Conflict as an Entry Point to Decision-Making

Aliya R. Dewey, Corey Allen, Katherine Boere

Forthcoming in F. De Brigard & W. Sinnott-Armstrong (Ed.),
Neuroscience & Philosophy, Vol. 2 at MIT Press.

10517 words

Decision-making is easy in the absence of conflict. After all, a set of possible choices *lacks conflict* when some choices are better than all others along *all* dimensions. For example, if a person is asked to choose between the option of receiving $100 at a 50% chance and the option of receiving $50 at a 25% chance, then it is trivial that they ought to choose the former option. The reward is better and the chances are better. In decision theory, choices like these are known as *Pareto improvements*. Decision-making is trivial in these cases because the correct decision is simply to select the choices that are Pareto improvements—i.e., the choices that are better along all dimensions.

By comparison, decision-making is difficult in the presence of conflict. After all, a set of possible choices has conflict when each choice is better than another along some dimensions and worse along other dimensions. For example, if a person is asked to choose between the option of receiving $100 at a 25% chance and the option of receiving $50 at a 60% chance, then it's far from obvious which option they should choose. The first option involves better (more) reward but worse (more) risk whereas the second option involves worse (less) reward but better (less) risk. In decision theory, choices like these are known as *Pareto efficient*—all choices that are worse along all dimensions have been ruled out of the choice set. Decision-making is non-trivial in these cases because there are no Pareto improvements to select.

Thus, decision-making under conflict requires *conflict resolution*: the selection of a single choice from a choice set with conflict. Conflict resolution requires some way to compare choices *across* different dimensions. A popular assumption in formal decision theory is that choices are compared across dimensions in a quantitative way—by assigning weights and mathematical relations between dimensions. For example, expected value is the standard way to reconcile reward and risk in decision theory. Reward and risk are weighted equally and multiplied against each other (this is known as a "risk-neutral" strategy). For example, the expected value of the first choice above is $100 × 25% = $25 whereas the expected value of the second choice above is $50 × 60% = $30. If expected value maximisation is the correct standard for decision-making, then the second choice is better "all-things-considered" and the person ought to make the second choice.

But that's a big "if"! So-called "risk-loving" agents may have a much stronger preference for more reward than for higher chances. This preference may be strong enough that they would insist (even on reflection) that the first choice is better. For example, perhaps an agent feels like they have enough money to live comfortably already, and they are only excited by the prospect of winning a relatively large sum at relatively steep odds. It is controversial to take a normative stand on whether such an agent demonstrates good or bad decision-making. These are issues best left for normative theory. We'll see in this chapter that it is surprisingly difficult to maintain neutrality on these controversial normative questions when neuroscientists and psychologists study conflict resolution.

Fortunately, conflict resolution isn't the only essential ingredient of decision-making under conflict. Decision-making also requires *conflict detection*: before conflict resolution can happen, a decision-maker must first recognise that there is conflict in the choice set—i.e., that the choice set is Pareto efficient. On the whole, conflict detection is much less controversial on normative grounds than conflict resolution. After all, good conflict detection simply involves identifying dimensions that are relevant to comparison[1] and then identifying when different choices are better along different dimensions. In this chapter, we'll argue that this difference helps explain (and justify) the growing interest in conflict detection within judgement and decision-making science—initially in the psychology of formal reasoning and then later in the neuroscience of moral decision-making.

In §1, we'll identify some of the key advantages that come with studying a cognitive process that is relatively uncontroversial on normative grounds. In particular, we will evaluate the more interesting prospect that studying conflict detection gives us more traction on more controversial processes like conflict resolution. We will see that there may be room for "triangulating" on conflict resolution by studying the less controversial processes (like conflict detection) that are upstream and downstream of it. In §2, we'll review the formal reasoning literature and show how normative controversies surrounding conflict resolution have driven (and justified) interest in conflict detection, arguing that it manifests the pattern we identified in §1. Then in §3, we'll review the moral decision-making literature and show how normative controversies surrounding conflict resolution have led to preliminary interest in conflict detection. Finally, we will also describe a planned study where we hope to identify the temporal basis of two forms of conflict detection and use those results to triangulate on the more elusive process of conflict resolution.

§1. Normativity in Explanation

At first, it may be surprising to say that philosophical controversies about what counts as *good* decision-making has anything at all to do with scientific investigation into *actual* decision-making. Famously, David Hume (1739) argues that there is an inferential gap between "ought" and "is", and that it is impossible to logically derive statements about what is from statements about what ought to be, and vice versa. Won't this is-ought gap shield the sciences of decision-making from any controversies in the philosophy of decision-making? In this section, we'll show that psychologists and neuroscientists alike generally explain success and failure in different ways, and that explanatory progress thereby depends on what behavioural outcomes we've categorised as successes and failures.

§1.1. Task-Based Explanation

The explanatory value of the distinction between success and failure is most apparent in experimental practice, specifically within the context of *task design*. Unfortunately, the importance of task design is relatively neglected in the philosophy of cognitive science in general and the philosophy of neuroscience in particular. (We'll discuss this point further in

[1] Identifying which dimensions are relevant to comparison can be difficult in real-life scenarios, but it tends to be much easier in carefully designed and controlled experiments (as we'll see later in this chapter).

§1.3.) This may be part of a broader bias among philosophers of cognitive science towards theory and explanation, and away from experimentation and practice. However, task design turns out to be essential to explanation in an experimental context. In fact, we'll see that strategic task design ultimately involves creating explanations prior to experimentation that the results of experimentation can be later scaffolded onto. Strategic task design is itself an explanatory task.

After all, the gold standard for strategic task design is a task that admits of a unique, determinate solution that can only be derived in one general kind of way (subject to certain minimal constraints). This ensures that there is only one non-accidental way to achieve success: by performing the unique derivation of the unique, determinate solution. For example, a popular task in the formal reasoning literature is the bat-and-ball task: "A bat and a ball cost $1.10. The bat costs $1.00 more than the ball. How much is the ball?" This is an elementary algebra problem, which admits of a unique, determinate solution: "the ball is $0.05". Moreover, there are a few ways of deriving this strategy: e.g., the substitution, elimination, and graphing methods are often taught in middle school mathematics courses. However, these different forms of derivation all share the same (homomorphic) structure and, in this way, all fall under a single *kind* of derivation.

Strictly speaking, there are ways of solving the task besides deriving the solution. For example, we could solve it through trial-and-error: plug arbitrary values into the equations (i.e., bat + ball = $1.10; bat = $1.00 + ball) and check for equality.[2] This strategy will end in success when we plug in the values 'bat = $1.05' and 'ball = $0.05' and find that equality holds in each equation. However, this strategy is generally very time intensive, so we can rule it out by appealing to temporal constraints. Alternatively, we could simply get lucky and guess the values 'bat = $1.05' and 'ball = $0.05' and find that equality holds in each equation. However, success is extremely improbable in this outcome, so we can rule it out using statistical analysis to demonstrate that success rates are *significantly* above chance.

If a participant (or group of participants) produces the correct solution to a task like the bat-and-ball task at a rate significantly above chance in a reasonable amount of time, we can *very confidently* infer that the participants must have derived the solution via a procedure that falls under the same general kind of way. This explanation of the successful behaviour is almost completely derivative from the *task analysis*—i.e., from an explanation of what it takes to solve the task. The task analysis itself is *a priori* (at least in the scientist's sense, if not the philosopher's: that it is known prior to running the experiment). Thus, we get an explanation of behaviour that is also *a priori*. All that the experimental results add is what percentage of responses are correct above chance and hence, how much of behaviour conforms to our *a priori*, task-derived explanation.

Nevertheless, this *a priori*, task-derived explanation is quite austere. It doesn't tell us which circuits or regions are responsible for performing each of the steps. It doesn't tell us how the brain decided what the task called for. It doesn't tell us how many false starts the brain might have taken before it found the correct response. It only tells us that the brain eventually did perform all the steps required to derive the unique, determinate solution. Therefore, a mechanistic, neurobiological explanation of the successful behaviour will have to go beyond what can be fully derived from the task analysis, of course. Nevertheless, the task analysis gives us an important head-start: it implies a task-based explanation that neurobiological

[2] We thank Sam McGrath for this example.

details can be scaffolded on top of. This is much better than having to start a mechanistic explanation from scratch. We'll see an example of this scaffolded mechanistic explanation for successful conflict detection in §2.4.

By comparison, the explanatory strategy for failure is quite different. If a participant (or group of participants) produces an incorrect solution to a task like the bat-and-ball task, we can only infer that the participants must *not* have derived the solution via a procedure that falls under the same general kind of way—or, perhaps, they did not report that solution after they successful derived it. But this doesn't give us much of a road-map for explaining failure. After all, there are arbitrarily many other things that participants could have done otherwise. This is the Anna Karenina principle: "All happy families are alike; each unhappy family is unhappy in its own way" (Tolstoy, 1878). In our case, all correct solutions are alike; each incorrect solution is incorrect in its own way. By itself, then, a task analysis doesn't give us an explanation for failure.

Fortunately, we can solve this problem with further task design. After all, failure is most difficult to explain when it is *at ceiling* (at 100%). When participants respond incorrectly in nearly all cases, we have no reason to believe that they are attempting any part of the appropriate kind of derivation for the unique, determinate task solution. We have no reason to believe that they've even understood the task. Total failure is unintelligible, for all intents and purposes. However, when participants respond incorrectly at a rate between floor (0%) and ceiling, we have much more reason to believe that they have understood the task and are attempting the appropriate kind of derivation for the unique, determinate task solution (Stanovich & West, 2000; Dewey, 2022). Thus, task difficulty has to be *well-calibrated*: it should elicit success and error rates between floor and ceiling (i.e., significantly above 0% and significantly under 100%).

Moreover, we should look for cases where error rates are significantly correlated with variable difficulty conditions (e.g., time pressure, cognitive load). In these cases, the best explanation seems to be that there is a process that aims to derive the correct solution but it is subject to some kind of interference from processes elicited by the difficulty conditions. Note that these explanations won't be fully derivative from the task analysis. After all, they both appeal to significant correlations that are found *after* experimental design during data analysis. Thus, error introduces an empirical element into *a priori*, task-based explanation. In this way, error is more interesting (and empirical) than success. Nevertheless, we still need some amount of success to scaffold empirical information onto. Thus, it is in this goldilocks zone between total failure and total success that we can reliably scaffold empirical information onto *a priori*, task-derived explanations to explain both success and failure.

Overall, then, the distinction between success and failure is essential to the task-based approach to explanation. It tells us which behaviours are correct and hence, amenable to *a priori*, task-derived explanations (if the tasks are designed to admit of unique, determinate solutions that are derivable by a single kind of procedure). Likewise, it tells us which behaviours are incorrect and hence, amenable to empirical explanations that appeal to interference from processes elicited by difficulty conditions (if the task difficulty is calibrated to ensure that success and failure are both significantly different from 0% and 100%).

§1.2. Triangulating Extension

However, the task-based approach to explanation has a serious caveat. Many tasks don't admit of unique, determinate solutions that can be solved in only one kind of way. Otherwise, strategic task design wouldn't be nearly so difficult as experimental researchers know it is. The bat-and-ball task is a rare example of a task that does satisfy these criteria. Accordingly, it elicits a kind of cognition that essentially has to do with deriving the unique, determinate solutions from tasks in conjunction with the known axioms of elementary algebra. In general, tasks that satisfy these criteria tend to ask subjects to follow difficult instructions and thereby test whether subjects succeed or fail at overcoming the difficulties of following the instructions.

By comparison, consider tasks that call for decision-making under conflict (i.e., selection from a Pareto-efficient set of choices). Such tasks elicit a kind of cognition that essentially has to do with achieving conflict resolution in the absence of an exogenous standard specified in the task instructions. To do this, cognition must endogenously generate its own standards—or, perhaps, endogenously retrieve standards from memory. As a result, resolving the task indeterminacy would simply elicit a different kind of cognition—one that doesn't involve having to endogenously generate (or retrieve) standards. For example, we could give a decision-making task a determinate solution by asking participants to select the choice that maximises expected value, but then the task would elicit a completely different kind of cognition: reasoning about maximising expected value, not decision-making.

This is a serious caveat for the task-based approach to explanation: we cannot derive the full skeleton of a cognitive explanation from the task analysis. In particular, we can't conclude anything about the internal process of conflict resolution from an analysis of the task and from observing whether participants chose the option to receive \$50 at 60% odds or \$100 at 25% odds. After all, nothing internal to the task or task instructions specifies what must happen for *successful* conflict resolution to take place. It would be controversial to say that successful conflict resolution involves selecting the choice that maximises expected value. The only uncontroversial thing to say is that successful conflict resolution involves selecting *some* choice and thereby resolving conflict. And there are many ways to do that: participants could appeal to any number of arbitrary considerations to break the tie between the two options. Recall, for example, our risk-loving participant who selects \$100 at 25% odds because they aren't in need of money but receiving \$100 would feel thrilling to them.

However, it would be premature to conclude that the task analysis has *no* implications at all for cognitive explanation on open-ended tasks like decision-making tasks under conflict. After all, suppose we found evidence for conflict detection during decision-making under conflict. This evidence would have to show that the presence of conflict in a choice set (vs. the absence of conflict) makes a significant difference to a participant's decision-making. This difference could manifest in behavioural differences: e.g., increased response times in response to choice sets with vs. without conflict. Or it could manifest in neural differences: e.g., significant differences in neural activity at intermediate stages of processing in response to choice sets with vs. without conflict. More on this in the next two sections.

A task analysis can tell us what's required for the successful detection of *conflict*. In our example from the introduction, it would require participants to (a) identify the reward and risk values for each choice, (b) compare each choice along both dimensions, and then (c) register that no choice is better along both dimensions. Thus, if we find that some intermediate process responds in significantly different ways depending on the presence or absence of conflict in the choice set, we could very reliably conclude that all processing up to that point had achieved (a)–(c). Therefore, we can derive *part* of a skeleton for the cognitive explanation of successful decision-making from the task analysis.

Likewise, a task analysis can tell us *part* of what's required for the successful resolution of conflict. Recall from above that it is uncontroversial that successful conflict resolution involves selecting *some* choice and thereby resolving conflict. In our example from the introduction, it would be insufficient for participants to make a choice without indicating their choice to the experimenter in the appropriate way (e.g., by pressing the appropriate button). Therefore, a task analysis would imply that successful conflict resolution would, in part, require participants to translate their choice into the task-appropriate format for conveying that choice.

Overall, then, we get the skeleton of a cognitive explanation with a *black box* for conflict resolution—which is the only part of successful decision-making that can't be derived from the task analysis. This may still be immensely useful for experimental inquiry. After all, it will allow us to scaffold empirical information onto our skeletal explanation upstream and downstream of the black box. Moreover, we can use this information to *triangulate* on the internal structure of the black box itself—without direct guidance from the task analysis. For example, we could identify the neural inputs to and neural outputs from the black box. This would give us a neural specification of what the black box achieves over and above the ambiguous description of "conflict resolution".

It would also tell us what neural states potentially realise the black box and hence, indicate which states we ought to measure to "fill in" the black box. For example, suppose we find that conflict is detected in the dorsal anterior cingulate cortex (dACC) and is resolved by the time the premotor cortex joins decision-making. That would tell us that conflict resolution happens between the output of dACC and the input of premotor cortex, thereby helping us triangulate on the locus of conflict resolution. Moreover, this neural input-output characterisation of the black-box would help to make sense of any measurements we do take and how the measured states might contribute to the input-output transformations. For example, the dACC might encode information in a particular format and the premotor cortex might encode information in a different format, and then we could reliably infer that conflict resolution must also involve changing the format of the representation.

The general implication here is that even if tasks don't admit of unique, determinate solutions that can be derived in one kind of way, it may be possible for this indeterminacy to be localised to one step of the task response (e.g., conflict resolution). What counts as successful for this one step won't be derivable from the task analysis. Nevertheless, the rest of the task response may still be determined by the structure of the task. As a result, the task-based approach to explanation may still be possible for most of the task response. Moreover, the task-based approach to explanation may even help us *triangulate* on the internal structure of cognitive performance for the step or steps that aren't derivable from the task analysis. For this reason, we'll refer to this strategy as a "triangulating extension" for the task-based approach to explanation. There may be many such extensions to the basic task-based approach to explanation, but just the triangulating extension is relevant for our purposes.

In conclusion, the task-based approach to explanation (and its triangulating extension) that we just described differs in several ways from popular accounts of cognitive explanation in the philosophical literature. One important difference is that popular accounts of cognitive explanation in the philosophical literature—e.g., functional analysis *a la* Cummins (1983) and mechanistic explanation *a la* Craver (2007) and Bechtel (2008)—tend to abstract away from the explanatory role served by strategic task design. This same explanatory role is also served by the normative distinction between success and failure, which leads popular accounts of cognitive explanation to abstract away from success and failure. We leave it to other work to compare the task-based approach to explanation with other accounts of cognitive explanation.

§2. Conflict in Formal Reasoning

In the past two decades, there has been a surge of interest in "cognitive conflict" across the psychology and neuroscience of judgement and decision-making. On an objectivist view, *cognitive* conflict is just the cognitive *representation* of conflict (i.e., Pareto efficiency) between objective payoffs in the choice set ($100 at 25% and $50 at 65% in our intro example). In general, neuroscientists and psychologists don't specify constraints on what kind (or format) of representation cognitive conflict must involve. This view makes the experimentalist's job easy. They just have to create a task with objective payoffs that are conflicting in one condition and not in the other condition. Then observations of *any* significant differences in cognitive measures between the conflict and no-conflict conditions are taken to be reliably diagnostic of the presence of cognitive conflict.

Matters are more difficult on a subjectivist view. Here cognitive conflict is the cognitive representation of conflict (i.e., Pareto efficiency) between *subjective* payoffs in the choice set. Subjective payoffs may diverge significantly from objective payoffs. This makes the experimentalist's job harder. After all, they can create a task with objective payoffs that are conflicting in one condition and not in the other condition, yet the participant may value the outcomes differently, such that the choice sets are conflicting in both conditions or non-conflicting in both conditions. This is quite rare, though: for $100 at 25% and $50 at 65% to lack conflict in subjective payoffs, participants would have to either prefer the riskier option ceteris paribus or the lower money option ceteris paribus, which is implausible. Most of the time, then, experimenters simply proceed under the assumption that conflict between objective payoffs implies conflict between subjective payoffs.[3]

In this chapter, we'll focus on the rise of cognitive conflict in the formal reasoning (this section) and the moral decision-making (next section) literatures. In both cases, we'll propose an explanation for this growing interest in cognitive conflict. This interest initially starts with implausibly strong assumptions about what counts as successful task performance. These assumptions allow for the ordinary task-based approach to explanation. Normative backlash to these assumptions leads to temporary stalling in explanatory progress. Then it is discovered that conflict detection can resume explanatory progress by providing a way to triangulate on the difficult steps of judgement and decision-making. This application of the triangulating extension to conflict detection leads to a surge in conflict detection. Growing interest in conflict detection in one part of the judgement and decision-making literature then inspires interest in other parts of the literature, and the process continues.

§2.1. Heuristics & Biases

[3] If experimenters suspect that subjective and objective payoffs may be significantly misaligned, they have two options: (a) redesign the experimenter to ensure more alignment or (b) perform an experimental analysis of revealed preferences to ensure whether the conditions are different vis-a-vis conflict.

The formal reasoning literature rose to prominence in the 1960s and 1970s with a number of key studies by Wason, Kahneman, & Tversky that demonstrated that human participants were surprisingly incompetent at formal reasoning tasks.

One famous example is Wason's (1968) selection task, which asks subjects to identify which cards could falsify a conditional statement like "If there is an A on one side of the card, there is an even number on the other side of the card". Participants are generally competent at testing this statement consistent with *modus ponens*: they identify that cards with an A on their visible side should be flipped because the statement will be falsified if they have an odd number on their hidden side. By comparison, participants are surprisingly incompetent at testing this statement consistent with *modus tollens*: they almost always fail to identify that cards with an odd number on their visible side should be flipped because the statement will be falsified if they have an A on their hidden side. In fact, subjects only managed to flip odd-numbered cards at chance levels (around 4% according to Johnson-Laird & Wason, 1970).[4]

Another famous example is Tversky & Kahneman's (1983) Linda problem, which starts by attributing descriptions to Linda that are commonly seen as representative of feminists: e.g., Linda is single, outspoken, bright, majored in philosophy, concerned with issues of discrimination, participated in anti-nuclear demonstrations. Then participants are asked whether it is more likely that (a) Linda is a bank teller or (b) Linda is a bank teller and a feminist. Since a conjunction is only true if both conjuncts are true, the conjunction of two statements is less likely than each of the conjuncts. For this reason, it is more likely that Linda is a bank teller than that she is both. Nevertheless, most participants indicate that it is more likely that Linda is a bank teller and a feminist. Tversky & Kahneman describe this as the *conjunction fallacy*.

A final famous example is Kahneman & Frederick's (2005) bat-and-ball task, which we already discussed in §1. The bat-and-ball task states that a bat and a ball together cost $1.10 and that the bat is $1 more than the ball. Then it asks: how much is the ball? This task is easily solved with elementary algebra that most participants have learned in middle school: bat + ball = $1.10 and bat = $1.00 + ball, so substituting gives ($1.00 + ball) + ball = $1.10, simplifying gives 2 × ball = $0.10, and further simplifying gives ball = $0.05. Nevertheless, most participants indicate that the ball is $0.10—as if the task had stated that the bat is $1.00 *tout court* rather than $1.00 *more than the ball*.

Results like these create an interesting puzzle. They wouldn't be surprising if *modus tollens*, conjunction, and elementary algebra were complicated rules that participants couldn't understand, much less apply. But that patently isn't the case: these are all simple rules that participants easily apply on reflection. In the case of elementary algebra, the rules are relatively more complicated but participants have often received years of education that ensure they can competently follow the rules. This creates a puzzle: if participants are capable of following these rules, why don't they? What is the source of this *bias* against following the relevant rules, as specified by the task?

A popular answer to this question suggested that *heuristics* were the source of this bias. For the Wason selection task, a simple example is Evans & Lynch's (1971) suggestion that participants may be using a *matching heuristic* to solve the Wason selection task: they flip

[4] For this reason, the WST is generally regarded as having poor difficulty calibration.

cards with visible faces that "match" the faces mentioned in the conditional statement (the A-card and cards with even numbers on their faces).[5] For the famous Linda problem, Tversky & Kahneman (1983) suggested that participants may be using a *representativeness heuristic*: Linda's description would be more representative of someone who is both a bank teller and a feminist than it would be of someone who was just a bank teller. For the bat-and-ball task, Kahneman & Frederick (2005) suggested that participants may be using a *substitution heuristic*: they substitute the more difficult task with an easier task that requires fewer steps for simplification.

§2.2. Dual-Process Theory 1.0

The heuristic explanation for observable bias poses its own puzzle, though. Humans don't rely on heuristics for solving all tasks. Participants may rely on matching heuristics in Wason's selection task, but they are competent at using *modus tollens* (rather than a matching heuristic) to recognise that an enforcer of the rule "If drinking beer, then over 19" should check whether those under 19 (not over 19) are drinking beer (Cox & Griggs, 1982). Likewise, participants may rely on the substitution heuristic to solve the bat-and-ball task, but they are competent at solving the structurally-identical rubber-and-pencil task: "A rubber and pencil together cost 37 cents. The rubber costs 13 cents more than the pencil. How much does the pencil cost?" (Bourgeois-Gironde & Vanderhenst, 2009).

The puzzle, then, is why do humans rely on heuristics for some tasks but not other tasks, which may be extremely similar or even structurally identical? One influential response to this puzzle was *dual-process theory* (DPT). All forms of DPT agreed that human judgement and decision-making was performed by two kinds of processes: one that relied on heuristics (often described in neutral terms as "Type-1 processing") and one that didn't (often described in neutral terms as "Type-2 processing"). The simplest form of DPT (known as the "parallel-competitive model") proposed that these two processes competed in parallel, that the use of a heuristic indicated that Type-1 processing probably won the competition, and that the correct response indicated that Type-2 processing probably won the competition.

A popular objection to the parallel-competitive model is that the use of both Type-1 and Type-2 processing is less efficient and less reliable than the use of just Type-2 processing. Another form of DPT (known as the "default-interventionist model") aims to address this objection. It proposes that Type-1 processing generates the initial (or default) response and Type-2 processing only intervenes whenever necessary. The idea is that Type-1 processing is more efficient and Type-2 processing is more reliable, so cognition can strike an optimal compromise between the two by using Type-1 processing when it is reliable and resorting to Type-2 processing whenever Type-1 processing is unreliable. Nowadays, this kind of argument is known as *resource-rational analysis* (e.g., Lieder & Griffiths, 2020).

A popular objection to the default-interventionist model is that it is impossible for Type-1 processing to know when it is reliable and when it is unreliable unless Type-2 processing is responding in parallel and Type-1 processing can "check its answers" against Type-2 processing. Of course, that would take us back to the parallel-competitive model and the worry that Type-1 processing is redundant. Alternatively, we might say that Type-1 processing has the capacity to know when it is unreliable without consulting Type-2 processing, but then we have to explain how Type-1 processing can possess that capacity

[5] For a much more influential account, see Oaksford & Chater (2007).

without possessing the capacity to directly solve the task without using the heuristic. Between these two objections, DPT was mired in serious theoretical problems.

Ultimately, some began to question whether the right way to explain the presence or absence of heuristics in behaviour was to point to the presence or absence of heuristic processing (in the default-interventionist model), or the presence or absence of causal efficacy in heuristic processing (in the parallel-competitive model). Critics argued that we don't really explain a behavioural difference by positing a corresponding cognitive difference that is causally upstream of the behavioural difference (Gigerenzer, 2020). After all, we might insist that genuine cognitive explanation should be something like functional analysis or mechanistic explanation—it should point to interactions between organised networks of simpler processes or manifestations of simpler sub-capacities that coordinate to give rise to different behaviours.

§2.3. Great Rationality Debate

Before long, though, critics argued that there were deeper methodological problems with the studies that led to DPT. One early issue was raised by the philosopher L. Jonathan Cohen (1979, 1981). He notes that there is usually an interpretive gap between the task and the norm relevant for evaluating performance on the task, and that experimenters often neglect this gap when they make tacit assumptions about which norm is relevant for evaluating performance on the task. For example, the results we mentioned above involve evaluating performance on the Linda problem, Wason selection task, and bat-and-ball task using classical probability theory, logic, and algebra, respectively. But there is an interpretive gap here: the tasks don't contain instructions telling subjects that they have to follow classical norms and that they will be evaluated accordingly.

As a result, subjects could just as well decide to use non-classical forms of probability theory, logic, and algebra. Then it would be *uncharitable* to evaluate subjects by classical norms rather than non-classical norms. In fact, Cohen points out that participant performance often turns out to be correct when we do evaluate it using non-classical forms of probability theory, logic, and algebra. Charitable interpretation requires us to recognise this interpretive gap and respect participants enough to trust that they may identify alternative interpretations to the task than the interpretations that the experimenters intended. Cohen notes that the failure to do this may be indicative of disrespect and even elitism on the part of experimenters. After all, we have no principled basis for privileging classical norms over non-classical norms when evaluating participant performance.

Cohen (1981) ultimately takes a moderate position: we may be justified in attributing error to participant performance, but only once we have sufficient evidence that participants possess the relevant competence and that participants aim to exercise that competence on the task. (Recall that we made a similar argument in §1.1 when we argued that task difficulty must be well-calibrated.) But other critics have since taken a more radical version of this position: they argue that charitable interpretation requires us to explain behaviour by *rationalising* it. These critics aim to "reverse engineer" the relevant normative standard for evaluating behaviour under the assumption that participants are nearly optimally rational. Then cognition is inferred to use this normative procedure to select the appropriate behaviour. This approach is known as *rational analysis* (Anderson, 1990; Oaksford & Chater, 2007).

Rational analysis is broadly consistent with influential work by Gerd Gigerenzer and colleagues. They argued that so-called "heuristics" generally outperform reasoning using classical norms of probability and decision theory when they are used by bounded agents in natural environments with imperfect, incomplete information. For example, Gigerenzer & Goldstein (1996) argue that when deciding between two options, it is optimal for bounded agents to follow the *take-the-best rule*: when deciding between two options, choose the option that is favoured by the first discriminating cue. When the US American students are asked whether, e.g., Erlangen or Leipzig is larger, recognisability is generally the most salient cue: e.g., Leipzig is more recognisable than Erlangen. But recognisability is roughly correlated with size: larger cities tend to be more recognisable. Thus, Gigerenzer & Goldstein found that students were most effective in deciding which German cities were larger when they were instructed to select the option that is most recognisable: Leipzig.

The justification here isn't that the take-the-best rule is more efficient than an exact strategy. That is the sort of thing that Kahneman, Tversky, and others would say. The justification here is that the take-the-best rule and other so-called heuristics are *Pareto improvements* over classical norms of reasoning: they are both more efficient and more effective. For example, Gigerenzer & Goldstein (1996) showed that the take-the-best rule significantly outperformed several kinds of inference that took into account strictly more relevant information. They explained this counterintuitive result by pointing to the fact that more complicated reasoning has more ways of going wrong than simpler reasoning.[6] The upshot was that humans should use simpler rules than classical norms of reasoning and hence, they should be evaluated by the standards of these simpler rules rather than by classical norms of reasoning.

Suppose Cohen, Gigerenzer, and others are right that participant performance has been interpreted uncharitably and that charitable interpretation will show that participant performance on the Wason selection task, Linda problem, and bat-and-ball task are all correct. This makes a *significant* difference to cognitive explanation. For one, there is no puzzle that needs solving by appealing to heuristics. After all, recall that the puzzle is: if participants are capable of following these rules, why don't they? Now we get a different answer: participants simply interpret the task differently than the experimenters do and their interpretation calls for the application of different, non-classical rules. There's no need to appeal to heuristics.

For another, there is no puzzle vis-a-vis heuristics that needs solving by appealing to DPT. After all, recall that this puzzle is: why do humans rely on heuristics for some tasks but not other tasks? Again, we get a different answer now: there are no heuristics per se, so participants simply interpret different tasks as calling for different rules that are both simple and highly effective for bounded agents in natural environments with incomplete, imperfect information. Without a distinction between heuristics and non-heuristics, there is no reason to posit a distinction between heuristic and non-heuristic reasoning. For this reason, critics of classical norms generally advocate for single-process theories of judgement and decision-

[6] For example, consider the *linear regression rule*, which integrates all considerations available to the agent that might be indicative of size, but is susceptible to skewing from irrelevant information. For example, a US American who uses this rule might be led to conclude that Erlangen is larger than Leipzig by the following train of thought that is almost entirely correct: (a) guessing that the name 'Leipzig" sounds more "Slavic" than 'Erlangen' (correct), (b) guessing that Leipzig is in East Germany and Erlangen is in West Germany (correct), (c) knowing that West Germany is more populous than East Germany (correct), and (d) deciding that this counts in favour of Erlangen being larger than Leipzig (incorrect).

making (e.g., Kruglanski & Gigerenzer, 2011). This explanatory difference ultimately boils down to normative disagreements about what counts as success and error.

§2.4. Dual-Process Theory 2.0

These normative disagreements about what count as success and error in judgement and decision-making became intense and protracted. Conversational breakdown ensued, with both sides accusing each other of missing the point. This has created deep, lasting rifts between research programs: e.g., Gigerenzer and colleagues nowadays rarely interact with Kahneman and colleagues. Some proposed dispensing with normative language altogether (Elqayam & Evans, 2011)—a proposal that seems like a non-sequitur to others (Dewey, 2022). As such, it acquired colourful monikers like “the rationality wars” (Samuels et al., 2002) and “the Great Rationality Debate” (Stanovich, 2011). These controversies have made the processes of conflict resolution difficult to explain.

Nevertheless, clever progress has been made to break this gridlock. One important movement (not the only one, by any means) has been growing interest in *conflict detection*.[7] This movement started with a key neuroimaging study by De Neys et al. (2008). They reasoned that classic tasks from the heuristics-and-biases program might present a kind of *conflict*. For example, the bat-and-ball task affords two responses: (a) solving the task as-is or (b) substituting the task with a simpler one and then solving the simpler task. Classical norms of reasoning favour the first response, while non-classical norms of reasoning generally favour the second response. Unfortunately, De Neys and colleagues do evaluate the responses using classical norms—consistent with the older literature by Tversky, Kahneman, and colleagues that they are working in. Fortunately, though, this evaluation makes no difference to their analysis: they are interested in conflict and its detection, not its resolution.

Instead, De Neys et al. (2008) designed counterpart tasks that only afford one response. For the bat-and ball task, its “no-conflict” counterpart might have been something like this: “A bat and a ball cost $1.10. The bat costs $1.00 ~~more than the ball~~. How much is the ball?” This task is almost identical to the bat-and-ball task, except that it’s modified such that the intuitive response “$0.10” is the correct response. As a result, classical and non-classical norms of reasoning favour the same response: solving the task as-is. By creating conflict and no-conflict versions of the same task, De Neys et al. can use subtractive comparisons for conflict and no-conflict versions of the task to identify neural activity that is selective to the presence of conflict. In this way, they redirect our explanatory attention away from the process of conflict resolution and toward the process of conflict detection.

What they found is that conflict tasks elicit significantly more blood-oxygen-level-dependent (BOLD) activity in the bilateral dACC and right dorsolateral prefrontal cortex (dlPFC) than no-conflict tasks, suggesting the involvement of these regions in conflict tasks. The dACC is associated with conflict detection in other tasks (e.g., Chung et al., 2024) and the dlPFC is associated with response inhibition in other tasks too (e.g., Aziz-Safaie et al., 2024). De Neys et al. argue that there is only one kind of non-accidental way for the system to exhibit this sensitivity to conflict: processing upstream of the dACC and dlPFC must have (a) identified the classically-correct response and the non-classically-correct response, (b) evaluated each

[7] Skovgaard-Olsen et al. (2019) propose another strategy, which looks for consistency between the norm that a participant aims to follow when performing a task and the norm that a participant uses to evaluate others’ performance on the same task. For an evaluation of this strategy, see Dewey (2022).

response against both standards, and then (c) registered that no response was correct by both standards. Recall that this is just the *triangulating extension* to the task-based approach to explanation that we discussed in §1.2.

This study by De Neys et al. (2008) led to a surge of interest in conflict detection using conflict vs. no-conflict task comparisons and subtractive analysis. In particular, the finding that the dACC is an area of interest for conflict detection has been replicated several times: e.g., by Simon et al. (2015) for numerosity judgments in preschool children, Vartanian et al. (2018) for base rate judgments, and Mevel et al. (2019) for ratio comparison judgments. One exception is Andersson et al. (2020) for likelihood judgments about conjunctions (e.g., the Linda problem). Bago et al. (2018) report that electroencephalogram (EEG) shows conflict in base rate tasks is detected by early processing in medial parietal areas after 200 milliseconds (N200) and frontal areas after 300 milliseconds (P300). Precise localisation is difficult, of course, but these findings are consistent with early conflict detection in posterior parietal cortex (as in Mevel et al., 2019) and later conflict detection in anterior cingulate cortex.

Ultimately, this led to several prominent researchers in the formal reasoning literature to propose a revamped conception of dual-process theory, which they described as "Dual-Process Theory 2.0" (De Neys, 2017). Whereas DPT 1.0 ultimately distinguishes between default reasoning that uses heuristics and intervening reasoning that doesn't use heuristics, DPT 2.0 distinguishes between default reasoning that makes an initial attempt to resolve conflict that is liable to fail and intervening reasoning that draws on further cognitive resources to resolve more difficult cases of conflict. In later work, De Neys (2021) cedes that it may not even be helpful to draw a type-distinction between default and intervening reasoning of these kinds. This is further suggestive that De Neys and colleagues have shifted interest from (controversial) conflict resolution to (less controversial) conflict detection.

We should add that De Neys and colleagues continue to use classical norms to evaluate judgments and decisions. This is unfortunate and, we think, unnecessary. For example, De Neys et al. (2008) also reported that dlPFC was significantly more active when "correct" vs. "incorrect" responses to conflict tasks were selected. From this, they concluded that dlPFC may be responsible for selectively inhibiting incorrect responses. However, this speculation about conflict resolution strikes as problematic for all the same reasons as before. After all, the "incorrect" responses may very well be correct if we evaluate reasoning using non-classical norms. If that is the case, then a different explanation may be better: the dlPFC may be disposed to reasoning with classical vs. non-classical norms and greater BOLD activity may be correlated with stronger manifestations of that disposition. This re-opens the thorny debates of §2.3. Thus, we urge that these can be avoided by maintaining focus on the normatively-straightforward issue of conflict detection.

§3. Conflict in Moral Decision-Making

In §1, we argued that the task-based approach to explanation has serious caveats for explaining decision-making and that these caveats can be addressed by focusing explanation on an important part of decision-making: conflict detection. In §2, we traced this pattern through the formal reasoning literature—showing that the caveats of the task-based approach to explaining decision-making generated the Great Rationality Debate and then showing that conflict detection provided a neutral way to make explanatory progress on decision-making. In this section, we'll trace this pattern through the moral decision-making literature. We'll see that this pattern is a bit less obvious for moral decision-making, but this creates important opportunities for neuroscientists and philosophers alike to make explanatory progress on moral decision-making while maintaining neutrality on controversial ethical issues.

§3.1. Heuristics & Biases

While 20th century moral psychology was framed by the developmental paradigms of Piaget, Kohlberg, Turiel, and colleagues, 21st century moral psychology is much more influenced by the judgment and decision-making paradigms of Greene, Haidt, Cushman, and colleagues. Drawing inspiration from the dual-process theories of Kahneman, Tversky, and others, this work purports to show that deontological decisions are associated with emotional heuristic reasoning and consequentialist decisions are associated with controlled reflective reasoning.

One famous example is Greene et al.'s (2001) trolley dilemma tasks, which ask subjects to evaluate whether it is morally appropriate to sacrifice one life to save five lives when the act of sacrifice involved (a) pulling a lever from a distance versus (b) pushing a person in front of a trolley. Greene et al. showed that the emotionally salient prospect of pushing a person in front of a trolley elicited significantly stronger BOLD response in brain regions associated with emotional processing (vs. the former condition and the control condition). They also showed that the less emotionally salient prospect of pulling a lever and making the sacrifice from a distance elicited significantly stronger BOLD response in brain regions associated with working memory than the latter condition (but not compared to the control condition).

A result like this creates an interesting puzzle. It wouldn't be surprising if participants would reject general consequentialist principles like (a) it is appropriate to sacrifice the few to save the many, (b) our proximity to a sacrifice makes no difference to its moral status, and (c) actions without harmful consequences are morally permissible. But that isn't the case: participants do generally seem to accept principles like these. In fact, many consequentialists have noted this: if we start moral reasoning from general principles like these, we are led to consequentialist conclusions (e.g., Singer, 2005). Non-consequentialists generally agree: they accept that consequentialist principles are highly intuitive at the level of general principles but insist that they are highly counterintuitive at the level of specific cases. Thus, non-consequentialist philosophers generally argue that we can avoid consequentialist conclusions by doing moral reasoning at multiple levels of generality—i.e., by finding a *reflective equilibrium* between the level of general principles and the level of moral verdicts about specific cases (e.g., Rawls, 2005).

This creates a puzzle: if participants are capable of following general consequentialist principles, why don't they in certain cases, such as sacrificing one life to save five when the sacrifice involves pushing a person to their death in front of a trolley or when the harmless actions involve incest? If we side with consequentialism, we may prefer a more loaded question: what is the source of this *bias* against following general consequentialist principles? Notice that this is the *same kind* of puzzle as the puzzle that Kahneman, Tversky, Wason, and others found when they were studying participant performance on formal reasoning tasks: if participants are capable of following the general principles of elementary algebra, why don't they in certain cases, such as when a bat and a ball cost $1.10 and the bat costs $1 more than the ball?

Likewise, a popular solution to this puzzle suggested that *heuristics* were the source of this bias. Baron (1996), Haidt (2001), and Greene (2007, 2014) all suggested that empirical evidence indicated that *emotional heuristics* were the source of this non-consequentialist bias. Greene (2007) suggests that participants may be using emotional aversion to using personal force against another person as a heuristic for the fact that personal force generally causes unnecessary harm and hence, is likely to be morally inappropriate—even though that isn't true for the particular case of the trolley problem.

§3.2. Dual-Process Theory 1.0

However, recall that the heuristic explanation for observable bias poses its own puzzle: humans don't rely on heuristics for solving all tasks, so why do we rely on heuristics for some tasks but not other tasks? While there has been significant experimental interest in debiasing measures in the formal reasoning literature, there hasn't been so much interest in "debiasing measures" in the moral decision-making literature. We'll suggest in §3.3 that this is because consequentialist norms are much more controversial in the moral decision-making literature than classical norms initially were in the formal reasoning literature. Nevertheless, it should be obvious on reflection from everyday life that emotional, deontological responses are liable to be more or less involved in different cases of moral decision-making, and that this calls for cognitive explanation.

Recall how the parallel-competitive model generally solves the heuristic puzzle: participants go with the heuristic response when Type-1 processing outcompetes Type-2 processing and they go with the analytic response when Type-2 processing outcompetes Type-1 processing. Applied to moral decision-making, the explanation is more specific: participants give the deontological response when emotional heuristics outcompete the reflective processes that apply general consequentialist principles to particular cases and participants give the consequential response otherwise. Haidt and Greene both propose that this puzzle can be solved with DPT, citing inspiration from the formal reasoning literature and related literatures (see Haidt, 2001; Greene, 2007, 2014). Unlike the formal reasoning literature, though, the parallel-competitive model has proven to be much more popular than the default-interventionist model (see Greene, 2007).[8]

Dual-process theories are said to be supported by evidence that increased consequentialist decision-making is associated with impaired emotional processing and improved analytic reasoning: increased BOLD activity in the frontoparietal control network (Greene et al., 2001, 2004), positive affect from viewing comedy clips (Valdesolo & DeSteno, 2006), frontotemporal and prefrontal cortex lesion (Mendez et al., 2005; Koenigs et al., 2007; Moll & de Oliveira-Souza, 2007; Tassy et al., 2012; Rowley et al., 2018), increased cognitive control, working memory, and reasoning capacities (Greene et al., 2008; Moore et al., 2008; Suter & Hertwig, 2011; Cushman et al., 2012; Paxton et al., 2012; Baron et al., 2015; Patil et al., 2021; c.f., Royzman et al., 2015), increased reward sensitivity (Moore et al., 2011), psychopathy and other negative personality traits (Glenn et al., 2010; Bartels & Pizarro, 2011), subclinical depression (Yin et al., 2022), less visual imagery (Amit & Greene, 2012), less mortality thought (Trémolière et al., 2012), more class privilege (Côté et al., 2013), less self-awareness (Reynolds et al., 2019), less empathic concern (Crockett et al., 2010; Conway

[8] Haidt (2001) himself endorsed a distinctive variant of the default-interventionist model: that Type-1 processing responds by default using efficient, emotional heuristics that are biassed against consequentialism and Type-2 processing occasionally intervenes using appeals to general principles. What's distinctive about Haidt's variant of the model is that he thought Type-2 processing usually creates a post hoc justification for the heuristic decision—instead of making a decision that *overrides* the heuristic distinction.

& Gawronski, 2013; Gleichgerrcht & Young; 2013; Royzman et al., 2015; Zhang et al., 2020), and increased consideration of alternative actions (Mata, 2019).

§3.3. Great Morality Debate

Emboldened by these empirical successes, a number of prominent moral psychologists almost immediately endeavoured the ambitious project of drawing philosophical implications from their preliminary experimental results (e.g., Gazzaniga, 2005; Hauser, 2006; Greene, 2007, 2014; Gigerenzer, 2008, 2010; Haidt, 2012; Redish, 2022). This elicited swift backlash from moral philosophers. Some of this criticism targeted the project of drawing normative implications from empirical results—with varying degrees of pessimism and optimism. However, some of this criticism also concerned the ways in which normative assumptions had contaminated the interpretations of the empirical results themselves (e.g., Allman & Woodward, 2008; Berker, 2009; Dean, 2010; Kahane, 2012; Kumar & Campbell, 2012; Bruni et al., 2013; Bluhm, 2014; Königs, 2018; Paulo, 2018).

This latter portion of the critical response bears deep similarities with the Great Rationality Debate in the formal reasoning literature. Recall L. Jonathan Cohen's (1979, 1981) point that there is an interpretive gap between the task and the norm relevant for evaluating performance on the task. Several philosophers have pressed similar criticisms against moral psychologists. They argue that classic moral dilemma vignettes are underspecified, such that there is no determinate verdict about what utilitarianism or deontology would require in these tasks (Berker, 2009; Hueber, 2011; Christensen & Gomila, 2012). This leaves room for participants to "fill in the blanks" in any number of idiosyncratic ways, making it impossible to infer how participants actually resolved conflict and reached either a deontological or consequentialist verdict (Dewey, 2022).

However, there is an important difference between the moral decision-making and formal reasoning literatures. On the one hand, cognitive scientists in the formal reasoning literature have evaluated formal judgments as correct or incorrect consistent with classical norms of formal reasoning and confidently included these evaluations in their empirical work. On the other hand, cognitive scientists in the moral decision-making literature have been much warrier about evaluating moral decisions as correct or incorrect. They officially maintain evaluative neutrality in their empirical work and instead write philosophical work on the side that defends more speculative evaluations of responses as correct or incorrect—usually consistent with the consequentialist norms and contrary to deontological norms. Philosophers have often argued that these consequentialist commitments have contaminated the empirical work, but moral psychologists adamantly deny this (e.g., Paxton et al., 2018).

However, here's the argument for contamination. Suppose consequentialism is true (in a realist way). Then a task-based explanation of consequentialist responses would describe how participants manifest their capacity to track the relations by which facts about consequences *ground* moral facts (e.g., that an action's maximising resultant happiness makes it morally right). According to this explanation, deontological heuristics somehow interfere with this capacity to track the consequentialist facts that ground moral facts. Perhaps, we could add, these heuristics are interfering in virtue of the fact that they have high causal and motivational efficacy—both characteristic features of emotional processing. But their status as interfering doesn't derive from the fact that they might originate in emotion—it derives from the fact that they impair the capacity to track the consequentialist facts that in fact (ex hypothesi) ground moral facts.

Alternatively, suppose deontology is true (in a realist way). Then a task-based explanation of deontological responses would describe how participants manifest their capacity to track the relations by which facts about rules *ground* moral facts (e.g., that an action could be rationally willed as a universal law). According to this explanation, consequentialist principles interfere with this capacity to track the deontological facts that ground moral facts. For example, we could explain that appealing to consequentialist reasons is widely *and mistakenly* (ex hypothesi) regarded as stronger justification for moral decisions (at least in many cultures), which biases reflective reasoning towards consequentialist decisions on the basis that they will be easier to justify post hoc (for an explanation similar to this, see Royzman et al., 2015).

Finally, suppose we reject that either deontology or utilitarianism is true (in a realist way). Then task-based explanation becomes impossible. In lieu of task-based explanation, we might appeal to some rational norm in a sort of rationality-based explanation. For example, Gigerenzer (2008) suggests that it would be rational to prefer simple, context-based heuristics in moral decision-making for the same reasons that it is rational to prefer them in formal reasoning: they involve fewer steps and so are susceptible to fewer errors. Consistent with some sort of rational analysis (*a la* Anderson, 1990), Gigerenzer could say that participants make moral decisions by recognising relevant heuristics and selecting out the simplest, highest-confidence one.

It should be clear that each of these explanations is quite different, and the explanation most often repeated by Greene and colleagues is the one that would be most plausible if consequentialism were true. This is precisely what we should expect, given our discussions in §1 and §2: making assumptions about the correct way to derive a solution to the task (a task analysis) is a prerequisite for developing the skeleton of an explanation for what the participant's cognition must have done to non-accidentally find the correct solution to the task. Moral psychologists may insist that they keep their normative commitments separate from their cognitive explanations, but normative commitments are, in fact, indispensable for guiding cognitive explanations. The reality, we propose, is that Greene and colleagues endorse cognitive explanations that are tacitly guided by consequentialist interpretations of the task.

However, we don't quite mean to join the chorus of criticism raised by moral philosophers against cognitive scientists who study moral decision-making. We worry that moral philosophers have criticised cognitive explanations of moral decision-making for their tacit commitments to moral evaluation without realising that normative commitments are generally *indispensable* to cognitive explanation (per our arguments in §1). We also believe that moral neuroscientists and psychologists may be more prepared to admit that their normative commitments have contaminated their cognitive explanations, if they accepted that there was no other way that they could have developed their cognitive explanations. We take this to be the main upshot of comparing this debate with the Great Rationality Debate. To emphasise this comparison, we'll co-opt Stanovich's (2011) colourful terminology and describe this exchange as the "Great Morality Debate".

§3.4. Dual-Process Theory 2.0

The analogy between the Great Morality Debate and the Great Rationality Debate becomes clearest when we consider the progress that De Neys and colleagues have made to break this gridlock. Following De Neys et al. (2008) and others, Białek & De Neys (2016, 2017) proposed a shift in focus from (controversial) conflict resolution in moral decision-making to (less controversial) conflict detection. To do this, they noted that Greene et al.'s (2001) dilemmas present a Pareto efficient, conflicting set of options: (a) the so-called deontological response is better insofar as it doesn't require the decision-maker to enter a causal chain that leads to harm and (b) the so-called consequentialist response is better insofar as it ensures a result with fewer deaths (one instead of five). Thus, they describe this as a conflict task.

Next, they propose a no-conflict version of the task, where one option is better along both dimensions (it is a Pareto improvement) and the other option is worse along both dimensions. For example, they tell a subject that a trolley will strike and kill one worker unless a lever is pulled, redirecting the trolley onto another track, where it will kill five workers. In this case, the participant obviously should not pull the lever to (a) avoid entering a casual china that leads to harm and/or (b) to ensure a result with fewer deaths. Then Białek & De Neys use subtractive comparisons for conflict and no-conflict versions of the task to identify neural and behavioural activity that is selective to the presence of conflict. They find that conflict increases response time and decreases confidence ratings (Białek & De Neys, 2016), and that these effects aren't modulated by the presence of cognitive load (Białek & De Neys, 2017).

Białek & De Neys (2016, 2017) note that these results are interesting in part because they are consistent with the parallel-competitive model but contradictory with the default-interventionist model. However, we think there is a deeper reason why these results are interesting: they create space for applying the triangulating extension to the task-based approach to cognitive explanation. In particular, this kind of experimental design makes it possible to spatially localise the brain regions involved in conflict detection using methods like functional magnetic resonance imaging (fMRI) and temporally localise the intervals where conflict detection occurs using methods like EEG. We can use that information to triangulate on the morally controversial step of conflict resolution in moral decision-making—by directing our attention to the neural activity causally and temporally downstream of conflict detection—without having to make controversial assumptions about how moral conflict ought to be resolved (see §1.2).

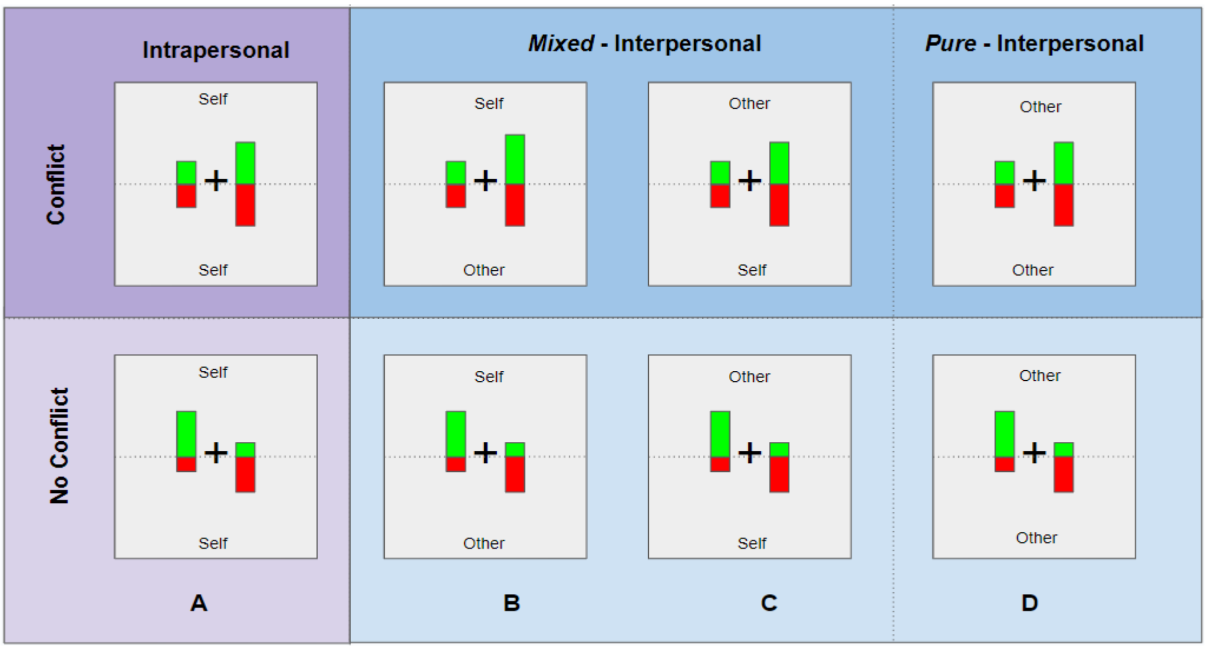

Figure 1. *The size of green bars represents the amount of reward for each option and the size of red bars represents the number of shocks for each option. Two options are presented in each case and the participant has to press a button to indicate which option they prefer. Choices are better along only one dimension in the conflict condition and better along both dimensions in the no-conflict condition.*

As far as we know, though, neural techniques have not yet been used to spatially or temporally localise conflict detection in moral decision-making tasks. In planned experimental research, our group is attempting to address this gap. Whereas Białek & De Neys (2016, 2017) examined conflict between deontological and consequentialist considerations using a conflict/no-conflict version of Greene et al.'s (2001) trolley dilemma paradigms, we preferred to examine conflict between altruistic and prudential considerations using a conflict/no-conflict version of Crockett et al.'s (2014, 2015, 2017) profit/pain dilemma paradigm. One reason for this change was that we suspect that the altruism/prudential distinction is much more likely to have a basis in different neural systems than the deontological/consequentialist distinction (*contra* Greene, 2007).

In particular, we've modified Crockett et al.'s (2017) paradigm to give participants two choices that allocate different painful shocks and monetary rewards between themselves and/or an anonymous (real) compatriot. In the conflict condition, one choice is better along the reward dimension (more reward) and worse along the pain dimension (more shocks) whereas the other choice is better along the pain dimension (fewer shocks) but worse along the reward dimension (less reward) (see Figure 1). In the mixed interpersonal conditions, which are the two experimental conditions, the shocks are given to one participant and the rewards are given to the other. Consistent with Crockett et al. (2014, 2015, 2017), we expect to see more altruism when the participant has to forego reward to spare their compatriot from more shocks than when the participant has to endure shocks to give their compatriot more reward.

This task design creates a distinction between two kinds of conflict: pain vs. reward and self vs. other. We are using EEG to help identify the spatial and especially temporal basis of both forms of conflict. First, we're interested in whether one kind of conflict is detected *before* another. Each consideration must be evaluated to some extent before conflict can be detected, so we can use the triangulating extension to task-based functional analysis from §1.2 to give explanations for any possible result. If pain vs. reward is detected first, e.g., we can reliably (but defeasibly) infer that pain and reward information are more quickly evaluated than self and other information (and vice versa). But if both forms of conflict are detected at the same time, then we can either infer that both considerations are processed with equal speed or perhaps that conflict detection processing is initiated after a fixed interval of time is given to allow for the evaluation of different considerations.

Second, we're interested in whether there are interactions between both forms of conflict detection or whether they are kept separate. We can answer this question by identifying whether the responses to both kinds of conflict are predictable from responses to each kind of conflict. The two forms of conflict are objectively independent, so we can use the triangulating extension to task-based analysis from §1.2 to infer that the processes responsible for conflict detection must also be doing something else. One potential explanation would be that these processes may be responding to some kind of feedback from the conflict resolution process—and, transitively, from the other conflict detection process that is feeding into the conflict resolution process. This would be a powerful example of the triangulating potential for the inference we described in §1.2. It would allow us to draw a concrete explanation for conflict resolution on the basis of results about conflict detection without having to make normative assumptions about conflict resolution.

§4. Conclusion

In conclusion, we've argued that conflict provides a helpful entry point for formal reasoning and moral decision-making for similar reasons. In particular, we've argued that task analysis provides a basic explanation for above-chance success onto which we can scaffold empirical, behavioural and neurobiological information in order to develop more detailed explanations for above-chance success and failure alike. We've argued that this task-based approach to explanation is especially difficult when we disagree on what counts as the correct way to solve a task. We found that this problem is the source of what Stanovich (2011) calls the "Great Rationality Debate" in the literature on formal reasoning and what we've called the "Great Morality Debate" in the literature on moral decision-making.

We found that De Neys and colleagues have managed to make headway in the Great Rationality Debate by shifting focus from formal reasoning as a whole process to conflict detection as a stage of formal reasoning. We argued that this strategy is so successful because it segments a normatively controversial process into sub-processes, some of which are controversial (e.g., conflict resolution) and some of which aren't (e.g., conflict detection). The task-based approach to explanation is unproblematic for conflict detection, so De Neys and colleagues make headway by taking a task-based approach to explanation and focusing it on conflict detection. We argued that this strategy could eventually help us triangulate on possible explanations for conflict resolution by helping localise it in time and space and by helping characterise its input and output conditions more precisely.

Thus, De Neys and colleagues have built a blueprint for task-based approaches to explaining other cognitive processes that are normatively controversial. We argued that a similar strategy would facilitate cognitive explanation for moral decision-making. In particular, we can segment moral decision-making into conflict detection and conflict resolution sub-processes and then take a task-based approach to explanation for conflict detection. We reviewed a pilot study that we're planning to run, which applies this basic approach to an earlier paradigm developed by Crockett et al. (2015, 2017). We hope this pilot study and any follow-up studies will give us deeper insight into the neural architecture of moral decision-making. More generally, though, we also hope that these studies will help refine our explanatory strategy for the neuroscience of cognition when the normative status of cognition is controversial.

References

Allman, J., & Woodward, J. (2008). What are moral intuitions and why should we care about them? A neurobiological perspective. *Philosophical Issues*, *18*, 164–185.

Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means: Visual imagery and moral judgment. *Psychological Science*, *23*(8), 861–868. https://doi.org/10.1177/0956797611434965

Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum Associates, Inc.

Andersson, L., Eriksson, J., Stillesjö, S., Juslin, P., Nyberg, L., & Wirebring, L. K. (2020). Neurocognitive processes underlying heuristic and normative probability judgments. *Cognition*, *196*, 104153. https://doi.org/10.1016/j.cognition.2019.104153

Aziz-Safaie, T., Müller, V. I., Langner, R., Eickhoff, S. B., & Cieslik, E. C. (2024). The effect of task complexity on the neural network for response inhibition: An ALE meta-analysis. *Neuroscience & Biobehavioral Reviews*, *158*, 105544. https://doi.org/10.1016/j.neubiorev.2024.105544

Chung, R. S., Cavaleri, J., Sundaram, S., Gilbert, Z. D., Del Campo-Vera, R. M., Leonor, A., Tang, A. M., Chen, K.-H., Sebastian, R., Shao, A., Kammen, A., Tabarsi, E., Gogia, A. S., Mason, X., Heck, C., Liu, C. Y., Kellis, S. S., & Lee, B. (2024). Understanding the human conflict processing network: A review of the literature on direct neural recordings during performance of a modified Stroop task. *Neuroscience Research*. https://doi.org/10.1016/j.neures.2024.03.006

Cohen, L. J. (1979). On the psychology of prediction: Whose is the fallacy? *Cognition*, *7*(4), 385–407. https://doi.org/10.1016/0010-0277(79)90023-4

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, *4*(3), 317–331. https://doi.org/10.1017/S0140525X00009092

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

Côté, S., Piff, P. K., & Willer, R. (2013). For whom do the ends justify the means? Social class and utilitarian moral judgment. *Journal of Personality and Social Psychology*, *104*(3), 490–503. https://doi.org/10.1037/a0030931

Cox, J. R., & Griggs, R. A. (1982). The effects of experience on performance in Wason's selection task. *Memory & Cognition*, *10*(5), 496–502. https://doi.org/10.3758/BF03197653

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. New York: Oxford University Press.

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(40), 17433–17438. https://doi.org/10.1073/pnas.1009396107

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, *111*(48), 17320–17325. https://doi.org/10.1073/pnas.1408988111

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, *20*(6), Article 6. https://doi.org/10.1038/nn.4557

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Ousdal, O. T., Story, G., Frieband, C., Grosse-Rueskamp, J. M., Dayan, P., & Dolan, R. J. (2015). Dissociable effects of serotonin and dopamine on the valuation of harm in moral decision making. *Current Biology*, *25*(14), 1852–1859. https://doi.org/10.1016/j.cub.2015.05.021

Cummins, R. (1983). *The nature of psychological explanation*. MIT Press.

Cushman, F., Murray, D., Gordon-McKeon, S., Wharton, S., & Greene, J. D. (2012). Judgment before principle: Engagement of the frontoparietal control network in condemning harms of omission. *Social Cognitive & Affective Neuroscience*, *7*(8), 888–895. https://doi.org/10.1093/scan/nsr072

Dean, R. (2010). Does neuroscience undermine deontological theory? *Neuroethics*, *3*(1), 43–60. https://doi.org/10.1007/s12152-009-9052-x

De Neys, W. (2017). *Dual Process Theory 2.0*. Routledge. https://doi.org/10.4324/9781315204550

De Neys, W. (2021). On dual- and single-process models of thinking. *Perspectives on Psychological Science*, *16*(6), 1412–1427. https://doi.org/10.1177/1745691620964172

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science*, *19*(5), 483–489. https://doi.org/10.1111/j.1467-9280.2008.02113.x

Dewey, A. R. (2022). Arbitrating norms for reasoning tasks. *Synthese*, *200*(6), 502. https://doi.org/10.1007/s11229-022-03981-8

Elqayam, S., & Evans, J. S. B. T. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, *34*(5), 233–248. https://doi.org/10.1017/S0140525X1100001X

Evans, J. S. B. T., & Lynch, J. S. (1973). Matching bias in the selection task. *British Journal of Psychology*, *64*(3), 391–397. https://doi.org/10.1111/j.2044-8295.1973.tb01365.x

Gazzaniga, M. S. (2005). Forty-five years of split-brain research and still going strong. *Nature Reviews Neuroscience*, *6*(8), 653–659. https://doi.org/10.1038/nrn1723

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. https://doi.org/10.1037/0033-295X.103.4.650

Gigerenzer, G. (2008). Moral intuition = fast and frugal heuristics? In W. Sinnott-Armstrong (Ed.), *Moral Psychology* (Vol. 2). MIT Press.

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science*, *2*(3), 528–554. https://doi.org/10.1111/j.1756-8765.2010.01094.x

Gigerenzer, G. (2020). How to explain behavior? *Topics in Cognitive Science*, *12*(4), 1363–1381. https://doi.org/10.1111/tops.12480

Gleichgerrcht, E., & Young, L. (2013). Low levels of empathic concern predict utilitarian moral judgment. *PLOS ONE*, *8*(4), e60418. https://doi.org/10.1371/journal.pone.0060418

Glenn, A. L., Raine, A., Yaralian, P. S., & Yang, Y. (2010). Increased volume of the striatum in psychopathic individuals. *Biological Psychiatry*, *67*(1), 52–58. https://doi.org/10.1016/j.biopsych.2009.06.018

Greene, J. D. (2007). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology*. MIT Press.

Greene, J. (2014). *Moral tribes: Emotion, reason and the gap between us and them*. Atlantic Books. http://ebookcentral.proquest.com/lib/uaz/detail.action?docID=1486561

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. https://doi.org/10.1126/science.1062872

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. https://doi.org/10.1016/j.neuron.2004.09.027

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Pantheon Books.

Hauser, M. (2006). *Moral minds: How nature designed our universal sense of right and wrong* (pp. xx, 489). Ecco/HarperCollins Publishers.

Huebner, B. (2011). Critiquing empirical moral psychology. *Philosophy of the Social Sciences*, *41*(1), 50–83. https://doi.org/10.1177/0048393110388888

Hume, D. (1739). *A Treatise of Human Nature*. London: John Noon.

Kahane, G. (2012). On the wrong track: Process and content in moral psychology. *Mind & Language*, *27*(5), 519–545. https://doi.org/10.1111/mila.12001

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In *The Cambridge Handbook of Thinking and Reasoning* (pp. 267–293). Cambridge University Press.

Klein, C. (2011). The dual track theory of moral decision-making: A critique of the neuroimaging evidence. *Neuroethics*, *4*(2), 143–162. https://doi.org/10.1007/s12152-010-9077-1

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

Königs, P. (2018). On the normative insignificance of neuroscience and dual-process theory. *Neuroethics*, *11*(2), 195–209. https://doi.org/10.1007/s12152-018-9362-y

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, *118*(1), 97–109. https://doi.org/10.1037/a0020762

Kumar, V., & Campbell, R. (2012). On the normative significance of experimental moral psychology. *Philosophical Psychology*, *25*(3), 311–330. https://doi.org/10.1080/09515089.2012.660140

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*. https://doi.org/10.1017/S0140525X1900061X

Mata, A. (2019). Social metacognition in moral judgment: Decisional conflict promotes perspective taking. *Journal of Personality and Social Psychology*, *117*(6), 1061–1082. https://doi.org/10.1037/pspa0000170

Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). An investigation of moral judgement in frontotemporal dementia. *Cognitive and Behavioral Neurology: Official Journal of the Society for Behavioral and Cognitive Neurology*, *18*(4), 193–197. https://doi.org/10.1097/01.wnn.0000191292.17964.bb

Mevel, K., Borst, G., Poirel, N., Simon, G., Orliac, F., Etard, O., Houdé, O., & De Neys, W. (2019). Developmental frontal brain activation differences in overcoming heuristic bias. *Cortex*, *117*, 111–121. https://doi.org/10.1016/j.cortex.2019.03.004

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

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*, *19*(6), 549–557. https://doi.org/10.1111/j.1467-9280.2008.02122.x

Moore, A. B., Stevens, J., & Conway, A. R. A. (2011). Individual differences in sensitivity to reward and punishment predict moral judgment. *Personality and Individual Differences*, *50*(5), 621–625. https://doi.org/10.1016/j.paid.2010.12.006

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198524496.001.0001

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., Silani, G., Cikara, M., & Cushman, F. (2021). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *Journal of Personality and Social Psychology*, *120*(2), 443–460. https://doi.org/10.1037/pspp0000281

Paulo, N. (2019). In search of Greene's argument. *Utilitas*, *31*(1), 38–58. https://doi.org/10.1017/S0953820818000171

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

Paxton, J. M., Bruni, T., & Greene, J. D. (2014). Are ‘counter-intuitive’ deontological judgments really counter-intuitive? An empirical reply to Kahane et al. (2012). *Social Cognitive and Affective Neuroscience*, *9*(9), 1368–1371. https://doi.org/10.1093/scan/nst102

Rawls, J. (2005). *Political Liberalism*. New York: Columbia University Press.

Redish, A. D. (2022). *Changing how we choose: The new science of morality*. Cambridge, MA: The MIT Press.

Reynolds, C. J., Knighten, K. R., & Conway, P. (2019). Mirror, mirror, on the wall, who is deontological? Completing moral dilemmas in front of mirrors increases deontological but not utilitarian response tendencies. *Cognition*, *192*, 103993. https://doi.org/10.1016/j.cognition.2019.06.005

Rowley, D. A., Rogish, M., Alexander, T., & Riggs, K. J. (2018). Counter-intuitive moral judgement following traumatic brain injury. *Journal of Neuropsychology*, *12*(2), 200–215. https://doi.org/10.1111/jnp.12117

Royzman, E. B., Leeman, R. F., & Sabini, J. (2008). “You make me sick”: Moral dyspepsia as a reaction to third-party sibling incest. *Motivation and Emotion*, *32*(2), 100–108. https://doi.org/10.1007/s11031-008-9089-x

Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of Julie and Mark: Unraveling the moral dumbfounding effect. *Judgment and Decision Making*, *10*(4), 296–313. https://doi.org/10.1017/S193029750000512X

Royzman, E. B., Landy, J. F., & Leeman, R. F. (2015). Are thoughtful people more utilitarian? CRT as a unique predictor of moral minimalism in the dilemmatic context. *Cognitive Science*, *39*(2), 325–352. https://doi.org/10.1111/cogs.12136

Samuels, R., Stich, S., & Bishop, M. (2002). Ending the rationality wars: How to make disputes about human rationality disappear. In R. Elio (Ed.), *Common Sense, Reasoning and Rationality* (pp. 236–268). Oxford University Press.

Simon, G., Lubin, A., Houdé, O., & Neys, W. D. (2015). Anterior cingulate cortex and intuitive bias detection during number conservation. *Cognitive Neuroscience*, *6*(4), 158–168. https://doi.org/10.1080/17588928.2015.1036847

Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics*, *9*(3/4), 331–352.

Stanovich, K. (2010). *Rationality and the reflective mind*. Oxford University Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–65. https://doi.org/10.1017/S0140525X00003435

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

Tassy, S., Oullier, O., Duclos, Y., Coulon, O., Mancini, J., Deruelle, C., Attarian, S., Felician, O., & Wicker, B. (2012). Disrupting the right prefrontal cortex alters moral judgement. *Social Cognitive and Affective Neuroscience*, *7*(3), 282–288. https://doi.org/10.1093/scan/nsr008

Tolstoy, L. (1878). *Anna Karenina*. The Russian Messenger.

Trémolière, B., Neys, W. D., & Bonnefon, J.-F. (2012). Mortality salience and morality: Thinking about death makes people less utilitarian. *Cognition*, *124*(3), 379–384. https://doi.org/10.1016/j.cognition.2012.05.011

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*(6), 476–477. https://doi.org/10.1111/j.1467-9280.2006.01731.x

Vartanian, O., Beatty, E. L., Smith, I., Blackler, K., Lam, Q., Forbes, S., & De Neys, W. (2018). The reflective mind: Examining individual differences in susceptibility to base rate neglect with fMRI. *Journal of Cognitive Neuroscience*, *30*(7), 1011–1022. https://doi.org/10.1162/jocn_a_01264

Johnson-Laird, P. N., & Wason, P. C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, *1*(2), 134–148. https://doi.org/10.1016/0010-0285(70)90009-5

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. https://doi.org/10.1080/14640746808400161

Yin, X., Hong, Z., Zheng, Y., & Ni, Y. (2022). Effect of subclinical depression on moral judgment dilemmas: A process dissociation approach. *Scientific Reports*, *12*(1), 20065. https://doi.org/10.1038/s41598-022-24473-2

Zhang, X., Wu, Z., Li, S., Lai, J., Han, M., Chen, X., Liu, C., & Ding, D. (2020). Why people with high alexithymia make more utilitarian judgments: The role of empathic concern and deontological inclinations. Experimental Psychology, 67(1), 23–30. https://doi.org/10.1027/1618-3169/a000474