Journal of Experimental Social Psychology 113 (2024) 104616

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

To beckon or not to beckon: Testing a causal-evaluative modelling approach to moral judgment: A registered report[☆]

Cillian McHugh [a,*], Kathryn B. Francis [b], Jim A.C. Everett [c], Shane Timmons [d,e]

[a] Department of Psychology, University of Limerick, Ireland
[b] School of Psychology, Faculty of Natural Sciences, Keele University, UK
[c] School of Psychology, Keynes College, University of Kent, UK
[d] Economic & Social Research Institute, Ireland
[e] School of Psychology, Trinity College Dublin, Ireland

ARTICLE INFO

Keywords:
Moral judgment
Causal evaluative modelling
Means vs side-effect
Trolley problems
Contextual influences
Deontology
Utilitarianism

ABSTRACT

Moral judgments are increasingly being understood as showing context dependent variability. A growing literature has identified a range of specific contextual factors (e.g., emotions, intentions) that can influence moral judgments in predictable ways. Integrating these diverse influences into a unified approach to understanding moral judgments remains a challenge. Recent work by Railton (2017) attempted to address this with a causal-evaluative modelling approach to moral judgment. In support of this model Railton presents evidence from novel variations of classic trolley type dilemmas. We present results from a pre-registered pilot study that highlight a significant confound and demonstrate that it likely influenced Railton's results. Building on this, our registered report presents a replication-extension of Railton's study, using larger more diverse samples, and more rigorous methods and materials, specifically controlling for potential confounds. We found that participants' judgments in sacrificial dilemmas are influenced by both direct personal force, and by whether harm occurs as a means or as a side-effect of action. We also show the relationship between a range of individual difference variables and responses to sacrificial moral dilemmas. Our results provide novel insights into the factors that influence people's moral judgments, and contribute to ongoing theoretical debates in moral psychology.

The need to account for the role of context in moral judgments has long been acknowledged (Basinger, Gibbs, & Fuller, 1995; Gilligan, 1977, 1993; Hofmann, Wisneski, Brandt, & Skitka, 2014; Schein, 2020). On-going research presents a growing list of contextual factors known to influence moral judgments - notable examples include emotions (Cameron, Payne, & Doris, 2013; Giner-Sorolla, 2018), intentionality and evitability (Christensen, Flexas, Calabrese, Gut, & Gomila, 2014; Christensen & Gomila, 2012), and how 'up close and personal' an action is (Greene, 2008; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). The diversity of contextual influences is not well accounted for by current theories of moral judgment (Hester & Gray, 2020; McHugh, McGann, Igou, & Kinsella, 2022; Schein, 2020).

Railton (2017) developed a social learning approach to moral judgment to better account for this context dependent variability. He supports his argument with data on people's responses to a range of novel *Trolley* type dilemmas. We have identified limitations with the empirical work presented by Railton (2017), that potentially undermine the conclusions that can be drawn. We propose a replication of Railton's core study, with modifications to the design, method, and sample, to provide a stricter test of the claims made.

1. Current theorizing about moral judgments: a snapshot

Theories of moral judgment have attempted to explain the underlying processes that lead to, and the factors that influence people's moral judgments. Focusing on sacrificial moral dilemmas, classic dual-process

[☆] This paper has been recommended for acceptance by Paul Conway.

[*] Corresponding author at: University of Limerick, Limerick, Ireland.
E-mail address: cillian.mchugh@ul.ie (C. McHugh).

https://doi.org/10.1016/j.jesp.2024.104616
Received 7 February 2022; Received in revised form 20 March 2024; Accepted 21 March 2024
Available online 8 April 2024

accounts have identified the characteristically deontological[1] judgments (principles or rules based) rejection of sacrifice as more intuitive/automatic or involving more emotional processing, while identifying the pro-sacrificial characteristically utilitarian judgments (maximizing benefit/minimizing harm) as involving more deliberation or 'cognition' (e.g., Byrd & Conway, 2019; Conway & Gawronski, 2013; Conway, Goldstein-Greenwood, Polacek, & Greene, 2018; Greene, 2016). Model-based approaches (Crockett, 2013; Cushman, 2013) present a similar pattern, though with a different focus. According to these approaches, judgments of *outcomes* involve model-based processing, while judgments of *actions* involving model-free processing (Crockett, 2013; Cushman, 2013). Other theorists have attempted to create a taxonomy of moral concerns (e.g., Graham et al., 2012; Graham, Haidt, & Nosek, 2009; Rozin, Lowery, Imada, & Haidt, 1999; Shweder, Much, Mahapatra, & Park, 1997), or attempted to identify a single factor that underlies all considerations (e.g., cooperation, see Curry, Jones Chesters, & Van Lissa, 2019; or harm, see Schein & Gray, 2018).

Despite the strengths of these approaches, many of their assumptions are based on the *content* of moral judgments, and this poses a challenge for accounting for the dynamism and context sensitivity observed in people's moral judgments. For instance, classic dual-process approaches cannot account for situations where it is the characteristically deontological rejection of the sacrifice that is supported by deliberation (Gamez-Djokic & Molden, 2016; Körner & Volk, 2014; McPhetres, Conway, Hughes, & Zuckerman, 2018; Pizarro & Bloom, 2003), or, conversely, the pro-sacrificial judgments that are grounded in affect (Gubbins & Byrne, 2014; Reynolds & Conway, 2018).

By mapping model-free and model-based processes onto judgments of *actions* and *outcomes* respectively (Crockett, 2013; Cushman, 2013), model-based approaches cannot adequately account for instances where the action and the outcome are the same but people make different judgments based on a means/side-effect distinction (also referred to as doctrine of double effect, e.g., Doris, 2010 ; Mikhail, 2000). The means/side-effect distinction (discussed below) is observed when causing harm as a means to an end is seen as worse than causing harm as a side effect even when the ends and the actions are the same (Doris, 2010; Mikhail, 2000). These differing responses based on a distinction between means and side-effects, when actions and outcomes remain constant (e.g., in the Loop vs Switch cases below) pose a challenge to model-based approaches.

Taxonomy approaches provide a framework for understanding individual differences in people's moral judgments; however they do not offer predictions regarding intrapersonal variability depending on context, e.g., judgments of the same act may vary depending on knowledge of the motives of actors involved, the relationship between the actors involved (for demonstration see, Andrejević, Feuerriegel, Turner, Laham, & Bode, 2020). Relatedly, attempts to explain all moral judgment as grounded in considerations of a single factor such as harm, are undermined by situations where harmless actions are regarded as wrong (Haidt, Björklund, & Murphy, 2000; McHugh, McGann, Igou, & Kinsella, 2017; McHugh, Zhang, Karnatak, Lamba, & Khokhlova, 2023), or by potentially harmful actions being regarded as *not* wrong (Alicke, 2012; McHugh, McGann, Igou, & Kinsella, 2020; Royzman & Borislow, 2022).

Other approaches draw a clear separation between claims about underlying cognitive processes and considerations regarding the content of specific judgments. For example, the CNI model of moral decision making (Gawronski, Armstrong, Conway, Friesdorf, & Hütter, 2017) has made significant advances in delineating multiple influences on moral judgments. Specifically, the CNI model provides an account of how moral judgments reflect sensitivity to consequences (C), moral norms (N), and a general preference for inaction (I) independent of consequences or norms. Importantly, this model does not make claims regarding the underlying processes (such as linking considerations of consequences to deliberation, or sensitivity to norms to automatic processing).

Similarly, some authors have highlighted the dynamism and context sensitivity observed in moral judgment and attempted to develop accounts of the underlying processes that do not rely on as heavily on content based assumptions (e.g., Bucciarelli, Khemlani, & Johnson-Laird, 2008; McHugh et al., 2022; Railton, 2017). These approaches make a clear distinction between the underlying processes, for which assumptions are *not* based on content considerations, and various contextual influences (that may include considerations of content). For example, McHugh et al. (2022) draw on the cognitive psychological accounts of categorization processes to present a model of the cognitive processes underlying *moral* categorization that does not rely on assumptions based on content. According to this approach, moral judgments are dynamic and context dependent, and McHugh et al. (2022) highlight a range of known contextual factors that influence moral judgments. In line with this approach, we define the decision-making context as encompassing all possible aspects of the experience of making a decision, and we use the term "contextual factor" to refer to any feature of the decision-making context that can vary. This may include features of the situation being judged (e.g., considerations relating to intentionality, and evitability), external factors not directly related to the situation (e.g., salient prior decisions), or factors relating to the decision maker (e.g., current mood, incidental emotions).

Drawing on an extensive review of both the morality and developmental literatures, Railton (2017) provides an account of moral judgment that is grounded in a moral learning approach. According to Railton, moral understanding emerges as a result of domain-general learning processes, and that moral judgments are grounded in causal-evaluative modelling of the situation, the agent, and the action and outcome. Prior experience generates expectations that inform and guide our perceptions and interactions with the world. This means that our moral judgments can be guided by a diverse range of contextual influences, including content, type of action, and social/relationship considerations.

2. Re-imagining the trolley problem

To support his argument, Railton (2017) presents a novel interpretation of the trolley problem. The trolley problem refers to the phenomenon in moral psychology, whereby people make different judgments in similar scenarios in which the eventual outcome of both scenarios is the same (see Kahane & Everett, 2022, for a recent overview of the trolley problem in moral psychology). Consider the following two scenarios:

> *Switch*: A runaway trolley is speeding down the track, its driver slumped over the controls, apparently unconscious. Ahead on the tracks are five workers, who do not see the trolley coming, and who soon will be struck and killed. You are standing next to a lever that operates a switch lying between the trolley and the workers. Pushing this lever would send the trolley onto a sidetrack. That would save the five workers, but there is a single worker on the sidetrack, who will be struck and killed. Should you push the lever to send the trolley down the sidetrack?

[1] There is ongoing debate about the appropriate terminology to be used here. While it is broadly agreed that the terms *deontological* and *utilitarian* are imperfect descriptors for responses in the dilemmas (see Conway et al., 2018; Everett & Kahane, 2020 and Kahane & Everett, 2022), there is less agreement about how imperfect they are and what precise terminology should be used. In this paper we will attempt a compromise, recognizing our own different views on the matter, following Greene (2016) by referring to "characteristically" utilitarian or deontological judgments, but also following Everett and Kahane (2020) by referring to the judgments as pro-sacrificial, and qualifying them within a sacrificial dilemma context.

> *Footbridge*. A runaway trolley is speeding down the track, its driver slumped over the controls, apparently unconscious. Ahead on the tracks are five workers, who do not see the trolley coming, and who soon will be struck and killed. You are standing on a footbridge over the track, next to a very large man. This man's weight is sufficient to stop the trolley, though your own is not. If he were to fall into the path of the trolley, that would bring it to a halt before hitting the five workers, saving their lives but killing him. Should you push the man off the footbridge into the path of the trolley?

The net outcomes in both *Switch* and *Footbridge* are the same, however, people are more likely to endorse action in *Switch* than in *Footbridge*. Railton (2017) characterizes existing explanations of this inconsistency (specifically the explanations of Cushman, 2013; and Greene, 2013) in terms of the relative use of direct muscular/personal force. The use of personal force in *Footbridge* means that people are unwilling to endorse action, even though it will save five lives, whereas for *Switch*, in the absence of personal force, people are willing to endorse action that harms one person in order to save five.

A third scenario, *Loop* is similar to *Switch*, but the side-track re-joins the main track, just before the location of the five workers; the five workmen are saved by the weight of the man on the side-track stopping the trolley (rather than the mere diversion), *Loop* reads as follows:

> *Loop*. A runaway trolley is speeding down the track, its driver slumped over the controls, apparently unconscious. Ahead on the tracks are five workers, who do not see the trolley coming, and who soon will be struck and killed. You are standing next to a lever that operates a switch lying between the trolley and the workers, and pushing the lever would send the trolley onto a side loop. This loop rejoins the main track just before the location of the workers. However, on that loop stands a single large worker, who would be struck and killed if you switched the trolley. His weight would bring the trolley to a halt before the loop rejoins the main track, saving the five workers. Should you push the lever to send the trolley onto the side loop?

Responses to *Loop* have been cited by some authors as evidence for the means/side-effect distinction, where people appear to make a distinction between harm as a *means* vs harm as a *side-effect*, and are less likely to endorse the former (Doris, 2010; Mikhail, 2007). We note that there remains some debate regarding this interpretation. For example, under its original conception *Loop* was intended as a counterexample to this distinction because the action in *Loop* is argued to be more acceptable than in *Footbridge* (Thomson, 1985). Furthermore, there is considerable disagreement regarding the nature of the means/side-effect distinction (e.g., is *means* defined as acting "because of" or acting "in order to"? For discussion on this point see Kamm, 2007). In addition to these philosophical concerns, concerns have been raised regarding the strength of the empirical evidence for the means/side-effect, with a recent meta-analysis reporting only a small effect for means/side-effect, and that this effect is moderated by the use of personal force (Feltz & May, 2017). Given these complexities, and that this effect is not directly discussed by Railton (2017), we present his argument first, before discussing the means/side-effect distinction as a potential limitation.

Railton creates a matrix of the intersection between features of the scenarios and the responses permissible according to current theorizing (see Fig. 1).[2] According to existing explanations, people should be willing to select an intervention that harms one to save five when this intervention does not involve personal force, and reject an intervention with the same payoff when the intervention involves personal force. Railton's analysis suggests that neither Greene's (2013), nor Cushman's (2013), approaches can account for responses that occupy the positions of *X* and *Y* in the matrix in Fig. 1. In contrast, Railton argues that such responses *are* possible according to his social learning approach (Railton, 2017). To support this, Railton develops several novel trolley-type scenarios and presents evidence that people do respond according to both *X* and *Y* in the matrix in Fig. 1.

2.1. *Permissible use of force*

According to Greene's (2013) and Cushman's (2013) approaches (as presented by Railton, 2017), the use of personal force harming one to save five is not a permissible intervention, while Railton's social learning approach allows for situations where this is permissible. To support his argument, Railton devised the following scenario, *Bus*.

> *Bus*: You are visiting a city where there have recently been terrorist suicide bombings. The terrorists target crowded buses or subway cars. To prevent anyone stopping them, they run up at the last moment when the bus or subway doors are closing, triggering their bomb as they enter. You are on a crowded bus at rush hour, just getting off at your stop. Next to you a large man is also getting off, and the doors are about to close behind the two of you. You spot a man with an overcoat rushing at the doors, aiming to enter just behind the exiting man. Under his coat you see bombs strapped to his chest, and his finger is on a trigger. If you were to push the large man hard in the direction of the approaching man, they both would fall onto the sidewalk, where the bomb would explode, killing both. You would have fallen back onto the bus, and the closing doors would protect you and the other occupants of the bus from the bomb. Alternatively, you could continue exiting the bus, and you and the large man would be on sidewalk, protected by the closing doors, as the bomb goes off inside the bus, killing the terrorist and five passengers. Either way, then, you will not be hurt. Should you push the large man onto the bomber?

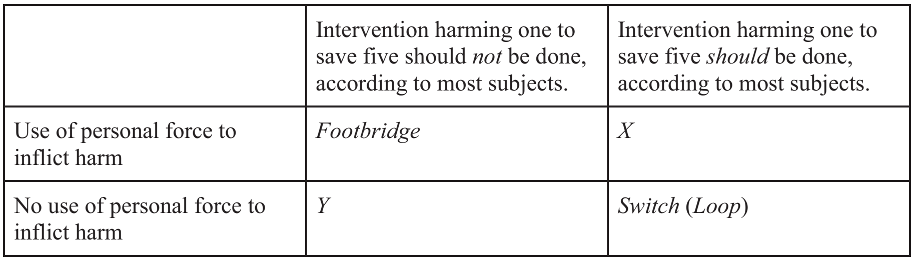
Railton's results suggest that responses to *Bus* are more similar to *Switch* than to *Footbridge*. That is, people appear to endorse the intervention in the case of *Bus* even though it involves the use of personal force causing harm. Thus, *Bus* presents a case that occupies the position of *X* in Fig. 1.

2.2. *Impermissible intervention without force*

If the classic explanations regarding personal force are correct, cases where an intervention that harms one to save five does not involve direct force, should be widely endorsed (Cushman, 2013; Greene, 2013; Railton, 2017). Railton proposes two scenarios to test this claim, *Wave* and *Beckon* which read as follows:

> *Wave*: A runaway trolley is speeding down the track, its driver slumped over the controls, apparently unconscious. Ahead on the tracks are five workers, who do not see the trolley coming, and who soon will be struck and killed. A wall prevents them from moving to their left to avoid the trolley, but there is space to their right. You are standing at some distance from the track, with no ability to turn the train. The workers are facing in your direction, and if you were to wave to their right with your arms, the five workers on the track would step off and escape injury. However, a single worker who is closer to you and standing to the left of the track, and who also does not see the trolley, will see you wave, and he will step onto the track, and immediately be hit and killed. Should you wave to the workers?
>
> *Beckon*: A runaway trolley is speeding down the track, its driver slumped over the controls, apparently unconscious. Ahead on the tracks are five workers, who do not see the trolley coming, and who

[2] We note that the top row of Fig. 1 reflects outcomes or endpoint judgments, and should not be interpreted as articulating prescriptive/proscriptive norms that a decision maker may be sensitive to.



	Intervention harming one to save five should *not* be done, according to most subjects.	Intervention harming one to save five *should* be done, according to most subjects.
Use of personal force to inflict harm	*Footbridge*	*X*
No use of personal force to inflict harm	*Y*	*Switch* (*Loop*)

Fig. 1. Responses according to current theorizing (adapted from Railton, 2017).

> soon will be struck and killed. You are standing at some distance from the track, with no ability to turn the train or warn the men. A large man, whose weight is sufficient to stop the trolley, is standing on the other side of the track, facing in your direction. He is unable to see the oncoming trolley owing to a traffic signal box that blocks his view up the track. If you would beckon to him, he would step forward onto the track, and be immediately struck and killed. This would halt the trolley and save the five workers. Should you beckon to the large man?

According to Railton's findings, people endorse the intervention in *Wave*, but not in *Beckon*. With these two cases Railton appears to have identified two scenarios with the same net outcome (harm one and save five), where the actions conducted are very similar (gesturing), and neither action involves personal force. Thus *Beckon* is presented as a scenario that occupies the position of *Y* in the matrix in Fig. 1, providing evidence for his social learning approach over existing dual-process approaches.

According to Railton's (2017) causal-evaluative modelling approach this difference between *Wave* and *Beckon* points to an underlying model at work in which a central focus is upon the nature of the agent or agency involved. That is, people's judgments are sensitive not just to actions and outcomes, but also to considerations of character, and type of person who would act/not act. To support this line of reasoning Railton (2017) notes that each action appears to be associated with different levels of trustworthiness. Railton asked participants to imagine their roommate had committed the action, and asked if they would trust their roommate the same, more, or less. Railton found that, for *Beckon* (and *Footbridge*), the majority of participants indicated they would trust their roommate less (79% and 78% respectively), while for *Wave* (and *Switch*) indicated they would trust their roommate the same (68% and 56% respectively). This perhaps suggests that beckoning (or pushing) someone to their death presents a greater betrayal of trust than simply waving to them (or flipping a switch). That is, beckoning is an inviting gesture conveying the potential for interpersonal closeness or establishing a social bond and a sense of togetherness. This social connection is brutally undermined when it ultimately leads to death-by-trolley. In contrast, waving does not convey the same sense of connection and thus when it also results in death the sense of betrayal is less stark. The potential importance of the role of trust is also apparent in the responses to *Bus*. For *Bus*, the use of force was generally seen as permissible, and a strong majority of participants indicated that if their roommate took the action they would either trust their roommate more (49%), or the same (39%). Railton thus presents trust, or perceptions of trustworthiness, as an important influence on participants' judgments of the permissibility of particular actions. We note that this line of argument is not necessarily limited to perceptions of trustworthiness and could be applied to a diverse range of traits that may convey different aspects of moral character across different moral dilemmas. For instance, in addition to appearing untrustworthy, someone who kills by beckoning may also be seen as deceitful or devious; someone chooses to push in *Footbridge* they may be viewed as cold or callous.

3. Limitations of Railton's method

There are two key limitations with the findings presented by Railton (2017) that pose a challenge to the conclusions that can be drawn. First, regarding his sample, Railton's participants were students of his introductory ethics class, limiting generalizability of the findings. In addition, his reported samples were small[3] (largest $N = 45$), limiting the statistical power, and increasing the possibility of error (or that observed findings are driven by outliers).

Second, Railton did not control for features of the scenarios that may act as confounding influences on participants' responding. Specifically, existing literature suggests that the different responses to *Wave* and *Beckon* can be explained by the means/side-effect distinction (Hauser, Cushman, Young, Kang-Xing Jin, & Mikhail, 2007; Mikhail, 2000; Sinnott-Armstrong, Young, & Cushman, 2010). The scenarios used by Railton cannot distinguish between the means/side-effect distinction, and Railton's proposed perceived trust explanation of the different responses between *Beckon* and *Wave*. That is, in *Wave*, the large man dies as a *side-effect* of waving at the five workers to step off the tracks, while in *Beckon* the large man dies as a *means* to save the five workers.

Existing evidence for this means/side-effect distinction comes from research involving *Loop*-type scenarios (though the strength of this evidence has come under scrutiny, see Feltz & May, 2017). Railton's participants appear to respond to *Loop* differently than participants in other studies; 90% of Railton's participants (total $N = 41$) endorsed the pro-sacrificial action in *Loop*, while previous research (e.g., Hauser et al., 2007) has found only 56% of participants (total $N = 2612$) endorse action in *Loop*. It is surprising that Railton found greater endorsement of action in *Loop* (90%) than in *Switch* (85%). The differences in permissibility between *Beckon* (42%) and *Wave* (87%) are not identical, but may be comparable to the differences between *Loop* (56%) and *Switch* (89%) observed by Hauser et al. (2007). Together with relative structures of *Beckon* and *Wave* this suggests that the differences in responding are likely due to this means/side-effect distinction rather than a difference in the type of gesture.

4. Alternative predictors of responding

In addition to addressing the methodological concerns detailed above, our study will also extend Railton's work by also including explicit consideration of individual difference variables that may influence people's responses to the dilemmas. Given that the outcome for all scenarios is the pro-sacrificial characteristically utilitarian response, the first measure we include is the Oxford Utilitarianism Scale (OUS, Kahane et al., 2018) which measures people's tendency to separately endorse the two dimensions of utilitarian psychology: instrumental harm (sacrificing one in order to save a greater number), and impartial beneficence (impartial concern for the well-being of everyone). Second, even if philosophically, classical utilitarianism and non-utilitarian

[3] Though Railton has continued to collect responses from his students and has accumulated four years' worth of responses (Railton, 2021).

deontological approaches are opposing ethical theories, it has been argued that psychologically endorsement of characteristically consequentialist and deontological principles need not be opposed (Conway et al., 2018; Plaks, Lv, Zhao, Staples, & Robinson, 2021; though see Everett & Kahane, 2020 for a more critical perspective on this). Given that it is at least possible that people's deontological tendencies may also influence their responding, independent of their utilitarian tendencies, we also included the Consequentialism Scale (Plaks et al., 2021) which is made up of two subscales, measuring deontological and utilitarian tendencies independently.

Previous research has found that people's responses to moral dilemmas are related to their tendency to engage in reflective thinking (e.g., Paxton, Ungar, & Greene, 2012), and as such we will include the Cognitive Reflection Test (Toplak, West, & Stanovich, 2014), to assess this. In addition, pro-sacrificial responses have been linked with self-reported preferences in thinking styles (Patil et al., 2020); to test this we also include the Comprehensive Thinking Styles Questionnaire (which has four subscales: Actively Open-minded Thinking, Close-Minded Thinking, Preference for Intuitive Thinking, Preference for Effortful Thinking; Newton, Feeney, & Pennycook, 2021).

5. The current research

Building on the logic of Railton's argument, we propose two studies for a stronger empirical test than that presented in Railton (2017). Railton builds on existing work that highlights the importance of personal force (Greene, 2016), presenting a strong argument for the combined influences of both trust, and the use of personal force as part of causal-evaluative modelling, on judgments of permissibility of an intervention that harms one individual to save five. However, his data come from small samples collected as part of in-class student feedback in introductory ethics classes (Railton, 2017, p. 183). Our proposed research will address two limitations of the data presented by Railton (2017). First our sample will be larger and more diverse, addressing issues of statistical power, error rate, and generalizability. Second, our study will include a range of scenarios carefully matched to control for the possible confounding influence of the means/side-effect distinction.

In Study 1, we propose a direct test of the differences between *Wave* and *Beckon*, and whether the differences described by Railton (2017), occur because of the type of action (waving vs beckoning), or due to the confounding influence of the means/side-effect distinction. In Study 2 we will additionally test for the influence of personal force on people's responses to moral dilemmas, along with the possible influence of trust (as suggested by Railton, 2017), and the possible predictive role of specific individual difference variables listed above.[4]

6. Study 1

The aim of Study 1 is to identify the source of the differences in responding between *Wave* and *Beckon*. We have identified two potential explanations for the differences reported by Railton (2017): (1) action type - people distinguish between waving vs beckoning; (2) means/side-effect - people distinguish between harming one individual as a *means* to save five, vs as a *side-effect* of saving five. We developed alternative versions of both *Wave* and *Beckon* that are matched for the influence of means/side-effect, that is, we developed a version of *Wave* where one individual is killed as a *means* to save five (where, in Railton's original version of *Wave*, one individual was killed as a *side-effect* of saving five), and we developed a version of *Beckon* where one individual is killed as a *side-effect* of saving five (where Railton's version of *Beckon* involved killing one individual as a *means* to save five). This allows us to test whether differences in responding emerge because people distinguish between the two types of actions (waving vs beckoning), or because people are making a distinction between causing harm as a means vs as a side-effect (in line with the means/side-effect distinction). Our basic hypothesis is that participants responding will vary depending on means/side-effect, and no variation will be observed for action type. We pre-registered these basic predictions at https://aspredicted.org/N85_2JV. Unpacking these predictions, we identified three sets of competing hypotheses:

H1. Action Type only[5]: Participants' responding will vary systematically depending on the type of action (waving vs beckoning);

H1a. Drawing on Railton (2017), H1a predicts that participants will be significantly more likely to endorse waving than beckoning.

H1b. Drawing on Railton (2017), H1b predicts no influence of means/side-effect, participants' responses will not vary depending on whether harm occurs as a means vs as a side-effect of intervention.

H2. Means/side-effect only: Participants' responding will vary systematically depending on whether harm occurs as a means vs as a side effect.

H2a. Drawing on previous research on the means/side-effect distinction, H2a predicts that participants will be significantly more likely to endorse action if harm occurs as a side-effect than if it occurs as a means.

H2b. Predicts that no influence of action type, participants' responses will not vary depending on whether the action involved is waving or beckoning.

H3. Both: Participants' responding will be influenced by both action type and means/side-effect.

6.1. Method

6.1.1. Design

Study 1 is a 2 × 2 within-subjects design. The first IV is means/side-effect with two levels *means* vs *side-effect*. The second IV is action type, with two levels *wave* vs *beckon*. The primary DV is action choice with two levels, act vs do not act. We will also test for the possible influence of the IVs on five additional measures: confidence (in their action choice decision), judgment (of how wrong they view the action), trustworthiness of three actors, imagining that each actor chose to act, the actors were: (a) a friend, (b) a romantic partner, (c) a politician.

6.1.2. Participants

A priori power analysis indicated that a sample of $N = 96$ was required to achieve 80% power with equivalence bounds of -0.3 and 0.3; a sample of $N = 215$ was required to achieve 80% power with equivalence bounds of -0.2 and 0.2, a sample of $N = 315$ was required to achieve 80% power with equivalence bounds of -0.15 and 0.15, and a sample of $N = 857$ was required to achieve 80% power with equivalence bounds of -0.1 and 0.1. We set our target sample at $N = 300$ A total sample of $N = 306$, (female = 146, male = 153, other = 7, $M_{age} = 25.7$, SD = 9.4, min = 18, max = 58) took part. Participants were students of University of Limerick and were recruited through an email circulated to all students on 3rd of November 2021. The link remained open until the target number of N = 300 was reached, at which time the data were downloaded and performance on the attention checks was assessed.

Investigation of responses to the attention checks revealed that $n =$

[4] The data, materials, and analysis scripts are available on this project's OSF page at https://osf.io/59quk/?view_only=18414ad4433a4145a718f7015c012e36.

[5] We note that while this reflects the way the scenarios are discussed in Railton (2017), it does not necessarily reflect the over-arching argument, that is, the proposed causal-evaluative model would also be consistent with a rejection of H1a and H1b. Study 1 addresses methodological concerns rather than theoretical.

86, failed the first attention check, and $n = 3$ failed the second attention check, taken together, only 1 participant failed both attention checks. As per our pre-registered exclusion criteria we excluded this participant from our analyses. This exclusion brought our total sample to $N = 305$ (female = 145, male = 153, other = 7, $M_{age} = 25.7$, $SD = 9.4$, min = 18, max = 58), remaining above the pre-registered target sample size ($N = 300$) so the survey was closed on the 5th November 2021. Sensitivity power analysis indicated that with alpha of 0.05 this would achieve 80% power with equivalence bounds of −0.168 and 0.168.

6.1.3. Materials

Four variants of the *Trolley* dilemma (listed in Appendix A) were used to manipulate the IVs. We use both *Wave* and *Beckon* as proposed by Railton (2017), along with a modified version of each to test for the influence of means/side-effect. Railton's original version of *Wave* is listed as *Wave (side-effect)* and we developed an alternative where the harm occurs as a means to save five: *Wave (means)*. Similarly, Railton's version of *Beckon* is listed as *Beckon (means)* and we developed an accompanying *Beckon (side-effect)*. These four scenarios allow for the influences of both IVs, action type (waving vs beckoning) and means/side-effect to be tested independently.

Our primary DV is action choice. This is recorded with a question "Should you wave/beckon to the workers/large man?" with a binary "yes"/"no" response.

In addition to action choice, we also recorded participants' confidence in their choice (1 = *Not at all confident*, 7 = *Extremely confident*), how right or wrong the action is (1 = *Wrong*, 4 = *Neutral*, 7 = *Right*), and how trustworthy (1 = *Not at all trustworthy*, 7 = *Extremely trustworthy*) they would rate each of a friend, a romantic partner, and a politician, if they heard these people chose to commit the act described (full wording of these is listed in Appendix B in the supplementary materials). These are included as dependent measures in tests for equivalence, however they are included as potential predictor variables in the overall regression, that is, does perceived trustworthiness of a third party committing the action predict participants' own endorsing of the action?

6.1.4. Procedure

Data collection was conducted using Qualtrics (Qualtrics, 2020). Participants read a scenario and responded to each of the measures. The action-choice measure was displayed on the same page as the scenario. The remaining measures were displayed on the following page. All participants responded to all scenarios, and the order of presentation was fully randomized. Participants took between 5 and 10 min to complete the entire study.

6.1.5. Analytic strategy

Testing for differences was conducted using chi-squared tests (for categorical measures, e.g., action choice), and ANOVAs for continuous measures. Tests for equivalence were conducted using two one-sided *t*-tests along with a standard t-test for differences using the TOSTER package in R (Lakens, 2017). A linear mixed effects model (with random intercepts varying by participant) was conducted to identify the predictors of action choice. Based on the analysis of our simulated data, and considerations of power, we anticipated inconsistent results for the equivalence tests. As such we test for equivalence at the level of d = 0.1; d = 0.15; and d = 0.2. Our conclusions will be informed by the results of these tests and the difference tests in combination.

6.2. Results

Below we report difference tests and equivalence tests for the possible effects of action type and means/side-effect. Our sample was sufficiently powered to detect equivalence at the level of $d = 0.2$, and as such our analysis below reflects this. In the supplementary analyses we additionally report tests for equivalence at the level of $d = 0.15$, and $d = 0.1$.

6.2.1. Effect of action-type

A chi-squared test for independence revealed no association between action type and participants' preference for action, $\chi^2(1, N = 1220) = 0.003$, $p = .953$, $V = 0.002$. The equivalence test was significant, $t(1218) = 3.38$, $p < .001$ given equivalence bounds of −0.1 and 0.1 (on a raw scale) and an alpha of 0.05. The null hypothesis test was non-significant, $t(1218) = -0.12$, $p = .907$, given an alpha of 0.05. Based on the equivalence test and the null-hypothesis test combined, we conclude no effect for action type on action choice. Fig. 2 shows participants' choice to act depending on the type of action (waving vs beckoning), there was no difference in participants responding depending on action type.

6.2.2. Effect of means/side-effect

A chi-squared test for independence revealed a significant association between participants' preference for action, and whether harm occurred as a means or as a side-effect of action $\chi^2(1, N = 1220) = 398.81$, $p < .001$, $V = 0.57$. The equivalence test was non-significant, $t(1074) = 20.93$, $p = 1.000$ given equivalence bounds of −0.08 and 0.08 (on a raw scale) and an alpha of 0.05. The null hypothesis test was significant, $t(1074) = 24.43$, $p < .001$, given an alpha of 0.05. Fig. 3 shows participants' choice to act depending on means/side-effect, participants were more likely to endorse action when harm occurred as a side-effect of than as a means to save five.

Having demonstrated that participants' preference for action does not differ between *Wave* and *Beckon*, we also tested for equivalence across the other measures, judgment, confidence, and trust; this is detailed in Table 1.

As a follow-up we also tested for equivalence across the other measures, judgment, confidence, and trust depending on means vs side-effect, see Table 2.

6.2.3. Combined effects of action-type and means/side-effect

Next we conducted a linear mixed model to test the combined effects of action-type and means/side-effect on action choice. Overall, the model significantly predicted action choice $\chi^2(3) = 661.29$, $p < .001$. Table 3 shows the that means/side-effect was the only significant predictor in the model, and action type (or scenario) did not predict action choice, nor did it interact with means/side-effect in predicting action choice.

We conducted an additional linear mixed model to test the range of possible influences on action-choice. Overall the model significantly predicted action choice $\chi^2(10) = 992.03$, $p < .001$. Table 4 shows the full model and the relevant predictors of action choice.

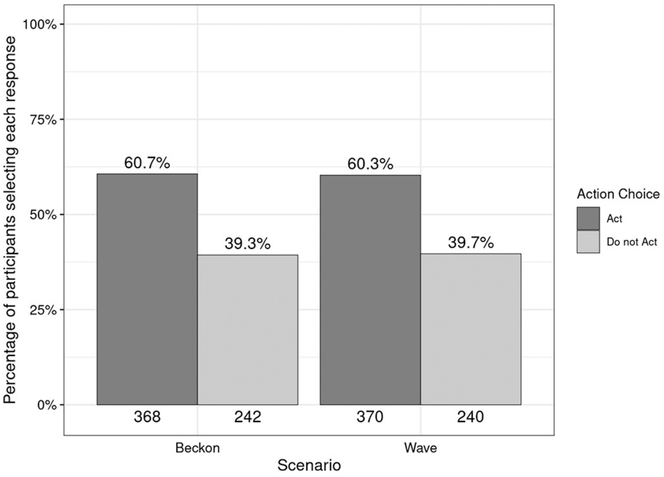


Fig. 2. Action type and action choice.

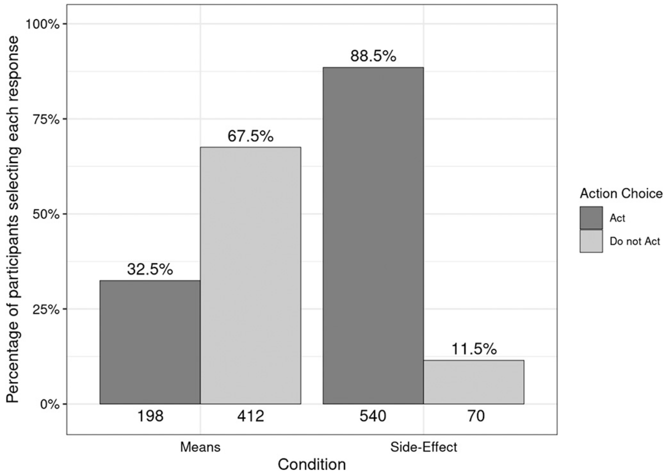


Fig. 3. Means/side-effect and action choice.

6.2.4. *Combined influences of action type and means/side-effect on other measures*

A series of 2 × 2 within-subjects ANOVAs tested combined influences of action-type and means/side-effect on the other measures taken.

A within-subjects ANOVA with confidence as the DV, and scenario and means/side-effect as IVs revealed no main effect for scenario, $F(1, 304) = 0.01$, $p = .920$, $\eta^2 < 0.001$; a significant main effect for means/side-effect, $F(1, 304) = 5$, $p = .026$, $\eta^2 = 0.02$; and no significant scenario × means/side-effect interaction $F(1, 304) = 0.69$, $p = .405$, $\eta^2 = 0.002$.

A within-subjects ANOVA with judgment as the DV, and scenario and means/side-effect as IVs revealed no main effect for scenario, $F(1, 304) = 3.88$, $p = .050$, $\eta^2 = 0.01$; a significant main effect for means/side-effect, $F(1, 304) = 377.5$, $p < .001$, $\eta^2 = 0.55$; and no significant scenario × means/side-effect interaction $F(1, 304) = 0.98$, $p = .324$, $\eta^2 = 0.003$.

A within-subjects ANOVA with trustworthiness of a friend as the DV, and scenario and means/side-effect as IVs revealed no main effect for scenario, $F(1, 304) = 0.002$, $p = .967$, $\eta^2 < 0.001$; a significant main effect for means/side-effect, $F(1, 304) = 246.4$, $p < .001$, $\eta^2 = 0.45$; and no significant scenario × means/side-effect interaction $F(1, 304) = 2.35$, $p = .126$, $\eta^2 = 0.01$.

A within-subjects ANOVA with trustworthiness of a partner as the DV, and scenario and means/side-effect as IVs revealed no main effect for scenario, $F(1, 304) = 0.02$, $p = .895$, $\eta^2 < 0.001$; a significant main effect for means/side-effect, $F(1, 304) = 236.26$, $p < .001$, $\eta^2 = 0.44$; and no significant scenario × means/side-effect interaction $F(1, 304) = 2.46$, $p = .118$, $\eta^2 = 0.01$.

A within-subjects ANOVA with trustworthiness of a politician as the DV, and scenario and means/side-effect as IVs revealed no main effect for scenario, $F(1, 304) < 0.001$, $p = 1.000$, $\eta^2 < 0.001$; a significant main effect for means/side-effect, $F(1, 304) = 208.43$, $p < .001$, $\eta^2 = 0.41$; and no significant scenario × means/side-effect interaction $F(1, 304) = 5.28$, $p = .022$, $\eta^2 = 0.02$.

Table 1
Equivalence tests for judgment, confidence, and trust depending on action type.

Measure	Test	M_{Wave}	SD_{Wave}	M_{Beckon}	SD_{Beckon}	t	df	p	95% CI
Confidence	Equivalence	5.03	1.64	5.03	1.61	3.55, −3.44	1218	<.001**, <.001**	[−0.149, 0.159]
	Difference	–	–	–	–	0.05	1218	.958	[−0.178, 0.188]
Judgment	Equivalence	4.07	1.86	4.17	1.84	2.63, −4.36	1218	.004*, <.001**	[−0.266, 0.082]
	Difference	–	–	–	–	−0.87	1218	.386	[−0.299, 0.116]
Trust Friend	Equivalence	4.58	1.63	4.58	1.68	3.51, −3.48	1217	< .001**, <.001**	[−0.155, 0.158]
	Difference	–	–	–	–	0.02	1217	.986	[−0.185, 0.188]
Trust Partner	Equivalence	4.53	1.69	4.52	1.75	3.54, −3.44	1217	<.001**, <.001**	[−0.157, 0.167]
	Difference	–	–	–	–	0.05	1217	.960	[−0.188, 0.198]
Trust Politician	Equivalence	3.97	1.78	3.97	1.82	3.49, −3.49	1217	<.001**, <.001**	[−0.169, 0.169]
	Difference	–	–	–	–	0	1217	1.000	[−0.202, 0.202]

Note. * = sig. at <.05; ** = sig. at <.001.

Table 2
Equivalence tests for judgment, confidence, and trust depending on means/side-effect.

Measure	Test	M_{Means}	SD_{Means}	$M_{Side\text{-}effect}$	$SD_{Side\text{-}effect}$	t	df	p	95% CI
Confidence	Equivalence	4.93	1.65	5.13	1.6	1.33, −5.66	1217	.092,<.001	[−0.355, −0.048]
	Difference	–	–	–	–	−2.17	1217	.031*	[−0.384, −0.019]
Judgment	Equivalence	3.22	1.72	5.02	1.5	−15.98, −22.97	1196	1.000, <.001	[−1.952, −1.648]
	Difference	–	–	–	–	−19.47	1196	<.001**	[−1.981, −1.619]
Trust Friend	Equivalence	3.92	1.67	5.25	1.35	−11.79, −18.78	1164	1.000, <.001	[−1.473, −1.186]
	Difference	–	–	–	–	−15.29	1164	<.001**	[−1.5, −1.159]
Trust Partner	Equivalence	3.85	1.72	5.2	1.42	−11.46, −18.44	1177	1.000, <.001	[−1.501, −1.204]
	Difference	–	–	–	–	−14.95	1177	<.001**	[−1.53, −1.175]
Trust Politician	Equivalence	3.37	1.74	4.57	1.64	−8.8, −15.79	1214	1.000, <.001	[−1.353, −1.034]
	Difference	–	–	–	–	−12.29	1214	< .001**	[−1.384, −1.003]

Note. * = sig. at <.05; ** = sig. at <.001.

Table 3
Combined influence of scenario and means/side-effect in predicting action choice.

Predictor	b	SE	df	t	p	2.5%	97.5%
(Intercept)	0.68	0.023	909	29.71	<.001**	0.64	0.73
Scenario	−0.02	0.025	909	−0.78	.438	−0.07	0.03
Means/side-effect	−0.58	0.025	909	−22.64	<.001**	−0.63	−0.53
Scenario × Means/side-effect	0.03	0.036	909	0.91	.361	−0.04	0.10

Note. * = sig. at <.05; ** = sig. at <.001.

Table 4
Combined influence of scenario and means/side-effect in predicting action choice.

Predictor	*b*	*SE*	*df*	*t*	*p*	2.5%	97.5%
(Intercept)	1.01	0.072	904.00	13.93	<.001**	0.86	1.15
Scenario	−0.02	0.023	904.00	−0.81	.416	−0.06	0.03
Means/side-effect	−0.35	0.026	904.00	−13.32	<.001**	−0.40	−0.30
Scenario × Means/side-effect	0.02	0.007	904.00	2.35	.019*	0.00	0.03
Confidence	−0.07	0.009	904.00	−7.82	<.001**	−0.09	−0.05
Judgment	0.03	0.018	904.00	1.71	.087	0.00	0.07
Trust: Friend	−0.10	0.017	904.00	−6.13	<.001**	−0.14	−0.07
Trust: Partner	0.00	0.010	904.00	−0.11	.910	−0.02	0.02
Trust: Politician	0.01	0.001	301.00	3.42	.001*	0.00	0.01
Age	−0.02	0.026	301.00	−0.58	.565	−0.07	0.04
Gender	0.02	0.032	904.00	0.58	.561	−0.04	0.08

Note. * = sig. at <.05; ** = sig. at <.001.

6.2.5. *Matched scenarios*

In addition to the aggregate analysis above, we directly compared corresponding versions of *Wave* and *Beckon* matched according to means/side-effect. First we compare *Wave (means)* with *Beckon (means)*, following this we compared *Wave (side-effect)* with *Beckon (side-effect)*.

Means. A chi-squared test for independence revealed no association between action type and participants' preference for action, $\chi^2(1, N = 610) = 0.19$, $p = .665$, $V = 0.02$. The equivalence test was significant, $t(608) = 1.95$, $p = .026$ given equivalence bounds of −0.09 and 0.09 (on a raw scale) and an alpha of 0.05. The null hypothesis test was non-significant, $t(608) = -0.52$, $p = .605$, given an alpha of 0.05. Based on the equivalence test and the null-hypothesis test combined, we conclude no effect for action type on action choice. Fig. 4 shows participants' choice to act depending on the type of action (waving vs beckoning), there was no difference in participants responding depending on action type. Having demonstrated that participants' preference for action does not differ between *Wave* and *Beckon*, we also tested for equivalence across the other measures, judgment, confidence, and trust (see Table 5).

Side-effect. A chi-squared test for independence revealed no association between action type and participants' preference for action, $\chi^2(1, N = 610) = 0.15$, $p = .703$, $V = 0.02$. The equivalence test was significant, $t(606) = -1.96$, $p = .025$ given equivalence bounds of −0.06 and 0.06 (on a raw scale) and an alpha of 0.05. The null hypothesis test was non-significant, $t(606) = 0.51$, $p = .612$, given an alpha of 0.05. Based on the equivalence test and the null-hypothesis test combined, we conclude no effect for action type on action choice. Fig. 5 shows participants choice to act depending on the type of action (waving vs beckoning), there was no difference in participants responding depending on action type. Having demonstrated that participants' preference for action does not differ between *Wave* and *Beckon*, we also tested for equivalence across the other measures, judgment, confidence, and trust (Table 6).

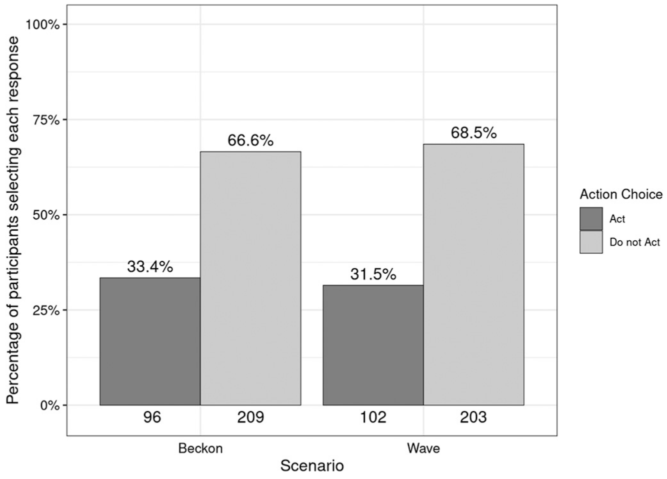


Fig. 4. Action Type and action choice.

6.3. *Discussion*

Study 1 demonstrates that the differences between *Wave* and *Beckon* described by Railton (2017) were most likely caused by participants making a distinction between causing harm as a *means* to an end vs as a *side-effect*, rather than making a distinction between *waving* vs *beckoning*. These findings support H2, and we reject H1 and H3. In Fig. 6, we have updated the figure proposed by Railton (2017) to include the results of Study 1. The *side-effect* versions of both *Wave* and *Beckon* have taken the place of *Switch* (and *Loop*), and the *means* versions of both have taken the place previously occupied by *Y* in Fig. 1. Thus, while the explanation for the observed pattern is different from that proposed by Railton (2017), the results are consistent with his claims. Importantly, these findings run counter to the predictions of classic dual-process approaches that focus on a distinction between "emotional" vs "cognitive" responses (e.g., Greene, 2008; Greene et al., 2001), or more contemporary dual-process theorizing relating linking judgments of actions and outcomes to model-free and model-based processes respectively (e.g., Crockett, 2013; Cushman, 2013), and therefore provide some evidence for his social learning approach to moral judgment. Study 1 also adds to the broader literature on the means/side-effect distinction (see Feltz & May, 2017), demonstrating that when actions and outcomes are held constant, participants do make a meaningful distinction between means and side-effects.

7. Study 2

Study 1 provided some evidence for Railton's social learning approach to moral judgment by demonstrating that people do respond according to *Y* in Fig. 1. We also found that this responding was driven by the means/side-effect distinction (people making a distinction between *means* vs *side-effect*). The aim of Study 2 is to build on this finding to (a) test for, and (b) attempt to explain cases where participants will respond according to *X*, that is, endorsing an intervention that uses personal force to harm one while saving five.

Given the importance of means/side-effect demonstrated in Study 1, we propose a test of the combined influences of personal force and means/side-effect in Study 2. This will serve the dual aims of disentangling the influences of personal force and means/side-effect (e.g., Feltz & May, 2017) while also providing a direct test of the claims of Railton (2017). As such, and in line with Railton (2017), we will include the *Switch* and *Footbridge* versions of the Trolley dilemma in Study 2. Furthermore, Railton (2017) demonstrated that for *Bus* his students do respond according to *X*, thus we will include *Bus*, as well an alternative

Table 5
Equivalence tests for judgment, confidence, and trust depending on action type.

Measure	Test	M_{Wave}	SD_{Wave}	M_{Beckon}	SD_{Beckon}	t	df	p	95% CI
Confidence	Equivalence	4.95	1.66	4.9	1.65	2.81, −2.13	608	.003*, .017*	[−0.175, 0.267]
	Difference	–	–	–	–	0.34	608	.732	[−0.217, 0.309]
Judgment	Equivalence	3.2	1.73	3.24	1.71	2.14, −2.8	608	.016*, .003*	[−0.276, 0.184]
	Difference	–	–	–	–	−0.33	608	.742	[−0.32, 0.228]
Trust Friend	Equivalence	3.95	1.65	3.89	1.7	2.93, −2.01	608	.002*, .022*	[−0.161, 0.286]
	Difference	–	–	–	–	0.46	608	.646	[−0.204, 0.329]
Trust Partner	Equivalence	3.88	1.7	3.82	1.75	2.94, −2	608	.002*, .023*	[−0.164, 0.295]
	Difference	–	–	–	–	0.47	608	.638	[−0.208, 0.339]
Trust Politician	Equivalence	3.41	1.73	3.33	1.76	3.05, −1.89	608	.001*, .030*	[−0.151, 0.315]
	Difference	–	–	–	–	0.58	608	.562	[−0.196, 0.359]

Note. * = sig. at <.05; ** = sig. at <.001.

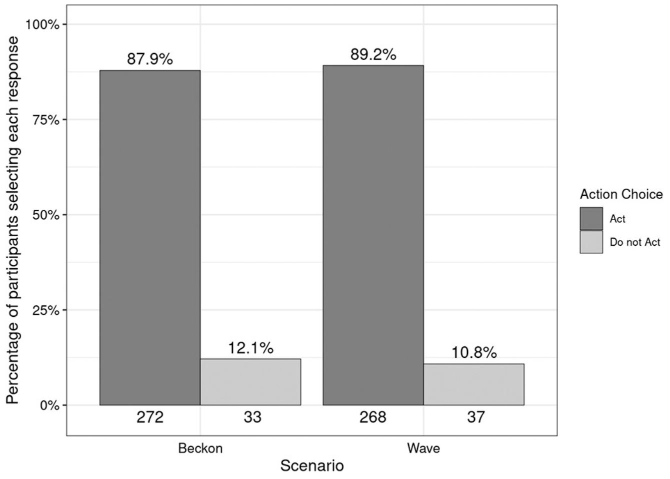


Fig. 5. Action type and action choice.

version of *Bus* specifically developed to control for means/side-effect. In addition, we developed two additional scenarios *Car* and *Escalator* to be combined with *Bus* and *Trolley* (*encompassing Footbridge, Switch, Wave* and *Beckon*). Using modified versions of these four dilemmas will enable us to systematically (and plausibly) manipulate force and means and test for their combined influence on responding. In addition, we will test for other dispositional influences on responding. Our hypotheses are as follows:

H1. Personal Force: Participants will be more willing to endorse actions that do not involve personal force compared to actions that do involve personal force.

H2. Means/side-effect: Participants will be more willing to endorse actions where harm occurs as a side-effect of saving five, than actions where harm occurs as a means to save five.

H3. : Perceived trustworthiness of an imagined actor (friend, romantic partner, politician) who commits an action predicts action choice.

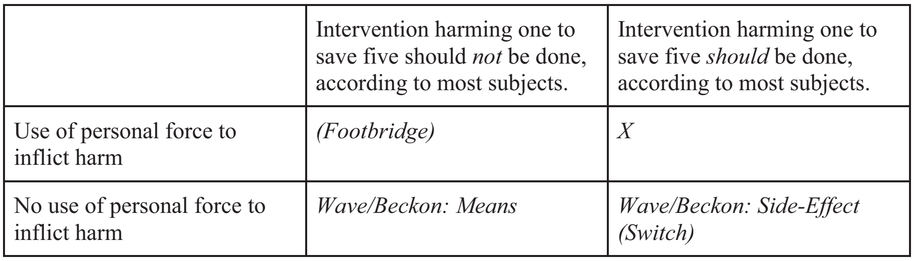
Unlike Study 1, we do not anticipate the effects of personal force and means/side-effect to be mutually exclusive. We additionally expect that responding will be predicted by dispositional variables measured (thinking styles, CRT, and scores on OUS and Complete Consequentialism Scale subscales). We do not make specific directional predictions relating to these variables, save participants scores on instrumental harm:

H4. Participants who score higher on instrumental harm will show a greater tendency to endorse action across all scenarios.

Table 6
Equivalence tests for judgment, confidence, and trust depending on action type.

Measure	Test	M_{Wave}	SD_{Wave}	M_{Beckon}	SD_{Beckon}	t	df	p	95% CI
Confidence	Equivalence	5.11	1.62	5.15	1.58	2.19, −2.75	607	.014*, .003*	[−0.249, 0.177]
	Difference	–	–	–	–	−0.28	607	.781	[−0.29, 0.218]
Judgment	Equivalence	4.95	1.54	5.09	1.46	1.34, −3.6	607	.091, <.001	[−0.338, 0.062]
	Difference	–	–	–	–	−1.13	607	.258	[−0.376, 0.101]
Trust Friend	Equivalence	5.22	1.34	5.28	1.35	1.93, −3.01	608	.027*, .001*	[−0.239, 0.121]
	Difference	–	–	–	–	−0.54	608	.589	[−0.273, 0.155]
Trust Partner	Equivalence	5.17	1.41	5.23	1.44	1.99, −2.95	608	.024*, .002*	[−0.246, 0.134]
	Difference	–	–	–	–	−0.48	608	.629	[−0.282, 0.171]
Trust Politician	Equivalence	4.52	1.64	4.61	1.65	1.85, −3.08	608	.032*, .001*	[−0.302, 0.138]
	Difference	–	–	–	–	−0.62	608	.539	[−0.344, 0.18]

Note. * = sig. at <0.05; ** = sig. at <0.001.



	Intervention harming one to save five should *not* be done, according to most subjects.	Intervention harming one to save five *should* be done, according to most subjects.
Use of personal force to inflict harm	*(Footbridge)*	*X*
No use of personal force to inflict harm	*Wave/Beckon: Means*	*Wave/Beckon: Side-Effect (Switch)*

Fig. 6. Results of Study 1 in the context of current theorizing (adapted from Railton, 2017).

7.1. Method

7.1.1. Design

This study is a 2 × 2 within-subjects design. The primary dependent variable is action choice with two levels (act vs do not act). In all cases, the action leads to a characteristically utilitarian outcome (sacrificing one innocent person to save five innocent people). Responses will be recorded using a "Yes" / "No" binary following a question "Should you [commit the named action]?" (See Appendix B). The first independent variable is use of direct force with two levels: *present* vs *absent*. The second independent variable is means/side-effect with two levels: *means* vs *side-effect*.

While the current approach expands on the work of Railton (2017), offering more a more rigorous test of the core hypotheses (larger sample size, systematic operationalization of independent variables across multiple descriptions) we note some key limitations as avenues for future research. In particular, we note a possible confound between the pro-sacrificial utilitarian tendencies and action tendencies: across all our scenarios, the characteristically utilitarian pro-sacrificial response involves action (see Gawronski et al., 2017). Thus, any observed preferences for or against the utilitarian sacrifice could reflect preferences for action/inaction, and our design cannot differentiate between these two tendencies. However, we note that this is beyond the scope of the current research. By holding both consequences, and the relationship between action/inaction and these consequences constant across scenarios, our study aims to investigate the independent influences of force/no force and means/side-effect on people's moral judgments. Future research can combine our insights with more complex designs to also investigate action/inaction and characteristically utilitarian/anti-utilitarian tendencies.

In addition to the primary dependent variable, participants will also rate their confidence in their action choice (1 = *Not at all confident*, 7 = *Extremely confident*), how right or wrong the action is (1 = *Wrong*, 4 = *Neutral*, 7 = *Right*), and how trustworthy (1 = *Not at all trustworthy*, 7 = *Extremely trustworthy*) they would rate each of a friend, a romantic partner, and a politician, if they heard these people chose to commit the act described. As in Study 1, these will be included as dependent measures in preliminary analyses, however they will also be included as potential predictor variables in the overall regression, that is, does perceived trustworthiness of a third party committing the action predict participants' own endorsing of the action?

Four additional potential predictor measures will be recorded: the Oxford Utilitarianism Scale (which has two subscales, Instrumental Harm and Impartial Beneficence; Kahane et al., 2018), the Complete Consequentialism Scale (containing two subscales, a deontological subscale, and a utilitarian subscale; Plaks et al., 2021), the Cognitive Reflection Test (Toplak et al., 2014), and the Comprehensive Thinking Styles Questionnaire (which has four subscales: Actively Open-minded Thinking, Close-Minded Thinking, Preference for Intuitive Thinking, Preference for Effortful Thinking; Newton et al., 2021).

7.1.2. Participants

Participants will be recruited from a range of research participation platforms including MTurk, Prolific, and Lucid. Participants will also be recruited from the student body at University of Limerick. Furthermore, additional Irish participants may be recruited as part of a larger study by Economic and Social Research Institute in Ireland. Power simulations (1000 simulations) indicated that a sample size of $N = 2200$ would be able to detect a medium interaction ($b = 0.3$) between means/side-effect and direct force with 92.50% power. Thus our target sample is 2200 participants (after exclusions). We will employ the same attention checks as in Study 1, participants who fail both attention checks will be excluded from analysis.

7.1.3. Materials

The entire survey will be programmed in Qualtrics (Qualtrics, 2020). Participants will be presented with each of the moral vignettes (*Trolley*, *Car*, *Bus*, *Escalator*) in random order. Each vignette has at least two versions designed to systematically manipulate our independent variables (force and means). The presentation of different versions of these vignettes will be randomized, such that all participants will be exposed to all experimental conditions across the four vignettes (see Table 7 for breakdown of the different versions of each vignette map onto the different experimental conditions).

Each vignette concludes with a question asking for participants to indicate if they should complete the action described at the end of the vignette (e.g., *Car*: "Should you continue with swerving your vehicle?"), responses to this question will be recorded using a binary yes/no response option. Following this, participants will be asked how trustworthy they would rate each of a friend, a romantic partner, and a politician if they learned these people committed the named act (see Appendix B).

The Oxford Utilitarianism Scale (OUS, Kahane et al., 2018) is a nine-item scale with two subscales that assess the degree to which people endorse Instrumental Harm (sacrificing one in order to save a greater number), or Impartial Beneficence (impartial concern for the well-being of everyone). All items are scored on a 7-point Likert scale (1 = *Strongly disagree*; 7 = *Strongly agree*). A sample Instrumental Harm items reads: "It is morally right to harm an innocent person if harming them is a necessary means to helping several other people"; while a sample Impartial Beneficence item reads: "From a moral perspective, people should care about the well-being of people who are especially close to them either physically or emotionally".

The Cognitive Reflection Test (CRT, Frederick, 2005; Thomson & Oppenheimer, 2016; Toplak, West, & Stanovich, 2011; Toplak et al., 2014) provides a measure of people's tendency to over-ride intuitive (habitual) responses and engage in deliberation to ensure accuracy in responding. We propose to use the seven item version of the CRT (Toplak et al., 2014). Participants are presented with a series of questions for which there is an answer that seems intuitive, but is incorrect (e.g., "A bat and a ball cost $1.10 in total. The bat costs a dollar more than the ball. How much does the ball cost?" Correct answer = 5 cents; intuitive answer = 10 cents). Previous work has found that people who score higher on the CRT are more likely to provide give pro-sacrificial responses to sacrificial moral dilemmas (Baron, Scott, Fincher, & Emlen Metz, 2015; Byrd & Conway, 2019).

The Comprehensive Thinking Styles Questionnaire (CTSQ, Newton et al., 2021), is a twenty-four-item questionnaire that measures people's thinking styles across four dimensions: Actively Open-minded Thinking, Close-Minded Thinking, Preference for Intuitive Thinking, Preference for Effortful Thinking, with a six item subscale for each dimension; all responses are recorded on a 6-point Likert scale (1 = *strongly disagree*; 6 = *strongly agree*). Actively Open-minded Thinking (AOT) is a tendency to engage in reflectivity, and a willingness to consider alternative opinions, explanations, and evidence that may contradict existing beliefs (Newton et al., 2021; Stanovich & Toplak, 2019). A sample item from the AOT subscale reads "Just because evidence conflicts with my current beliefs does not mean my beliefs are wrong." (all items reverse scored). Close-Minded Thinking (CMT) is a tendency to view things in absolute terms, and may be seen as similar to dogmatism (Newton et al., 2021). A sample item from the CMT subscale is "Either something is true or it is

Table 7
Breakdown of vignettes and experimental conditions.

	Means	Side-effect
Force	*Trolley (Footbridge)* *Bus* *Escalator*	– *Bus* *Escalator*
No Force	*Trolley (Wave/Beckon)* *Car* –	*Trolley (Wave/Beckon/Switch)* *Car* *Escalator*

false; there is nothing in-between.”. A sample item from the Preference for Intuitive Thinking (PIT) subscale reads “I believe in trusting my hunches”. A sample item from the Preference for Effortful thinking (similar to Need for Cognition, see Cacioppo & Petty, 1982; Newton et al., 2021) reads “Thinking is not my idea of an enjoyable activity.” (all items reverse scored).

7.1.4. Procedure

Participants will be provided with a link to the survey. On clicking the link, participants will be presented with an information sheet providing details of what their participation will entail. Following this, participants will be presented with the consent form. The main survey will only become accessible when participants have provided their explicit consent to take part. If participants do not consent to take part, they will be diverted to the end of the survey.

Once consent has been granted, participants will complete a moral judgment task based on the scenarios in Appendix A. For each scenario, participants will be asked the associated questions outlined in Appendix B. Participants will be presented with a version of each of the four different scenarios (*Trolley*, *Bus*, *Escalator*, *Car*). The order of presentation of scenarios will be randomized. When the moral judgment task has been completed participants will complete the additional measures (in randomized order).

7.1.5. Analysis plan

First we will test for relationships between the different variables. We will test for simple associations between the primary dependent variable (action choice) and the experimental conditions using a chi-squared test. Exploratory independent samples *t*-tests will test for simple effects on participants’ confidence in their action choice, and the various trustworthiness judgments. We will conduct factorial ANOVAs to test the combined influences of personal force and means on responding. We will also test for relationships between the various dependent variables using correlation analyses, and regression (where action choice is the outcome variable). We will run correlational analyses to test for relationships between the predictor variables (OUS, CRT, and the subscales of the CTSQ). We will also test for relationships between these and the various dependent variables, this will be done for the entire sample and for each condition individually.

The primary analysis will be a series of linear mixed effects models, with participant included as a random factor, action choice as the dependent variable and experimental conditions, predictor variables, and other dependent measures entered as separate blocks. While testing for influences on action choice is the primary test of interest, we will additionally run separate regression models with each of the dependent variables as outcome variables. We will also test and control for order effects in our analyses (Sample analysis using simulated data is available on the OSF page).

7.2. Results

7.2.1. Participants

An initial sample of $N = 2346$ took part. Participants and were recruited through convenience/snowball sampling in Ireland ($n = 294$), the student population at the University of Limerick using the SONA credit system ($n = 536$), UK based Prolific participants ($n = 335$), and USA based Prolific participants ($n = 1181$). After the exclusion of participants who failed both attention checks, we were left with a final sample of $N = 2213$, (Irish convenience / snowball: $n = 213$, Irish SONA: $n = 506$, UK Prolific: $n = 332$, USA Prolific: $n = 1162$). Convenience/snowball participants were not reimbursed. SONA participants were awarded course credit for their participation. Prolific participants were paid £3.15 Sterling for their participation. The age and gender breakdown for the final sample is as follows, female = 862, male = 1250, other = 29, $M_{age} = 35$, $SD = 14.6$, min = 18, max = 94.

7.2.2. Effect of force

To test our first hypothesis, we conducted a chi-squared test and found a significant association between the presence/absence of direct force and participants’ preference for action, $\chi^2(1, N = 8852) = 89.04$, $p < .001$, $V = 0.1$ Participants were more likely to recommend action when action did not involve direct force, see Fig. 7, thus overall our first hypothesis was supported. This association between direct force and a preference for inaction held for both the *Trolley*, $\chi^2(1, N = 8852) = 156.31$, $p < .001$, $V = 0.13$, and *Escalator*, $\chi^2(1, N = 8852) = 181.4$, $p < .001$, $V = 0.14$, dilemmas (we also conducted follow-up logistic regressions to provide a more robust test and again the effects of force held for both *Trolley*, $p < .001$, and *Escalator*, $p < .001$, see Supplementary Analyses). All versions of *Bus* involved direct force, and all versions of *Car* involved no direct force. This meant it was not possible to test for the effect of direct force for either scenario individually.

7.2.3. Effect of means/side-effect

To test our second hypothesis, we conducted a chi-squared test and found a significant association between means/side-effect and participants’ preference for action, $\chi^2(1, N = 8852) = 477.72$, $p < .001$, $V = 0.23$. Participants were more likely to recommend action when harm occurred as a side-effect than when harm occurred as a means to save five, providing overall support for our second hypothesis, see Fig. 8. This effect of means vs side-effect held for all scenarios. *Trolley*: $\chi^2(1, N = 8852) = 156.31$, $p < .001$, $V = 0.13$; *Escalator*: $\chi^2(1, N = 8852) = 181.4$, $p < .001$, $V = 0.14$; *Bus*: $\chi^2(1, N = 8852) = 207.11$, $p < .001$, $V = 0.15$; *Car*: $\chi^2(1, N = 8852) = 257.58$, $p < .001$, $V = 0.17$. We note that these results may be confounded by unequal cell sizes for means/side-effect for the individual scenarios, and therefore advise caution when interpreting these scenario level results. To mitigate this, we also conducted a series of logistic regressions and found significant effects for means for each of *Trolley*, $p < .001$, *Bus*, $p < .001$, and *Escalator*, $p < .001$, but not for *Car*, $p = .489$ (see Supplementary Analyses).

7.2.4. Combined effects of force and means/side-effect

To assess the combined effects of direct force and means on action choice, we computed a series of mixed effects models, with participants entered as random effects and the variables of interest entered as fixed effects. The combined model with both direct force and means included was a better fit (AIC = 10,900.7, BIC = 10,936.1, Log-likelihood = −5445.4) than the models containing either direct force (AIC = 11,569.9, BIC = 11,598.3, Log-likelihood = −5781, Log ratio = 671.2, $p < .001$) or means (AIC = 11,029.4, BIC = 11,057.8, Log-likelihood = −5510.7, Log ratio = 130.7, $p < .001$) only. A final model included a force × means interaction term. Investigation of the AIC suggested the

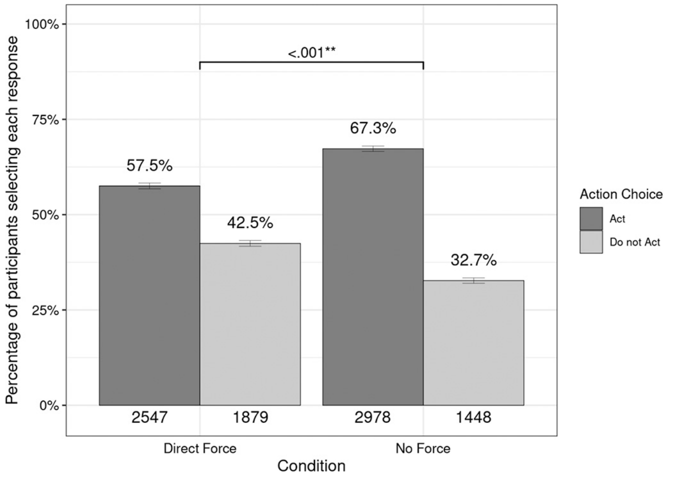


Fig. 7. Differences in action choice depending on whether it involved direct force.

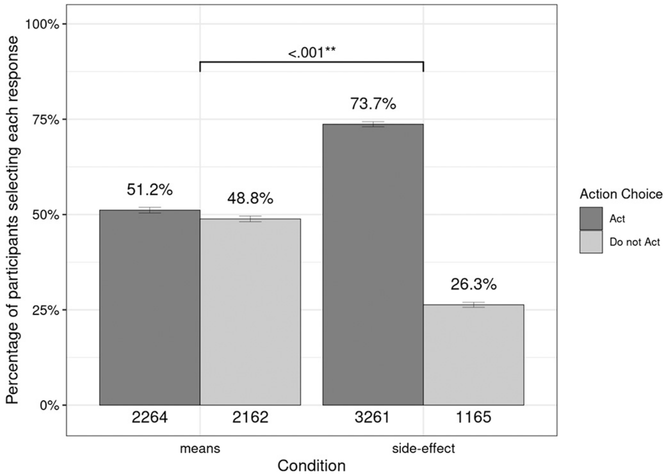


Fig. 8. Differences in action choice depending on whether harm occurred as a means vs side-effect.

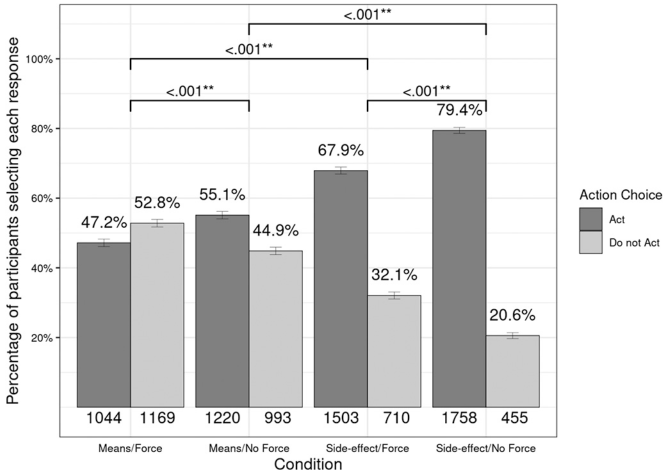


Fig. 9. Variation in action choice depending on both force/no force and means/side-effect.

inclusion of this interaction term provided a better fit, however, this was not reflected in the BIC (AIC = 10,898.3, BIC = 10,940.8, Log-likelihood = −5443.1, Log ratio = 4.4, p = .035). Below we present follow-up analyses to attempt to interpret this possible interaction, however, given this AIC vs BIC uncertainty, we advise caution in interpreting these analyses. Table 8 shows that each force, means, and the force × interaction are all significant predictors in the model, providing further support for both our first and second hypotheses. Table 8 also suggests that Means is the strongest predictor in the model.

To investigate the possible interaction between means and direct force, we split the data first by means/side-effect, and then by direct force/no force. Using these split data we conducted a series of tests to investigate (a) if the effect of force/no force varies depending on means/side-effect, and (b) if the effect of means/side-effect varies depending on the presence/absence of direct force. We report each in turn.

When the data were split according to means/side-effect there was a significant effect of direct force in both the means, $\chi^2(1, N = 4426) = 27.69, p < .001, V = 0.08$, and side effect conditions, $\chi^2(1, N = 4426) = 75.16, p < .001, V = 0.13$. Interestingly, the effect size was larger in the side-effect condition than in the means condition. This suggests that when harm occurs as a means to save five, people are less concerned about whether or not the harm involved direct force.

Similarly, when the data were split according to force/no force there was a significant effect of means in both the direct force, $\chi^2(1, N = 4426) = 193.99, p < .001, V = 0.21$, and no force conditions, $\chi^2(1, N = 4426) = 295.98, p < .001, V = 0.26$. The effect size was larger in the no force condition.

The combined influences of Force and Means are displayed in Fig. 9. When harm was caused as a means, *and* involved direct force, participants preferred inaction over action. Conversely, when harm was a side-effect, and no direct force was involved participants preferred action over inaction. The presence of either direct force or harm as a means reduced participants' willingness to endorse action, and this effect appears to be stronger for means than for force (see Table 8, and Fig. 9).

7.2.5. *Effects on other variables*

Having shown the effects of both force and means on action choice, we tested their influences on other variables of interest: confidence in their own action choice, moral judgment of acting, trust in a friend / partner / politician who chose to act. First, we conducted a series of t-tests and equivalence tests to test for differences in these variables depending on either direct force, or means. Following this, we conducted a series of factorial ANOVAs to test for the combined influences of force and means on participants' responding. The results of the t-tests are displayed in Tables 9 and 10 (Table 9: Force, Table 10: Means).

With the exception of confidence, all scores were significantly higher in the no force condition than in the force condition. Participants rated action in the absence of force as more morally right, and would place higher trust in a friend, a partner, and a politician if they chose to act when acting did not involve force than if it involved force.

Again, with the exception of confidence, all scores were significantly higher in the means condition than in the side-effect condition. Acting was rated as more morally right if harm occurred as a side-effect than as a means. A friend, a partner, and a politician were trusted more if they acted when harm occurred as a side-effect than as a means to save five.

To test the combined influences of means and force on participants judgments we conducted a series of factorial ANOVAs. The results of these ANOVAs are reported below, and can be interpreted by referring to Fig. 10 below.

A within-subjects factorial ANOVA with confidence as the DV, and force and means as IVs revealed no significant main effect for force, $F(1{,}2212) = 0.35, p = .552, \eta^2 = 0$, ($M_{Force} = 4.5, SD_{Force} = 1.8, M_{No\text{-}Force} = 4.5, SD_{No\text{-}Force} = 1.8$). There was a significant main effect for means, $F(1, 2212) = 5.53, p = .019, \eta^2 = 0$ with participants reporting significantly more confidence in the side-effect condition ($M = 4.5, SD = 1.8$) than in the means condition ($M = 4.5, SD = 1.8$). We also found a significant force × means interaction $F(1, 2212) = 51.2, p < .001, \eta^2 = 0.02$, ($M_{Means:Force} = 4.6, SD_{Means:Force} = 1.8, M_{Means:No\text{-}Force} = 4.4, SD_{Means:No\text{-}Force} = 1.8, M_{Side\text{-}Effect:Force} = 4.4, SD_{Side\text{-}Effect:Force} = 1.8, M_{Side\text{-}Effect:No\text{-}Force}$

Table 8
Combined influences of direct-force and means in predicting action choice.

Predictor	*b*	*SE*	*df*	*t*	*p*	2.5%	97.5%
(Intercept)	0.79	0.010	6636	79.76	<.001**	0.77	0.81
Force	−0.12	0.012	6636	−9.61	<.001**	−0.14	−0.09
Means	−0.24	0.012	6636	−20.28	<.001**	−0.27	−0.22
Force × Means	0.04	0.017	6636	2.11	.035*	0.00	0.07

Note. * = sig. at $p < .05$; ** = sig. at $p < .001$; Action choice variable: 0 = Inaction, 1 = Action, Force variable: 0 = No Force, 1 = Force, Means variable: 0 = Side-Effect, 1 = Means.

Table 9
Variation in confidence, judgments, and trust, depending on force.

Measure	Test	M_{Force}	SD_{Force}	$M_{No\ Force}$	$SD_{No\ Force}$	t	df	p	95% CI
Confidence	Equivalence	4.51	1.76	4.52	1.78	2.24, -3.08	8849	.013*, .001*	[−0.078, 0.046]
	Difference	–	–	–	–	−0.42	8849	.674	[−0.09, 0.058]
Judgment	Equivalence	3.59	1.71	4.09	1.65	−11, -16.61	8839	1.000, <.001	[−0.551, −0.434]
	Difference	–	–	–	–	−13.8	8839	<.001**	[−0.562, −0.422]
Trust Friend	Equivalence	4.45	1.49	4.7	1.39	−4.82, -11.35	8800	1.000, <.001	[−0.298, −0.197]
	Difference	–	–	–	–	−8.09	8800	<.001**	[−0.308, −0.188]
Trust Partner	Equivalence	4.52	1.56	4.78	1.45	−4.75, -10.99	8799	1.000, <.001	[−0.305, −0.199]
	Difference	–	–	–	–	−7.87	8799	<.001**	[−0.315, −0.189]
Trust Politician	Equivalence	3.78	1.64	4.02	1.56	−4.34, -10.21	8831	1.000, <.001	[−0.304, −0.192]
	Difference	–	–	–	–	−7.28	8831	<.001**	[−0.314, −0.181]

Note. * = sig. at <.05; ** = sig. at <.001.

Table 10
Variation in confidence, judgments, and trust, depending on means.

Measure	Test	M_{Means}	SD_{Means}	$M_{Side\text{-}effect}$	$SD_{Side\text{-}effect}$	t	df	p	95% CI
Confidence	Equivalence	4.48	1.77	4.55	1.77	2.37, -5.61	8850	.009*, <.001**	[−0.123, 0.001]
	Difference	–	–	–	–	−1.62	8850	.105	[−0.135, 0.013]
Judgment	Equivalence	3.51	1.69	4.17	1.64	−14.24, -22.72	8843	1.000, <.001	[−0.712, −0.595]
	Difference	–	–	–	–	−18.48	8843	<.001**	[−0.723, −0.584]
Trust Friend	Equivalence	4.33	1.47	4.82	1.38	−11.26, -21.17	8810	1.000, <.001	[−0.541, −0.441]
	Difference	–	–	–	–	−16.22	8810	<.001**	[−0.551, −0.432]
Trust Partner	Equivalence	4.39	1.54	4.91	1.43	−11.63, -21.11	8798	1.000, <.001	[−0.57, −0.466]
	Difference	–	–	–	–	−16.37	8798	<.001**	[−0.58, −0.456]
Trust Politician	Equivalence	3.67	1.62	4.13	1.56	−8.92, -17.8	8841	1.000, <.001	[−0.507, −0.396]
	Difference	–	–	–	–	−13.36	8841	<.001**	[−0.518, −0.385]

Note. * = sig. at < .05; ** = sig. at <.001.

= 4.6, $SD_{Side\text{-}Effect:No\text{-}Force}$ = 1.8). Participants were more confident in their judgments when both means and direct force were either both present or both absent (no force and side-effect), than when only one was present (e.g., means-no force or side-effect-force).

A within-subjects factorial ANOVA with judgment as the DV, and force and means as IVs revealed a significant main effect for force, F(1, 2212) = 287.64, $p < .001$, $\eta^2 = 0.12$, participants rated action as more favourable in the no force condition (M = 4.1, SD = 1.6) than in the force condition (M = 3.6, SD = 1.7). There was also a significant main effect for means, F(1, 2212) = 630.75, $p < .001$, $\eta^2 = 0.22$ with judgments significantly more favourable in the side-effect condition (M = 4.2, SD = 1.6) than in the means condition (M = 3.5, SD = 1.7). We also found a significant force × means interaction F(1, 2212) = 44.75, $p < .001$, η^2 = 0.02, ($M_{Means:Force}$ = 3.3, $SD_{Means:Force}$ = 1.7, $M_{Means:No\text{-}Force}$ = 3.7, $SD_{Means:No\text{-}Force}$ = 1.6, $M_{Side\text{-}Effect:Force}$ = 3.8, $SD_{Side\text{-}Effect:Force}$ = 1.7, $M_{Side\text{-}Effect:No\text{-}Force}$ = 4.5, $SD_{Side\text{-}Effect:No\text{-}Force}$ = 1.5).

A within-subjects factorial ANOVA with trust in a friend as the DV, and force and means as IVs revealed a significant main effect for force, F (1, 2212) = 110.24, $p < .001$, $\eta^2 = 0.05$, participants reported higher trust in the no force condition (M = 4.7, SD = 1.4) than in the force condition (M = 4.4, SD = 1.5). There was also a significant main effect for means, F(1, 2212) = 526.64, $p < .001$, $\eta^2 = 0.19$ with significantly higher trust ratings in the side-effect condition (M = 4.8, SD = 1.4) than in the means condition (M = 4.3, SD = 1.5). There was also a significant force × means interaction F(1, 2212) = 12.56, $p < .001$, $\eta^2 = 0.01$, ($M_{Means:Force}$ = 4.2, $SD_{Means:Force}$ = 1.5, $M_{Means:No\text{-}Force}$ = 4.4, $SD_{Means:No\text{-}Force}$ = 1.4, $M_{Side\text{-}Effect:Force}$ = 4.7, $SD_{Side\text{-}Effect:Force}$ = 1.4, $M_{Side\text{-}Effect:No\text{-}Force}$ = 5, $SD_{Side\text{-}Effect:No\text{-}Force}$ = 1.3).

A within-subjects factorial ANOVA with trust in a partner as the DV, and force and means as IVs revealed a significant main effect for force, F (1, 2212) = 107.58, $p < .001$, $\eta^2 = 0.05$, participants reported higher trust in the no force condition (M = 4.8, SD = 1.4) than in the force condition (M = 4.5, SD = 1.6). There was also a significant main effect for means, F(1, 2212) = 549.9, $p < .001$, $\eta^2 = 0.2$ with significantly higher trust ratings in the side-effect condition (M = 4.9, SD = 1.4) than in the means condition (M = 4.4, SD = 1.5). There was also a significant force × means interaction F(1, 2212) = 7.89, p = .005, $\eta^2 = 0$, ($M_{Means:Force}$ = 4.3, $SD_{Means:Force}$ = 1.6, $M_{Means:No\text{-}Force}$ = 4.5, $SD_{Means:No\text{-}Force}$ = 1.5, $M_{Side\text{-}Effect:Force}$ = 4.8, $SD_{Side\text{-}Effect:Force}$ = 1.5, $M_{Side\text{-}Effect:No\text{-}Force}$ = 5.1, $SD_{Side\text{-}Effect:No\text{-}Force}$ = 1.4).

A within-subjects factorial ANOVA with trust in a politician as the DV, and force and means as IVs revealed a significant main effect for force, F(1, 2212) = 112.22, $p < .001$, $\eta^2 = 0.05$, participants reported higher trust in the no force condition (M = 4, SD = 1.6) than in the force

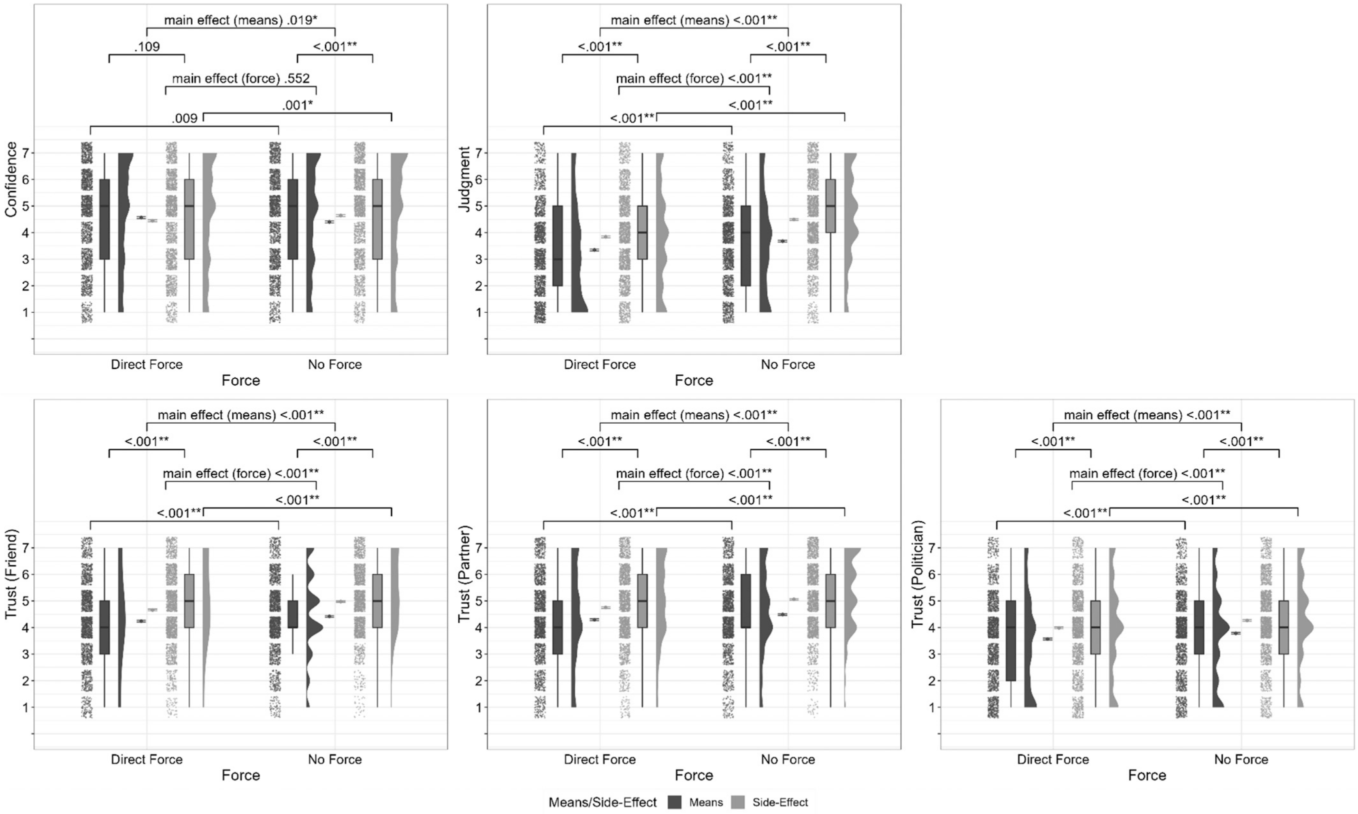


Fig. 10. Variation in confidence, judgment, and trust (friend/partner/politician) depending on both force/no force and means/side-effect.

condition ($M = 3.8$, $SD = 1.6$). There was also a significant main effect for means, $F(1, 2212) = 429.88$, $p < .001$, $\eta^2 = 0.16$ with significantly higher trust ratings in the side-effect condition ($M = 4.1$, $SD = 1.6$) than in the means condition ($M = 3.7$, $SD = 1.6$). There was no significant force × means interaction $F(1, 2212) = 2.2$, $p = .138$, $\eta^2 = 0$, ($M_{Means:Force} = 3.6$, $SD_{Means:Force} = 1.6$, $M_{Means:No\text{-}Force} = 3.8$, $SD_{Means:No\text{-}Force} = 1.6$, $M_{Side\text{-}Effect:Force} = 4$, $SD_{Side\text{-}Effect:Force} = 1.6$, $M_{Side\text{-}Effect:No\text{-}Force} = 4.3$, $SD_{Side\text{-}Effect:No\text{-}Force} = 1.5$).

7.2.6. *Combined influence of all variables on action choice*

Our third hypothesis was that perceived trustworthiness of an imagined actor (friend, romantic partner, politician) would predict action choice, and our fourth hypothesis was that participants who score higher on instrumental harm would show a greater tendency to endorse action. Our final analysis was a mixed effects model to test both hypotheses, as well as allowing us to conduct an exploratory test of whether action choice is predicted by the other dispositional variables measured (thinking styles, CRT, and scores on OUS and Complete Consequentialism Scale subscales).

In addition to the mixed effects models reported above (testing hypotheses 1 and 2) we computed a further mixed effects model, with participants entered as random effects and the remaining variables of interest entered as fixed effects. The inclusion of these additional predictor variables improved the fit (AIC = 7966.3, BIC = 8128.3 Log-likelihood = −3960.2, Log ratio = 2488.1, $p < .001$) when compared to the model that included direct force, means and the direct force × means interaction (AIC = 10,420.4, BIC = 10,462.7, Log-likelihood = −5204.2). The full model is presented in Table 11.

Trust in all three imagined actors (friend, partner, politician) who committed the action was a significant predictor in the model, with increased trust positively predicting action, supporting our third hypothesis. Higher scores on instrumental harm also positively predicted action, supporting our fourth hypothesis. Additional predictors can be found in Table 11, including Age (+), Impartial Beneficence (+), the Deontology sub scale of the Complete Consequentialism Questionnaire (−), Cognitive Reflection (−), and an Intuitive Thinking Style (−).

7.3. *Discussion*

Study 2 was a replication-extension of the empirical work presented by Railton (2017). Using the *Bus* dilemma developed by Railton, as well as newly developed *Escalator* and *Car* dilemmas we showed, consistent with Railton (2017), (a) that there are situations in which participants do endorse an intervention (harming one to save five) that involves personal force to inflict harm (position *X* in Fig. 1), and, (b) there are situations where participants reject an intervention (harming one to save five) when there is no use of personal force to inflict harm (position *Y* in Fig. 1). In Fig. 11 we have revised Fig. 1 to include the scenarios that lead to action/inaction in positions *X* and *Y* (see Table 12 for an overview of the key features of the different scenarios to aid in interpreting Fig. 11). These findings are consistent with the critique of dual-process approaches presented by Railton (2017), and arguably provide evidence in support of alternative approaches that can better account for the variability and complexity of moral judgment, including his causal-evaluative modelling/social learning approach (Railton, 2017), but also categorization approaches (McHugh et al., 2022), and mental models approaches (Bucciarelli et al., 2008). Furthermore, we extend on the work by Railton (2017), by providing novel insights into the unique and shared influences of specific factors on participants' moral decision making in sacrificial dilemmas (including features of the scenario – *means/side-effect*, *force/no-force* – and participant level individual differences).

Despite the consistency with Railton's (2017) argument, we note that our findings only provide evidence against a very narrow view of dual-process approaches, a view that places a strong emphasis on personal force (e.g., Greene, 2008; Greene et al., 2001). However, more recent

Table 11
Combined influence of scenario, means, and other measures in predicting action choice.

Predictor	*b*	*SE*	*df*	*t*	*p*	2.5%	97.5%
(Intercept)	0.05	0.049	6333	1.03	.302	−0.05	0.15
Force	−0.05	0.011	6333	−4.43	<.001**	−0.07	−0.03
Means	−0.14	0.011	6333	−12.37	<.001**	−0.16	−0.12
Judgment	0.06	0.004	6333	17.92	<.001**	0.06	0.07
Confidence	−0.03	0.003	6333	−12.21	<.001**	−0.04	−0.03
Trust: Friend	0.06	0.008	6333	7.59	<.001**	0.04	0.07
Trust Partner	0.04	0.007	6333	5.11	<.001**	0.02	0.05
Trust Politician	0.02	0.004	6333	5.07	<.001**	0.01	0.03
Age	0.00	0.000	2102	−4.80	<.001**	0.00	0.00
Gender	−0.01	0.011	2102	−1.11	.266	−0.03	0.01
OUS IB	0.03	0.006	2102	4.69	<.001**	0.02	0.04
OUS IH	0.05	0.005	2102	9.56	<.001**	0.04	0.06
CCS Deontology	−0.02	0.005	2102	−3.54	<.001**	−0.03	−0.01
CCS Utilitarianism	0.00	0.005	2102	−0.88	.379	−0.01	0.01
CRT	−0.01	0.003	2102	−3.27	.001*	−0.01	0.00
TS AOT	0.00	0.006	2102	0.03	.979	−0.01	0.01
TS Closed Minded	0.00	0.006	2102	0.73	.463	−0.01	0.01
TS Intuitive	−0.01	0.006	2102	−2.03	.042*	−0.02	0.00
TS Effortful	0.00	0.005	2102	0.12	.908	−0.01	0.01
Scenario Order	0.01	0.003	6333	1.88	.060	0.00	0.01
Force x Means	0.02	0.015	6333	1.20	.230	−0.01	0.05

Note. * $p < .05$; ** $p < .001$; Action choice variable: 0 = Inaction, 1 = Action, Force variable: 0 = No Force, 1 = Force, Means variable: 0 = Side-Effect, 1 = Means; OUS = Oxford Utilitarianism Scale, OUS IB = Impartial Beneficence, OUS IH = Instrumental Harm; CCS = Complete Consquentialism Scale; CRT = Cognitive Reflection Test; TS = Thinking Styles, AOT = Actively Open Minded Thinking.

	Intervention harming one to save five should *not* be done, according to most subjects.	Intervention harming one to save five *should* be done, according to most subjects.
Use of personal force to inflict harm	*Trolley: Footbridge (means)* *Escalator (means)*	*(X)* *Bus (means & side-effect)* *Escalator (side-effect)*
No use of personal force to inflict harm	*(Y)* *Trolley: Wave (means)* *Trolley: Beckon (means)*	*Trolley: Wave (side-effect)* *Trolley: Beckon (side-effect)* *Trolley: Switch (side-effect)* *Escalator (side-effect)* *Car (means & side-effect)*

Fig. 11. Results of Study 2 in the context of current theorizing (adapted from Railton, 2017).

interpretations suggest that any factor that might increase emotional aversion to causing harm (including but not limited to personal force) should lead to reduced willingness to causing harm (e.g., Reynolds & Conway, 2018). In this view, our findings may not necessarily be seen as evidence against dual-process approaches. For example, one reading of our findings may be that they provide suggestive evidence that participants view causing harm as a means to save five as more emotionally aversive than causing harm as a side-effect. Under this interpretation, our findings would be fully consistent with dual-process approaches. We did not record participants' emotional responses, and therefore we cannot rule this possibility in or out.

Overall, we found that participants were more willing to endorse actions that did not involve personal force than actions that did involve personal force, supporting our first hypothesis (and supporting classic dual-process approaches e.g., Greene et al., 2001; Greene, 2008). We also found that participants were more willing to endorse actions where harm occurred as a side-effect than if harm occurred as a means (to save five), supporting our second hypothesis. Interestingly, the effect of *means/side-effect* was larger than the effect of *force*. The effect of *means* was stronger in the absence of *force*, and similarly the effect of *force* was stronger in the absence of *means* (i.e., in the *side-effect* condition).

Replicating the results of Study 1, we showed that participants in both *Wave* and *Beckon* favored action when harm occurred as a *side-effect* of saving the five workers, but did not favor action when harm occurred as a *means* to save the five workers. Thus, the *means* versions of both scenarios occupy the cell originally denoted by *Y*. Again, we note that this is similar to Railton's demonstration, but with the added illustration that participants' distinctions between harm as a *means* vs harm as a *side-effect* is driving the effect.

Regarding position *X*, in line with Railton's findings, we found that for *Bus*, participants did endorse an intervention that involved personal force to inflict harm. Interestingly, for *Bus*, the overall preference for action held ("according to most subjects") for both the *means* and *side-effect* conditions, though it was significantly lower in the *means* condition (55.9% endorsing action in the *means* condition, vs 71.8% endorsing harm in the side-effect condition). We also found that for *Escalator*, when harm (involving personal force) occurred as a *side-effect* the majority of participants endorsed action, but if it occurred as *means* the majority of participants did not endorse action.

We also found that perceived trustworthiness of an imagined actor (friend, romantic partner, politician) who committed the action predicted action choice, supporting our third hypothesis. This finding is consistent with Railton's (2017) argument that decisions regarding whether or not to act in these sacrificial dilemmas may be sensitive to participants modelling of the perceived trustworthiness of *the kind of person who would endorse action*. However, our experimental design is not suited for testing the directionality of this relationship. We recommend future research should further investigate this hypothesis.

Table 12
Overview of key similarities and differences between different versions of each scenario.

Scenario	Means	Force	Premise	Action	Outcome
Footbridge (Trolley)	Means	Force	Trolley hurtling towards 5 workers	**Push** a man off a bridge	**Weight of man** stops the **trolley** saving 5
Switch (Trolley)	Side-effect	No Force	Trolley hurtling towards 5 workers	**Flip a switch** to divert the trolley	Trolley **diverts** to another track killing 1 worker
Wave (Trolley)	Means	No Force	Trolley hurtling towards 5 workers	**Wave** to a solitary worker	Worker steps onto track, **his weight** stops the **trolley**
Wave (Trolley)	Side-effect	No Force	Trolley hurtling towards 5 workers	**Wave** to 5 workers (another worker sees)	5 workers **step off** the track, another worker steps on and is killed
Beckon (Trolley)	Means	No Force	Trolley hurtling towards 5 workers	**Beckon** to a solitary worker	Worker steps onto track, **his weight** stops the **trolley** (killing him)
Beckon (Trolley)	Side-effect	No Force	Trolley hurtling towards 5 workers	**Beckon** to 5 workers (another worker sees)	5 workers **step off** the track, another worker steps on and is killed
Car	Means	No Force	Driving a car that will collide with 5 pedestrians	**Swerve** to hit a car pulling out	**Weight of other car** stops **car**, driver of other car is killed
Car	Side-effect	No Force	Driving a car that will collide with 5 pedestrians	**Swerve** to hit a parked car (discovering passenger after maneuver started)	Car is **diverted** into parked car; passenger of parked car is killed
Bus	Means	Force	Terrorist bombing a bus with 5 passengers	**Push** a man to obstruct the bomber	5 remain safely on the bus bomber is **impeded by man** (he is killed)
Bus	Side-effect	Force	Terrorist bombing a bus with 5 passengers	Push a man out of the way (onto sidewalk) **obstruct** the doorway for the bomber	5 remain safely on the bus bomber is **prevented from alighting bus** and is left on sidewalk with the other man (he is killed)
Escalator	Means	Force	Terrorist bomber on an escalator, will kill 5 at top	**Push** another person down the escalator	**Bomber is impeded by person** from reaching the top (the person dies)
Escalator	Side-effect	Force	Terrorist bomber on an escalator, will kill 5 at top	Push another person down the escalator to access and **press the "Emergency Stop" button**	The **escalator stops**, and the bomber is prevented from reaching the top, the other person dies from the blast
Escalator	Side-effect	No Force	Terrorist bomber on an escalator, will kill 5 at top	**Press the "Emergency Stop" button**	The **escalator stops**, and the bomber is prevented from reaching the top, another person at the bottom dies from the blast

Finally, we tested a range of individual differences that may predict action choice in sacrificial dilemmas. Supporting our fourth hypothesis, we found that participants who scored higher on the Instrumental Harm subscale of the Oxford Utilitarianism Scale (OUS, Kahane et al., 2018) were more likely to endorse action. We did not make directional predictions regarding the remaining measures. We found that endorsing action was positively predicted by the Impartial Beneficence subscale of the OUS, i.e., participants who scored higher in impartial beneficence were more likely to endorse action. This finding is interesting because even though these are two distinct constructs, they both predict action in our study, in the same direction. This is consistent with previous work demonstrating that participants who score highly on both Impartial Beneficence and Instrumental Harm are more inclined to endorse characteristically utilitarian actions (e.g., Kahane et al., 2018; Körner, Deutsch, & Gawronski, 2020).

We found that action choice was negatively predicted by the Deontology subscale of the Complete Consequentialism Scale (CCS, Plaks et al., 2021), i.e., participants who scored higher on this scale were less likely to endorse action. A similar negative relationship was found for the cognitive reflection test (CRT, Toplak et al., 2014), and a Preference for Intuitive thinking, as measured by the Comprehensive Thinking Styles Questionnaire (Newton et al., 2021), that is, participants who scored higher on each of these measures were also less likely to endorse action. Interestingly, the Utilitarianism subscale of the CCS (Plaks et al., 2021), as well as the Actively Open-minded Thinking, Closed Minded Thinking, and Preference for Effortful Thinking subscales of the Comprehensive Thinking Styles Questionnaire, were not related to action choice.

It is interesting that the deontology subscale of the CCS (Plaks et al., 2021) is associated with inaction while the utilitarianism subscale does not predict action/inaction. On the one hand, this is partially consistent with classic faming of these sacrificial dilemmas, where inaction is considered a "characteristically deontological" response (e.g., Greene, 2016). On the other hand, however, this classic framing suggests that endorsing action is a "characteristically utilitarian" response (Greene, 2016), it is therefore surprising that the utilitarianism subscale of the CCS does not predict action. Our findings show that responding is likely more complex than suggested by this characterization, and that preference for action may be more accurately measured using a more nuanced characterization (e.g., such as the two dimensions of the OUS, Kahane et al., 2018).

8. General discussion

Across two studies we demonstrated interesting variability and context sensitivity in participants' responses to sacrificial moral dilemmas. Our findings provide evidence in support of approaches to moral judgment that aim to account for the dynamism and context sensitivity of people's moral judgments (e.g., Bucciarelli et al., 2008; McHugh et al., 2022; Railton, 2017). We note, that while our findings may be seen as evidence against a narrow "personal-force" focused view of dual-process approaches (e.g., Greene et al., 2001; Greene, 2008; in line with the argument presented by Railton, 2017), our findings remain consistent with dual-process approaches to moral judgment more broadly, where people may be sensitive to a range of factors that might increase emotional aversion to causing harm (e.g., Reynolds & Conway, 2018). We show that, when considering an intervention that harms one to save five, people are sensitive not just to the level of personal force, but also whether harm occurs as a means or as a side-effect of saving five. Our results suggest that people are more sensitive to this means/side-effect distinction than they are to the presence/absence of personal force. Our findings also provide insights into the relationship between action choice in these sacrificial dilemmas and a range of other measures.

8.1. Implications, limitations, and future directions

Our studies offer novel understandings into different influences on people's responses to sacrificial moral dilemmas. Specifically, we unpack the unique, and the shared influences of personal force, and of means/side-effects, as well as a range of other individual difference measures on action choice. These findings contribute to on-going theoretical debates in moral psychology, providing support for approaches that aim to account for the observed variability (e.g., Bucciarelli et al., 2008; McHugh et al., 2022; Railton, 2017). Our findings do not directly refute dual-process approaches (e.g., Reynolds & Conway, 2018) beyond a narrow "direct-force" focused view (e.g., Greene et al., 2001; Greene, 2008; see Railton, 2017). Future research should identify testable competing predictions between these various context-sensitive and dual-process approaches in order to further advance theory.

A key finding of this work is that participants appear to be more sensitive to the means/side-effect distinction than they are to the presence/absence of personal force (though they are sensitive to both). This is most clearly shown in the *Bus* and *Escalator* cases, and suggests that the combination of *force* and *means* is the important difference between *Footbridge* and *Switch* rather than just the presence of personal force (e.g., Greene, 2008).

We note some variability in our results that suggests that *means* and *personal force* may not be the only features of the scenarios that influence participants responses. For example, the majority response for *Bus* (*Bus* always involved personal force) was to endorse action, even in the *means* condition. In contrast, for *Escalator*, when harm both involved *personal force* and occurred as a *means* to save five, the majority response was to reject action. This suggests there are features of the scenarios beyond the means/side-effect distinction that are also influencing participants' responding. It is possible, in line with the CNI/CNIS approaches (Gawronski et al., 2017; Skovgaard-Olsen & Klauer, 2023) that these scenarios may generate different action/inaction biases, or that the norms of what is considered acceptable are different between the two scenarios, e.g., prior experience with pushing and jostling at the doorway to public transport may make pushing someone out of the bus seem less aversive than pushing someone down an escalator (an action that participants are unlikely to much prior experience with). Follow-up work should attempt to better explain this variation.

Building on this possibility that other features of the scenarios influenced participants' judgments (beyond means/side-effect, and level of personal force), one such feature is the plausibility of the scenarios (e.g., Körner, Joffe, & Deutsch, 2019). In particular, all *means* versions of each the Trolley dilemmas (including *Footbridge*, and versions of *Beckon*, and *Wave*) require participants to believe that the weight of a person was sufficient to stop the runaway trolley. Participants may find this implausible and it is possible that this implausibility that is influencing the judgments – that is, Körner et al. (2019) report that when moral dilemmas contain more implausible aspects, participants are more likely to endorse characteristically deontological responses. Our findings regarding the *means* versions of the Trolley dilemmas used here are consistent with this link between implausibility and characteristically deontological responding, and thus we cannot definitively rule out this as a possible explanation for some of the responses observed in our study. The *means* versions of the other dilemmas in our study (*Car*, *Escalator*, and *Bus*) were intended to be more plausible, e.g., the weight of a car stopping another car, or the weight of a man stopping another man (see Table 12), however we did not record plausibility, future research should investigate this.

Table 12 details key similarities and differences between the different scenarios that may influence participants judgments, potentially impacting our results. In *Bus* and *Escalator*, the threat is a bomb attached to a terrorist, whereas in *Car*, and all *Trolley* dilemmas the threat is a vehicle. In *Car*, participants are asked to imagine they are in control of the vehicle, while in each of the *Trolley* dilemmas they are onlookers. Additionally, in *Car*, the person who dies to save five is in another car, while the five are pedestrians. In contrast, in all versions of *Trolley* all deaths (the 1 and the 5) occur due to the trolley colliding directly with the victims. In both *Bus* and *Escalator*, the terrorist also dies, posing an additional potential confound.

A further challenge to interpreting our findings is the absence of any significant effect for means/side-effect in the *Car* scenario. One interpretation is that this results from limitations with the materials used, that is, the two versions of *Car* scenario did not effectively convey a clear distinction between *means* and *side-effect*. While this is plausible, it is not necessarily the only possible explanation. It is also possible that the moral norms in car-driving contexts are better defined than in the other scenarios. If, for example, there was a strong norm within car-driving contexts for being proactive in minimizing overall harm caused by the car, it is possible that this strong norm could eliminate participants' sensitivity to harm as a means vs as a side-effect. Future research should investigate these possibilities.

9. Conclusion

We conducted two studies that had the dual aims of (a) addressing methodological limitations with the studies described by Railton (2017) to provide a robust test of the hypotheses presented, and (b) to build on Railton (2017) to provide clarity on the specific factors that are influencing participants decision making in different contexts. Our findings are generally consistent with the arguments made by Railton (2017), and provide novel insights into specific factors that can affect moral judgments, helping us to better understand their dynamism and context-sensitivity (Bucciarelli et al., 2008; McHugh et al., 2022). We also examined a range of individual difference variables and their relationship with participants' responses in sacrificial dilemmas. Together these results contribute to ongoing theoretical debates in moral psychology, while also providing avenues for future research.

Open practices

Pre-registration for Study 1 is available here: https://aspredicted.org/N85_2JV. Post-peer review pre-registration for Study 2 is available here: https://doi.org/10.17605/OSF.IO/UERPT

Full materials for both Study 1, and Study 2 are available as downloadable qualtrics files on this paper's project page on the OSF at: https://osf.io/59quk/?view_only=18414ad4433a4145a718f7015c012e36.

All anonymized data, scripts for simulations and analyses are also available on this paper's project page on the OSF (https://osf.io/59quk/?view_only=18414ad4433a4145a718f7015c012e36).

Author note

All procedures performed in studies involving human participants were approved by institutional research ethics committee and conducted in accordance with the Code of Professional Ethics of the Psychological Society of Ireland, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study. The authors declare that there are no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. All authors consented to the submission of this manuscript.

CRediT authorship contribution statement

Cillian McHugh: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Kathryn B. Francis:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Jim A.C. Everett:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Shane Timmons:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

The Authors affirm that they have no competing interests.

Data collection was jointly funded by Seed Funding from the Faculty of Education and Health Sciences University of Limerick, and by Research Funding from the Department of Psychology at Keele

University.

Jim Everett gratefully acknowledges support by the Leverhulme Trust (PLP-2021-095).

Data availability

All anonymized data, scripts for simulations and analyses are also available on this paper's project page on the OSF (https://osf.io/59quk/?view_only=18414ad4433a4145a718f7015c012e36)

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2024.104616.

References

Alicke, M. D. (2012). Self-injuries, harmless wrongdoing, and morality. *Psychological Inquiry, 23*(2), 125–128. https://doi.org/10.1080/1047840X.2012.666720

Andrejević, M., Feuerriegel, D., Turner, W., Laham, S., & Bode, S. (2020). Moral judgements of fairness-related actions are flexibly updated to account for contextual information. *Scientific Reports, 10*(1), 17828. https://doi.org/10.1038/s41598-020-74975-0

Baron, J., Scott, S., Fincher, K., & Emlen Metz, S. (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. https://doi.org/10.1016/j.jarmac.2014.09.003

Basinger, K. S., Gibbs, J. C., & Fuller, D. (1995). Context and the measurement of moral judgement. *International Journal of Behavioral Development, 18*(3), 537–556. https://doi.org/10.1177/016502549501800309

Bucciarelli, M., Khemlani, S., & Johnson-Laird, P. N. (2008). The psychology of moral reasoning. *Judgment and Decision making, 3*, 121–139.

Byrd, N., & Conway, P. (2019). Not all who ponder count costs: Arithmetic reflection predicts utilitarian tendencies, but logical reflection predicts both deontological and utilitarian tendencies. *Cognition, 192*, Article 103995. https://doi.org/10.1016/j.cognition.2019.06.007

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cameron, C. D., Payne, B. K., & Doris, J. M. (2013). Morality in high definition: Emotion differentiation calibrates the influence of incidental disgust on moral judgments. *Journal of Experimental Social Psychology, 49*(4), 719–725. https://doi.org/10.1016/j.jesp.2013.02.014

Christensen, J. F., Flexas, A., Calabrese, M., Gut, N. K., & Gomila, A. (2014). Moral judgment reloaded: A moral dilemma validation study. *Frontiers in Psychology, 5*, 1–18. https://doi.org/10.3389/fpsyg.2014.00607

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews, 36*(4), 1249–1264. https://doi.org/10.1016/j.neubiorev.2012.02.008

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), 216–235. https://doi.org/10.1037/a0031021

Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition, 179*, 241–265. https://doi.org/10.1016/j.cognition.2018.04.018

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences, 17*(8), 363–366. https://doi.org/10.1016/j.tics.2013.06.005

Curry, O. S., Jones Chesters, M., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation' with a new questionnaire. *Journal of Research in Personality, 78*, 106–124. https://doi.org/10.1016/j.jrp.2018.10.008

Cushman, F. A. (2013). Action, outcome, and value a dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292. https://doi.org/10.1177/1088868313495594

Doris, J. M. (Ed.). (2010). *The moral psychology handbook*. Oxford University Press.

Everett, J. A. C., & Kahane, G. (2020). Switching tracks? Towards a multidimensional model of utilitarian psychology. *Trends in Cognitive Sciences, 24*(2), 124–134. https://doi.org/10.1016/j.tics.2019.11.012

Feltz, A., & May, J. (2017). The means/side-effect distinction in moral cognition: A meta-analysis. *Cognition, 166*, 314–327. https://doi.org/10.1016/j.cognition.2017.05.027

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42.

Gamez-Djokic, M., & Molden, D. (2016). Beyond affective influences on deontological moral judgment: The role of motivations for prevention in the moral condemnation of harm. *Personality and Social Psychology Bulletin, 42*(11), 1522–1537. https://doi.org/10.1177/0146167216665094

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology, 113*, 343–376. https://doi.org/10.1037/pspa0000086

Gilligan, C. (1977). In a different voice: Women's conceptions of self and of morality. *Harvard Educational Review, 47*(4), 481–517. https://doi.org/10.17763/haer.47.4.g6167429416hg5l0

Gilligan, C. (1993). *In a different voice*. Harvard University Press.

Giner-Sorolla, R. (2018). A functional conflict theory of moral emotions. In K. J. Gray, & J. Graham (Eds.), *Atlas of moral psychology* (pp. 81–87). The Guilford Press.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2012). *Moral foundations theory: The pragmatic validity of moral pluralism (SSRN scholarly paper ID 2184440)*. Social Science Research Network.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology volume 3: The neurosciences of morality: Emotion, brain disorders, and development* (pp. 35–79). The MIT press.

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. The Penguin Press.

Greene, J. D. (2016). Why cognitive (neuro) science matters for ethics. In S. M. Liao (Ed.), *Moral brains: The neuroscience of morality* (pp. 119–149). Oxford University Press.

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 (New York, N.Y.), 293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Gubbins, E., & Byrne, R. M. J. (2014). Dual processes of emotion and reason in judgments about moral dilemmas. *Thinking & Reasoning, 20*(2), 245–268. https://doi.org/10.1080/13546783.2013.877400

Haidt, J., Björklund, F., & Murphy, S. (2000). *Moral dumbfounding: When intuition finds no reason*. Unpublished Manuscript. University of Virginia.

Hauser, M. D., Cushman, F. A., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language, 22*(1), 1–21. https://doi.org/10.1111/j.1468-0017.2006.00297.x

Hester, N., & Gray, K. (2020). The moral psychology of Raceless, genderless strangers. *Perspectives on Psychological Science, 15*(2), 216–230. https://doi.org/10.1177/1745691619885840

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science, 345*(6202), 1340–1343. https://doi.org/10.1126/science.1251560

Kahane, G., & Everett, J. A. C. (2022). Trolley dilemmas from the philosopher's armchair to the psychologist's lab. In H. Lillehammer (Ed.), *The trolley problem* (pp. 134–157). Cambridge University Press. https://www.cambridge.org/core/books/trolley-problem/6DEAFA4B5A1389EDFF34CB25E3328EE7.

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review, 125*(2), 131–164. https://doi.org/10.1037/rev0000093

Kamm, F. M. (2007). *Intricate ethics: Rights, responsibilities, and permissible harm*. Oxford University Press.

Körner, A., Deutsch, R., & Gawronski, B. (2020). Using the CNI model to investigate individual differences in moral dilemma judgments. *Personality and Social Psychology Bulletin, 46*(9), 1392–1407. https://doi.org/10.1177/0146167220907203

Körner, A., Joffe, S., & Deutsch, R. (2019). When skeptical, stick with the norm: Low dilemma plausibility increases deontological moral judgments. *Journal of Experimental Social Psychology, 84*, Article 103834. https://doi.org/10.1016/j.jesp.2019.103834

Körner, A., & Volk, S. (2014). Concrete and abstract ways to deontology: Cognitive capacity moderates construal level effects on moral judgments. *Journal of Experimental Social Psychology, 55*, 139–145. https://doi.org/10.1016/j.jesp.2014.07.002

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 1*, 1–8. https://doi.org/10.1177/1948550617697177

McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. L. (2017). Searching for moral dumbfounding: Identifying measurable indicators of moral dumbfounding. *Collabra: Psychology, 3*(1), 1–24. https://doi.org/10.1525/collabra.79

McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. L. (2020). Reasons or rationalizations: The role of principles in the moral dumbfounding paradigm. *Journal of Behavioral Decision Making, 33*(3), 376–392. https://doi.org/10.1002/bdm.2167

McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. L. (2022). Moral judgment as categorization (MJAC). *Perspectives on Psychological Science, 17*(1), 131–152. https://doi.org/10.1177/1745691621990636

McHugh, C., Zhang, R., Karnatak, T., Lamba, N., & Khokhlova, O. (2023). Just wrong? Or just WEIRD? Investigating the prevalence of moral dumbfounding in non-Western samples. *Memory & Cognition, 51*(5), 1043–1060. https://doi.org/10.3758/s13421-022-01386-z

McPhetres, J., Conway, P., Hughes, J. S., & Zuckerman, M. (2018). Reflecting on God's will: Reflective processing contributes to religious peoples' deontological dilemma responses. *Journal of Experimental Social Psychology, 79*, 301–314. https://doi.org/10.1016/j.jesp.2018.08.013

Mikhail, J. (2000). *Rawls' Linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 'a theory of justice'*. Phd Dissertation. Cornell University [SSRN Scholarly Paper, Social Science Research Network] https://papers.ssrn.com/abstract=766464.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143–152. https://doi.org/10.1016/j.tics.2006.12.007

Newton, C., Feeney, J., & Pennycook, G. (2021). The comprehensive thinking styles questionnaire: A novel measure of intuitive-analytic thinking styles. *PsyArXiv*. https://doi.org/10.31234/osf.io/r5wez

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., … Cushman, F. (2020). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000281. No Pagination Specified-No Pagination Specified.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*(1), 163–177. https://doi.org/10.1111/j.1551-6709.2011.01210.x

Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: A comment on Haidt (2001). *Psychological Review, 110*(1), 193–196. https://doi.org/10.1037/0033-295X.110.1.193

Plaks, J. E., Lv, J., Zhao, M., Staples, W., & Robinson, J. S. (2021). Using conflict negativity to index psychological tension between impartiality and status-upholding principles. *Social Neuroscience, 16*(5), 500–512. https://doi.org/10.1080/17470919.2021.1953133

Qualtrics. (2020). *Qualtrics*.

Railton, P. (2017). Moral learning: Conceptual foundations and normative relevance. *Cognition, 167*, 172–190. https://doi.org/10.1016/j.cognition.2016.08.015

Railton, P. (2021). Moral learning—natural and artificial moral competence. In *Moral learning—natural and artificial moral competence. Engineering and reverse-engineering morality; Workshop at Cogsci 2021, Virtual*. https://sites.google.com/view/engineering-morality/home.

Reynolds, C. J., & Conway, P. (2018). Not just bad actions: Affective concern for bad outcomes contributes to moral condemnation of harm in moral dilemmas. *Emotion, 18*(7), 1009–1023. https://doi.org/10.1037/emo0000413

Royzman, E. B., & Borislow, S. H. (2022). The puzzle of wrongless harms: Some potential concerns for dyadic morality and related accounts. *Cognition, 220*, Article 104980. https://doi.org/10.1016/j.cognition.2021.104980

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology, 76*(4), 574–586. https://doi.org/10.1037/0022-3514.76.4.574

Schein, C. (2020). The importance of context in moral judgments. *Perspectives on Psychological Science, 15*(2), 207–215. https://doi.org/10.1177/1745691620904083

Schein, C., & Gray, K. J. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The 'big three' of morality (autonomy, community, divinity) and the 'big three' explanations of suffering. In A. M. Brandt, & P. Rozin (Eds.), *Morality and health* (pp. 119–169). Routledge.

Sinnott-Armstrong, W., Young, L., & Cushman, F. A. (2010). Moral intuitions. In J. M. Doris (Ed.), *The moral psychology handbook* (pp. 206–245). Oxford University Press.

Skovgaard-Olsen, N., & Klauer, K. C. (2023). Invariance violations and the CNI model of moral judgments. *Personality and Social Psychology Bulletin, 01461672231164888*. https://doi.org/10.1177/01461672231164888

Stanovich, K. E., & Toplak, M. E. (2019). The need for intellectual diversity in psychological science: Our own studies of actively open-minded thinking as a case study. *Cognition, 187*, 156–166. https://doi.org/10.1016/j.cognition.2019.03.006

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*(6), 1395. https://doi.org/10.2307/796133

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision making, 11*(1), 99–113.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition, 39*(7), 1275. https://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning, 20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729