social Psychology, 63,* 320-330.

Banks, A. (2017). Comparing dual process theories: evidence from event-related potentials. In W. De Neys (Ed.), *Dual Process Theory 2.0*. Oxon, UK: Routledge.

Baron, J. (2017). Utilitarian vs. deontological reasoning: method, results, and theory. In J.-F. Bonnefon & B. Trémolière (Eds.), *Moral inferences* (pp. 137–151). Hove, UK: Psychology Press.

Baron, J., & Gürçay, B. (2017). A meta-analysis of response-time tests of the sequential two-systems model of moral judgment. *Memory & Cognition*, *45*(4), 566–575.

Baron, J., 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*(3), 265–284.

Białek, M., & De Neys, W. (2016). Conflict detection during moral decision-making: evidence for deontic reasoners' utilitarian sensitivity. *Journal of Cognitive Psychology*, *28*(5), 631–639.

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

Bonnefon, J.-F. (2016). The Pros and Cons of Identifying Critical Thinking with System 2 Processing. *Topoi*, 1–7.

Bonnefon, J.-F., & Trémolière, B. (Eds.). (2017). *Moral Inferences*. Oxon, UK: Routledge.

Botvinick, M. M. (2007). Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(4), 356–366.

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.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089.

De Neys, W. (2012). Bias and conflict a case for logical intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38.

De Neys, W. (2017). Bias, conflict, and fast logic: Towards a hybrid dual process future? In W. De Neys (Ed.), *Dual Process Theory 2.0*. Oxon, UK: Routledge.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load. *Experimental Psychology (Formerly Zeitschrift Für Experimentelle Psychologie)*, *54*(2), 128–133.

De Neys, W., & Verschueren, N. (2006). Working memory capacity and a notorious brain teaser: The case of the Monty Hall Dilemma. *Experimental Psychology*, *53*(2), 123–131.

Dolgin, E. (2011). World's most expensive drug receives second approval for deadly blood disease. *Nature Medicine*. Retrieved from http://blogs.nature.com/spoonful/2011/09/soliris.html

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278.

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

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking & Reasoning*, *15*(2), 105–128.

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in cognitive sciences*, *11*, 322-323.

Greene, J. (2013). *Moral tribes: emotion, reason and the gap between us and them*. New York, NY: Penguin Press.

Greene, J. D. (2009). The cognitive neuroscience of moral judgment. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (4th ed., pp. 987–999). Cambridge, MA: MIT Press.

Greene, J. D. (2015). Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. *The Law & Ethics of Human Rights*, *9*(2), 141–172.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144-1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, *6*(12), 517–523.

Gürçay, B., & Baron, J. (2017). Challenges for the sequential two-system model of moral judgement. *Thinking & Reasoning*, *23*(1), 49–80.

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

Handley, S. J., & Trippas, D. (2015). Chapter Two-Dual Processes and the Interplay between Knowledge and Structure: A New Parallel Processing Model. *Psychology of Learning and Motivation*, *62*, 33–58.

Hao, J., Liu, Y., & Li, J. (2015). Latent Fairness in Adults' Relationship-Based Moral Judgments. *Frontiers in Psychology*, *6*, 1871.

Kahane, G. (2012). On the wrong track: Process and content in moral psychology. *Mind & Language*, *27*(5), 519–545.

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*, *10*(5), 551–560.

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

Kant, I. (1785). *Groundwork for the metaphysics of morals*. New Haven, CT: Yale University Press.

Koop, G. J. (2013). An assessment of the temporal dynamics of moral decisions. *Judgment and Decision Making*, *8*(5), 527.

Kruglanski, A. W. (2013). Only one? The default interventionist perspective as a unimodel—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, *8*(3), 242–247.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmerTest'.

London, J. A. (2012). How should we model rare disease allocation decisions? *Hastings Center Report*, *42*(1), 3.

Marewski, J. N., & Hoffrage, U. (2015). Modeling and aiding intuition in organizational decision making. *Journal of Applied Research in Memory and Cognition*, *4*, 145–311.

Mega, L. F., & Volz, K. G. (2014). Thinking about thinking: implications of the introspective error for default-interventionist type models of dual processes. *Frontiers in Psychology*, *5*.

Mill, J. S., & Bentham, J. (1987). *Utilitarianism and other essays*. Harmondsworth, UK: Penguin.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*(4), 621–640.

Moore, A. B., Stevens, J., & Conway, A. R. (2011). Individual differences in sensitivity to reward and punishment predict moral judgment. *Personality and Individual Differences*, *50*(5), 621–625.

Newman, I., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief -bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1154–1170.

Nichols, S. (2004). Folk concepts and intuitions: From philosophy to cognitive science. *Trends in Cognitive Sciences*, *8*(11), 514–518.

Patil, I., Cogoni, C., Zangrando, N., Chittaro, L., & Silani, G. (2014). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social Neuroscience, 9*, 94-107.

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

Pennycook, G. (2017). A perspective on the theoretical foundation of dual process models. In W. De Neys (Ed.), *Dual Process Theory 2.0*. Oxon, UK: Routledge.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, *80*, 34–72.

Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. *Psychonomic Bulletin & Review*, *19*(3), 528–534.

Rosas, A. (2017). On the Cognitive (Neuro) science of Moral Cognition: Utilitarianism, Deontology, and the "Fragmentation of Value." In A. Ibáñez, L. Sedeño, & A. García (Eds.), *Neuroscience and Social Science* (pp. 199–215). Springer.

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*(2), 165–184.

Schellens, G. (2015). Alexion deal with Belgian government got public. Retrieved from http://bbibber.blogspot.be/2015/03/alexion-deal-with-belgian-government.html

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.

Sloman, S. (2015). Opening editorial: The changing face of cognition. *Cognition*, *135*, 1–3.

Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 10393–10398.

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

Stanley, M. L., Dougherty, A. M., Yang, B. W., Henne, P., & De Brigard, F. (2018). Reasons probably won't change your mind: the role of reasons in revising moral decisions. *Journal of Experimental Psychology: General*. Advance online publication.

Stanovich. K. (in press). Miserliness in human cognition: the interaction of detection, override and mindware. *Thinking & Reasoning*.

Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology*, *4*, 250.

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244.

Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140.

Thompson, V. A., & Newman, I. (2017). Logical intuitions and other conundra for dual process theories. In W. De Neys (Ed.), *Dual Process Theory 2.0*. Oxon, UK: Routledge.

Thompson, V. A., Pennycook, G., & Trippas, D., & Evans, J. S. B. T. (in press). Do smart people have better intuitions? *Journal of Experimental Psychology: General*.

Tinghög, G., Andersson, D., Bonn, C., Johannesson, M., Kirchler, M., Koppel, L., & Västfjäll, 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), 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.

Trémolière, B., & De Neys, W. (2013). Methodological concerns in moral judgement research: Severity of harm shapes moral decisions. *Journal of Cognitive Psychology*, *25*(8), 989–993.

Trémolière, B., De Neys, W., & Bonnefon, J. F. (2012). Mortality salience and morality: Thinking about death makes people less utilitarian. *Cognition*, *124*(3), 379-384.

Trémolière, B., De Neys, W., & Bonnefon, J.-F. (2018). Reasoning and moral judgment: A common experimental toolbox. In L. J. Ball & V. A. Thompson (Eds.), *The Routledge International Handbook of Thinking and Reasoning*. Oxon, UK: Routledge.

Trippas, D., & Handley, S. (2017). The parallel processing model of belief bias: review and extensions. In W. De Neys (Ed.), *Dual Process Theory 2.0*. Oxon, UK: Routledge.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971.

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

Wiech, K., Kahane, G., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2013). Cold or calculating? Reduced activity in the subgenual cingulate cortex reflects decreased emotional aversion to harming in counterintuitive utilitarian judgment. *Cognition*, *126*(3), 364–372.

Table 1

Initial and final average percentage (SD) of utilitarian responses in study 1-4.

		Conflict		No-conflict	
		Initial	**Final**	**Initial**	**Final**
Study 1	No family - Moderate ratio	79.7% (40.3)	84.5% (36.2)	90.3% (29.6)	95.4% (20.9)
Study 2	Family - Moderate ratio	21.2% (40.9)	17.5% (38.1)	94.5% (22.8)	96.4% (18.6)
Study 3	No family - Extreme ratio	81.7% (38.7)	89.8% (30.3)	94.2% (23.4)	93.8% (24.2)
	Family - Extreme ratio	28.4% (45.2)	31.4% (46.5)	95.6% (20.1)	97.5% (15.7)
Study 4	Family - Moderate ratio	49.4% (50.1)	50.9% (50.1)	84.8% (36)	89% (31.3)
	Greene - Personal harm	64.1% (48)	65.3% (47.7)	-	-
Overall average		55.2% (49.7)	57.9% (49.4)	92.2% (26.8)	94.5% (22.7)

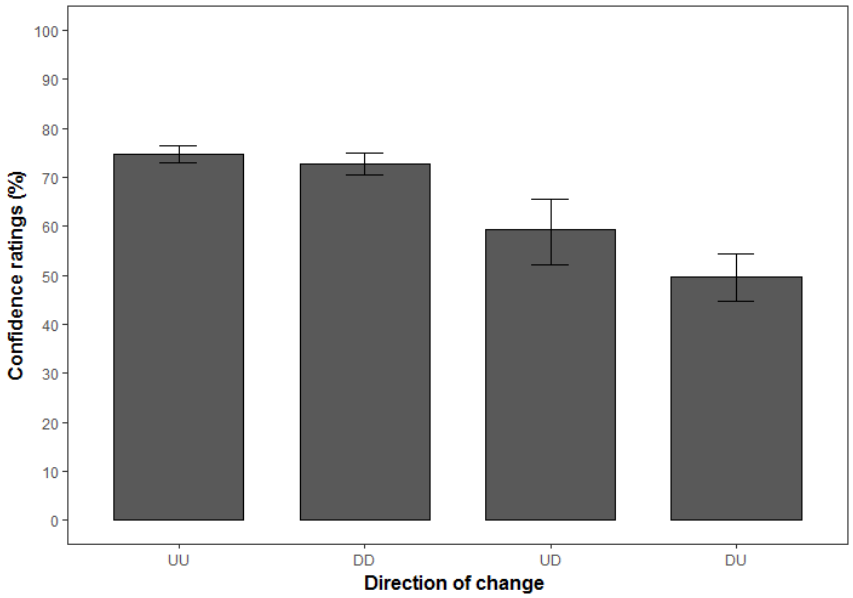
Table 2

Frequency of direction of change categories in study 1-4 for conflict problems. Raw number of trials are in brackets.

		Direction of change category				Non-correction rate
		UU	**DD**	**UD**	**DU**	**UU/(DU+UU)**
Study 1	No family - Moderate ratio	73.9% (258)	9.7% (34)	5.7% (20)	10.6% (37)	87.5%
Study 2	Family - Moderate ratio	12.7% (45)	74% (262)	8.5% (30)	4.8% (17)	72.6%
Study 3	No family - Extreme ratio	78.9% (332)	7.4% (31)	2.9% (12)	10.9% (46)	87.8%
	Family - Extreme ratio	22.8% (77)	63% (213)	5.6% (19)	8.6% (29)	72.6%
Study 4	Family - Moderate ratio	38.8% (106)	38.5% (105)	10.6% (29)	12.1% (33)	76.3%
	Greene - Personal harm	55.8% (182)	26.4% (86)	8.3% (27)	9.5% (31)	85.4%
Overall average		48.5% (1000)	35.5% (731)	6.7% (137)	9.4% (193)	83.8%

Note. U = utilitarian. D = Deontological.

A. Initial confidence



B. Final confidence

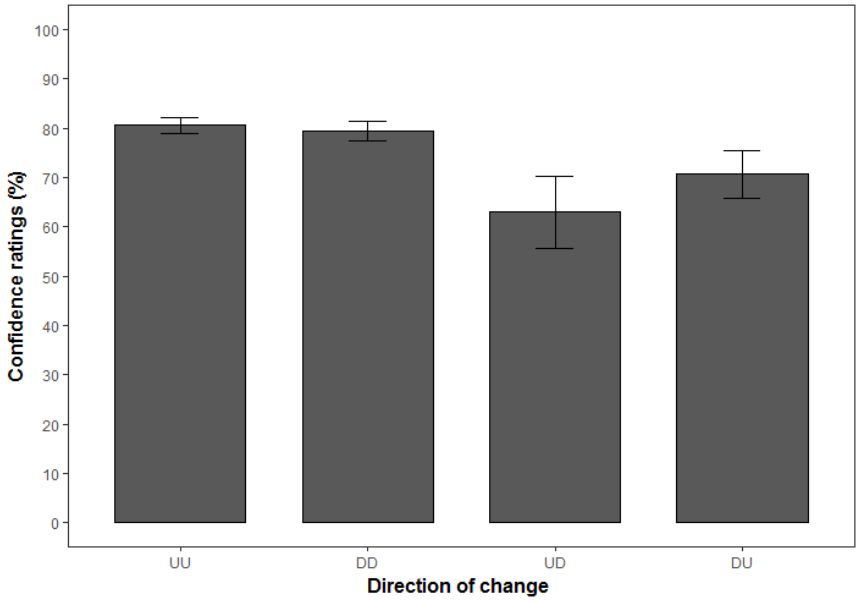
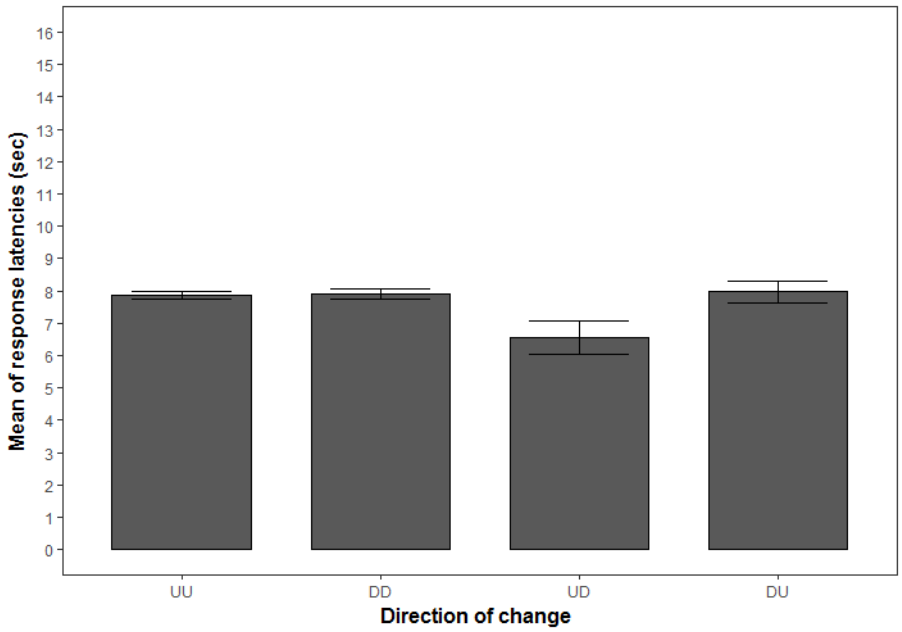


Figure 1. Mean initial (A.) and final (B.) conflict problem response confidence ratings as a function of direction of change category averaged across Study 1-3. Error bars are 95% confidence intervals.

A. Initial response time



B. Final response time

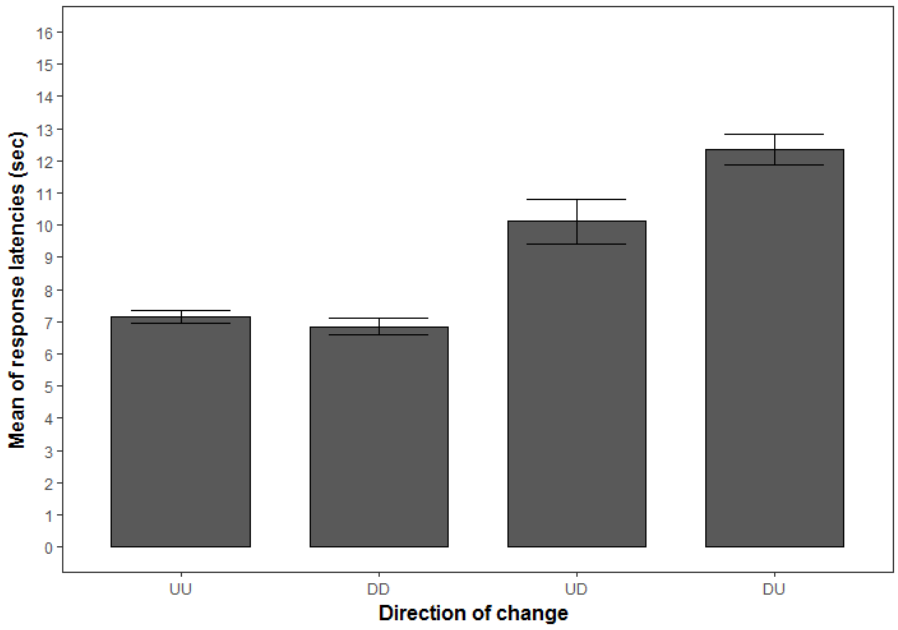


Figure 2. Mean initial (A.) and final (B.) conflict problem response times as a function of direction of change category averaged across Study 1-3. Error bars are 95% confidence intervals.

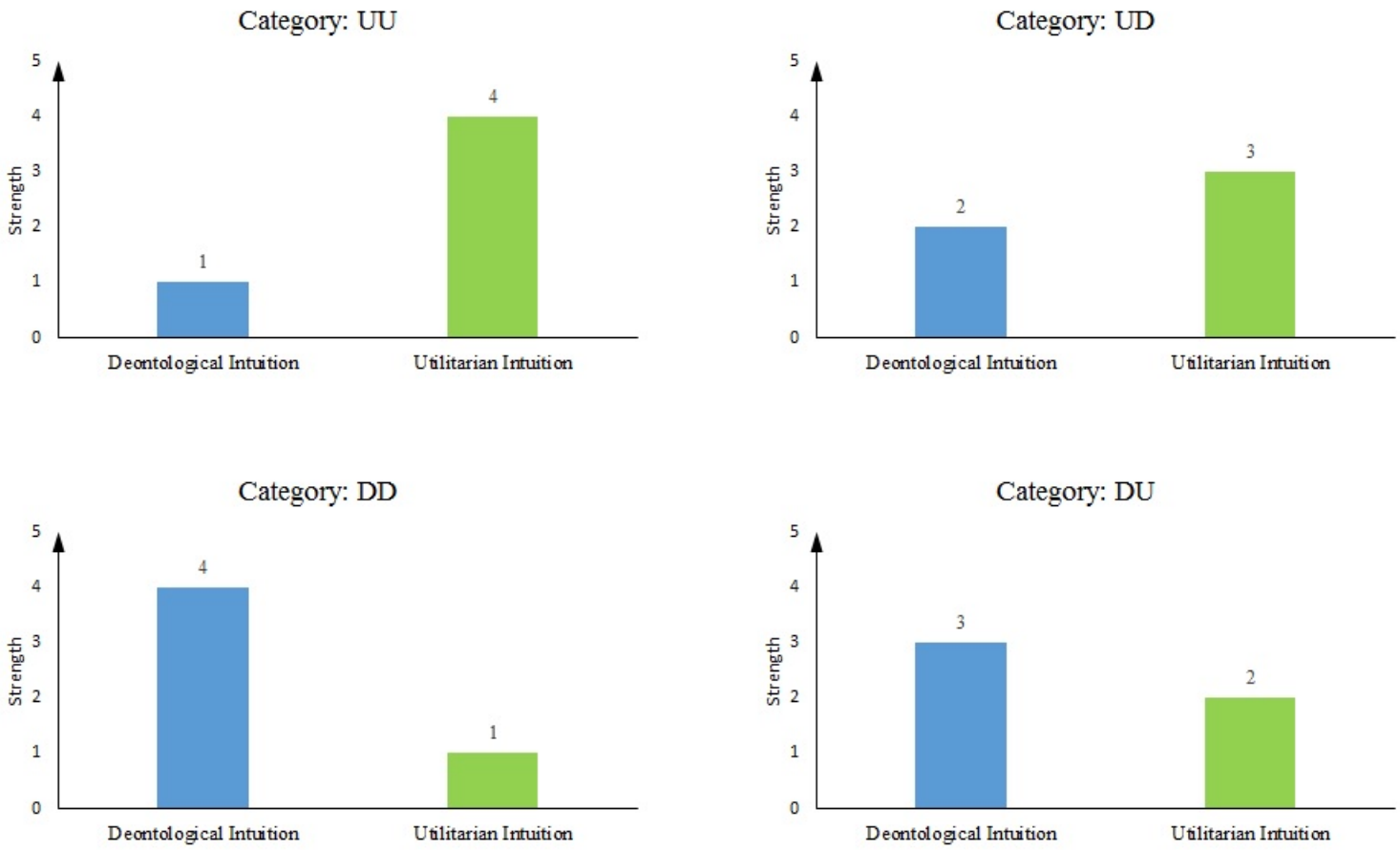


Figure 3. Illustration of a hybrid dual process model of moral cognition. Possible absolute (which one of the two intuitions is strongest?) and relative (how pronounced is the activation difference between both intuitions?) strength differences between the utilitarian and deontological intuition in the different direction of change categories. The figure shows the strength of the utilitarian and deontological intuition for each direction of change category in (imaginary) activation strength "units". Note: U = Utilitarian, D = Deontological.

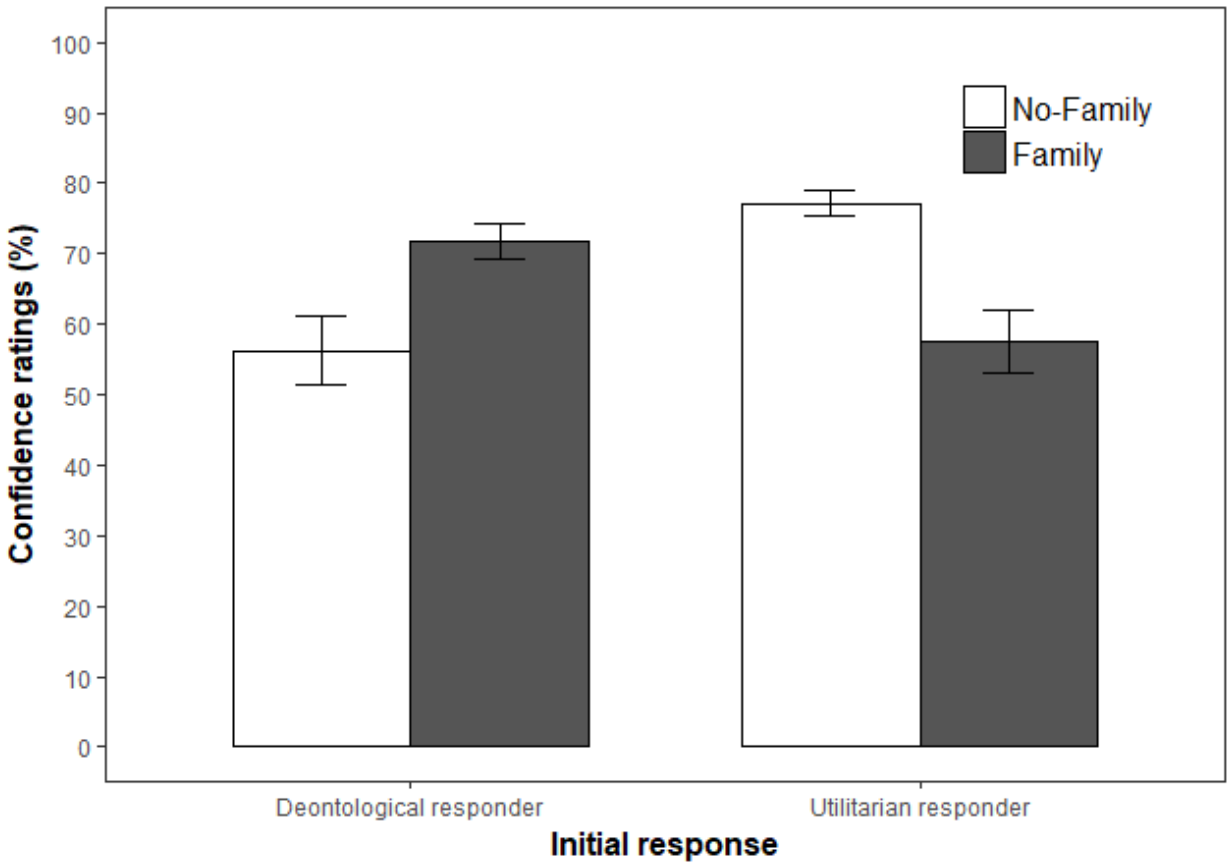


Figure 4. Mean initial conflict problem response confidence ratings for initial utilitarian and deontological responses as a function of the family member manipulation across Study 1-3. Error bars are 95% confidence intervals.

SUPPLEMENTARY MATERIAL

A. Scenario contents used in Study 1-4

We first present the "No family member" versions (Study 1, Study3) followed by the "Family member" versions (Study 2, Study 3), and the "Greene personal harm" dilemmas (Study 4). Numbers used in the "Extreme" kill-save ratio condition (Study 3) are in brackets next to the scenario name.

No family member versions (Study 1, Study3)

Iceberg (3/55)
You are radar operator overseeing vessel movement near Greenland. Due to sudden ice movement a boat carrying 11 passengers is about to crash into an iceberg. If nothing is done, all passengers will die.

The only way to save the 11 passengers is for you to order the captain to execute an emergency maneuver that will sharply alter the course of the boat.

However, the maneuver will cause the boat to overrun a life raft carrying 3 people that is floating next to the iceberg and out of sight of the captain. The 3 people on the life raft will be killed if you order the captain to execute the maneuver, but the 11 people on the boat will be saved.
Would you order to execute the manoeuvre?

Gas explosion (4/60)
You are the inspector of gas lines in a city. For some reason, the gas in a pipeline, which is running under a building, got on fire. If the fire reaches the building, it will explode, and will kill 12 people in it.

You realize that the only way to stop the explosion of the building is to close the pipeline by pushing on a button, and divert the fire to a side-pipeline. If you do so, the 12 people will be saved.

However, above the side-line is another building with 4 people in it. If you push the button and divert the fire into the sideline, this building will explode and kill the 4 people in it, but the 12 in the building above the main line will be saved.
Would you push the button?

Fumes (5/65)
You are the late-night watchman in a hospital. Due to an accident in the building next door, there are deadly fumes rising up through the hospital's ventilation system. The fumes are directly and quickly heading towards a room with 13 patients in it. If you do nothing the fumes will rise up into this room and cause their deaths.

The only way to avoid the deaths of these patients is to hit a certain switch, which will cause the fumes to bypass the room and enter a second room instead.

However, you realize that 5 patients are in the second room. These 5 patients will be killed if you hit the switch and let the fumes bypass the first room, but the 13 patients in the first room will be saved.
Would you hit the switch?

Airplane (5/65)
You are a military base commander. A missile has just been mistakenly fired at a commercial airliner. If you do nothing, the missile will reach the airliner and 13 people on the airliner will die.

You realize that the only way to save these people, is to alter the course of the commercial airliner. In this case, the missile will pass by the airliner and the 13 people inside will be saved.

However, if you alter the course of the commercial airliner, the missile will hit another airliner with 5 people inside which is flying right behind it. These 5 people who are travelling on this airliner will be killed if you alter the other's course, but the 13 people in the commercial airliner will be saved.
Would you alter the commercial airliner's course?

Submarine (4/60)
You are responsible for the mission of a submarine. You are leading this operation from a control center on the beach. An onboard explosion has damaged the ship and collapsed the only access corridor between the upper and lower levels of the ship. As a result of the explosion, water is quickly approaching to the upper level of the ship. If nothing is done, 12 people in the upper level will be killed.

You realize that the only way to save these people is to hit a switch in which case the path of the water to the upper level will be blocked and it will enter the lower level of the submarine instead.

However, you realize that 4 people are trapped in the lower level. If you hit the switch, the 4 people in the lower level (who otherwise would survive) will die, but the 12 people in the upper level will be saved.
Would you hit the switch?

Mine (3/55)
Due to an accident there are 11 miners stuck in one of the shafts of a copper mine. They are almost out of oxygen and will die if nothing is done. You are the leader of the rescue team.

The only way for you to save them is to activate an emergency circuit that will transfer oxygen from a nearby shaft into the shaft where the 11 miners are stuck.

However, your team notices that 3 miners are trapped in the nearby shaft. If you activate the emergency circuit to transfer the oxygen, these 3 miners will be killed, but the 11 miners will be saved.

Would you activate the emergency circuit?

Metro (2/50)
You are a late-night metro controller in a station. A metro with automated driver is heading towards the station. Right before the station, the track gets blocked by a huge rock. If nothing is done, the train crashes into the rock and 10 passengers will die.

You realize that the only way to save these people is to pull a lever and divert the metro to a short emergency side-track. However, there is another metro coming right after it. If you divert the first metro to the side-track the second one will crash into the rock, as there is only place on the side track for one metro. If you do nothing, the first metro crashes, and the second one will be automatically directed into the side-track.

However, you notice that 2 passengers are travelling in the second metro. If you pull the lever and let the first metro divert to the side-track, these 2 passengers will die, but the 10 people in the first metro will be saved.
Would you pull the lever?

Volcano (2/50)
After a volcano eruption deadly hot lava is heading towards a nearby village. You are directing the rescue operations. There are 10 people standing on the roof of a house. If nothing is done, these 10 people will inevitably be killed by the lava stream.

The only way to save these people is to order the construction of an emergency barrier that will divert the lava stream into an old river bed.

However, you suddenly receive the information that right along the old river bed 2 people are standing on the roof of a barn. If you order the construction of the emergency barrier and divert the lava to save the 10 people on the roof of the house, the 2 people on the roof of the barn will inevitably be killed.
Would you order the construction of the emergency barrier?

FILLER ITEM (same in both Family and No-family conditions)

Trolley (2/50)

You are a railway controller. There is a runaway trolley barrelling down the railway tracks. Ahead, on the tracks, there are 6 workmen. The trolley is headed straight for them and they will be killed if nothing is done.

You are standing some distance off in the train yard, next to a lever. If you pull this lever, the trolley will switch to a side track and you can save the 6 workmen on the main track.

You notice that there are no workmen on the side track. So there will be no workmen who will be killed if you pull the lever and change the tracks but the 6 workmen on the main track will be saved.

Would you pull the lever?

Family member versions (Study 2, Study 3)

Iceberg (3/55)
You are radar operator overseeing vessel movement near Greenland. Due to sudden ice movement a boat carrying 11 passengers is about to crash into an iceberg. If nothing is done, all passengers will die.

The only way to save the 11 passengers is for you to order the captain to execute an emergency maneuver that will sharply alter the course of the boat.

However, the maneuver will cause the boat to overrun a life raft carrying your father and 2 other people that is floating next to the iceberg and out of sight of the captain. Your father along with the 2 other people will be killed if you order the captain to execute the maneuver, but the 11 people on the boat will be saved.
Would you order to execute the manoeuvre?

Gas explosion (4/60)
You are the inspector of gas lines in a city. For some reason, the gas in a pipeline, which is running under a building, got on fire. If the fire reaches the building, it will explode, and will kill 12 people in it.

You realize that the only way to stop the explosion of the building is to close the pipeline by pushing on a button, and divert the fire to a side-pipeline. If you do so, the 12 people will be saved.

However, above the side-line is another building with your mother and 3 other people in it. If you push the button and divert the fire into the sideline, this building will explode and kill your mother and the 3 other people, but the 12 in the building above the main line will be saved.
Would you push the button?

Fumes (5/65)
You are the late-night watchman in a hospital. Due to an accident in the building next door, there are deadly fumes rising up through the hospital's ventilation system. The fumes are directly and quickly heading towards a room with 13 patients in it. If you do nothing the fumes will rise up into this room and cause their deaths.

The only way to avoid the deaths of these patients is to hit a certain switch, which will cause the fumes to bypass the room and enter a second room instead.

However, you realize that your sister is in the second room along with 4 other patients. Your sister and the 4 other patients will be killed if you hit the switch and let the fumes bypass the first room, but the 13 patients in the first room will be saved.
Would you hit the switch?

Airplane (5/65)
You are a military base commander. A missile has just been mistakenly fired at a commercial airliner. If you do nothing, the missile will reach the airliner and 13 people on the airliner will die.

You realize that the only way to save these people, is to alter the course of the commercial airliner. In this case, the missile will pass by the airliner and the 13 people inside will be saved.

However, if you alter the course of the commercial airliner, the missile will hit another airliner with your sister and 4 other people inside which is flying right behind it. Your sister and the 4 other people who are travelling on this airliner will be killed if you alter the course, but the 13 people in the commercial airliner will be saved.
Would you alter the commercial airliner's course?

Submarine (4/60)
You are responsible for the mission of a submarine. You are leading this operation from a control center on the beach. An onboard explosion has damaged the ship and collapsed the only access corridor between the upper and lower levels of the ship. As a result of the explosion, water is quickly approaching to the upper level of the ship. If nothing is done, 12 people in the upper level will be killed.

You realize that the only way to save these people is to hit a switch in which case the path of the water to the upper level will be blocked and it will enter the lower level of the submarine instead.

However, you realize that your brother and 3 other people are trapped in the lower level. If you hit the switch, your brother along with the 3 other people in the lower level (who otherwise would survive) will die, but the 12 people in the upper level will be saved.
Would you hit the switch?

Mine (3/55)
Due to an accident there are 11 miners stuck in one of the shafts of a copper mine. They are almost out of oxygen and will die if nothing is done. You are the leader of the rescue team.

The only way for you to save them is to activate an emergency circuit that will transfer oxygen from a nearby shaft into the shaft where the 11 miners are stuck.

However, your team notices that your father and 2 other miners are trapped in the nearby shaft. If you activate the emergency circuit to transfer the oxygen, your father and the 2 other miners will be killed, but the 11 miners will be saved.

Would you activate the emergency circuit?

Metro (2/50)
You are a late-night metro controller in a station. A metro with automated driver is heading towards the station. Right before the station, the track gets blocked by a huge rock. If nothing is done, the train crashes into the rock and 10 passengers will die.

You realize that the only way to save these people is to pull a lever and divert the metro to a short emergency side-track. However, there is another metro coming right after it. If you divert the first metro to the side-track the second one will crash into the rock, as there is only place on the side track for one metro. If you do nothing, the first metro crashes, and the second one will be automatically directed into the side-track.

However, you notice that your brother and 1 other passenger are travelling in the second metro. If you pull the lever and let the first metro divert to the side-track, your brother and the 1 other passenger will die, but the 10 people in the first metro will be saved.
Would you pull the lever?

Volcano (2/50)
After a volcano eruption deadly hot lava is heading towards a nearby village. You are directing the rescue operations. There are 10 people standing on the roof of a house. If nothing is done, these 10 people will inevitably be killed by the lava stream.

The only way to save these people is to order the construction of an emergency barrier that will divert the lava stream into an old river bed.

However, you suddenly receive the information that right along the old river bed your mother and 1 other person are standing on the roof of a barn. If you order the construction of the emergency barrier and divert the lava to save the 10 people on the roof of the house, your mother and the 1 other person on the roof of the barn will inevitably be killed.
Would you order the construction of the emergency barrier?

Greene high-conflict personal harm dilemmas (Study 4)

Submarine
You are the captain of a military submarine traveling underneath a large iceberg. An onboard explosion has caused you to lose most of your oxygen supply and has injured one of your crew who is quickly losing blood. The injured crew member is going to die from his wounds no matter what happens.

The remaining oxygen is not sufficient for the entire crew to make it to the surface. The only way to save the other crew members is to shoot dead the injured crew member so that there will be just enough oxygen for the rest of the crew to survive.

Is it morally acceptable for you to kill the fatally injured crew member in order to save the lives of the remaining crew members?

Lawrence of Arabia
You are the leader of a small army that consists of warriors from two tribes, the hill tribe and the river tribe. You belong to neither tribe. During the night a hill tribesman got into an argument with a river tribesman and murdered him. The river tribe will attack the hill tribe unless the murderer is put to death.

The hill tribe refuses to kill one of its own warriors. The only way for you to avoid a war between the two tribes that will costs hundreds of lives is to publicly execute the murderer by cutting off his head with your sword.
Is it morally acceptable for you to cut off this man's head in order to prevent the two tribes from fighting a war that will cost hundreds of lives?

Sacrifice
You, your husband, and your four children are crossing a mountain range on your return journey to your homeland. You have inadvertently set up camp on a local clan's sacred burial ground.

The leader of the clan says that according to the local laws, you and your family must be put to death. However, he will let yourself, your husband, and your three other children live if you yourself will kill your oldest son.
Is it morally acceptable for you to kill your oldest son in order to save your husband and your other three children?

Crying Baby
Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables.

Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death.
Is it morally acceptable for you to smother your child in order to save yourself and the other townspeople?

B. Stability index

Table S1

Frequency of stability index values on conflict items in Study 1-4. The raw number of participants for each value is presented between brackets.

Study		Stability index value					Average
		>33	**50%**	**66.7%**	**75%**	**100%**	**Stability**
Study 1	No family - Moderate ratio	2.9% (3)	14.4% (15)	17.3% (18)	23.1% (24)	42.3% (73)	83.8%
Study 2	Family - Moderate ratio	5.7% (6)	10.5% (11)	11.4% (12)	8.6% (9)	63.8% (67)	85%
Study 3	No family - Extreme ratio	3.2% (4)	11.3% (14)	12.1% (15)	14.5% (18)	58.9% (73)	84.5%
	Family - Extreme ratio	5.8% (6)	17.5% (18)	11.7% (12)	12.6% (13)	52.4% (54)	80.3%
Study 4	Family - Moderate ratio	4.5% (4)	19.1% (17)	22.5% (20)	4.5% (4)	49.4% (44)	78.9%
	Greene - Personal harm	6.2% (6)	31.2% (30)	12.5% (12)	19.8% (19)	30.2% (29)	71.1%
Overall average		4% (29)	13.3% (105)	13.1% (89)	14.7% (87)	54.6% (311)	80.1%

C. Conflict detection analysis on combined Study 1-3

For each direction of change category one may ask whether reasoners are faced with two competing intuitions at the first response stage. We can address this question by looking at the contrast between conflict and control problems. If conflict problems cue two conflicting initial intuitive responses, people should process the problems differently than the no-conflict problems (in which such conflict is absent) in the initial response stage and show lower confidence when solving the conflict problems (Bialek & De Neys, 2017, see also footnote 3). Therefore, we contrasted the confidence ratings for the initial response on the conflict problems with those for the initial response on the no-conflict problems for each of the four direction of change categories. Note that we used only the dominant no-conflict "UU" category in which participants refused to sacrifice more people to save less. We refer to this category as "baseline". The rare responses in the other no-conflict direction of change categories were not cued by utilitarian or deontological considerations and cannot be interpreted unequivocally. To avoid spurious conclusion in this exploratory analysis we combined the data from our three studies to get the most general and robust test.

Table S2 shows the results. Visual inspection of Table S2 (bottom) indicates that overall there is a general trend towards a decreased initial confidence when solving conflict problems for all direction of change categories. However, this effect is larger for the "UD" and "DU" cases in which reasoners subsequently changed their initial response. This suggests that although reasoners might be experiencing some conflict between competing intuitions in all cases, this conflict is more pronounced in the "UD" and "DU" case.

We ran a separate analysis for each of the four direction of change conflict problem categories on the combined data from Study 1-3. In the analysis, the confidence for the initial response in a given direction of change category in question was contrasted with the initial response confidence for no-conflict "UU" responses which served as our baseline. We will refer to this contrast as the conflict factor. The conflict factor was entered as fixed factor, and participants were entered as random factor. Results showed that conflict improved model fit significantly for each of the four direction of change categories (UU, χ^2 (1) = 21.4, $p < 0.0001$, b = -6.76; DD, χ^2 (1) = 25.3, $p < 0.0001$, b = -9.1; DU, χ^2 (1) =17.96, $p < 0.0001$, b = -17.04; UD, χ^2 (1) = 43.4, $p < 0.0001$, b = -26.9). Hence, the conflict detection analysis on the confidence data indicates that by and large participants showed decreased response confidence (in contrast with the no-conflict baseline) after having given their first, intuitive response on the conflict problems in all direction of change categories. This supports the hypothesis that just like utilitarian responders, deontological responders were being faced with two conflicting intuitive responses when solving the conflict dilemmas (Bialek & De Neys, 2016, 2017).

A contrast analysis[8] that contrasted the conflict effects on the change (i.e., "UD" and "DU") and no-change ("UU" and "DD") indicated that the trend towards larger effects for the change categories did not reach significance, $Z = -0.98$, p (one-tailed) = 0.16, ($r = 0.14$ for no-change and $r = 0.18$ for change group). Nevertheless, the trend suggests that although reasoners might be generating two intuitive responses and are being affected by conflict between them in all cases, this conflict is more pronounced in cases where people subsequently change their answer. In line with our absolute confidence level findings on the conflict problems (see Figure 1), this tentatively suggests that it is the more pronounced conflict experience that makes them subsequently change their answer (Bago & De Neys, 2017; Thompson et al., 2012).

As we noted in footnote 3, our conflict detection analysis focused on the confidence data because these have been shown to be more reliable than latency data in the moral reasoning case (Bialek & De Neys, 2017). Nevertheless, for completeness the interested reader finds an overview of the latency data in Table S3. Visual inspection of the table indicates that there were few consistent initial conflict detection effects (i.e., longer initial response times on conflict than no-conflict problems) in the latency data.

[8] For this contrast analysis, we first calculated the r effect sizes out of t-values (Rosnow & Rosenthal, 2003). As a next step we used Fisher r-to-z transformation to assess the statistical difference between the two independent r-values. We used the following calculator for the z-transformation and p-value calculation: http://vassarstats.net/rdiff.html

Table S2

Average confidence ratings and confidence contrast difference between the no-conflict baseline and conflict problems as a function of response stage and direction of change category. Numbers in brackets are standard deviations of the means for initial and final responses, and standard errors for initial and final conflict contrast.

Study		Direction of Change	Initial response	Final response	Initial conflict contrast	Final conflict contrast
Study 1	No family - Moderate ratio	Baseline	78.5% (21.3)	85.7% (17.7)	-	-
		UU	76.2% (20.9)	80.7% (19.7)	2.3% (1.8)	5% (1.6)
		DD	57.7% (25.9)	68.5% (21.8)	20.8% (4.6)	17.2% (3.9)
		UD	71% (27.1)	62.9% (35)	7.5% (6.2)	22.8% (7.9)
		DU	47.2% (30.2)	77.4% (25.8)	31.3% (5.1)	8.3% (4.4)
Study 2	Family - Moderate ratio	Baseline	79.5% (24.7)	88.3% (19.5)	-	-
		UU	55.3% (26.9)	61.4% (24.7)	24.2% (4.2)	26.9% (3.8)
		DD	74.1% (25.4)	80.7% (23)	5.4% (2.1)	7.6% (1.8)
		UD	57.1% (27)	66.5% (28.7)	22.4% (5.1)	21.8% (5.3)
		DU	51.3% (26.1)	50.9% (24.2)	28.2% (6.5)	37.4% (6)
Study 3	No family - Extreme ratio	Baseline	80.9% (23.1)	87.6% (18.7)	-	-
		UU	78.8% (24.1)	85.6% (19.9)	2.1% (1.7)	2% (1.4)
		DD	76.3% (24.1)	80% (22.9)	4.6% (4.5)	7.6% (4.2)
		UD	61.2% (39.6)	54.3% (39.2)	19.7% (11.5)	33.3% (11.3)
		DU	48.5% (30.5)	78.5% (25.9)	32.4% (4.6)	9.1% (3.9)
	Family - Extreme ratio	Baseline	84.4% (21.3)	93.1% (15)	-	-
		UU	61.7% (27.9)	70.3% (27.8)	22.7% (3.6)	22.8% (3.3)
		DD	73% (28.6)	79.3% (25.8)	11.4% (2.4)	13.8% (1.9)
		UD	46.5% (33.7)	63.2% (36.1)	37.9% (8.8)	29.9% (8.3)
		DU	53.1% (20.5)	61.9% (25.3)	31.3% (4.0)	31.2% (4.8)
Overall average		Baseline	80.8% (22.7)	88.6% (18)	-	-
		UU	74.7% (24.5)	80.6% (22.2)	6.1% (1.1)	8% (1)
		DD	72.7% (26.8)	79.4% (24.2)	8.1% (1.3)	9.2% (1.1)
		UD	59.3% (31.1)	63% (33.3)	21.5% (3.6)	25.6% (3.7)
		DU	49.5% (27.7)	70.8% (27.2)	31.3% (2.5)	17.8% (2.4)

Note. U = utilitarian. D = Deontological.

Table S3

Average response times and response time contrast difference between the no-conflict baseline and conflict problems as a function of response stage and direction of change category. Means were calculated on log-transformed data and were back-transformed prior to the subtraction. Numbers in brackets are (geometric) standard deviations of the means for initial and final responses, and standard errors for the initial and final conflict contrast.

Study		Direction of Change	Initial response	Final response	Initial conflict contrast	Final conflict contrast
Study 1	No family - Moderate ratio	Baseline	7.72s (1.49)	6.94s (2.33)	-	-
		UU	8s (1.5)	7.11s (2.35)	-0.28s (0.12)	-0.17s (0.19)
		DD	8.02s (1.6)	7.77s (2.48)	-0.3s (0.29)	-0.83s (0.44)
		UD	7.48s (1.43)	7.69s (3)	0.24s (0.33)	-0.75s (0.68)
		DU	7.3s (1.82)	12.7s (2.17)	0.42s (0.31)	-5.76s (0.38)
Study 2	Family - Moderate ratio	Baseline	8.02s (1.43)	7.1s (2.5)	-	-
		UU	8.7s (1.43)	7.83s (2.24)	-0.68s (0.23)	-0.73s (0.36)
		DD	8.23s (1.45)	6.84s (2.51)	-0.21s (0.12)	0.26s (0.21)
		UD	5.42s (2.29)	16.71s (2.1)	2.6s (0.43)	-9.61s (0.41)
		DU	9.47s (1.2)	16.82s (1.94)	-1.45s (0.3)	-9.72s (0.49)
Study 3	No family - Extreme ratio	Baseline	7.32s (1.58)	7.87s (2.43)	-	-
		UU	7.62s (1.53)	7.33s (2.3)	-0.3s (0.12)	0.54s (0.17)
		DD	4.98s (2.24)	4.62s (2.4)	2.34s (0.41)	3.25s (0.45)
		UD	6.2s (2.62)	4.79s (2.37)	1.12s (0.76)	3.08s (0.7)
		DU	8.09s (1.73)	11.87s (2.49)	-0.77s (0.27)	-4s (0.39)
	Family - Extreme ratio	Baseline	7.63s (1.42)	7.51s (2.62)	-	-
		UU	7.95s (1.61)	6.3s (2.52)	-0.32s (0.2)	1.21s (0.32)
		DD	8.03s (1.48)	7.1s (2.73)	-0.4s (0.13)	0.41s (0.24)
		UD	7.87s (1.67)	9.77s (2.76)	-0.24s (0.39)	-2.26s (0.65)
		DU	7.86s (1.67)	10.61s (2.81)	-0.23s (0.32)	-3.1s (0.54)
Overall average		Baseline	7.65s (1.49)	7.37s (2.47)	-	-
		UU	7.86s (1.52)	7.16s (2.34)	-0.21s (0.07)	0.22s (0.11)
		DD	7.9s (1.54)	6.84s (2.6)	-0.25s (0.08)	0.53s (0.13)
		UD	6.54s (2.02)	10.11s (2.74)	1.11s (0.23)	-2.74s (0.31)
		DU	7.97s (1.69)	12.36s (2.4)	-0.32s (0.15)	-4.99s (0.22)

Note. U = utilitarian. D = Deontological.

D. Supplementary confidence analysis

Given the core hybrid model principles one can expect that changes in the strength levels of competing intuitions should lead to predictable consequences. Just as the family member manipulation can be assumed to affect the strength of the postulated deontological intuition, the kill-save ratio manipulation can—in theory—be assumed to affect the strength of the postulated logical intuition (i.e., stronger logical intuition with a more extreme kill-save ratio). However, our overall utilitarian response rate (Table 1) already indicated that the impact of the kill-save manipulation in the current studies was less marked than that of the family member manipulation. Extremer kill-save ratios did not lead to a significantly higher initial utilitarian response rate. This questions whether the kill-save ratio manipulation successfully affected the strength of the postulated utilitarian intuition. Nevertheless, for completeness and consistency, we also tested the impact of the kill-save ratio extremity on response confidence. If extremer kill-save ratios increase the strength of the utilitarian intuition, the key prediction is again that utilitarian and deontological responders' response confidence should show opposite effects. Figure S1 plots the average initial response confidence as a function of the kill-save extremity across our studies. As the figure shows, there was a slight trend in the expected direction: Making the utilitarian intuition "stronger" (extreme vs moderate kill-save condition), increased initial confidence for utilitarian responders but decreased it for deontological responders (i.e., deontological responders are more likely to doubt their deontological decision when the utilitarian intuition is stronger). However, statistical testing showed that the interaction trend was not significant, χ^2 (3) = 0.03, p = 0.86. Obviously, it is possible that adopting more extreme kill-save ratios (e.g., 1/5000 vs 1/5, see Trémolière & Bonnefon, 2014) might result in stronger effects of the kill-save ratio manipulation on the utilitarian response rate and response confidence.

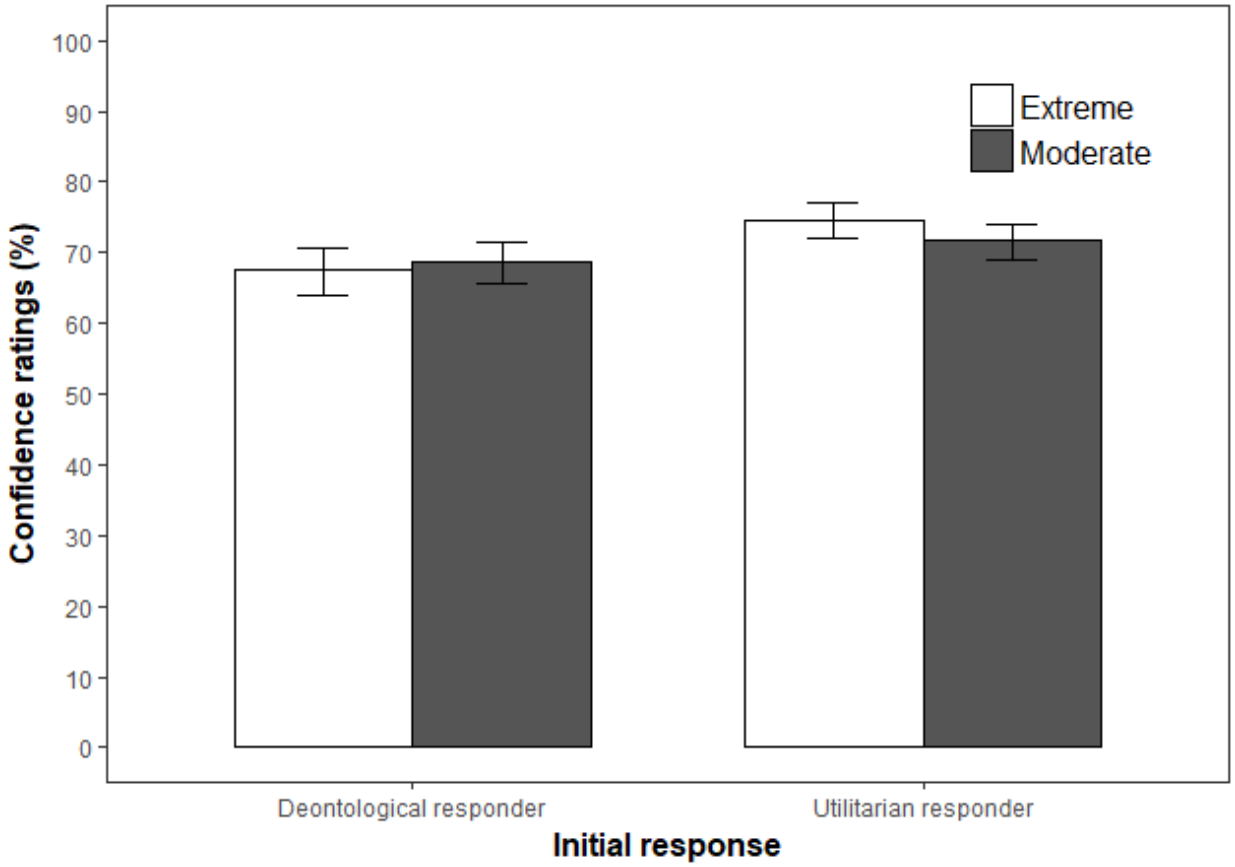


Figure S1. Mean initial conflict problem response confidence ratings for initial utilitarian and deontological responses as a function of the kill-save ratio (bottom) manipulations across Study 1-3. Error bars are 95% confidence intervals.

E. Justification study

Here we report an exploratory study in which people were given moral dilemmas and were asked to give a justification after both their initial and final response. We were specifically interested in the rate of proper utilitarian justifications that explicitly mentioned the greater good (e.g., "I opted for this decision because more people will be saved").

Method

Participants

A total of 120 Hungarian students (95 female, Mean age = 20.3 years, SD = 1.4 years) from the Eotvos Lorand University of Budapest were tested. 93.3% of the participants reported high school as highest completed educational level, while 6.7% reported having a post-secondary education degree. Participants received course credit for taking part.

Material

We adopted the same material and design as in our main studies. Half of the participants received "Family" versions and the other half "No family" versions. We used the moderate kill-save ratios in all versions. Since the primary goal was to study participant's response justifications we made a number of procedural changes to optimize the justification elicitation. Given that explicit justification might be hard (and/or frustrating) we opted to present only half of the main study problems (i.e., two conflict and two no-conflict versions). These items were chosen randomly from the main study problems. The procedure followed the same basic two-response paradigm as in the main studies with the exception that cognitive load was not applied and participants were not requested to enter their response confidence so as to further simplify the task design. As in the main studies, the initial response deadline was set to 12 s. Note that previous work from our team that contrasted deadline and load treatments indicated that a challenging response deadline may suffice to minimize System 2 engagement in a two-response paradigm (see Bago & De Neys, 2017).

After the initial and final response people were asked the following justification question: "*Why did you choose this response option? Please try to justify why you opted for the answer you selected."* There was no time restriction to enter the justification. Whenever participants missed the response deadline for the reasoning problem, they were not presented with the justification question, but with a message which urged them to make sure to enter their response before the deadline on the next item.

Justification analysis. To analyse participants' justifications we defined 3 main justification categories on the basis of an initial screening. Although our key interest lies in the rate of utilitarian justifications, the categorization gives us some insight into the variety of justifications participants spontaneously produce. The three justification categories along with illustrative examples are presented below.

Utilitarian. People made reference to the greater good or, in some cases to the less negative consequences (e.g., "People are all equal, the least people should die", "If I do this, fewer people will die", "Because more people will be saved")

Feeling/Intuition. People referred to a gut feeling, intuition or their sentiments towards the family member in question. (e.g., "Because I would feel guilty for the death of those people", "I just can't kill my brother", "I don't know, this is what my heart would say").

Other. All responses that could not be readily categorized as Utilitarian or Feeling/intuition (e.g., "There must be a possibility to divert both airplanes", "For the same reason as before", "I don't risk the life of humans").

Exclusion criteria. Trials on which the response deadline was missed (24.3% of all trials) were discarded. Therefore, in total, 454 trials (out of 600) were analysed.

Results and discussion

By and large, people's dilemma choices were consistent with the results of our main studies. The overall non-correction rate was again high and reached 82.4%. Table S4 gives a detailed overview. But the central question of this study concerned the response justifications. Table S5 presents an overview of the justification results on the critical conflict items. Our primary interest lies in the utilitarian responders; could they justify their initial utilitarian conflict response by referring to the greater good, or do they require further deliberation? As Table S5 shows, there is an overall increase in utilitarian justifications in the final response stage compared to the initial response stage (7.7% increase). This difference was especially clear in the family condition (23.2% increase) in which the emotional averseness of the utilitarian option was highest.

But it is also clear that the data are noisy. This is evidenced by the relatively high number of "Other" responses, and by the fact that participants sometimes referred to "Utilitarian" justifications even

when giving a deontological response, for example. As we already noted (see footnote 5), we cannot exclude that participants use the justification phase to deliberate about their initial response which would inflate utilitarian justifications overall. Furthermore, the percentage of discarded trials in which the initial response deadline was missed was quite high (i.e., 24.3%—about 3 times higher than what we observed in Study 1-3 with similar deadline). This might indicate that the mere fact that people were asked to justify their answer triggered additional reflection throughout the study. In line with this hypothesis we also found that average initial response latencies were about 1 s longer in the justification study vs main studies (8.8 s vs 7.8 s). Taken together this indicates that the findings should be interpreted with some caution. The study might overestimate the overall likelihood of utilitarian justifications. Nevertheless, the results present some preliminary evidence for the idea that such justifications are more likely after deliberate reasoning.

Table S4

Initial and final average percentage (SD) of utilitarian responses in Justification study.

	Conflict		**No-conflict**	
	Initial	**Final**	**Initial**	**Final**
No family	72.4% (45)	77.6% (41.9)	90.2% (29.9)	91.5% (28.1)
Family	26.4% (44.4)	17.2% (38)	94.7% (22.6)	93.6% (24.6)
Average	47.9% (50.1)	45.4% (49.9)	92.6% (26.2)	92.6% (26.2)

Table S5

Frequency of different types of justifications for conflict items (raw number of justifications in brackets).

Condition	Justification	Initial response		Final response	
		Utilitarian	Deontological	Utilitarian	Deontological
No family	Utilitarian	84.6% (44)	35% (7)	83.1% (49)	46.7% (7)
	Feeling/Intuition	3.8% (2)	5% (1)	-	6.7% (1)
	Other	11.5% (6)	60% (12)	16.9% (10)	46.7% (7)
Family	Utilitarian	43.5% (10)	1.7% (1)	66.7% (10)	-
	Feeling/Intuition	26.1%(6)	85% (51)	13.3%(2)	70.3% (45)
	Other	30.4% (7)	13.3% (8)	20% (3)	29.7% (19)
Overall	Utilitarian	72% (54)	10% (8)	79.7% (59)	8.9% (7)
	Feeling/Intuition	10.7% (8)	65% (52)	2.7% (2)	58.2% (46)
	Other	17.3% (13)	25% (20)	17.6% (13)	32.9% (26)