**Title: Why moral judgments change across variations of trolley-like problems**

Short title: *Why moral judgments change*

 Dale J. Cohen1, Philip T. Quinlan\*2

1 Department of Psychology, University of North Carolina at Wilmington; Wilmington, NC, USA.

2 Department of Psychology, University of York; Heslington, York, UK.

\*Corresponding author information: Philip Quinlan, Department of Psychology, University of York, Heslington, York, YO10 5DD, UK. (e-mail: philip.quinlan@york.ac.uk).

**Abstract:**

In the standard “trolley problem,” respondents must decide whether to save a condemned group of individuals by sacrificing a safe bystander. Although respondents often are willing to sacrifice the bystander in some circumstances (e.g., by pulling a lever), they are loath to sacrifice the bystander in others (e.g., by pushing the bystander off a footbridge). This difference in responding has been explained via a Dual Process theory of moral judgements (DPT). DPT, however, is a classic boxes-and-arrows model that only makes directional predictions. Meehl (1967, Phil. Sci, 34, 103-115) cautioned against theories that only make directional predictions, explaining that they are notoriously difficult to falsify. Meehl (1967) argued that researchers should follow the lead of Physics and develop computational models that make functional and point predictions. Here, we use a value-based, computational cognitive model of decision-making (Psychological Value Theory) to predict precisely both the speed and kind of response in trolley-like problems in three experiments. We show that this model accounts for the changes in choices across variations of the trolley problem with a response bias parameter.

**Keywords:**

trolley problems; sacrificial moral dilemmas; computational modelling; Psychological Values Theory; Dual Process Theory

**Data availability statement:**

All data have been made publicly available on the Github and can be accessed at https://github.com/ccpluncw/ccpl\_data\_responseBias2022.git. This study was not preregistered.

The analysis was conducted in part using R packages written by the first author (fourth quarter, 2021). Those packages can be installed from the following locations:

<https://github.com/ccpluncw/ccpl_R_chValues.git>

<https://github.com/ccpluncw/ccpl_R_chMorals.git>

<https://github.com/ccpluncw/ccpl_R_RRW.git>

<https://github.com/ccpluncw/ccpl_R_smartGridSearch.git>

**Acknowledgements:**

Dale J. Cohen, Department of Psychology, University of North Carolina Wilmington. Philip T Quinlan, Department of Psychology, University of York; Heslington, York, UK. We thank the members of the Cohen Cognition and Perception Laboratory for their assistance and support. The current work received no funding. There are no real or potential conflicts of interest that may impact the current research.

Correspondence concerning this article should be addressed to Philip Quinlan, Department of Psychology, The University of York, Heslington, York, YO10 5DD, UK.

Author Contributions –

PQ Writing – Original Draft Preparation, Review & Editing; Conceptualization

DJC - Original Draft Preparation, Review & Editing; Conceptualization; Data Curation; Formal Analysis; Investigation; Methodology; Project Administration; Resources; Software; Validation; Visualization.

**Why moral judgments change across variations of trolley-like problems**

In attempting to understand the cognitive and neural systems underlying moral judgements (Greene, 2009), researchers often study how people respond to sacrificial moral dilemmas. In sacrificial moral dilemmas, one group of individuals is to be killed by default (here referred to as the *condemned*) and a second is safe by default (here referred to as the *bystander*). For example, in the standard version of the classic trolley problem (Foot, 1967), a runaway trolley will kill five individuals (i.e., the condemned) if it continues on its current track and you (as the respondent) choose todo nothing. However, you can divert the trolley onto a siding by pulling a lever with the consequence that the trolley will kill a single individual (the bystander). A key research issue has been to understand the circumstances that influence the likelihood of the respondent choosing to save the condemned by sacrificing the bystander (here termed a “*do something*” response) vs. choosing to do nothing and let the condemned be killed (here termed a “*do nothing*” response; see e.g., Białek & De Neys, 2016; Bostyn et al., 2018; Conway & Gawronski, 2013; Greene et al., 2001; Greene, & Haidt, 2002; Hauser et al., 2007; Koenigs et al., 2007). To examine these circumstances, a particularly influential variation of the trolley dilemma, known as the *Footbridge version*, has typically been employed[[1]](#endnote-1). In this version, rather than pulling a lever to divert the trolley, you must consider pushing the bystander off a footbridge to stop the trolley. Whereas people are generally keen to endorse acting in the standard version, they are far less likely to endorse acting in the Footbridge version (e.g., Awad et al., 2020; Chu & Lui, 2023; Claessens et al., 2020; Greene et al., 2001; Hauser et al., 2007; Moore et al., 2011).

Traditionally, sacrificial moral dilemmas are designed so that the condemned group contains more people than bystanders. This design is intentional because a Utilitarian ethics system provides a motivation for acting (i.e., it is morally just to sacrifice the few to save the many; e.g., Bostyn et al., 2018; Conway & Gawronski, 2013; Kahane et al., 2015; Rosen, 2005; Trémolière & Bonnefon, 2014). Given this, the “do something” response is often characterized as being consistent with Utilitarian principles (e.g., Conway & Gawronski, 2013; Koenigs et al., 2007; Moll & de Oliveira-Souza, 2007; Trémolière & Bonnefon, 2014) and is typically referred to as the “utilitarian” response. Furthermore, the cognitive process leading to this response is often described as being slow, effortful and, consequentialist (e.g., Greene et al., 2008; Paxton & Greene, 2010; Suter & Hertwig, 2011).

In the Footbridge version, the Utilitarian motivation for action is the same as the standard version (sacrifice one to save the many). Therefore, researchers disregard a consequentialist explanation for the shift in responding. Greene et al. (2001) proposed that the act of imagining the personal force required by the “do something” response in the Footbridge version produces an automated negative emotional response that interferes with the Utilitarian “do something” response. Greene et al. (2004) later clarified that this negative emotional response comes about from an evolved social/emotional decision process. This decision process is proposed to be fast, effortless, and non-consequentialist. For these reasons, the “do nothing” response is often characterized as being consistent with Deontological principles (i.e., killing innocent people is categorically wrong; Bago & De Neys, 2019; Johnson & Cureton, 2018; 1994; Paxton & Greene, 2010; Sullivan, 1994; Trémolière et al., 2018) and is typically referred to as the “deontological” response.

This theoretical framework has been widely adopted as the standard explanation of the choices made in sacrificial moral dilemmas and has come to be known as the *Dual Process Theory* of moral judgment (henceforth, DPT; e.g., Greene, 2009; Greene & Young, 2020; Paxton & Greene, 2010). DPT states that the participant’s response reflects a competition between the slow, effortful consequentialist process and the fast, effortless non-consequentialist process. The “do nothing” response occurs when the social/emotional decision process is activated and outcompetes the consequentialist process. Otherwise, the consequentialist process produces the “do something” response. Because the consequentialist process is hypothesized to be activated by default, it is argued that changes in the proportion of the “do something” and “do nothing” responses is a function of whether and how strongly the social/emotional decision process is activated.

If the social/emotional decision process is the explanatory mechanism for the shift in response patterns, then to predict these shifts, one must identify the conditions that activate it (Greene et al., 2009). DPT, however, is a classic arrows-and-boxes account (Quinlan & Dyson, 2008) in which various cognitive subprocesses are organized as boxes in an information processing flow diagram. Despite the obvious utility of this kind of approach, serious concerns arise over the lack of precise predictions that such accounts afford. For example, DPT contains no detailed specification of the underlying cognitive and social processes, how or when the social/emotional decision process is activated, or of how the competition between the two reasoning systems comes about (e.g., Is it a race model, a sequential model, etc.? see e.g., De Neys, 2022; Quinlan & Cohen, 2023). As such, DPT makes broad directional predictions about reaction time (RT) and/or choice, rather than precise functional forms or point predictions.

In the absence of clear predictions, many DPT researchers have used a data driven approach to identify when the social/emotional decision process is activated. Here, researchers identify shifts in responding and then infer that the conditions that brought about the shift activated the social/emotional decision process. As should be evident, this logical analysis assumes the validity of DPT rather than tests it. Originally, Greene et al. (2001) proposed that the Footbridge version was “… intuitively ‘up close and personal’ (and putatively more emotional)” (p. 2106) whereas the standard version was “intuitively impersonal (and putatively less emotional)” (p. 2106)[[2]](#endnote-2). More recently, researchers have concluded that the social/emotional decision process is activated when the respondent (as the agent) contemplates *intentionally* sacrificing the bystander using *personal force* (e.g., Greene, et al., 2009; Hauser et al., 2007). Greene et al. (2009) defines personal force as when the force, “that directly impacts the other is generated by the agent’s muscles,” (p. 365). Indeed, Greene (2023) recently confirmed personal force as a necessary activator of the social/emotional process, stating, “…the “non-negotiable” deontological response is thought to be triggered by the use of ‘personal force’ …” (p., 1; here, Greene refers to the choice to “do nothing” as the “deontological” response).

Despite the significant influence that this framework for thinking has had on the field, here, we present an alternative explanation for the shift in the likelihood of respondents choosing the “*do something*” vs. the “*do nothing*” option in sacrificial moral dilemmas. We base our account in Psychological Value Theory (PVT), a quantitatively specified theory of value and the decision process leading to choice (for a detailed description of PVT please see Cohen et al., 2022; for a short tutorial on PVT, see https://osf.io/3xu98).

**Psychological Value Theory**

Meehl (1967) outlined the dangers inherent in arrows-and-boxes models with directional predictions such as the case with DPT, explaining that they are notoriously difficult to falsify. Meehl (1967) argued that researchers should follow the lead of Physics and develop computational models that make functional and point predictions. Unlike directional predictions, the likelihood of falsifying functional and point predictions increases with increased power. Here we adopt Meehl’s (1967) model-based approach to understanding the shifts in the tendency towards and away from acting across variations of sacrificial moral dilemmas. PVT is a computational cognitive model of value and choice. PVT makes *a priori* functional and point predictions about the likelihood of respondents choosing the “*do something*” response vs. the “*do nothing*” response, and the time it takes them to make those responses. We have successfully applied PVT to predict choice and RT in sacrificial moral dilemmas across a variety of conditions (Cohen & Ahn, 2016; Cohen et al., 2022; Cohen et al., 2023).

A key assumption of PVT is that moral judgments are not a special type of decision, rather, they are value-based preferential choices in which the decision-maker strives to choose the option with the greatest Psychological Value. Cohen et al. (2022) define Psychological Value as the “importance, worth or usefulness of an item to the observer” (p. 2). Within the account, it is assumed that Psychological Value is *perceived*, and therefore subject to perceptual variability (Ashby & Lee, 1993). Perceptual variability is the continuous fluctuation of the perceived stimulus, even though the physical stimulus, environment, and observer remain constant. As such, the perceived Psychological Value of an item (e.g., a car) is best represented as a distribution of quantities on an abstract value continuum, rather than a single quantity.

Because Psychological Value is described by a distribution, one can quantify the similarity between the Psychological Values of two options by their distributional overlap. Figure 1 illustrates this concept. It presents the Psychological Value distributions of two options (a lower valued option, termed *LVO*, and a higher valued option, termed *HVO*) and their overlapping regions. A large overlapping region indicates highly similar options, whereas a small overlapping region indicates less similar options. The overlap of the Psychological Value distributions of the options is the main predictor of choice and response time.



Figure 1. A schematic representation of the Psychological Value distributions for a thief and an astronaut.

Cohen and colleagues use a magnitude estimation task to measure Psychological Value distributions independent of choice. Magnitude estimation tasks have been shown to provide valid assessments of the magnitude of perceptions (Cohen & Lecci, 2001; Marks & Algom, 1998; Stevens, 1956). In this task, participants are asked to estimate their personal value of a series of probes. Participants are told that personal value is not necessarily monetary value and that one’s first report card may have a low monetary value but a high personal value. Participants are shown a standard stimulus (a chimpanzee) with an assigned personal value of 1000. Participants are told to estimate their personal value of each probe in relation to the standard stimulus. So, if the probe is twice as valuable, assign it a 2000, etc. In this way, Cohen and colleagues collected the Psychological Values of hundreds of items (people, non-human animals, objects, concepts, etc.).

PVT assumes the perception of Psychological Value is relatively stable across individuals (see Cohen et al., 2023). As such, an individual’s Psychological Value distribution of a stimulus should be well approximated by the population’s Psychological Value distribution of that stimulus. Therefore, we estimate the population’s Psychological Value distribution by asking many people to provide a single magnitude estimation of that stimulus. The resulting distribution serves as the Psychological Value distribution for that stimulus. If, the population’s Psychological Value distribution approximates that of the individual, then the population’s Psychological Value distribution should accurately predict individual choice in a preferential decision task. Cohen and colleagues (Cohen & Ahn, 2016; Cohen et al. 2022; Cohen et al., 2023) confirmed this prediction.

PVT assumes that the decision-maker is trying to identify and choose the option with the higher Psychological Value (the HVO). When the Psychological Value distributions of the options overlap, identifying the HVO becomes inherently uncertain – and therefore inherently error prone (i.e., choosing the LVO when attempting to identify and choose the HVO). To model this uncertainty, Cohen et al. (2022) developed a variant of the classic two-dimensional sequential sampling process, termed the Robust Random Walk (RRW).

Similar to traditional sequential sampling processes, the RRW accumulates evidence stochastically over time (see Figure 2). The process initiates at a point (the *start point*) on the ordinate between two ‘decisional’ boundaries: an upper boundary representing “choose option A” and a lower boundary representing “choose option B.” At each stage in the walk, a random value is taken from each option’s Psychological Value distribution, and whichever is greater dictates the direction of travel. The process curtails when either boundary is reached. The length of the path is an index of the time to reach a decision, and the boundary reached signifies the choice made. In this way, the RRW simultaneously predicts RT and choice as a function of the overlap of the two options’ Psychological Value distributions. When the overlap of the Psychological Value distributions of the two options is large, the time to decision is protracted and the decision is highly error prone (see Figure 2 lower panel). When the distributions overlap minimally, the decision time will be short, and the errors will be minimal (see Figure 2 upper panel; see Supplemental Information for a more detailed account).



Figure 2. Schematic representation of two random walk decisions. Upper panel provides an example of when the competing options have relatively dissimilar psychological values, and the lower panel when the competing options have highly similar psychological values. PVT assumes decision-makers are attempting to identify and choose the HVO. In this diagram, Option A is the HVO. Therefore, Option A is the correct response and Option B is the incorrect response.

The RRW uses the measured overlap of the Psychological Value distributions of the options as the predictor variable that “drives” the random walk. This permits the RRW to predict the functional forms and magnitude of RT and choice responses *a priori*. This also strongly constrains the predictions (the model forbids most patterns of data), thus producing a higher likelihood of falsification (Meehl, 1967). Traditional sequential sampling procedures, in contrast, use the participants’ responses to estimate latent distributions that drive the random walk (Andrejević, et al., 2022; Hutcherson et al., 2015; Son et al., 2019). Because traditional sequential sampling processes estimate the latent distributions from responses, they are less constrained than the RRW and do not predict responses a priori.

To test PVT’s predictions, Cohen and colleagues (2016, 2022, 2023) presented participants with sacrificial moral dilemmas and the participant had to indicate whether they would choose to “do something,” which saves the condemned by sacrificing the bystander, or “do nothing,” which allows the condemned to die. On every trial the the Psychological Value (as measured by the magnitude estimation task) of the condemned and bystander varied. Choice and RT were recorded. Across a series of experiments (Cohen et al., 2016, 2022, 2023), the RRW predicted novice participants’ RT and choice extremely accurately (*r2* ≈ 0.9). Furthermore, the overlap of the condemned (the person or group who will be killed by default) and bystander (the person or group who is safe by default) Psychological Value distributions predicted RT and choice far more accurately that the *number* *of lives saved* in the condemned vs. bystander groups.

Cohen and colleagues’ (2016, 2022, 2023) experiments provide strong evidence that sacrificial moral dilemmas are value-based decisions in which decision-makers weigh the relative Psychological Values of the condemned and bystander. In contrast to DPT, PVT conceptualizes the moral decision making as a competition between the Psychological Value of the options and, therefore, both the “do something” and “do nothing” responses manifest from a single, *consequentialist* process.

Although PVT has successfully modeled moral judgment in sacrificial moral dilemmas, it has yet to address the shift in the response patterns between different versions of sacrificial moral dilemmas. Within sequential sampling procedures (such as the RRW), there are theoretically two potential sources of a shift in the likelihood of respondents choosing the “*do something*” response vs. the “*do nothing*” response in sacrificial moral dilemmas: the position of the start point (referred to as a *start point* *bias)*, and the evidence accumulation speed (referred to as an *evaluation bias*; Ratcliff, 1985; Ratcliff & McKoon, 2008; Zhao et al., 2019). When a start point bias is present, the position of the start point of the evidence accumulation is closer to one boundary than another. To model a bias in favor of the “do nothing” option (hereafter referred to as an *inaction bias*), the start point is positioned closer to the “do nothing” boundary than the “do something” boundary. The shorter distance from the start point to the “do nothing” boundary relative to the “do something” boundary results in faster, and more frequent, “do nothing” responses than is the case with an unbiased start point (see Figure 3). When an evaluation bias is present, the speed that the evidence accumulates is faster towards one boundary than the other. To model an evaluation bias that favors the “do nothing” response in the RRW, the likelihood of the evidence accumulating towards the “do nothing” boundary would increase relative to that accumulating towards the “do something” boundary. This results in faster, and more frequent, “do nothing” responses.

One way to conceptualize the difference between a start point and evaluation bias in sacrificial moral dilemmas is to imagine a foot race between the condemned and the bystander, with the loser being ultimately sacrificed. Here, the speed of each racer corresponds to their Psychological Value. A start point bias towards the “do nothing” boundary is analogous to giving all bystanders a head start of a constant size (e.g., 10 meters). Across all races, the head start increases the likelihood of the bystander winning. An evaluation bias advantaging the “do nothing” option is analogous to giving all bystanders carbon plate running shoes that increase the wearer’s speed by a constant amount (say, 2 mph). The carbon plate running shoes, like the head start, increase the likelihood of the bystander winning. For both types of response biases, the impact of the bias on the eventual outcome of the race is a function of the unaided speed difference between the condemned and bystander. When the condemned and bystander are equally fast (have similar Psychological Values), the response biases will have their greatest impact. The response biases will have decreasing influence as the speed difference between the condemned and bystander increases. For example, a fast bystander (one with a high Psychological Value) will win the race against a slow condemned individual (one with a low Psychological Value), even without a head start or carbon plate sneakers. Similarly, a very fast condemned individual (one with a high Psychological Value) will win the race against a very slow bystander (one with a low Psychological Value) despite a head start or carbon plate sneakers. Although the two sources of bias influence the probability of winning in similar ways (i.e., the probability of saving the HVO), they are differentiated by subtle differences in the patterns of RT.

 Although the interpretation of a start point bias is relatively straightforward (e.g., Ratcliff & Rouder, 1998), there are at least two potential interpretations of an evaluation bias. The first interpretation is analogous to a biased criterion in Signal Detection Theory (SDT; Ratcliff, 1985; Ratcliff & McKoon, 2008). Recall, at each iteration of the RRW, a sample is randomly selected from each of the two options’ Psychological Value distributions and a difference score is calculated (sDiff = Sample A – Sample B). If the difference score exceeds an evaluation criterion (sDiff > Criterion), then the RRW takes a step in the direction of Boundary A, otherwise it takes a step in the direction of Boundary B. By default, the evaluation criterion is set at 0 (unbiased) and the larger of the two samples determines the direction of evidence accumulation. If the participant has a bias in favor of Option A, then they would set their evaluation criterion below 0, thus increasing the likelihood of taking a step in the direction of Boundary A. Similarly, if the participant has a bias in favor of Option B, then they would set their evaluation criterion above 0, thus increasing the likelihood of taking a step in the direction of Boundary B. In this way, the evaluation criterion influences the speed of evidence accumulation.

A second interpretation of an evaluation bias is as a value shift (e.g., Zhao et al., 2019). Within the context of the RRW, the speed of evidence accumulation is a function of the overlap of the two options’ Psychological Value distributions. As such, an evaluation bias is equivalent to systematically increasing or decreasing the overlap of the options. For example, a biased evaluation criterion favoring the “do nothing” option will produce the same pattern of data as systematically increasing the value of all bystanders by a constant (or, equivalently, decreasing the value of all condemned individuals). Increasing the value of the bystander, increases the rate of evidence accumulation toward the “do nothing” option (see Figure 2).

Critically, the two interpretations of the evaluation bias are currently indistinguishable by the RRW. Nevertheless, the criterion shift in SDT has been shown to be sensitive to changes in the cost/benefit analyses associated with making different types of errors, whereas the relative positions of the underlying distributions are not (Martín-Guerrero et al., 2016; Perales et al., 2005). Therefore, if an evaluation bias is predictably influenced by changes in the cost/benefit analyses associated with making different types of errors, it is likely a function of a shift in the evaluation criterion. Otherwise, the evaluation bias is likely a function of a value shift.



Figure 3: An illustration of a start point bias. In the upper panel, the start point is equidistant from each response threshold. Therefore, it is unbiased. In the lower panel, the start point is closer to the “do nothing” threshold than the “do something” threshold. This is a pre-choice bias favoring the “do nothing” response. Here, the “do nothing” responses will be faster and more error prone than the “do something” responses.

In an experimental setting, the cost/benefit of response outcomes is typically manipulated with a *payoff matrix* (Swets, 1977). A payoff matrix is a table that identifies the costs/benefits of each potential outcome of a choice. Typically, the relation between the cost of missing the target, termed a *miss*, is manipulated relative to that of a falsely choosing a non-target, termed a *false alarm*. To understand how the cost of a miss relative to a false alarm impacts decision making, it is illustrative to consider cancer detection (for an analysis, see Swets, 1998). When cancer is suspected, the doctor must determine if there is sufficient evidence to warrant a biopsy. If the threshold for this evidence is set too low, there will be an unnecessary high use of resources because most irregularities will not be cancerous – a *false alarm*. If the threshold is set too high, there will be people whose cancer is not caught early enough to treat, and they may die – a *miss*. Here, the optimal threshold is function of the cost/benefit analysis of the errors.

Both start point and criterion induced evaluation biases are sensitive to changes in payoff matrices (Ratcliff, 1978 Ratcliff, 1985; Ratcliff & McKoon, 2008; Zhao et al., 2019). For example, there is extensive evidence that changing the payoff matrix influences the criterion in SDT to maximize reward (e.g., Macmillan & Creelman, 2005; Maddox, 2002) and, by extension, criterion induced evaluation biases (e.g., Ratcliff, 1985). There is also evidence that manipulating payoff matrices will induce a start point bias (e.g., Voss et al., 2004). In a series of experiments, Voss et al. (2004) manipulated the payoff matrix associated with target identification and modeled the result of that manipulation with a drift diffusion process. It was confirmed that, “The starting point moved toward the threshold that was connected with the rewarded answer. No other parameter was affected by this manipulation.” (p. 1215).

The logic of a payoff matrix can be directly applied to sacrificial moral dilemmas. Because PVT assumes the participant is attempting to save the HVO, the HVO is the *target*. When the participant fails to save the HVO, that is assumed to be an error. There are two ways that one can make an error: 1) by failing to save the condemned HVO by inaction and 2) by saving the LVO through action (see Figure 4). When the participant fails act to save the condemned HVO (chooses to “do nothing”), it is termed a *miss* because the participant missed the target. When the participant acts (chooses to “do something”) to save the condemned LVO, it is termed a *false alarm* because the participant erroneously identified the LVO as the target. The cost/benefit analysis of different types of errors is, by definition, a form of consequentialist reasoning. As such, if varying the cost of a false alarm relative to a miss shifts the general tendency towards action/inaction, we will be providing evidence that invoking a non-consequentialist process is unnecessary to explain these shifts.



Figure 4: The different response types of errors one can make in sacrificial moral dilemmas

Here we report three experiments in which we change the cost of a false alarm relative to a miss in variations of the trolley problem. We vary the cost of a false alarm by manipulating the degree of pain and suffering inflicted on the actors in the scenario. It is well established that watching the pain of another will activate the brain areas associated with experiencing pain in the observer’s brain (e.g., Hein, et al., 2010 Singer & Lamm, 2009). Although there are likely differences in the psychological representation of experienced vs empathetic pain, they both are aversive states (e.g., Vlaev et al., 2009; Zaki et al., 2016) demonstrated that people assign economic value to aversive states (i.e., pain) and this value is subject to the dynamics of economic theory. Therefore, people’s willingness to help a suffering actor should be subject to an economic analysis. Hein et al. (2010) confirmed this prediction by demonstrating that one’s motivation to reduce the pain and suffering of others is a function of the value they place on the suffering other and the personal cost such help will inflict on themselves. As such, there is good evidence that manipulating the pain and suffering inflicted on the actors in our scenarios should effectively influence the cost of a false alarm relative to a miss.

We use the same general methods in all three experiments. On each trial, the participant is presented with a unique sacrificial moral dilemma. In each dilemma, one individual is to be killed by default and a second is safe by default. The participant was instructed to decide whether to act (thus saving the condemned individual by sacrificing the bystander) or do nothing. In all experiments both the agent of death and the method of killing was varied factorially across the trials. To predict responses, we used a subset of the Psychological Value distributions garnered from the previous work (Cohen et al., 2022). These Psychological Value are input into the RRW to predict participants’ response choices and times.

In Experiment 1, we present participants with sacrificial moral dilemmas in which the agent of death (you vs. an anonymous other) and the method of death (strangulation vs unspecified) were manipulated. We model response time (RT) and choice with the RRW using previously collected estimates of Psychological Value of each option as the predictor variable. Furthermore, our experimental design and analysis will allow us to identify the *source* of inaction biases. Specifically, we will model inaction biases as either start point or evaluation bias. By modelling the process, the RRW can identify the source and quantify the degree to which inaction biases are a function of the agent of death, the method of death, and/or a general inaction bias that is present across all conditions. The RRW will fail to accurately model responses if (i) responses to sacrificial moral dilemmas are not value-based choices, (ii) the measurements of Psychological Value are inaccurate, (iii) PVT’s model of the decision process is inaccurate, or (iv) inaction biases are not the result of a start point or evaluation biases. If the RRW successfully models response choice and RT, then we can be confident that all four of these conditions were met. As such, Experiment 1 serves as both a validation of the model’s ability to predict responses and a baseline for Experiments 2 and 3.

In Experiments 2 and 3, we assess whether manipulating the cost of a false alarm relative to that of a miss will have predictable influences on the inaction response bias. In Experiment 2, we attempt to decrease participants’ tendency toward inaction by *increasing* the cost of a miss. We do so by stating that the condemned individual will be tortured to death over three days. If, as explained above, people weigh the cost of letting another suffer, then increasing the suffering of the condemned individual through torture should increase the cost of “doing nothing” when the condemned is the HVO (here a miss). To minimize misses, the bias towards inaction should be reduced relative to Experiment 1. If either a start point or evaluation bias successfully models response choice bias (as we predict), then the estimated start point/evaluation parameter values in Experiment 2 will reduce relative to that in Experiment 1 (where the condemned individual is not tortured).

If the defining feature of a generalized shift in responses is the payoff matrix, rather than personal force and intentions, one should be able to develop a scenario that both contains intentional personal force and *increases* participants “do something” responses. In Experiment 3, we do exactly that. We attempt to increase participants’ biases toward action (i.e., decrease the inaction bias) by *decreasing* the cost of a false alarm. We do so by changing the method of death from strangulation to administering a humane lethal injection. The humane lethal injection reduces the pain and suffering inflicted on the bystander. As such, the cost of erroneously killing a higher valued bystander (i.e., a false alarm) is also reduced. Importantly, administering a humane lethal injection is an intentional action that requires personal force analogous to that the footbridge scenario. In the footbridge dilemma, the agent pushed the bystander off an overpass and the train kills the bystander. When one administers a humane lethal injection, the agent pushed the needle into the bystander’s body and pushes the plunger to inject the poison, which kills the bystander. In this way, the predictions of DPT are pitted against those of PVT. Again, if the start point/evaluation biases successfully model responses, then the estimated start point/evaluation parameter values in Experiment 3 will reduce relative to that in Experiment 2 (where the bystander is strangled) or may even shift to a bias towards action (see Figure 5).



Figure 5. The different payoff matrices for Experiment 2 (upper panel) and Experiment 3 (lower panel).

If successful, this series of experiments will demonstrate that participants bias toward inaction can be manipulated by changing the cost of a miss relative to that of a false alarm. Furthermore, it will demonstrate that inaction/action biases and can be successfully modelled by the RRW. Such a finding would support the PVT account of the cognitive processes underlying responses to sacrificial moral dilemmas and, as such, provide an alternative explanation to that of Dual Process Theory. All procedures were approved by the UNCW IRB, protocol #16-0210.

**Experiment 1**

Experiment 1 is an attempt to validate PVT’s ability to predict responses and quantify the changes in inaction biases resulting from variations in sacrificial moral dilemmas. Recall, it has been demonstrated that when intentional personal force is required to sacrifice the bystander, participants are less likely to choose the “do something” response. From DPT, it is hypothesized that the required intentional personal force activates a non-consequentialist, social/emotional decision process that competes with the default consequentialist decision process. From PVT, in contrast, it is hypothesized that response biases, such as the inaction bias, are the result of a consequentialist process that weighs the cost of a miss relative to the cost of a false alarm. Such a process is generally modelled in sequential sampling processes (such as the RRW) as start point or evaluation bias.

Because the distinction between personal and impersonal is subject to debate and has changed over time, we use more precise terms here. Here, we manipulated the *agent of death* as either the respondent themselves or some unnamed other. We term the dilemmas *personalized* when the the respondent is asked to imagine being placed in the position of the agent, and *impersonalized* when an anonymous other is the agent. We assume the personalized dilemma is more “intuitively ‘up close and personal’” (see Greene et al., 2001) than the impersonalized dilemmas. We also manipulated the *method of death* as strangling or some unspecified method of killing. Here, method of death manipulates personal force: whereas strangulation necessitates personal force, the unspecified condition does not.

If the *agent of death* manipulation is successful, participants should exhibit a stronger inaction bias in the *personalized* trials than the *impersonalized* trials. If the *method of death* manipulation is successful, participants should exhibit a stronger inaction bias in the strangle trials than the unspecified trials. If PVT successfully models the data from Experiment 1, then the best fit model will serve as a baseline against which the manipulations of other experiments will be compared.

**Methods**

***Participants***

Eighty-twonaïve participants volunteered and received course credit for their participation. Although it would have been preferable to collect demographic data, we did not. Nevertheless, the UNCW undergraduate student body is 64% female, 78% white, 9% Hispanic, 4% African American, 8% one or more other races, and 1% unknown. The average student age is 21 with 70% of students between the ages of 18 and 22.

The power calculations were completed in steps. First, the effect size for the relation between Psychological Value and the DVs is huge (e.g., *r2*>0.85; Cohen et al., 2022). With 120 trials per participant, power ≈ 1.0 to detect the effect in an individual participant. Additionally, the effect size for the inaction bias is large (e.g., *d ≈ 0.8*; Hauser, et al., 2007). To be cautious, we ran a power analysis assuming a medium effect size, two-tailed alpha = 0.05, and power = 0.9. Although the minimum sample size needed is N = 44, we rounded up to N = 50. Final sample size was determined by (a) setting a minimum number of participants (50), (b) estimating the time necessary to collect that number of participants, (c) posting all available experimental slots for the time estimated in b, and (d) running all participants who signed up.

***Stimuli***

Participants sat at a desk in front of a 24-in LED color monitor Mac with a 72-Hz refresh rate (resolution 1920×1200 pixels) controlled by a Macintosh Mini in a dark private testing room. White noise played at a low volume from ceiling speakers to mask any ambient noise. Participants entered responses on an Apple keyboard. The experiments were controlled by bespoke JAVA software developed by the authors running under MacOS.

On each trial, a sacrificial moral dilemma was displayed. The scenario was such that a hypothetical respondent could act to save one person (Item 1), but this action resulted in another person being killed (Item2). For every trial, “Item1” and “Item2” were randomly chosen from a list of 23 possible stimuli. The Psychological Values distributions of the 23 stimuli were collected by Cohen et al., (2022). Furthermore, the 23 individuals were identical to the human stimuli used in Cohen et al.’s, (2022) choice experiments, with the exception that we removed “an adult with a deadly contagious disease.” We removed that stimulus because of the large number of characters required for the description. Table S1 in the Supplemental Materials describes the 25th, 50th, and 75th percentile of each stimuli’s Psychological Value distribution.

***Procedure***

All participants were tested individually in a small windowless testing cubicle. The experiment was run on a Mac desktop. An example of a typical scenario is shown below as presented on a trial. Cohen et al., (2022) demonstrated that the pattern of results from the simplified scenario (see below) is identical to that of elaborate scenarios.

Through circumstances out of your control,

Item1

is about to be killed, but

Item2

will not be affected. You have the opportunity to save Item1.

However, if you save Item1, Item2 will be killed.

Would you save

Item1

causing

Item2

to be killed?

Both the Agent of Death and the Method of Death varied over trials (see Supplementary Materials for scenarios). The Agent of Death referred to who would have to kill Item2 to save Item1 – either the actual participant (personalized) or some anonymous other (impersonalized). The Method of Death referred to how the Agent of Death must kill Item2 to save Item1 – either by strangling or by some unspecified means. In the example scenario, the Agent of Death is *impersonalized* and the means of killing is *unspecified*. In the impersonalized-strangled scenario, “will be killed” was changed to read “will be strangled.” In the personalized-unspecified scenario, “will be killed” was changed to read “you must kill.” In the personalized-strangled scenario, “will be killed” was changed to read “you must strangle.” The two levels of Agent of Death and of Method of Death were combined factorially (within-participants), giving rise to four different kinds of sacrificial moral dilemmas.

Participants were instructed to respond under standard RT conditions, “d” and “k” were the designated response keys. Allocation of a particular key to a particular response was randomized over participants. To exclude reading time from response time, we used a masked presentation type developed by Cohen and Ahn (2016). At the start of each trial, a red fixation cross was presented for 500 ms. Next, a masked form of the moral dilemma appeared on the screen. Each character in Item1 was replaced by a teal “+,” and each character in Item2 replaced by a teal “=.” At the onset of the text a progress bar on the left of the screen began to descend and unfolded at a rate of 150 ms per word. The Items were unmasked when bar reached its limit. The unmasked dilemma remained on the screen until the participant made a response.

In each experiment there were eight practice trials prior to the experimental trials. The practice trials were identical to experimental trials. Once the practice trials were initiated testing continued for forty minutes. In those 40 minutes, participants ran an average of 186 trials (SD = 31). A self-timed break was provided every 13 minutes. At the conclusion of the experimental session, a dialog box appeared that thanked the participant for their participation and indicated that they should leave the testing room.

**Results**

We used the Psychological Values collected by Cohen et al., (2022) to predict the RT and response choices in the moral dilemmas. Using the technique described by Cohen and Ahn (2016), we calculated a non-parametric measure of the overlap of each pair of probes (Cohen et al., 2022), termed *Overlap*. Overlap ranged from 0 (no overlap, thus highly dissimilar values) to 1 (complete overlap, thus identical values). To ensure stable estimates of the probability of choosing the higher valued option, *p*(HVO), and RT, Overlap was rounded to the nearest 0.1, termed as O0.1 (see Cohen & Ahn, 2016).

Following on from previous work (see Cohen et al., 2023), the following inclusion/exclusion criteria were applied to filter plausibly inattentive responses. The RT Coefficient of Variation (CVRT=SDRT/MRT) is arguably a measure of attention fluctuation that corrects for average RT (because SD systematically varies with RT). Therefore, we calculated the overall CVRT for each participant and removed the 5% of participants with the greatest CVRT (five participants were so removed). Individual outlier trials were removed via deletion of the fastest and slowest 2.5% of trials for each level of distributional overlap. Removing outliers by quantile does not require an assumption that the data are distributed normally (as is the case with SD). Finally, we removed individual participants who responded at or below chance (three participants removed). We analyzed the remaining participants’ datasets (74). To remove the effects of learning from RT, we fitted a learning function to each participant’s data and used the residual as our measure of RT (*RTres* after Cohen & Ahn, 2016).

**Psychological Value**

A defining property of PVT is that response choice and RT are a function of the overlap of the two options’ Psychological Value distributions. Little overlap should produce fast and accurate responses (where accurate is defined as choosing the HVO), whereas a large overlap should produce slow and error prone responses. Specifically, the overall probability of choosing the HVO, *p*(HVO) should follow an exponential decay as a function Overlap:

$p\left(HVO\right)= 0.5\* \left(1-\left(O\_{0.1}^{b}\right)\right)+0.5$ (1)

where *b* describes the shape of the exponential decay. Furthermore, the average RT should follow a linear as a function Overlap (O0.1). We summarized the data by calculating the mean RT and *p*(HVO) for each level of O0.1. The data support PVT’s predictions (see Figure 6). There was a significant exponential decay function relating *p*(HVO) and Overlap, *t* = 17.7, p < 0.001, *r2* = 0.97; *beta* = 1.27. There was also a significant linear function relating RT and Overlap, *F*(1,8) = 56.58, *p* < 0.001, *r2* = 0.88; slope = 0.25; intercept = -0.13.



Figure 6: Graphical summaries of the behavioural data for the three experiments. Upper panels show *p*(HVO) as a function of distributional overlap. Lower panels *RTres* as function of distributional overlap.

**Inaction Biases**

Overall, our participants had a low probability of choosing to act, *p*(action) = 0.38, SD = 0.46. We designed the experiment so that the HVO is the condemned individual about 50% of the time. Therefore, an unbiased participant who is attempting to identify and save the HVO (i.e., consequentialist reasoning) would choose to act about 50% of the time. To assess whether our participants had an inaction bias consistent with PVT’s predictions, we conducted a one sample *t*-test that assessed whether p(action) is significantly different that 0.5. The data revealed a significant inaction bias with a large effect size, *t*(73) = -8.1, *p* < 0.001, Cohen’s *d* = 0.95.

To assess whether the agent and method of death manipulations influenced the size of the inaction bias, we calculated a Mixed Model ANOVA with participant identified as a random variable. There was a significant effect of agent of death, *F*(1, 71.7) = 6.24, *p* = 0.015, such that the probability of action in the personalized condition (M = 0.36, SD = 0.16) was lower than that in the impersonalized condition (M = 0.39, SD = 0.15). There was no significant influence of method of death, *F*(1, 70.7) = 0.21, *p* = 0.65.

**Robust Random Walk (RRW)**

The above analyses have demonstrated participants’ RT and choices are a function of the overlap of the Psychological Value distributions the options and that the participants had an inaction bias. Here, we model the data using our RRW procedure. We implemented a RRW procedure that instantiated the assumptions of PVT (Busemeyer et al., 2019; Krajbich, 2019; Krajbich et al., 2010). Unlike traditional value-based sequential sampling procedure (henceforth VSSP) (see e.g., Busemeyer et al., 2019), the RRW uses the direct measures of value Overlap (described above) to drive the evidence accumulation, rather than estimating value indirectly (for detailed description, see Cohen et al., 2022). This provides strong constraints on the ability of the model to fit the data.

To fit the model, we first coded each trial to identify whether the option condemned to death by default was the HVO (as determined by the Psychological Values of Cohen et al., 2022), termed DefaultHVO, or the LVO, termed DefaultLVO. By doing so, we can identify whether participants have a start point or evaluation bias towards the “do nothing” or “do something” options. In all cases, *p*(HVO) will be a function of Overlap (i.e., the participants’ ability to identify the HVO decreased as the overlap of the options’ Psychological Value distributions increases). However, if participants have a bias toward “do nothing,” then participants will be faster and more accurate (i.e., *p*(HVO) will be greater) on the DefaultLVO trials than the DefaultHVO trials. If participants have a bias toward “do something,” then the reverse will be true. To test the hypothesis that the start point or evaluation bias can be influenced by both the agent and death method, we also coded for agent (personalized vs. impersonalized) and death method (unspecified vs. strangled). We summarized the data by calculating the probability of choosing the HVO (i.e., *p*(HVO)) and the mean RT by Overlap x Default HVO/LVO x Agent x Death Method. To remove unstable estimates, we excluded conditions that had fewer than 50 trials (i.e., there were few incorrect responses for some low overlap (high accuracy) trials).

We compared the ability of five models to predict the data (see the Supplementary Materials for complete model specifications). All models had four free parameters: non-decision time, *Ter*; 0.5\*boundary separation, *b*; the standard deviation of the Gaussian noise added to the drift, *nSD*; and the Information Accrual Bias (*IAB*). The *IAB* weights the influence of the sample by the time since the start of the process (see Cohen et al., 2022 for details). The function is defined relative to two parameters *dA* and *dB*. Here, we fix *dA* at 0.2 and let *dB* fit to the data. A positive d*B* indicates that more recent information carries more decisional weight. When *dA* is fixed and *dB* is positive, this indicates that perceptually recent information carries more decisional weight than perceptually distant information. All other parameters were fixed at 0.

Inferential tests revealed two inaction biases: an overall bias and a bias resulting from the Agent manipulation. We ran several models to identify whether these biases are best explained by shifts in the start point and/or evaluation criteria. We ran the following models:

1. *The Unbiased Model*: The unbiased model contains only the common parameters. The Unbiased model fits the data as a function of Psychological Value alone. Therefore, it provides a measure of the model fit when no bias parameters are included. This model will be superior, if neither a shift in the start point or evaluation criteria can accommodate the inaction biases we identified as present in the data.
2. The *Agent SE Overall SE Model*: In addition to the common parameters, this model includes a start point bias parameter for all trial types, SPoverall, plus a start point bias parameter to capture a change in start point as a function of the agent of death, SPagent. Therefore, there will be a start point shift toward either the “do nothing” or “do something” boundary that is different for personalized and impersonalized trials. A positive start point parameter (SP) indicates a shift toward the “do nothing” boundary and a negative SP parameter indicates a shift toward the “do something” boundary.
3. The *Agent EC Overall SE Model*: This model is identical to the *Agent SE Overall SE* Model, with the exception that the Agent of Death bias is coded for a shift in the evaluation bias rather than the start point bias, ECagent. The evaluation bias parameter assumes that the Psychological Value distributions are Gaussian with equal variance. With this assumption, we can estimate the shift in the evaluation criterion in SD units, where 0 is unbiased and a negative evaluation bias parameter (EC) indicates a shift in the evaluation criterion to favor the “do nothing” response and a positive EC parameter indicates a shift in the evaluation criterion to favor the “do something” response.
4. The *Agent SE Overall EC Model:* This model is identical to the *Agent SE Overall SE* Model, with the exception that the overall inaction bias is coded for a shift in the evaluation bias rather than the start point bias, ECoverall.
5. The *Agent EC Overall EC Model:* This model is identical to the *Agent SE Overall SE* Model, with the exception that the overall inaction bias and the Agent of Death biases are coded for a shift in the evaluation bias rather than the start point bias, ECoverall & ECagent.

Because sequential sampling models are stochastic, they can be run with the same parameter values and get slightly different fits. To quantify the variability of the fits with a single set of parameters, we ran the best fit model 40 times and calculated the average predicted RT and *p*(HVO) for each overlap (see Supplemental Materials for figures with the individual 40 runs). To determine the best fit model, we compared the BIC of each model. When the BIC’s were essentially equivalent (within 1 or 2 points), we accepted the simpler model as the better model based on parsimony.

Table 1 presents the *r*2 and BIC for each RRW model run. The data from Experiment 1 were best fit by the Agent EC + Overall EC Biased Model (*r2*=0.87; see Figure 7). This supports the conclusion that the inaction bias is a function of a shift in the evaluation bias, rather than the start point. Although, the model fit the RT data well (*r2*=0.75), there were indications that participants responded to the personalized trials faster than the impersonalized trials. This suggested that there was an additional effect of Agent of Death on the boundary parameter. The boundary parameter specifies how much information is required before a choice is made. The closer boundaries are together, the less information is required before a response, and therefore responses are quicker. Critically, closer boundaries also result in more errors. So, boundary placement is often used to model speed/accuracy tradeoffs (e.g., Ratcliff & Kang, 2021). We added a boundary parameter for the Agent of Death manipulation, BAgent, to all the bias models and ran them again. The data revealed that the addition of the BAgent parameter meaningfully improved the fit. The Agent B + Agent EC + Overall EC Biased Model outperformed all other models (*r2*=0.89). This model improved the fit of the RT data (*r2*=0.82).

Table 2 presents the parameter values for the best fitting model in Experiments 1-3. In Experiment 1, the Agent B + Agent EC + Overall EC Biased Model revealed an evaluation criterion shift of 0.46 standard deviations favoring the “do nothing” response for all trials and an additional 0.14 standard deviation shift favoring the “do nothing” response for the personalized trials (totaling a 0.6 standard deviation shift for the personalized trials). Furthermore, in the personalized condition, participants set their boundaries about 16% closer together than in the impersonalized condition. This suggests that participants required less information in the personalized condition before they felt confident enough to respond. Thus, the personalized condition was faster for two reasons: 1) the evidence accumulation was faster in the personalized than impersonalized condition and 2) the boundaries were closer together in the personalized than impersonalized condition. This suggests that participants were more confident in their answers in the personalized than the impersonalized condition.

**Table 1** BIC and *r2* for each RRW model by experiment. The best fit model will have the lowest BIC and is identified in the “Best Fit” column.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | Model Fit | Best Fit |
|  | Model |  | *r2* | BIC |
| **Experiment 1** |  |  |  |  |  |
| Unbiased Model |  | 0.52 | 414 |  |
| Agent SE + Overall SE Biased Model |  | 0.78 | 292 |  |
| Agent EC + Overall SE Biased Model |  | 0.81 | 248 |  |
| Agent SE + Overall EC Biased Model |  | 0.86 | 265 |  |
| Agent EC + Overall EC Biased Model |  | 0.87 | 160 |  |
|  |  |  |  |  |
| Agent B + Agent SE + Overall SE Biased Model |  | 0.80 | 278 |  |
| Agent B + Agent EC + Overall SE Biased Model |  | 0.84 | 226 |  |
| Agent B + Agent SE + Overall EC Biased Model |  | 0.89 | 142 |  |
| Agent B + Agent EC + Overall EC Biased Model |  | 0.89 | 131 | **✓** |
|  |  |  |  |  |
|  |  |  |  |  |
| **Experiment 2** |  |  |  |  |  |
| Unbiased Model |  | 0.65 | 310 |  |
| Overall EC Biased Model |  | 0.76 | 245 |  |
| Agent EC + Overall EC Biased Model |  | 0.78 | 234 |  |
| Agent B + Agent EC + Overall EC Biased Model |  | 0.83 | 198 | **✓** |
| **Experiment 3** |  |  |  |  |  |
| Unbiased Model |  | 0.86 | 172 |  |
| Method SE + Agent EC + Overall EC Biased Model |  | 0.88 | 160 |  |
| Method EC + Agent EC + Overall EC Biased Model |  | 0.88 | 151 | **✓** |
| Agent B + Method SE + Agent EC + Overall EC Biased Model |  | 0.88 | 163 |  |
| Agent B + Method EC + Agent EC + Overall EC Biased Model |  | 0.88 | 161 |  |

**Table 2.** Model fits and model parameter values broken down according to experiment.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Free Parameters |  |  | *r2* |
| Exp. | Ter | b | dB | nSD | ECoverall | ECAgent | BAgent | ECMethod |  | Final Model | 3-fold validity |
| 1 | -1.35 | 161 | 0.15 | 1.86 | -0.46 | -0.14 | -26 | NA |  | 0.89 | 0.82 |
| 2 | -1.24 | 178 | 0.14 | 1.88 | -0.16 | -0.16 | -28 | NA |  | 0.83 | 0.74 |
| 3 | -1.74 | 154 | 0.11 | 2.54 | -0.13 | -0.09 | NA | 0.11 |  | 0.88 | 0.80 |

**3-Fold Cross Validation**

To assess the skill of the model to fit the data, we ran a 3-fold cross validation analysis (Bischl et al., 2012; see Supplementary Materials for details). In a 3-fold cross validation analysis, the data are randomly (without replacement) split into three equal sized samples. On each fold, one sample is designated the *Test* dataset and the remaining samples are the *Fit* dataset. First, the best fit model from Experiment 1 (i.e., Agent B + Agent EC + Overall EC Biased Model) is fit to the *Fit* dataset. Then the model parameter values identified on the Fit dataset are fixed and used to predict the *Test* dataset. The model fit statistic (*r2*) of the *Test* dataset is saved, and the process is repeated for each fold. As can be seen in Table 2, the average *r2* for the *k*-fold cross validation analysis is close to that fit to the entire dataset (though predictably lower because they were fit to 1/3 the data), demonstrating the skill of the model to fit the data.



**Figure 7**. The PVT computational model fit to the behavioral data from Experiment 1. The behavioral data are displayed as symbols, whereas the computational model fit is displayed as lines. The top half of the figure contains three graphs of the data from the personalized trials and the bottom half contains three graphs of the data from the impersonalized trials. In each half, the top graph illustrates how the probability of choosing the higher valued option, *p*(HVO) (i.e., accuracy) changes with Overlap of the Psychological Value of the two probes in each scenario. The bottom graphs show how RT changes with Overlap of the Psychological Value of the two probes in each scenario: the left graph displays the “correct” response data, whereas the right graph displays the “incorrect” response data. The computational model fit all conditions simultaneously. In Experiment 1, participants revealed a large inaction bias. This bias is visible when the data are viewed as a function of the Psychological Value of the item that will be killed by default. The DefaultHVO symbols represent the data when the HVO will be killed by default, so one must “do something” for the trial to be coded as “correct.” The DefaultLVO symbols represent the data when the LVO will be killed by default, so one must “do nothing” for the trial to be coded as “correct.” The gap between the two reveals the “do nothing” bias. That is, participants were slower and less accurate when the HVO is killed by default, relative to when the LVO is killed by default. The “do nothing” bias is larger in the personalized trials than the impersonalized trials.

**Discussion**

In Experiment 1, we presented participants with sacrificial moral dilemmas in which the agent and method of death was varied. Regression analyses revealed a strong relation between responses (separately, for both RT and *p*(HVO)) and the overlap of the Psychological Value distributions of the condemned vs. bystander. Furthermore, an analysis of the probability of choosing the “do something” option revealed a large, generalized inaction bias and a smaller, additional inaction bias resulting from the agent of death manipulation. Because our participants had a low probability of choosing to act, DPT researchers would typically conclude that our manipulation activated the social/emotional decision process, and that process out competed the consequentialist process (Białek & De Neys, 2016; Bostyn et al., 2018; Conway & Gawronski, 2013; Greene et al., 2001; Greene, & Haidt, 2002; Hauser et al., 2007; Koenigs et al., 2007).

We also modeled the data with the RRW. The RRW precisely specifies the process underlying participants’ responses based on the assumptions of PVT. The best fit model, the Agent B + Agent EC + Overall EC Biased Model, accounted for 89% of the variance in our three measured DVs: *p*(HVO), RTHVO, and RTLVO. The data revealed that participants’ response choices and RTs were extremely well predicted by the distributional overlap of the Psychological Values of the options (see also Figure 7). The ability of the RRW to fit both RT and response choice data simultaneously with few parameters speaks to the validity of the single process model and the underlying normed Psychological Value data. The Agent B + Agent EC + Overall EC Biased Model identified two sources of inaction bias: a generalized response bias toward inaction and a more specific response bias toward inaction when the decision maker was required to sacrifice the bystander to save the condemned individual. These data validate the hypothesis that inaction biases can be accurately modelled within the PVT framework as an evaluation bias favouring the “do nothing” response. These findings provide strong evidence that participants were engaging in consequentialist reasoning and that the “do something” and “do nothing” responses manifest from this consequentialist process alone (see also Cohen & Ahn, 2016; Cohen et al., 2021; Cohen et al. 2023).

Although Experiment 1 demonstrated that PVT can successfully predict responses in sacrificial moral dilemmas and account for changes in the inaction bias by varying the evaluation bias parameter, the information that drives that change is not specified by the model. Nevertheless, there is other excellent evidence showing that response biases arise because the decision-maker weighs the cost of a miss relative to that of a false alarm (Link & Heath, 1975; Macmillan & Creelman, 2005; Ratcliff, 1978). Experiment 2 attempts to provide evidence for this. The intention was to increase the cost of a miss relative to a false alarm and assess whether that increase predictably influences participants’ inaction bias.

**Experiment 2**

In Experiment 2, we attempted to increase the probability of action by increasing the cost of a miss. In this respect, we changed the dilemmas so that the condemned individual was to be tortured to death over several days by default. The pain and suffering associated with torture should increase respondents’ aversion to allowing the torture of a HVO (increasing the cost of a miss) and hence the tendency to “do something” should increase.

**Methods**

***Participants***

Sixty-eight naïveparticipants volunteered and received course credit for their participation. Sample size was determined similar to Experiment 1.

***Stimuli***

The stimuli in Experiment 2 were identical to Experiment 1, with the following exception. Instead of the Item1 individual being *killed*, each scenario specified that this individual would be *tortured to death over the course of several days* if the respondent did nothing. In this way the participant now had to consider that doing nothing would result in a very extreme form of painful death. Critically, relative to the scenarios in Experiment 1, we changed nothing about the participant’s required actions to save the condemned individual.

***Procedure***

The procedure in Experiment 2 was identical to Experiment 1. In the 40 minute experimental session, participants ran an average of 146 trials (SD = 27).

**Results**

We implemented the same data filtering procedure as Experiment 1. In addition to removing individual outlier trials, we removed four participants according to the CVRT criteria and two participants who responded below chance. We analyzed the remaining 62 participants’ datasets.

**Psychological Value and Inaction Biases**

We conducted the same regression and ANOVA analyses in Experiment 2 as we did in Experiment 1. Again, the data support PVT’s predictions (see Figure 6). There was a significant exponential decay function relating *p*(HVO) and Overlap, *t*=18.9, *p* < 0.001, *r2* = 0.98; beta = 1.52. There was also a significant linear function relating RT and Overlap, *F*(1,8) = 45.5, *p* < 0.001, *r2* = 0.85; slope = 0.24; intercept = -0.12.

Participants had an overall inaction bias, *p*(action) = 0.44, SD = 0.47. This bias was significantly lower than an unbiased *p*(action) = 0.5, *t*(61) = -3.5, *p*< 0.001, Cohen’s *d* = 0.45. The inaction bias in Experiment 2 was significantly less than that identified in Experiment 1, *t*(123.2) = -2.4, *p* = 0.018, Cohen’s *d* = 0.42.

To assess whether the agent and method of death manipulations influenced the size of the inaction bias, we calculated a Mixed Model ANOVA with participant identified as a random variable. There was a significant effect of agent of death, *F*(1, 59.2) = 9.79, *p* = 0.003, such that the probability of action in the personalized condition (M = 0.41, SD = 0.16) was lower than that in the impersonalized condition (M = 0.46, SD = 0.18). There was no significant influence of method of death, *F*(1, 56.3) = 0.92, *p* = 0.34.

**RRW**

In Experiment 2, we *increased the cost of a miss* by changing the scenario so that the condemned individual was tortured to death. The above analyses confirmed that this change reduced the inaction bias in Experiment 2 relative to Experiment 1. If moral judgments are the result of the same processes as other decisions, then if we can increase the cost of a miss, we will see a reduction in the evaluation bias parameter favoring the “do nothing” response.

Here, we modeled the data using our Robust Random Walk procedure. We modeled the overall and agent of death inaction biases as a function of a shift in the evaluation bias. We did so because Experiment 1 revealed that the evaluation bias best accommodated these biases in the data. The best fit model from Experiment 2 was, again, the Agent B + Agent EC + Overall EC Biased Model (*r2*=0.83; see Tables 1 & 2 and Figure 8). Critically, participants exhibited an evaluation criterion shift of 0.16 standard deviations favoring the “do nothing” response for all trials and an additional 0.16 standard deviation shift favoring the “do nothing” response for the personalized trials (totaling a 0.32 standard deviation shift for the personalized trials). This 47% reduction in the “do nothing” evaluation bias parameter relative to Experiment 1 supports our prediction that increasing the cost of a miss will increase the probability of action. Once again, in the personalized condition, participants set their boundaries about 16% closer together than in the impersonalized condition. The results of the 3-fold cross validation analysis confirm the robust nature of this result and are presented in Table 2 (and see Supplementary Materials).



**Figure 8**. The PVT computational model fit to the behavioral data from Experiment 2. The formatting of the figure is the same as for Figure 7. In Experiment 2, participants revealed a reduced bias toward the “do nothing*”* choice. Moreover, participants were slower and less accurate when the HVO is killed by default, relative to when the LVO is killed by default. As is visible, the inaction bias is larger in the personalized trials than the impersonalized trials.

**Discussion**

Experiment 2 tested the hypothesis that the inaction bias is a function of the cost of a miss relative to that of a false alarm. We increased the cost of a miss by stating that the condemned individual would be tortured for three days prior to death. We hypothesized that the pain and suffering involved in the torture would increase the cost of failing to save a high valued individual (a miss), relative to saving a lower valued individual (a false alarm). If response biases are influenced by the cost of a miss relative to that of a false alarm, this manipulation should reduce the inaction bias seen in Experiment 1.

 The results of Experiment 2 confirmed these predictions. Specifically, the generalized inaction bias reduced by about 65%, while the inaction bias resulting from the agent of death manipulation stayed about the same. This result is particularly striking because the agent and method of death manipulations were identical in Experiments 1 and 2. Thus, we were able to reduce the inaction bias independent of the intentions or actions of the actor who sacrifices the condemned individual (the hypothesized cause of the inaction bias by DPT researchers, see e.g., Greene et al., 2004).

In Experiment 3, we attempt to provide additional evidence that payoff matrices influence the inaction response bias.

**Experiment 3**

In Experiment 3, we decrease the cost of a false alarm relative to that in Experiment 2. Rather than ask the respondent whether they would strangle the bystander to save the condemned individual (Experiment 2), we asked whether they would humanely kill the bystander with a lethal injection. The pain and suffering that the bystander experiences with a humane lethal injection (Experiment 3) is less than that the bystander experiences with strangulation (in Experiment 2). Therefore, the cost of a false alarm (killing a higher valued bystander in error) in Experiment 3 is lower than that of Experiment 2. Because we keep the cost of a miss the same in Experiments 2 and 3 (i.e., torturing the condemned individual for three days until death), the lower cost of a false alarm in Experiment 3 should further reduce the participants’ inaction bias relative to Experiment 2. According to PVT, such a reduction should *decrease* the evaluation bias favoring the “do nothing” response. Furthermore, this effect should manifest from the method of death manipulation.

**Methods**

***Participants***

Ninety-eight naïveparticipants volunteered and received course credit for their participation. Sample size was determined similar to Experiment 1.

***Stimuli***

The stimuli in Experiment 3 were identical to Experiment 2, with the following exception. The Method of Death with respect to the demise of Item 2 was altered. Now the choice was between some unspecified means of killing vs. administrating a humane lethal injection (rather than strangulation). In this way the intention was to have the participant consider the possibility of acting to save someone from a very extreme and painful death by killing another person humanely.

***Procedure***

The procedure in Experiment 3 was identical to Experiment 1. In the 40-minute experimental session, participants ran an average of 133 trials (SD = 30).

**Results**

We implemented the same data filtering procedure as Experiment 1. In addition to removing individual outlier trials, we removed five participants according to the CVRT criteria and seven participants who responded below chance. We analyzed the remaining 86 participants’ datasets.

**Psychological Value and Inaction Biases**

We conducted the same regression and ANOVA analyses in Experiment 3 as we did in Experiments 1 and 2. Again, the data support PVT’s predictions (see Figure 6). As Figure 6 shows there was a significant exponential decay function relating *p*(HVO) and Overlap, *t*=18.81, *p* < 0.001, *r2* = 0.98; beta = 1.46, and a significant linear function relating RT and Overlap, *F*(1,8) = 152.9, *p* < 0.001, *r2* = 0.95; slope = 0.32; intercept = -0.17.

Participants overall inaction bias, *p*(action) = 0.48, SD = 0.46, was not significantly different than an unbiased p(action) = 0.5, *t*(85) = -0.82, *p*= 0.42, Cohen’s *d* = 0.09. When comparing the inaction biases of Experiments 2 and 3, we split the data according to the levels of the method of death manipulation. When the method of death was “unidentified,” the scenarios in Experiment 2 and 3 were identical. Therefore, in this condition the inaction biases should not significantly differ across experiments. The data support this hypothesis: Experiments 2 and 3 did not significantly differ, *t*(137.8) = -1.34, *p*= 0.18, Cohen’s *d* = 0.22. When we identified the method of death, the scenarios differed in Experiment 2 (strangulation) and 3 (humane lethal injection). If reducing the cost of a false alarm relative to a miss influenced the inaction biases, then the inaction bias in the Experiment 3 humane lethal injection condition should be reduced relative to the Experiment 2 strangulation condition. The data support this hypothesis: the inaction bias was significantly reduced in humane lethal injection condition of Experiment 3 relative to the strangulation condition of Experiment 2, *t*(144.9) = -2.31, *p*= 0.02, Cohen’s *d* = 0.38. These data support the conclusion that Experiment 3 eliminated the inaction bias in the humane injection condition.

To assess whether the agent and method of death manipulations influenced the size of the inaction bias, we calculated a Mixed Model ANOVA with participant identified as a random variable. There was a significant effect of agent of death, *F*(1, 70) = 6.37, *p* = 0.01, such that the probability of action in the personalized condition (M = 0.47, SD = 0.20) was lower than that in the impersonalized condition (M = 0.50, SD = 0.19). There was also a significant influence of method of death, *F*(1, 76.4) = 10.44, *p* = 0.002, such that the probability of action in the *unidentified* condition (M = 0.47, SD = 0.19) was lower than that in the *human injection* condition (M = 0.50, SD = 0.19). These data support the conclusion that decreasing the cost of a false alarm reduces the inaction bias.

**RRW**

In Experiment 3, in addition to retaining the increased cost of a miss implemented in Experiment 2, we reduced the cost of false alarm by having the agent save the condemned individual from the torture by sacrificing the bystander with a humane lethal injection. As such, we predict the evaluation bias favoring the “do nothing” response to reduce even further.

Similar to Experiment 2, we modeled the overall and agent of death inaction biases as a function of an evaluation bias. However, we modeled the method of death inaction bias both as a shift in the start point and an evaluation bias. We did so because there was no method of death inaction bias identified in Experiment1 1 and 2. The best fit model from Experiment 3 was the Method EC + Agent EC + Overall EC Biased Model (*r2*=0.88; see Tables 1 & 2 and Figure 9). Participants exhibited an evaluation criterion shift of 0.13 standard deviations favoring the “do nothing” response for all trials and an additional 0.09 standard deviation shift favoring the “do nothing” response for the personalized trials. Critically, participants exhibited an evaluation criterion shift of 0.11 standard deviations favoring the “do something” response when the method of death was a humane lethal injection. Thus, the total inaction bias was only a 0.1 standard deviation shift for the personalized, lethal injection trials and a 0.02 standard deviation shift for the impersonalized, lethal injection trials (virtually eliminating the inaction bias). The results of the 3-fold cross validation analysis confirm the robust nature of this result and are presented in Table 2 (and see Supplementary Materials).



 **Figure 9**. The PVT computational model fit to the behavioral data from Experiment 3. The formatting of the figure is the same as for Figures 2 and 3. As is visible, there is a small “do nothing*”* bias is in the unspecified trials, but no “do nothing*”* bias is in the injection trials.

**Discussion**

Experiment 3 provides further evidence that the inaction bias is a function of the cost of a miss relative to that of a false alarm. We decrease the cost of a false alarm by sacrificing the bystander with a humane injection. We hypothesized that the pain and suffering involved in the humane injection would decrease the cost of sacrificing a high valued individual (a false alarm) relative to allowing a high valued individual to endure three days of torture and eventual death (a miss).

As predicted, the manipulation reduced the inaction bias present in Experiments 1 and 2. Specifically, the generalized inaction bias was reduced and the inaction bias resulting from the agent of death manipulation was virtually eliminated. Critically, there was a *small bias towards action* for the humane injection condition. This bias towards action cancelled out the small generalized inaction bias, resulting in, effectively, unbiased performance.

DPT researchers propose that when the agent of death uses intentional personal force, the inaction bias will be increased. In Experiment 3, the agent intentionally administered the injection by puncturing the bystanders’ body with a needle and injecting the fluid into the bystander. If one accepts that this manipulation is a form of intentional personal force, then the current results contradict the predictions of DPT. Some may, however, argue that injecting poison into the body does not meet the conditions of intentional personal force because the poison does the killing rather than the needle itself (which penetrates the body). This objection, however, is inconsistent with DPT’s interpretation of the Footbridge dilemma. In the Footbridge dilemma, the bystander is sacrificed by being pushed off the overpass. In this scenario, the push does not kill the bystander – the train does. If one accepts that the action in the Footbridge dilemma constitutes intentional personal force, then consistency requires that administering a humane lethal injection also constitutes intentional personal force.

**General Discussion**

In 2001, Greene et al. posed the following question, “What makes it morally acceptable to sacrifice one life to save five in the trolley dilemma but not in the footbridge dilemma” (p. 2105). This question spurred decades of research on sacrificial moral dilemmas and provided the foundational evidence in favor of DPT (Awad et al., 2020; Bago & De Neys, 2019; Bialek & De Neys 2016; Chu & Liu, 2023; Conway & Gawronski, 2013; De Neys, 2022); Greene, 2009; Greene et al., 2009; Greene & Haidt, 2002; Hauser, et al., 2007; Koenigs, et al., 2007; Trémolière, & Bonnefon, 2014). Greene and colleagues (Greene et al., 2009; Greene & Haidt, 2002; Greene et al., 2004; Greene et al., 2001; Greene & Young, 2020) hypothesized that a slow, rational consequentialist decision process that favors a “do something” response competes with a fast, non-consequentialist Social/Emotional decision process that favors a “do nothing” response. Within this system, “intentional personal force” activates the Social/Emotional decision process, which explains why participants are reluctant to act in the Footbridge dilemma. Here, we present a radically different answer to Greene’s question. We propose that moral judgments recruit a general purpose, value-based decision mechanism that attempts to choose the option with the highest Psychological Value. Although the Psychological Value of the options is the primary predictor of choice, response biases (e.g., a start point or evaluation bias) can also influence the choice. Response biases are sensitive to the cost of a miss relative to a false alarm, and in the Footbridge dilemma the cost of a false alarm was likely higher than the Trolley dilemma.

Previous research (Cohen & Ahn, 2016; Cohen, et al., 2022, Cohen et al., 2023) has demonstrated that the overlap of the Psychological Value distributions of the condemned and the bystander were the primary predictors of responses in sacrificial moral dilemmas, as predicted by PVT. Here, we extended these results by demonstrating that inaction biases across varieties of sacrificial moral dilemmas are accurately modeled with PVT as changes in the evaluation bias parameter of the RRW. Furthermore, across three experiments, we demonstrate that these response biases are predictably influenced by the cost of miss relative to a false alarm.

In Experiment 1, we presented participants with sacrificial moral dilemmas that varied the agent and method of death. We found a large, inaction bias across all conditions and an additional smaller inaction bias resulting from the agent of death manipulation. The inaction bias was larger when the participant was the agent than when the agent was an anonymous other. PVT successfully predicted RT and choice in all conditions with a single process model, accounting for 89% of the variance. The inaction biases were well accommodated by the evaluation bias parameter. In Experiments 2 and 3, we assessed whether the inaction bias was predictably influenced by the cost of a miss relative to that of a false alarm. In Experiment 2, we increased the cost of a miss, whereas in Experiment 3, we decreased the cost of a false alarm. As predicted, both manipulations reduced the inaction bias, and the changes in responding were well accounted for by shifts in the evaluation bias parameter. These data provide strong evidence that inaction biases are a function of a cost/benefit analysis associated with making different types of errors.

PVT models choice as a sequential sampling procedure. In sequential sampling procedures (such as the RRW), response biases are modelled as shifts in the start point or evaluation criterion (Ratcliff, 1985; Ratcliff & McKoon, 2008; Zhao et al., 2019). There is robust evidence in the extant literature that response biases manifest from participants weighing the cost of a miss relative to a false alarm (Link & Heath, 1975; Macmillan & Creelman, 2005; Ratcliff, 1978; Ratcliff & McKoon, 2008; Zhao et al., 2019). To test the validity of this hypothesis, we varied the cost of a miss relative to a false alarm and assessed whether the inaction bias changed consistent with the model predictions. In our scenarios, the cost/benefit analysis of a false alarm vs. a miss is influenced by the pain and suffering of the two individuals in the scenario. There is a high cost of a miss when the condemned individual is being tortured, and a lower cost of a false alarm when the bystander will be sacrificed in a more humane manner. As predicted, the inaction biases decreased in both Experiments 2 and 3, accordingly. This provides evidence that the inaction biases in sacrificial moral dilemmas are sensitive to changes in a payoff matrix, similarly to other types of response biases in other decision tasks.

Some DPT theorists may claim that a theoretical Social/Emotional decision process is driving the evaluation criterion bias identified in the present research. This is an open empirical question. In our opinion, there are several reasons to reject such a hypothesis. First, adding a theoretical Social/Emotional decision process increases the complexity of the model without providing a concomitant predictive advantage. Thus, on the basis of parsimony, such a hypothesis should be rejected. Second, a defining feature of the Social/Emotional decision is that it is non-consequentialist. Here, we show that the evaluation criterion bias is sensitive to changes in a payoff matrix. Sensitivity to the relative costs of a miss and false alarm demonstrates that the evaluation criterion bias is a consequentialist process. As a consequentialist process, it is incompatible with a non-consequentialist, Social/Emotional decision process.

Finally, DPT posits that the theoretical Social/Emotional decision process competes with the consequentialist process. Response biases, in contrast, are an inherent part of PVTs random walk procedure. There are three key parameters of a random walk procedure that capture response biased: start points, evaluation criterion, and boundaries. These three parameters are necessary parts of all random walk procedures. Their values lay on a continuum. Traditionally, one point on that continuum may be labelled “unbiased.” For example, when the evaluation criterion equals 0, it is considered unbiased. Importantly, the “unbiased” value is not *categorically different* from all the other values. Rather, all values simply represent points on a continuum. As such, start points, evaluation criterion, and boundaries neither “compete with,” nor “short circuit,” PVT’s random walk procedure. PVT’s random walk procedure could not function without them.

There is robust evidence that including “intentional personal force” in a sacrificial moral dilemma will increases the likelihood that participants will choose to “do nothing” (Greene et al., 2009). We claim that “intentional personal force” *per se* is not the critical feature driving the inaction bias. Rather, the underlying feature driving the inaction bias is the cost of a miss relative to a false alarm. As such, under the right circumstances, intentional personal force should decrease the likelihood that participants will choose to “do nothing.” To support this claim, in Experiment 3 we manipulated the relative costs of a miss vs. a false alarm such that intentional personal force decreases the cost of a false alarm (the humane injection condition). As predicted, the inaction bias *decreased*. This provides strong support for the conclusion that the cost of a miss relative to a false alarm drives the inaction bias, rather than “intentional personal force” per se.

Traditional sequential sampling procedures have been used (sparingly) to study moral judgment (Andrejević, et al., 2022; Pärnamets et al., 2015; Son et al., 2019; Yu, et al., 2021). Traditional sequential sampling procedures, like utility theory, use participants’ responses to estimate the latent distributions hypothesized to be driving those responses. Although such research is important and informative, it does not address the fundamental question of what the latent distributions are representing. As such, it only informs our understanding of the decision processes rather than what is the underlying factor driving those decisions.

PVT, in contrast, is a theory of value and choice. It describes the features of Psychological Value, measures those features independent of choice, and then uses those measurements to drive the RRW. By doing so, PVT examines both the decision process and the source and features of the data driving that decision. Because PVT accurately predicted response choices and time, PVT provides a useful process model of how sacrificial moral dilemmas are solved. PVT describes a system that will fail to accurately predict RT and response choice if, (i) sacrificial moral dilemmas are not value based tasks, (ii) PVT’s definition and measurement of Psychological Value are invalid, *and/or* (iii) PVT’s model of the decision process (RRW) is invalid. As such, accurate predictions provide strong evidence for the validity of the entire system (the sequential sampling decision process and the source and features of the data driving that process). Importantly, because PVT uses measurements of Psychological Value to drive the RRW (rather than infer the latent distributions from the responses), it is a more tightly constrained model than traditional sequential sampling procedures.

PVT is also a tool that can be used by other researchers. Specifically, there are likely a myriad of factors that might influence participants’ responses to sacrificial moral dilemmas. PVT can be used to identify whether these factors are influencing the values of the items, or various bias parameters. Thus, rather than identifying *that* a factor influences judgments, PVT provides a method for identifying *how* that factor influences judgments.

There are several limitations in the present research. First, our sample consisted of college students in a Southeastern University in the USA. We have, however, demonstrated that Psychological Value Theory generalizes across cultures (Cohen et al., 2023). Second, the evaluation criterion bias is indistinguishable from a value shift in the RRW. As such, we cannot rule out the possibility that the effects were the result of shifts in valuation. There is, however, extensive evidence that evaluation criterion biases are sensitive to the cost of a miss relative to a false alarm, whereas perceptual evidence (such as a value shift) is not (Macmillan & Creelman, 2005). For this reason, we conclude that the inaction bias is caused by an evaluation criterion bias rather than a shift in valuation. Finally, like other moral judgment studies, we use hypothetical scenarios in highly controlled tasks. These laboratory studies are a necessary first step in understanding how moral judgments are made in real world contexts. It would be beneficial to extend the laboratory findings to more externally valid contexts.

In sum, our data support the conclusion that moral judgments are not special – rather they are value-based preferential choices that follow the same decision processes as any other value-based preferential choice. As such, we remove the need to appeal to *post‑hoc* morality specific constructs as the bases for an explanation. By specifying a precise, quantitative single process model, namely, PVT, we provide a robust computational foundation for the future study of moral judgment in general. We believe that such a model is far more useful in endeavoring to understand moral judgments than are the attempts to develop the arrows-and boxes accounts offered by DPT. Finally, we call upon the time-honored notion of simplicity in theory development to favor our approach (Sober, 2015). PVT is based on a single well specified decision-making process that is has a long tradition and general applicability in decision science. This account explains the same behavioral data as DPT but with fewer putative processes. Why a more complex account should be favored over a less complex account is something that remains both puzzling and troubling.

**References**

Andrejević, M., White, J. P., Feuerriegel, D., Laham, S., & Bode, S. (2022). Response time modelling reveals evidence for multiple, distinct sources of moral decision caution. *Cognition, 223*, 105026. doi: <https://doi.org/10.1016/j.cognition.2022.105026>.

Ashby, F. G., & Lee, W. W. (1993). Perceptual Variability as A Fundamental Axiom of Perceptual Science. In S. C. Masin (Ed.), *Advances in psychology* (Vol. 99, pp. 369-399): North Holland.

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J. F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences of the United States of America, 117*(5), 2332-2337. doi: <https://doi.org/10.1073/pnas.1911517117>

Bago, B., & De Neys, W. (2019). The intuitive greater good: Testing the corrective Dual Process Model of moral cognition. *Journal of Experimental Psychology: General, 148*(10), 1782-1801. doi: <http://dx.doi.org/10.1037/xge0000533>

Bialek, M., & De Neys, W. (2016). Conflict detection during moral decision-making: evidence for deontic reasoners’ utilitarian sensitivity. *Journal of Cognitive Psychology, 28*(5), 631-639. doi: <http://dx.doi.org/10.1080/20445911.2016.1156118>

Bischl, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation, 20*(2), 249-275. doi: <https://doi.org/10.1162/EVCO_a_00069>

Bostyn, D. H., Sevenhart, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science, 29*(7), 1084-1093. doi: <http://dx.doi.org/10.1177/0956797617752640>

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends in Cognitive Sciences, 23*, 251-263. doi: <http://dx.doi.org/10.1016/j.tics.2018.12.003>

Chu, Y. Y., & Liu, P. (2023). Machines and humans in sacrificial moral dilemmas: Required similarly but judged differently? *Cognition, 239*. doi: 10.1016/j.cognition.2023.105575

Claessens, S., Kyritsis, T., & Atkinson, Q. D. (2020). Revised analysis shows relational mobility predicts sacrificial behavior in Footbridge but not Switch or Loop trolley problems. *Proceedings of the National Academy of Sciences of the United States of America, 117*(24), 13203-13204. doi: 10.1073/pnas.2005012117

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General, 145*(10), 1359-1381. doi: <http://dx.doi.org/10.1037/xge0000210>

Cohen, D. J., Cromley, A. R., Freda, K. E., & White, M. (2022). Psychological value theory: The psychological value of human lives and economic goods. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(12), 2015–2050. doi: <http://dx.doi.org/10.1037/xlm0001047>

Cohen, D. J., & Lecci, L. (2001). Using magnitude estimation to investigate the perceptual components of signal detection theory. *Psychonomic Bulletin & Review, 8*(2), 284-293. doi: <https://doi.org/10.3758/BF03196163>

Cohen, D. J., Quinlan, P. T., & Liu, X. Y. (2023). Moral judgments are value-based decisions driven by culturally stable valuations and culturally variable decision biases. *Social Psychological and Personality Science*, 1-11. doi: 10.1177/19485506231199527

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. doi: <http://dx.doi.org/10.1037/a0031021>

De Neys, W. (2022). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences, 46*. doi: 10.1017/s0140525x2200142x

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

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

Greene, J. D. (2023). The dual-process theory of moral judgment does not deny that people can make compromise judgments. *Proceedings of the National Academy of Sciences of the United States of America, 120*(6), e2220396120. doi: <https://doi.org/10.1073/pnas.2220396120>

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364-371. doi: <http://dx.doi.org/10.1016/j.cognition.2009.02.001>

Greene, J. D., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences, 6*(12), 517-523. doi: [https://doi.org/10.1016/S1364-6613(02)02011-9](https://doi.org/10.1016/S1364-6613%2802%2902011-9)

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*(3), 1144-1154. doi: <http://dx.doi.org/10.1016/j.cognition.2007.11.004>

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389-400. doi: <https://doi.org/10.1016/j.neuron.2004.09.027>

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105-2108. doi: <http://dx.doi.org/10.1126/science.1062872>

Greene, J. D., & Young, L. (2020). The cognitive neuroscience of moral judgment and decision-making. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (Vol. 6, pp. 1005-1015). Cambridge, MA: The MIT Press.

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

Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural Responses to Ingroup and Outgroup Members' Suffering Predict Individual Differences in Costly Helping. *Neuron, 68*(1), 149-160. doi: 10.1016/j.neuron.2010.09.003

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A Neurocomputational Model of Altruistic Choice and Its Implications. *Neuron, 87*(2), 451-462. doi: 10.1016/j.neuron.2015.06.031

Johnson, R., & Cureton, A. (2018). Kants’ moral philosophy. In E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*.

Kahane, G., Everett, J. A. C., Earp, B. D., & Savulescu, J. (2015). ‘Utilitarian’ judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134*, 193-209. doi: <http://dx.doi.org/10.1016/j.cognition.2014.10.005>

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M. D., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446*(7138), 908-911. doi: <http://dx.doi.org/10.1038/nature05631>

Krajbich, I. (2019). Accounting for attention in sequential sampling models of decision making. *Current Opinion in Psychology, 29*, 6-11. doi: <https://doi.org/10.1016/j.copsyc.2018.10.008>

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience, 13*(10), 1292. doi: <https://doi.org/10.1038/nn.2635>

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika, 1*, 77-105. doi: <http://dx.doi.org/10.1007/BF02291481>

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). New York, NY: Lawrence Erlbaum Associates.

Maddox, W. T. (2002). Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Analysis of Behavior, 78*(3), 567-595. doi: 10.1901/jeab.2002.78-567

Marks, L. E., & Algom, D. (1998). Psychophysical scaling. *Measurement, judgment and decision making* (pp. 84-178). Cambridge, MA: Academic Press.

Martín-Guerrero, T. L., Ramos-Álvarez, M. M., & Rosas, J. M. (2016). Payoff affects tasters’ decisions but it does not affect their sensitivity to basic tastes. *Food Quality and Preference, 48*, 11-16. doi: <https://doi.org/10.1016/j.foodqual.2015.07.016>

Meehl, P. E. (1967). Theory- testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115. doi: <https://www.jstor.org/stable/186099>

Moll, J., & de Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences, 11*(8), 319-321. doi: <https://doi.org/10.1016/j.tics.2007.06.001>

Moore, A. B., Lee, N. Y. L., Clark, B. A. M., & Conway, A. R. A. (2011). In defense of the personal/impersonal distinction in moral psychology research: Cross-cultural validation of the dual process model of moral judgment. *Judgment and Decision Making, 6*(3), 186-195. doi: <https://doi.org/10.1017/S193029750000139X>

Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J., & Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences of the United States of America, 112*(13), 4170-4175. doi: 10.1073/pnas.1415250112

Paxton, J. M., & Greene, J. D. (2010). Moral reasoning: Hints and allegations. *Topics in Cognitive Science, 2*(3), 511-527. doi: <http://dx.doi.org/10.1111/j.1756-8765.2010.01096.x>

Perales, J. C., Catena, A., Shanks, D. R., & González, J. A. (2005). Dissociation between judgments and outcome-expectancy measures in covariation learning: A signal detection theory approach. *Journal of Experimental Psychology: Learning Memory and Cognition, 31*(5), 1105-1120. doi: 10.1037/0278-7393.31.5.1105

Quinlan, P. T., & Cohen, D. J. (2023). *A critical appraisal of the Dual Process account of moral judgment*. <https://osf.io/preprints/psyarxiv/62rgf>.

Quinlan, P. T., & Dyson, B. J. (2008). *Cognitive Psychology*. Harlow, England: Pearson.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59-108. doi: <http://dx.doi.org/10.1037/0033-295X.85.2.59>

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review, 92*(2), 212-225. doi: 10.1037/0033-295x.92.2.212

Ratcliff, R., & Kang, I. (2021). Qualitative speed-accuracy tradeoff effects can be explained by a diffusion/fast-guess mixture model. *Scientific Reports, 11*(1). doi: 10.1038/s41598-021-94451-7

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation, 20*(4), 873-922. doi: <http://dx.doi.org/10.1162/neco.2008.12-06-420>

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*(5), 347-356. doi: <https://doi.org/10.1111/1467-9280.00067>

Rosen, F. (2005). *Classical utilitarianism from Hume to Mill*. New York, NY: Routledge.

Singer, T., & Lamm, C. (2009). The Social Neuroscience of Empathy. In M. B. Miller & A. Kingstone (Eds.), *Year in Cognitive Neuroscience 2009* (Vol. 1156, pp. 81-96).

Sober, E. (2015). *Ockham’s Razors. A user’s manual*. Cambridge, UK: Cambridge University Press.

Son, J. Y., Bhandari, A., & FeldmanHall, O. (2019). Crowdsourcing punishment: Individuals reference group preferences to inform their own punitive decisions. *Scientific Reports, 9*. doi: 10.1038/s41598-019-48050-2

Stevens, S. S. (1956). The direct estimation of sensory magnitudes-loudness. *The American Journal of Psychology, 69*(1), 1-25. doi: <http://www.jstor.org/stable/1418112>

Sullivan, R. (1994). *An introduction to Kant’s ethics*. Cambridge, MA: Cambridge University Press.

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition, 119*(3), 454-458. doi: <http://dx.doi.org/10.1016/j.cognition.2011.01.018>

Swets, J. A. (1977). Signal detection theory applied to vigilance. *Vigilance: Theory, operational performance, and physiological correlates* (pp. 705-718). Boston, MA: Springer.

Swets, J. A. (1998). Increasing the Accuracy of Mammogram Interpretation. BBN Technologies Cambridge: MA.

Trémolière, B., & Bonnefon, J. F. (2014). Efficient kill–save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality and Social Psychology Bulletin, 40*(7), 923-930. doi: <https://doi.org/10.1177/0146167214530436>

Trémolière, B., De Neys, W., & Bonnefon, J. F. (2018). Reasoning and moral judgment: A common experimental toolbox. In L. J. Ball & V. A. Thompson (Eds.), *International Handbook of Thinking and Reasoning* (pp. 575-589). Abingdon, UK: Routledge.

Vlaev, I., Seymour, B., Dolan, R. J., & Chater, N. (2009). The price of pain and the value of suffering. *Psychological Science, 20*(3), 309-317. doi: 10.1111/j.1467-9280.2009.02304.x

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32*(7), 1206-1220. doi: 10.3758/bf03196893

Yu, H. B., Siegel, J. Z., Clithero, J. A., & Crockett, M. J. (2021). How peer influence shapes value computation in moral decision-making. *Cognition, 211*. doi: 10.1016/j.cognition.2021.104641

Zaki, J., Wager, T. D., Singer, T., Keysers, C., & Gazzola, V. (2016). The Anatomy of Suffering: Understanding the relationship between nociceptive and empathic pain. *Trends in Cognitive Sciences, 20*(4), 249-259. doi: 10.1016/j.tics.2016.02.003

Zhao, W. J., Diederich, A., Trueblood, J. S., & Bhatia, S. (2019). Automatic biases in intertemporal choice. *Psychonomic Bulletin & Review, 26*(2), 661-668. doi: 10.3758/s13423-019-01579-9

**Open Practices Statement**

All data have been made publicly available on the Github and can be accessed at https://github.com/ccpluncw/ccpl\_data\_responseBias2022.git. This study was not preregistered.

The analysis was conducted in part using R packages written by the first author (first quarter, 2024). Those packages can be installed from the following locations:

<https://github.com/ccpluncw/ccpl_R_chValues.git>

<https://github.com/ccpluncw/ccpl_R_chMorals.git>

<https://github.com/ccpluncw/ccpl_R_RRW.git>

<https://github.com/ccpluncw/ccpl_R_smartGridSearch.git>

Footnotes

1. The standard and footbridge dilemmas are simply examples of a class of dilemmas used by researchers to study moral judgment. That is, researchers have used many variations of the standard and footbridge dilemmas (as we do in the present experiments). [↑](#endnote-ref-1)
2. Prior to testing participants, Greene et al. (2001) provided an independent group of respondents with a set of moral dilemmas so that these could be classified as being either ‘personal’ or ‘impersonal’. Respondents were instructed to use three criteria: they were asked to consider (i) “whether the action in question could ‘reasonably be expected to lead to serious bodily harm’, (ii) “whether this harm would be ‘the result of deflecting an existing onto a different party.’, and (iii) “whether the resulting harm would ‘befall a particular person or a member or members of a particular group of people.’”. As a result, “The moral dilemmas of which the coders said that the action in question (a) could reasonably be expected to lead to serious bodily harm (b) to a particular person or a member or members of a particular group of people (c) where this harm is not the result of deflecting an existing threat onto a different party were assigned to the ‘moral-personal’ condition and the others were assigned to the ‘moral-impersonal’ condition.” (all quoted material Greene et al., 2001, Note 9). [↑](#endnote-ref-2)