Judgment and Decision Making, Vol. 16, No. 2, March 2021, pp. 566–595

Consequences, norms, and inaction: Response to Gawronski et al. (2020)

Jonathan Baron* Geoffrey P. Goodwin†

Abstract

In response to arguments made by Gawronski et al. (2020), responding to Baron and Goodwin (2020), we concentrate on four issues. First, the CNI design requires substantial numbers of "perverse" responses to congruent items — those in which both consequences and norms both favor action, or both favor inaction — and these responses depend on the ambiguity of the items concerning which norms apply or which consequences are worse. Effects of outside variables, such as psychopathy, may result from the effect of such variables on the interpretation of ambiguous items, rather than from their effect on sensitivity to norms or consequences. Second, the CNI design may not be so useful at measuring general action/inaction biases. Third, the order of the two processes in the model could in fact affect the conclusions drawn (even though it does not do so in most studies done so far). Fourth, the conclusions drawn do in fact depend on the items' susceptibility to reinterpretation (owing to their ambiguity): the tests proposed for item validity are too weak, since they require only that a majority of subjects agree with the experimenters' classification, even though a minority could affect the conclusions drawn. We illustrate some of our points with an analysis of the psychopathy study of Luke and Gawronski (2020).

Keywords: utilitarianism, deontology, omission bias, psychopathy

*Department of Psychology, University of Pennsylvania, 3720 Walnut St., Philadelphia, PA, 19104. Email: baron@upenn.edu. ORCID: 0000-0003-1001-9947.

†Department of Psychology, University of Pennsylvania.

We thanks Andreas Glöcker and Michael Laakasuo for helpful comments.

*Department of Psychology, University of Pennsylvania, 3720 Walnut St., Philadelphia, PA, 19104. Email: baron@upenn.edu. ORCID: 0000-0003-1001-9947.

†Department of Psychology, University of Pennsylvania.

1 Introduction

We thank Gawronski et al. (2020, henceforth Getal) for their comments. They make a number of useful points. In particular, they clarify the main contribution of the CNI model, which is, in our words, to separate two possible influences on responding in the usual "sacrificial dilemmas". Recent research on moral judgment has tended to look at dilemmas posing a conflict between some sort of moral rule or norm, on the one hand, and clearly defined consequences, such as numbers of deaths, on the other. Some of this research investigates manipulations such as time pressure or cognitive load; other research investigates individual differences. In both cases, especially the latter, any observed effects could occur on either horn of the dilemma, or both. In particular, some people may be or become (due to a manipulation) more or less sensitive to the consequences, the numbers. But they could also be or become more or less concerned about the norm or rule, such as "Do no harm," where "do" is often interpreted as requiring action or direct causation. Gawronski et al. (2017, henceforth GACFH) propose to separate these possible effects by varying both norms and consequences in sets of dilemmas. In addition, they propose that a bias towards inaction is a third factor that determines subjects' responses, which is separate from sensitivity to norms or consequences.

GACFH achieved this objective by creating new vignettes in which the standard links between action, consequences, and norm violation are varied. In a standard sacrificial dilemma, an action violates a norm (usually by creating some specific harm) in order to improve the overall consequences, and contrasts with inaction that does not violate a norm at the expense of worsening the overall consequences. To this basic case, GACFH add three new cases. In one "reversed" case, action no longer violates a norm, but it creates worse consequences than inaction, which itself does violate a norm. There are also two congruent cases. In one, action violates a norm and creates worse consequences than inaction, which also does not violate a norm. In the other, action does not violate a norm and improves the consequences compared to inaction which also violates a norm. The result is a three-dimensional space with eight separate choices (created by the crossing of three variables each with two levels). This decision space is displayed in Table 1 below, along with the options favored by each of three decision principles, which prioritize the consequences, norm following, and inaction, respectively.

TABLE 1: The separation of consequences, norms, and inaction in GACFH's experimental designs, along with the choice option favored by each decision principle (indicated with √).

Vignette type	Choice options	Decision Principle		
		Maximize consequences	Conform with norms	Favor inaction
Standard sacrificial dilemma (NormOmit.-ConsAct)	1. Action violates a norm but creates better consequences.	√		
	2. Inaction does not violated a norm but creates worse consequences.		√	√
Reversed sacrificial dilemma (NormAct.-ConsOmit	1. Action does not violate a norm, but creates worse consequences.		√	
	2. Inaction violates a norm, but creates better consequences.	√		√
Congruent case 1 (NormOmit.-ConsoOmit)	1. Action violates a norm and creates worse consequences.			
	2. Inaction does not violated a norm and creates better consequences.	√	√	√
Congruent case 2 (NormAct.-ConsoAct)	1. Action does not violate a norm and creates better consequences.	√	√	
	2. Inaction violateds a norm and creates worse consequences.			√

1.1 Our main objections

As we noted in our original critique, the general idea of trying to separate out these three factors in this way has merit in principle, but in practice, it creates several problems.

Before we comment further on points of agreement and disagreement, we would like to re-state our main concerns, which we apparently did not communicate entirely clearly in our initial paper. In the standard sacrificial dilemma, which is, in our notation, NormOmit.ConsAct, the norm favors omission and the consequences favor action. As shown above, the CNI design adds three other dilemmas: NormAct.ConsOmit, the opposite of the original conflict, and two congruent dilemmas in which norms and consequences both favor action, NormAct.ConsAct, or omission, NormOmit.ConsOmit. By looking at all four types, we can do two things that we cannot do with the standard dilemma alone. First, we

can ask how much of the response to the standard dilemma is determined by a bias toward action or omission as such. Second, we can ask whether various individual differences and experimental manipulations affect sensitivity to norms, consequences, or action/omission bias. For example, we might find that some individual difference predicts sensitivity to norms but not sensitivity to consequences. (By "sensitivity" to a factor, we mean the extent to which this factor predicts responses.)

In Section 2 we discuss the measurement of action/omission bias. We suggest that much of what counts as bias in the CNI model is likely to consist of the endorsement of specific deontological rules, rather than being an entirely separate factor. We suggest alternative ways to measure generalized bias, which would apply to all cases regardless of the rules involved.

Turning to the second function of the CNI design, the possibility of separating effects of individual difference factors (such as gender, or psychopathy) on sensitivity to norms and consequences, our objection was, and is, that this function depends on the use of "perverse" responses to the congruent items, in which subjects go against both norms and consequences, despite those factors both supposedly pointing in the same direction. Importantly, it is likely that most of these responses result from interpretations of the items that disagree with the experimenters' intentions in creating them. This was the main point of the new data we presented. That is, individual difference factors may primarily affect how subjects' interpret the congruent cases rather than their sensitivity to norms and consequences. If this is the case, it poses a major challenge to GACFH's interpretation of their data.

In the next section, we use an example to discuss what happens when the congruent responses are clear enough so that idiosyncratic interpretations are largely absent. The result is that the CNI model can no longer assess three parameters. It can, however, sometimes assess two parameters: action/inaction bias and *relative* sensitivity to norms and consequences. For this to happen, the norm involved must be neutral between acts and omissions. In the usual sacrificial dilemmas, by contrast, the norm is specific to action, and this makes it difficult, as best, to separate the influence of the norm itself from a bias toward inaction.

A related concern is that the perverse responses can distort the measurement of C and N, the parameters that indicate sensitivity to consequences and norms, respectively. In Section 5 we present some numerical examples of possible effects of perverse responses on measures of C (consequences) and N (norms). In one of them, a difference between conditions in the proportion of perverse responses could switch the relative magnitudes of the N and C parameters. This is, of course, an artificially constructed example, but it is unclear how we could remove such alternative explanations from reported results based on the CNI design. Note that any factor that appears to affect only norms, or only consequences, in the CNI model must also affect responses to the congruent cases in some way. How can we tell whether this effect is the result of sensitivity to norms or consequences, as the model implies, or to some factor that leads to different interpretations of some type of case?

Putting this point another way, the CNI model assumes that N and C are constant across all the items. Thus, the action-opposing norms in the congruent cases and the action-opposing norms in the incongruent cases have the same strength. Likewise for the consequences. But it is clear, as we argued originally and again here, that this assumption is false, and not just a little false.

Getal make several replies to our critique. We feel that some of them are reasonable and require little in the way of rebuttal. In particular, we acknowledge that the CNI model fits the data pretty well.[1] Getal present more data showing this, acknowledging and responding to our concern about testing generality across sets of items as well as across subjects (or across observations, as apparently done in GACFH). This demonstration answers much of our argument about inconsistency between scenarios.

However, it does not answer a more general concern that we did not emphasize enough, or at all. A model can fit quite well and still systematically mis-estimate parameters of interest. In particular, consistency of fit across cases can occur if many cases mis-estimate in the same way. We illustrate here how that can happen.

1.2 An example of the role of congruent items

Suppose we want to apply the CNI design to the following example (based on Ubel et al., 1996):

> A health provider has a limited budget for screening tests. It plans to offer **all its members** a simple screening test suitable for a population at low risk of colon cancer. The test would save **1000 lives** in the next 5 years.
>
> A new, more expensive, test has just become available. It can be offered to **half of the members, selected at random**. The two tests differ in effectiveness. The new test would save **1100 lives** in the next 5 years.
>
> Which one should be chosen?

The deontological rule here is to treat everyone equally by giving everyone the old test. The utilitarian response is to choose the new test because it saves more lives. (A substantial proportion of subjects in Ubel et al., 1996, study chose the old test in a similar case.) We could introduce action and omission by stipulating that the plan was to give one of the tests but the other test just became available, and so the question is whether to switch to the new one or do nothing. By counterbalancing which test is new, we could change the association of the norm with action or omission, as required by the CNI design. Note that this is a completely symmetric switch, unlike the cases actually used in GACFH's research, where undoing a previous plan is often required only in half of the items. Moreover, the norm does not itself imply anything about whether the outcome is caused by action or omission.

[1]However, in these tests of model fits, no data on alternative models were included. Many models could fit well, and good fit alone need not provide strong evidence for the underlying assumptions (Roberts & Pashler, 2000). In Section 6.1, we present an example of a simple alternative model.

We also need two congruent cases. To accomplish that, we could switch which option leads to 1000 saved and which leads to 1100 saved, so now we have the new test given to half the population and saving fewer lives (1000) than the old test (1100). (Again, this is a completely symmetric switch, unlike those commonly used in the CNI design.) The two congruent cases would emerge from varying which test was the older, default option.

Suppose that we run this experiment and find that almost nobody chooses perverse responses and that 50% choose according to the equality norm and 50% according to consequences. If we apply the CNI model, we find that C is .5, N is 1, and we cannot estimate I. The estimate of I require a reasonable number of perverse responses. When the items are constructed clearly, so as to produce uniform assumptions of what the norms and consequences actually entail, then there is no reason why subjects would choose a response that is inconsistent with both norms and consequences.

A natural conclusion is that we need different congruent cases, cases that are more difficult, so that we have something to work with, since the CNI design requires a reasonable number of perverse responses. One option would be to modify the action condition so that it requires overriding an earlier decision by someone who will likely feel unappreciated and useless as a result. Note, however, that this would add another norm to the mix, probably one that is not as powerful as the original norm of equal treatment. We could also make the consequences more ambiguous by saying that the more effective test has some unpleasant side effects. Both of these maneuvers could "work" in the sense that they produce some perverse responses to the cases that we defined as congruent. For instance, some people might feel a conflict between the two norms, equality and regard for others' feelings, and decide that the latter was more important. Some other people might interpret the consequences differently than originally intended, by incorporating hurt feelings and the test's side effects into their deliberations. Of course, a majority of subjects are unlikely to be affected by these modifications, which allows us to claim that our manipulation of norms and consequences is still valid, since the differences among relevant conditions will be in the predicted direction. But it is no longer so clear that these factors, as originally conceived, are the sole drivers of people's responses.

Similarly, what do we conclude if we find group differences? Perhaps one group of subjects is more likely than the other to look for reasons why "obvious" or "no brainer" cases are actually not so obvious. ("Why are you asking questions like this? There must be something you are looking for.") Maybe this is what some people think a good subject should do. In addition, there may be subjects who, with sufficient reflection or repeated trials of the same item, would agree with the intended interpretation but whose thinking process is more variable such that they require a clear difference for a consistent response (see Section 7). And, if the manipulations of norms and consequences are systematically different, we might find that different groups of people require clear differences in one factor or the other. Thus, we have created a situation that allows us to apply the CNI model, but at the same time we have opened a small Pandora's box of alternative explanations. Perhaps

this situation is not so far from the typical uses of the CNI design.

In Appendix B, we report some results from a new experiment based on the Ubel, et al., cases described above, in which the norms and consequences are the same in all four versions. Indeed, perverse responses largely disappear. We do note, however, that the incongruent cases alone can be used to estimate two parameters: the relative balance of consequences and norms, and action/inaction bias. We can do this because we can switch the assignment of action and omission to the two outcomes, while maintaining exactly the same norm. In addition, we use two different cases, and we find that the estimate of each parameter in one case correlates with its estimate in the other case; these correlations indicate some generality in the parameters.

2 Action/inaction bias

By contrast, the CNI design is not very good at measuring action/inaction bias. For one thing, if it were a good measure, we would expect the four item types to show a consistent effect of action/omission. All four types are affected by the bias to the same extent: when neither consequences nor norms drive the response, action bias increases action responses. Thus, we expect them to correlate more highly than otherwise because they are all influenced by the bias. But they do not correlate very well. Using GACFH's data, we looked at action choices in each of the four cases (NormOmit.ConsAct, NormOmit.ConsOmit, NormAct.ConsAct, NormAct.ConsOmit) in several studies in GACFH. If we consider these action choices as a four-item test, its reliability is essentially zero: sometimes negative, sometimes positive but with 0 in the 95% C.I.[2] In particular, action choices in the two congruent cases were negatively correlated. This lack of consistency does not show that the method is useless, since the sources of negative correlation could cancel out. Rather, the point is that there are other very large sources of variance aside from bias, so it is unlikely that action bias affects all four items in a way that is accurately measurable.

Perhaps a better way to assess the role of action/omission bias is to use cases in which norms do not favor either option and the consequences are identical. For example, one could use a vaccination case in which the rates of harm for vaccination and non-vaccination are the same. Of course, when people favor action or omission in such cases, they may be following case-specific norms, such as "vaccination is good" or "vaccination is bad". However, it may be that a multi-item test with many such cases would reveal a clear general preference for action or omission, in different people.[3] That is, people may be generally more, or less, likely to favor action over a wide variety of situations. Would such a test correlate with action preferences in the standard dilemmas? It is at present unclear. An additional strategy

[2]In fact, the correlations among the four measures were determined both by norms and consequences. The most negative correlations occurred when norms favored different options and consequences favored different options. Norms seemed to play a larger role.

[3]Some efforts have been made in this direction, but they do not involve multiple cases, e.g., McCullogh et al. (2012).

would be to ask subjects general questions after they answer each dilemma in order to probe the explicit endorsement of action biases, such as, "In cases like this, it is generally better to [do something, do nothing]."

The question of generality raises an important question about when an action/inaction bias should count as a deontological rule. We argued originally that any such bias should count, because it is a consideration other than consequences. Although this argument is supported by many philosophers, it requires some qualification and, we now admit, leaves room for a different interpretation.

First, note that most, perhaps all, of the results listed as potential "biases" in the literature on moral judgment seem to be domain general, across both moral and non-moral situations. And many common biases in judgments and decisions have been found in the moral domain (e.g., Baron & Szymanska, 2010; Shallow, Iliev & Medin, 2011). Of interest here, the tendency to favor harmful omissions over directly harmful acts, the central finding in sacrificial dilemmas, is found in self-interested decisions as well as decisions that affect others (e.g., Ritov & Baron, 1990). Similar arguments have been made by other researchers (e.g., Cushman, 2015). Of course, some deontological rules are very specifically moral, but it seems that many of these are also culturally limited in their relevance and therefore not used much in experiments on culturally diverse populations (e.g., rules concerning observance of the Sabbath, or the subservience of women). Thus, from the perspective of those interested in "heuristics and biases" or just "cognitive biases", domain-general biases are just as interesting as those that are specific to moral judgments.

Moreover, as we noted in our comment, one reason to be interested in the utilitarian/deontological distinction is that we are interested in explaining why decisions are non-optimal in terms of consequences alone. If the consequences of some decision are bad, then one reason is that the decision itself did not aim at maximizing the outcome but was affected by something else. Any sort of "something else" thus counts as deontological, for this purpose. Note, however, that a bias toward action, when action itself yields the best consequences, would not necessarily count as a deontological principle by this standard. It would do so only if the bias were the result of some general disposition that would yield non-optimal choices elsewhere, for example when inaction yields the best consequences. Likewise a bias toward inaction would count if it yields inaction when action yields the best consequences.

Getal distinguish general biases toward action or inaction from specific normative moral rules such as "Do no harm." For the purpose of understanding non-optimal decisions, this distinction is not needed. However, it is certainly a reasonable question to ask whether action/inaction biases that affect moral judgments in sacrificial dilemmas are general across moral and non-moral domains. And we admit that it is also reasonable to adopt (explicitly) a definition of "deontology" that limits this term to factors specific to moral judgments.

Getal say that biases and deontological rules are "confounded" with action/inaction biases, but also note that these biases may in fact be deontological principles in their own

right, as we argued. Getal see the role of biases as an important issue in the classification of norms.[4] And, as they acknowledge, their N parameter does not capture norms that simply favor or oppose action, which are, in our classification, deontological because they refer to properties of the choice that can lead people to choose the option with worse consequences.

The empirical issue may be settled by a test of generality of action/omission biases across both moral and non-moral cases. To our knowledge, such a test has not been done.

In sum, we do not think of action/inaction bias as a confound but as a potential deontological principle or feature, with much the same status as other sorts of features of deontological rules. It is not the only feature of deontological principles, nor is it an inevitable feature of deontological principles; but it is one characteristic feature among many. Furthermore, if the action/inaction bias applies to non-moral cases as well, it may still be opposed to consequentialist considerations and thus of interest in explaining why consequences are not optimal. But we acknowledge that other definitions are reasonable, and that further empirical studies may be worth conducting.

3 Misunderstanding the CNI model

We pointed out that many of the items other than the standard (NormOmit.ConsAct) contained ambiguities, which made it easier for subjects to interpret them differently from the experimenters' intention, or (we might have added, see Section 7) to be sufficiently unsure of their interpretation to lead to some randomness in responding. Getal seem to think we have misunderstood something. We are puzzled by this response:

> For the *C* parameter, scores significantly greater than zero indicate that responses were affected by the manipulation of consequences in a manner such that participants showed a response pattern that maximizes the greater good. For the *N* parameter, scores significantly greater than zero indicate that responses were affected by the manipulation of moral norms such that participants showed a response pattern that is congruent with both proscriptive and prescriptive norms.

This comment seems to misunderstand the problem, since our critique is that responses that were apparently affected by the C or N factors may have in fact been influenced by something else entirely. "Significantly greater than zero" does not imply that other influences are absent.

3.1 Unconscious effects and dual-process theory

Getal say: "Baron and Goodwin's (2020) critique is based on the tacit background assumption that an experimental manipulation of moral norms is construct-valid only if the

[4]Other issues are also important, such as whether norms are considered absolute (Baron & Ritov, 2009) or "prima facie" (Ross, 1930) and whether they are agent-relative or limited in scope (Baron & Miller, 2000).

behavioral effect of this manipulation is driven by conscious thoughts about moral norms." This critique is misguided in our view.

For one thing, Getal argue that we ignore "one of the most significant contributions to moral psychology in the 21st century: the idea that norm-congruent judgments may not necessarily be the product of conscious thoughts about norms." They cite Greene's dual-process model as evidence.

The claim that unconscious effects of norms follows from some sort of dual-process model may be consistent with some extreme claims in the literature (such as those of Haidt, 2001, which have been questioned even for the extremely unusual examples he studied by Royzman, Kim & Leeman, 2015), but we cannot find any claim in Greene's writing that the effects of norms (or consequences) are unconscious. It is quite clear from a variety of results that the dilemmatic quality of the sorts of items studied in the literature becomes consciously apparent early on. People who read an entire item with understanding immediately realize the presence of a conflict (Białek & De Neys, 2017). The idea of a corrective (or sequential) dual-system model is that system-1 produces an intuitive response, and, after that, system-2 sometimes corrects it. But the system-1 response is completely conscious. It is simply un-mediated, or immediate, in the same way that our response to "2+2" is immediate, as distinct from our response to "21 times 43", which, for most of us, requires a sequences of steps. The intuitive response to "2+2" is certainly not unconscious and unavailable. Thus, for a moral dilemma in which the action is killing a comatose patient in order to use his organs to save five others, the first reaction may well come after full understanding of the dilemma, and it will still be "Yikes! That's awful." This response will of course be consciously available, and it will result from full understanding of the relevant moral norm. If a subject is asked to identify a relevant norm in this case, it will not be at all difficult (some variant on "Do not kill innocent people"). Nor would it be difficult to identify the consequences of each course of action. Yet, despite all of these details being transparent and accessible, subjects who read the scenario will still see that it is a dilemma, albeit an easy one for many.

For what it is worth, there is also considerable evidence against the corrective dual-system model for moral judgment (Baron & Gürçay, 2017; Gürçay & Baron, 2017; Koop, 2013), as we noted in our comment).

3.2 What defines norms and consequences?

Getal argue that: ". . . if norm-congruent behavior would qualify as an effect of norms only if it is mediated by conscious thoughts about norms, conscious thoughts about norms would not explain norm-congruent behavior, because the two would be conceptually identical."

This argument is particularly puzzling. If "conscious thought" means identification of what the relevant norms are, and whether they favor action or inaction, these are not equivalent to judgments favoring action in the dilemma. The norms can still be overridden by consequences. So "norm-congruent behavior" need not result merely from a conscious

recognition of the relevant norms. For example, a subject in the NormAct.ConsAct case might judge that the norm favors inaction, but that the consequence, which favors action, outweighs the norm, so that action is preferred, which is "norm-incongruent behavior" from the perspective of the experimenters.

Getal also explain how they derive the three new cases from the basic case (NormOmit.ConsAct), as an argument that they, not the subjects, determine the implications of the norms and consequences for each item. The implication seems to be that they, the experimenters, know what norms are in play, and the subjects do not know. But, how do the experimenters know that their implementation of their plan was correct? We have argued that the norms in several cases are ambiguous, as are the consequences. So we too must be engaging in some sort of distortion of the true norms.[5] If the experimenters know the right answers about the norms and consequences for each item, they seem to have some sort of insight that others, including us and the subjects, lack.

Getal want to define norms and consequences not in terms of anyone's judgments but in terms of "behavioral effects". What they seem to mean is that the N and C parameters should each be positive: "if participants did not respond in a manner 'as if' they agreed with the assumptions underlying our operationalization of moral norms, the N parameter should not significantly differ from zero." But this seems to be a weak test. All it requires is for a majority of subjects either to agree with the experimenters' classification, or to make judgments that coincide with it, even if those judgments are made for entirely different reasons. Moreover, the presence of a substantial minority who deviate from the experimenters' classification could affect the conclusions drawn from each experiment.

Getal also argue that, "if participants did not respond in a manner 'as if' they agreed with the assumptions underlying our operationalization of moral norms, the CNI model should be unable to provide accurate descriptions of the data, and thus show poor model fit across studies." Yet it is unclear how different interpretations of the norms involved could lead the model to fail. If the distortions we describe happen across several items, the model could fit quite well.[6]

Likewise, the suggestion that classification should be based on "intrinsic aspects of the focal action" runs into trouble when several conflicting aspects exist. Which are "intrinsic"?

The problematic fact, as outlined in our original critique, is that many of the dilemmas used by Getal contain more than one norm and more than one consequence, for example, a norm against overriding a superior and a norm requiring the prevention of harm; or a consequence of harm and a consequence of weakening lines of authority. Many of the consequence manipulations do not so much remove bad consequences as weaken them, so

[5]Recently, J.B. was trying to remember how each item was classified. He thought he could do this by reading the text of the items (shown in the Appendix to this paper). After trying this with a few items, he gave up. Some items were so ambiguous that he could not determine on his own what the "right answer" according to norms, and sometimes consequences, was supposed to be. He had to fall back on the table in GACFH.

[6]We mistakenly emphasized differences among cases in making this point originally. Here we call attention to another point we made, which is that some of these distortions are systematic, applying to all cases. For example, the reversed norm is usually weaker than the original norm.

that their seriousness becomes a matter of subjective judgment. Without these ambiguities — which are largely absent from traditional dilemmas concerning vaccines, trolleys, or truth telling — most perverse responses would probably disappear. It would then become more difficult to estimate the I parameter. We return to this issue in Section 5.

The perverse responses could have the following possible causes:

- Inattention. We agree with Getal on this point, and also that there are other ways to check for inattention.
- Non-serious responding, intentional sabotage, or anti-social responses.
- Action/inaction biases, which is what the CNI model assumes.
- The subjects, even on reflection, disagree that the items are congruent. This is the result of ambiguity in the items, which allows subjects to interpret the items differently.
- Random processes that lead subjects to produce responses that they would probably correct on reflection. These responses are more likely when the differences between the two options in norms or consequences are small (see Section 7).

Note that "random processes" do not require that subjects systematically re-interpret the norms and consequences. When two competing responses differ only a little in the strengths of the arguments for each, responses will be less consistent. Thus, very clear dilemmas, in which there was no ambiguity and no room for misinterpretation of both the norm and the consequence, would produce very few perverse responses. Even fewer perverse responses would occur if the differences between choice options in terms of their norms and consequences were large in all items. In these situations, although the perverse responses could serve as a measure of inattention, the N and C parameters would depend on the ordering in the model. The C parameter (which is now first) would be nearly identical to the result in the standard cases, and the N parameter would be close to 1 (since the N parameter represents the conditional probability of control by norms given that the response is *not* controlled by consequences, which would be most of the remaining responses).

4 Ordering of C and N

Getal now agree with us that the ordering of C and N in the model is arbitrary. They correctly criticize our analysis for not actually fitting the CNI model. So they report the results of such fits with both orderings in Table 4. Indeed, the results in their paper are ordinarily the same in every case.[7]

[7]We are not concerned with p-values here, as they depend on sample size and on how they are estimated. We argue that most of the original ones are deceptively low because they ignore subject variance, item variance, and within-subject "random slopes".

Yet this is an empirical fact and not a necessary one. The numbers in their Table 4 do change, and, in principle, it is possible that in a close case they could reverse direction.

Getal also argue that their ordering (C first) is consistent with a corrective dual-process model, on the assumption that whether the correction occurs is independent of the results of the N process. Putting this in temporal terms (which Getal admittedly resist doing), it is as if the subject decides in advance whether to proceed to system-2 or not. This is, we admit, possible, although it is inconsistent with current views of the way in which system-2 corrections occur (Glöckner & Betsch, 2008; Ackerman & Thompson, 2017; Thompson, Turner & Pennycook, 2011; Pennycook, Fugelsang & Koehler, 2015).

Perhaps more seriously, suppose we are correct in claiming that most perverse responses to the congruent items result from reasonable interpretations of these items as not actually congruent. These interpretations arise from the ambiguity of the items themselves. If the items were unambiguous, so that perverse responses were rare, then the interpretation of any results will depend largely on the ordering of N and C in the model. For example, suppose that an experiment manipulated only consequences, e.g., by changing the number of deaths (Trémolière & Bonnefon, 2014) and nothing else, or only norms, e.g., by changing the directness of the killing (Green et al., 2009). Both of these manipulations would affect the responses to the incongruent dilemmas. But both effects would be attributed almost entirely to C if C came first in the tree, and almost entirely to N if N came first.

Why assume any ordering at all? Why bother to have to demonstrate time after time that the conclusions are independent of the ordering? A variety of models can analyze situations in which effects are exerted in parallel (e.g., Section 7.)

5 The logic of the CNI model

The CNI model depends only on response probabilities (not response times), so the correction for response bias depends on the existence of fair numbers of perverse responses. If these responses are largely eliminated by the use of unambiguous dilemmas, then there is very little to correct.

To see this, consider how the C and N parameters (also known as U and D, respectively) are calculated. As we noted, each parameter is, in effect, estimated twice, thus allowing for a test of consistency within the four cases of each scenario.[8]

One estimate of the C parameter is the difference between NormOmit.ConsAct (an incongruent case) and NormOmit.ConsOmit (congruent). The latter condition controls for the possibility that the choice of action in the former is the result of a bias toward action. The other estimate of the C parameter is the difference between NormAct.ConsAct and NormAct.ConsOmit. Here, the former, congruent, action response would have a probability less than 1 if there is a bias toward inaction. A large difference would indicate sensitivity

[8]Our separation of the model into two PD-type models was admittedly a misguided attempt to put these estimates into two groups. But here we look at the estimates themselves.

to consequences. Presumably, with a linear probability assumption, the best estimate of the C parameter would be the average of these two differences.

In both cases, if there are no perverse responses, the C parameter would be estimated as NormOmit.ConsAct and 1−NormAct.ConsOmit. In the standard paradigm, only NormOmit.ConsAct would be used. In principle, the second test would be useful, but it depends heavily on NormAct being equivalent in strength to NormOmit. As we have argued, it is difficult to find norms that can be matched in this way.[9]

Our concern is that the CNI approach, as used, includes cases that are not well matched. Many cases involve multiple conflicting norms and multiple conflicting consequences. The conclusions could depend on the particular cases used. Only through luck, or perhaps stringent pretesting, could a particular sample of cases arrive at the correct balance of strength between NormAct and NormOmit, and we have argued that the method of constructing NormAct cases by undoing someone else's choice is open to systematic error, as is the method of constructing ConOmit cases by making the consequences of omission less obviously bad.

To get an idea of how the asymmetry comes about, consider how the items are generated (from Getal): "In a first step, we identified pairs of morally relevant actions and inactions that have the same outcome (e.g., killing Person A and letting Person A die both result in the loss of Person A's life). Whereas the identified action within each action-inaction pair was conceptually linked to a proscriptive norm (i.e., killing someone is morally prohibited), the opposite of the identified inaction was conceptually linked to a prescriptive norm (i.e., saving someone's live is morally prescribed)." The difficulty here is these prescriptive norms are generally weaker and more limited in their application, as we originally argued. (Appendix A lists all the original items.) For example, if we were obliged to save all lives, we could do nothing else without shirking this obligation. The obligation not to kill (intentionally) does not have this problem. With the exception of case 1 (admittedly a problematic case), four of the actions that are supposed to be prescribed involve overturning (or trying to overturn) someone else's decision.[10] Case 4 involves changing the status-quo (preventing death from a natural heart attack when the death is desired). Thus, these five cases involve competing normative considerations that are not present in the case of prohibitions (except for Case 1, which involves undoing); the prescribed action to overturn a prior decision conflicts either with a norm against interfering or with a status-quo effect. This conflict weakens the strength of the prescriptive norm. Some subjects may simply decide that the competing norm is more important.

Getal write further that, "In a second step, we generated hypothetical consequences of the two actions that involve costs for the well-being of others that are either greater or smaller than the benefits of each action." Again, this method suggests that the differences

[9]Such cases could be studied on their own. Arguably, some dilemmas used by Baron et al., 2015, such as #7 and #9 in Study 2, are of this form, as well as some used in Study 1. And we had one of these in Baron and Goodwin (2020).

[10]In case 5, involving taking a student out of quarantine, presumably someone else has decided to put the student in quarantine, although this prior decision is not explicit.

in consequences in the base case (NormOmit.ConsAct) and the newly generated case are not matched in terms of the "difficulty" of overcoming a norm to draw the utilitarian conclusion.[11] In each base case, there is a clear and unambiguous difference in consequences between the two options, as in the usual sacrificial dilemmas in which the question is whether to kill one person to save many others, or do nothing, in which case the others will die. The modification of the consequences involves making them less severe but still bad, e.g., instead of many people dying, many people get stomach aches. So it becomes easier to decide to allow these bad consequences to happen, but they are still bad. From a utilitarian perspective, the task is now a trade-off requiring judgment. How many stomach aches is as bad as one death? Surely the answer is "a lot more than what is implied by the scenario", but the number need not be infinite. In contrast, no judgment is required to conclude that five deaths is worse than one death.

The case of death vs. stomach aches (from Case 5, "Immune deficiency") is very clear, but some of the others are not so clear. The Appendix shows the details of what was varied, with the consequences in bold font. In each of the newly generated cases, the supposedly better consequence is not as different from the alternative as it is in the original cases from which these new cases were derived. This lack of difference could lead to perverse responses to the congruent cases, either because of random factors affecting the resolution of close judgments or because of interpretations that would differ from those intended even after reflection.

For example, in Case 2, the NormOmit.ConsOmit (congruent) item asks about killing a comatose patient who will (presumably) die anyway in order to transplant organs into five other people to prevent unspecified "health problems". A subject could respond perversely (favor action) either because the norm against killing a comatose patient who will die anyway is weak or non-existent (especially if the norm against killing an individual is justified by the consequences for that individual alone), or because the sorts of health problems that are treated with organ transplants (such as needing dialysis or being severely disabled with heart or lung disease) can be quite serious.

In the NormOmit.ConsOmit item Case 4, the action was assisting a patient who begs for help in committing suicide when the patient would recover from the extreme pain that was causing his agony. Presumably the patient knows he will recover. Although paternalism would favor not providing the help, the consequences of going against the patient's wishes at least raises some concern.[12]

In the NormOmit.ConsOmit version of Case 6, the vaccine will save the same number as it kills. Thus, the consequences of acting and not acting are at first blush the same, which is a smaller difference than in the NormOmit.ConsAct version, in which the vaccine will cause dozens of deaths but also save hundreds of lives.

[11] See Baron and Gürçay (2017) for evidence of the effect of this sort of variation.

[12] There could also be a norm against going against someone's wishes. Such norm would have the same effect as the consequence.

The NormAct.ConsAct congruent items usually have weaker norms because of the switched action, and they have similar problems with the consequences.

Getal say, "Although we agree that the asymmetry between proscriptive and prescriptive norms is fundamentally important, it is irrelevant for the construction of CNI model dilemmas to the extent that the proximal outcomes of a given action-inaction pair can be held constant (e.g., loss of the same life as a result of killing or letting die)." However, if consequences are held constant, while norm strength varies, then this would seem to distort the estimate of how sensitive people really are to the relevant norms and consequences — it is inaccurate to say that people are more "sensitive to consequences" if they choose in favor of the consequences when those consequences are opposed by a weak norm, and less sensitive if they choose differently when the consequences are opposed by a strong norm. Subjects' actual sensitivity to consequences may be identical across the two cases, since the comparison is unbalanced. Moreover, Getal's objection does not address the other fundamental problem we have raised, which is that across the new vignettes, the alternative consequences vary, as do the alternative norms. Thus, the inferential problem results both from the difference in the norms between the two alternatives presented, as well as the difference in the consequences.

If the two different kinds of congruent dilemmas are not symmetric with regard to the strength of their norms and their consequences, to the point where subjects could respond as if they disagree with the experimenters about the interpretation of each dilemma, then this asymmetry could have two effects. One is to distort the measure of inaction bias, so that it is affected by the interpretation of the dilemmas as well as by the bias itself. The other is to affect the measures of norm and consequence effects, so that (for example) group differences in these effects could be enlarged or obscured by group differences in the interpretation of cases that allow some ambiguity, as we illustrate now.

Consider a numeric example of how distortion can result from ambiguous cases of the sort we discuss. First, the following table is a reminder of how the probability of action depends on the parameters in the four cases of interest.

W	NormOmit.ConsAct	=	C +		(1-C)(1-N)A
X	NormOmit.ConsOmit	=			(1-C)(1-N)A
Y	NormAct.ConsAct	=	C +	(1-C)N +	(1-C)(1-N)A
Z	NormAct.ConsOmit	=		(1-C)N +	(1-C)(1-N)A

Here, the letters W-Z indicate shorthand for how we will refer to these formulae. C and N are the parameters of the model. A is the probability of action responses when the last step is reached, that is, when neither the C nor the N process is activated. (In the CNI model, the I parameter is the probability of inaction, so it is $1-A$.)

Thus, estimates of the C and N parameter are the averages of two estimates of each parameter, and the estimate of the A parameter is the average of four estimates. In listing these we include N_{raw}, which is the N parameter uncorrected for the fact that N is evoked only if C is not evoked.

C = (W-X + Y-Z)/2
N_{raw} = (Y-W + Z-X)/2
N = N_{raw}/(1-C)
A = (W-C + X + Y-C-N_{raw} + Z-N_{raw})/(4*(1-C)*(1-N))

Here, the separate estimates are indicted by spaces around "+", and the denominator includes the number of estimates averaged (2 or 4), as well as any correction required for prior steps in the model.

To examine effects of lack of symmetry, consider first a "classic" case in which people favor action in .5 of the W and Z cases (the two conflicting cases, NormOmit.ConsAct) and NormAct.ConsOmit, respectively). We assume that they also make .1 perverse responses in the two congruent cases (NormOmit.ConsOmit and NormAct.ConsAct) so that X is .1 and Y is .9.

From these assumptions we calculate[13] the parameters shown in row a of the following table.[14]

	W	X	Y	Z	C	N_{raw}	N	A
a	.5	.1	.9	.5	.40	.40	.67	.50
b	.3	.1	.7	.5	.20	.40	.50	.25
c	.5	.1	.7	.3	.40	.20	.33	.25
d	.3	.1	.5	.3	.20	.20	.25	.17
e	.6	.1	.9	.4	.50	.30	.60	.50
f	.6	.3	.8	.4	.40	.20	.33	.50

Row a is the case just described. In row b, we assume that ConsAct is ambiguous. (Recall that this is likely if the consequences of inaction are only a little better than the consequences of omission in the ConsAct case, while the corresponding difference in consequences is stark in the matched ConsOmit case.) The two places where this would matter are W and Y in the table above. W would be lower than .5. Assume that W is now .3 and Y is .7, a drop by .2 for both. From these values of W−Z, we can now re-estimate the apparent values of C, N, and A, as shown on the right side of row B. (Note that the "true" values are unchanged.) This ambiguity leads us to conclude that both N and C are reduced from the classic case (row a), and that the bias toward action is lower. (Hence the bias toward inaction, I, is higher). Note that, in the classic case, N_{raw} and C were equal, but now C is

[13]Using the following R code:

```
C <- function(W,X,Y,Z) (W-X + Y-Z)/2
Nraw <- function(W,X,Y,Z) (Y-W + Z-X)/2
N <- function (W,X,Y,Z) Nraw(W,X,Y,Z)/(1-C(W,X,Y,Z))
A <- function(W,X,Y,Z) (W-C(W,X,Y,Z) + X +
  Y-C(W,X,Y,Z)-Nraw(W,X,Y,Z) +
  Z-Nraw(W,X,Y,Z))/(4*(1-C(W,X,Y,Z))*(1-N(W,X,Y,Z)))
```

[14]Note that the equal responses to the conflicting cases W and Z suggest some sort of equality of C and N, this is true only for C and N_{raw}, which is why we display both of these. N is increased because it is corrected for the fact that it can be invoked only if C is not invoked.

lower than N_{raw}, and the decline in C is greater than the decline in N (either in absolute difference or proportionally); if some group were more sensitive to the ambiguity, then we might conclude that this group differed from the original less-sensitive group more in C than in N.

Row c shows the effect of weakness in NormAct. (Recall that this could happen systematically if it is difficult to make up prescriptive norms that are as strong as prohibitions.) This would reduce the action responses in Y and Z, the two cases that include NormAct. Again, we reduce the proportions by .2. This weakness leads us to include that N is now lower than C, rather than higher (as it is in the classic case), and that the bias toward action is lower.

Finally, row d combines both effects. All parameters are now lower. While N remains higher than C, it is not much higher.[15]

Thus, an apparent effect of an experimental manipulation or group difference on N, C, or I (which is 1−A) could result from how people interpret the cases, which, as a result of how they were constructed, could differ in their susceptibility to misinterpretation (i.e., interpretation different from the researchers' intentions). Note that, if these effects happened similarly to several cases, the CNI model could fit just as well in rows b, c, and d, as in row a.

Rows e and f show the effect of simply increasing the number of perverse responses (from .1 to .2, symmetrically for both congruent cases). In row e, N is larger than C; in row f, C is larger. This is a consequence of the ordering of N and C in the model. Thus, in principle, if groups differ only in their tendency to make perverse responses, they will appear to differ more in N than in C, unless the order of these two components is reversed.

6 An example: Psychopathy

To illustrate some of these effects in action, we discuss some of the data from Luke and Gawronski (in press), which was cited by Getal to illustrate the benefits of the CNI design for the analysis of individual and group differences. In this case, the measure of interest was a scale measuring psychopathy, which has been found in several studies to correlate with more utilitarian choices in the standard dilemmas. Application of the CNI design led to the conclusion that psychopaths (those with higher scores on the scale) were lower in both C and N, and lower in I (hence higher in A, action bias, as we define it). We agree that this is one of the more interesting applications of the CNI design, as (unlike many reported in GACFH), the effects of the external variable of interest are quite large.

We looked at the data and found that psychopaths gave a much higher number of perverse responses in the congruent cases. Figure 1 shows the best fitting lines relating action responses to NormAct.ConsAct (red) and NormOmit.ConsOmit (black/gray) cases,

[15]The results in the table can be checked by using the values of C, N, and A in the original model equations.

with each circle representing a subject.[16] At the highest levels of psychopathy, subjects do not distinguish the two congruent cases, suggesting that they are completely insensitive to both consequences and norms as defined by those cases. The apparent reduction in both C and N with psychopathy may be the result of some factor that increases the number of perverse responses.

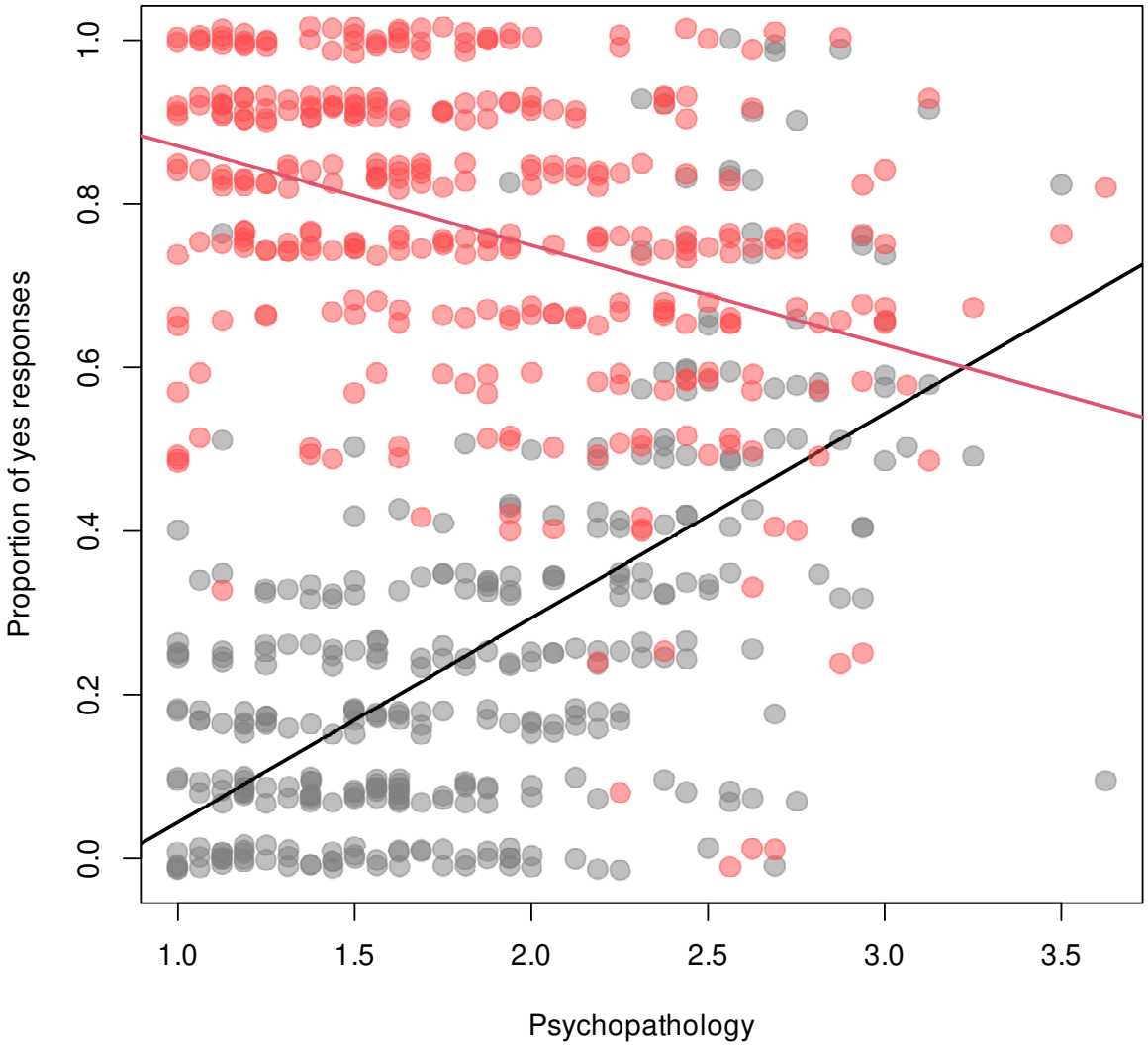

FIGURE 1: Action responses to congruent cases as a function of psychopathy score. Red indicates NormAct.ConsAct; black/gray indicates NormOmit.ConsOmit.

We measured the overall utilitarian vs. deontological tendency, UvsD, as W−Z, the action responses to the two incongruent dilemmas, that is NormOmit.ConsAct minus NormAct.ConsOmit. When we regressed UvsD, on psychopathy, the coefficient was positive (.11, using the scales as indicated in Figure 1) and significant (p=.001), as might be expected from previous results. However, when we included the proportion of perverse responses in

[16]We used only the conditions in which subjects answered, "Would you find it acceptable . . . " rather than "Would society find it acceptable . . . ".

the model, the coefficient declined to −.01.[17] This leaves us in a somewhat unclear situation with respect to the nature of the effect of psychopathy on moral judgment. We are not yet convinced that psychopathy reduces sensitivity to consequences in the standard dilemmas, in which the consequences are much more clearly ascertainable.[18]

6.1 An alternative model

We used the same data to examine model fits.[19] As an alternative, we tested a single somewhat over-simplified model based on the argument we made in Section 5. Specifically, we assumed that the standard case (NormOmit.ConsAct, W) was clear and thus did not evoke any responses based on the ambiguity of norms or consequences. We also assumed that action responses to the NormOmit.ConsOmit congruent cases were entirely the result of the ambiguity in consequences, so that the proportion of these responses (X) measured the effect of that ambiguity for each subject. And we assumed that inaction responses to the NormAct.ConsAct congruent cases measured (as 1−Y) the effect of ambiguity in the norms (of the sort that might result from requiring that someone else's decision be overridden). The final case, NormAct.ConsOmit potentially involves both kinds of ambiguity. Because it is difficult to say how these two sources would interact, we did not use this case to estimate the two ambiguity effects, but we did predict the results of this case by assuming that the combined effect would simply be the difference in the two ambiguity parameters. Although mean-squared error of prediction was greater in our model (.010) than in the CNI model (.004), our model was more accurate by other criteria that we think are reasonable. Specifically, the mean absolute error (not squared) was lower (.035 vs. .042), as was the median error (0 vs. .031 absolute); our alternative model fit 77% of the subjects exactly, compared to 9% for the CNI model. In sum, other models might fit the data as well as or better than the CNI model.

7 Tree models vs. drift-diffusion

A more general problem, which we did not emphasize in our original comment, is that the multinomial-processing-tree model might just not apply to moral judgment tasks of the sort at issue. An alternative type of model is the drift-diffusion model (DDM, e.g., Ratcliff, 1978), which has been applied to moral judgment tasks (Baron & Gürçay, 2017). The

[17]We used regression rather than correlation because we interpret UvsD as a difference in probabilities. Correlations are affected not only by the effect on such a difference but also by the relative error, which could, for example, be higher when the range of variation in the true effect is lower. Luke and Gawronski report correlations. This approach is also inconsistent with the analyses used in GACFH, which examines actual differences in the parameters.

[18]Nor are we convinced that psychopathy is correlated with utilitarianism in other sorts of dilemmas that do not involve pitting numbers against some action that is chosen because it is emotionally repugnant.

[19]The data had enough items to allow model fits for each subject, and the C, N, and I parameters for each subject were included.

general idea is that responses are generated by the accumulation of "evidence" or "reasons" for each possible response, over time (in this case seconds or minutes). When a threshold or boundary is reached, the response is made. The model applies to moral conflicts between consequences and other norms in the following way. As the subject considers the problem, each possible response accumulates strength in a predictable but somewhat random way, and the difference between their strengths eventually reaches a boundary (like a drunk wandering down a football field — eventually reaching one of the side boundaries through a series of random steps). Thus, any manipulation of the strength of norms or consequences will be reflected in the difference between the two strengths, and thus in the probability of reaching a boundary on one side or the other, as well as in the response time (RT). Time pressure can cause the boundaries on both sides to tighten, so that the random process plays a greater role and the responses become less consistently predictable. Models of this sort are useful in many cognitive tasks that involved two choices.

Tree models, by contrast, assume that, in cases involving conflict, the subject attends to the evidence only on one side (whichever comes first in the model). In this case, the evidence on the other side has no effect at all. Such results are observed in multi-attribute choice tasks, in which subjects must actively sample the information (e.g., by opening a small window with a pointer on a computer); in some cases, a subject will sample only some of the relevant attributes before making a choice, so that the values of the other attributes have no effect at all when this happens.

In the case of moral dilemmas, tree models would make sense when some attributes are considered absolute (e.g., Baron & Spranca, 1997). For example, someone who had an absolute norm against active killing might simply stop reading the dilemma if this information came first, before the information about the number of people that could be saved by killing one. The latter attribute would not be considered. (The opposite pattern is more difficult to imagine. Even a strict utilitarian would want to know about the action required, since it would have utilities of its own.) Yet evidence from a wide variety of studies of moral dilemmas, in which otherwise identical dilemmas with different numbers are presented to each subject, suggest that such neglect is extremely rare. Subjects who do this would favor the deontological response in every single case, even when the number of people saved is enormous, and this is rare (Trémolière & Bonnefon, 2014).

When subjects do not ignore information about either horn of the dilemma, then the critical variable is the difference between the strengths of the opposing responses. Any manipulation of norms or consequences will manifest itself only through the difference. The CNI model will not tell us about the relative sensitivity of the two relevant processes. It may still fit the data well, but, as we have argued, the use of the CNI model to look for effects on N or C will either depend on a fairly large number of perverse responses, or, if the number of perverse responses is small, on whether N or C comes first in the tree.

It is difficult to apply the DDM to moral judgments, since this requires sufficient data from each subject. One way to apply it might be to use a small number of norms, crossed

with a larger number of consequences described numerically, generating enough situations so that they can be repeated a few times without subjects remembering what they said before to the same case. With such a design, it would be possible to use response times to distinguish various parameters, such as the drift rate difference between "yes" and "no", and the starting-point bias relative to the boundaries, which could be interpreted as an action/inaction bias. The distance of the boundaries would reflect willingness to respond quickly at the expense of errors (i.e., responses that would be different if more time were allowed for the accumulation of evidence).

8 Conclusion

The four-fold design developed by GACFH does accomplish something. It uses both "proscriptive" and "prescriptive" norms, in the same study. This could be useful, although trying to match them, as we argued, is difficult. Yet some efforts have been made to study prescriptive norms. Spranca et al. (1991) began with dilemmas in which the norm was to say something to prevent harm to another person, as opposed to staying silent (an omission). Many other papers have followed in this tradition (e.g., Baron, Gürçay & Luce, 2018).[20] We agree that the study of prescriptive norms should continue.

In most of the literature, the norms and consequences have been clear. This is possible because there was no need to fill out the complete design. A major topic of research, mostly concerning proscriptive norms, has concerned the nature of the norms themselves. Dilemmas have been used to carry out controlled studies of the nature of particular norms, such as the norm prohibiting the killing of humans, looking at factors such as perceived causality and protected values (Baron & Ritov, 2009; Greene et al., 2009).[21]

Getal have provided helpful answers to some of our concerns. In particular, we are close to agreement on whether action/inaction biases are deontological, and on the importance of testing generality across cases.

Two other major concerns remain, both of which arise from the difficulty of constructing the full set of examples needed to apply the CNI model. When the model is used to estimate effects of experimental manipulations or group differences on sensitivity to norms or consequences, the results may be confounded by the effects that different manipulations or individual differences have on the interpretation of the congruent cases.

We do not accept the view that the researchers' intention determines the correct answers. The dilemmas must be understood in the same way by the subjects.

The second concern is that, if the congruent cases were so clear as to avoid interpretations that disagreed with the experimenters' intentions, then all the perverse responses would result from inattention or some sort of lack of cooperation. These responses would be

[20]Likewise, they are certainly not the first to use realistic dilemmas (e.g., Ritov & Baron, 1990; Baron et al., 2018).

[21]Some studies have examined the nature of prescriptive norms as well as proscriptive norms (e.g., Spranca et al., 1991; Baron & Ritov, 2009).

used primarily to eliminate subjects, not to change the estimates of sensitivity to norms and consequences. The basic dilemmas would suffice for that.

We do agree that there are questions that remain to be answered about just what properties of "norms" and "consequences" are relevant to which people.

References

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Science, 8*, 607–617.

Baron, J. (1996). Do no harm. In D. M. Messick & A. E. Tenbrunsel (Eds.), *Codes of conduct: Behavioral research into business ethics*, pp. 197–213. New York: Russell Sage Foundation.

Baron, J. & Goodwin, G. P. (2020). Consequences, norms, and inaction: A comment. *Judgment and Decision Making, 15*(3), 421–442.

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

Baron, J. & Miller, J. G. (2000). Limiting the scope of moral obligation to help: A cross-cultural investigation. *Journal of Cross-Cultural Psychology, 31*, 705–727.

Baron, J., & Ritov, I. (2009). Protected values and omission bias as deontological judgments. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral Judgment and decision making*, Vol. 50 in B. H. Ross (series editor), *The Psychology of Learning and Motivation*, pp. 133–167. San Diego, CA: Academic Press.

Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes, 70*, 1–16.

Baron, J., & Szymanska, E. (2010). Heuristics and biases in charity. In D. Oppenheimer & C. Olivola (Eds). *The science of giving: Experimental approaches to the study of charity*, pp. 215–236. New York: Taylor and Francis.

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

Cushman, F. (2015). From moral concern to moral constraint. *Current Opinion in Behavioral Sciences, 3*, 58–62.

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

Gawronski, B., Conway, P., Hütter, M., Luke, D. M., Armstrong, J., & Friesdorf, R. (2020). On the validity of the CNI model of moral decision-making: Reply to Baron and Goodwin (2020). *Judgment and Decision Making, 15*.

Glöckner, A., & Betsch, T. (2008). Modelling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making.

Judgment and Decision Making, 3, 215–228.

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*, 364-371.

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

Koop, G. J. (2013). An assessment of the temporal dynamics of moral decisions. *Judgment and Decision Making, 8*, 527–539.

Luke, D. M., & Gawronski, B. (in press). Psychopathy and moral dilemma judgments: A CNI model analysis of personal and perceived societal standards. *Social Cognition*.

McCulloch, K. C., Li, H., Hong, S., & Albarracin, D. (2012), Naïve definitions of action and inaction: The continuum, spread, and valence of behaviors. *European Journal of Social Psychology, 42*, 227–234.

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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.

Ritov, I., & Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making, 3*, 263–277.

Roberts, S. & Pashler, H. (2000) How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358–367.

Ross, W. D. (1930). *The right and the good.* (Reprinted 2002 by Oxford University Press, Oxford.)

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*, 296–313.

Shallow, C., Iliev, R., & Medin, D. (2011). Trolley problems in context. *Judgment and Decision Making, 6*, 593–601.

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

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*, 923–930.

Ubel, P. A., DeKay, M. L., Baron, J., & Asch, D. A. (1996). Cost effectiveness analysis in a setting of budget constraints: Is it equitable? *New England Journal of Medicine, 334*, 1174–1177.

Appendix A: Cases from GACFH. Bold font indicates manipulation of consequences (ConsAct vs. ConsOmit).

1. Abduction

NormOmit.ConsAct (W)	NormOmit.ConsOmit (X)
You are the president of your country. A guerilla group operating in a conflict zone has abducted a journalist from your country and threatens to behead him if your government does not pay a ransom of one million dollars. The group will use the ransom money to **buy weapons for their guerilla war**, which will cause the deaths of many people. Congress has approved payment of the ransom, but you have the power to veto the payment. Is it acceptable in this case to veto the ransom payment?	You are the president of your country. A guerilla group operating in a conflict zone has abducted a journalist from your country and threatens to behead him if your government does not pay a ransom of one million dollars. The group will use the ransom money to **buy food for their families**, who live in an area that has been plagued by several draughts. Congress has approved payment of the ransom, but you have the power to veto the payment. Is it acceptable in this case to veto the ransom payment?
NormAct.ConsAct (Y)	NormAct.ConsOmit (Z)
You are the president of your country. A guerilla group operating in a conflict zone has abducted a journalist from your country and threatens to behead him if your government does not pay a ransom of one million dollars. The group will use the ransom money to **buy food for their families**, who live in an area that has been plagued by several draughts. As the president, you have the power to approve payment of the ransom. Is it acceptable in this case to approve the ransom payment?	You are the president of your country. A guerilla group operating in a conflict zone has abducted a journalist from your country and threatens to behead him if your government does not pay a ransom of one million dollars. The group will use the ransom money to **buy weapons for their guerilla war**, which will cause the deaths of many people. As the president, you have the power to approve payment of the ransom. Is it acceptable in this case to approve the ransom payment?

2. Transplant

NormOmit.ConsAct (W)	NormOmit.ConsOmit (X)
You are a surgeon in a small hospital. One day, your hospital receives five badly hurt patients from a car accident. The patients all need organ transplants or they will die. You have no spare organs, but there is a patient who has been in a coma for several weeks and it seems unlikely that he will wake up again. You could terminate his life support and take his organs for the five accident victims, **so that their lives will be saved**. Is it acceptable in this case to terminate the patient's life support to take his organs?	You are a surgeon in a small hospital. One day, your hospital receives five badly hurt patients from a car accident. These patients all need organ transplants or they will have serious health problems for the rest of their lives. You have no spare organs, but there is a patient who has been in a coma for several weeks and it seems unlikely that he will wake up again. You could terminate his life support and take his organs for the five accident victims, **so that they won't suffer from health problems**. Is it acceptable in this case to terminate the patient's life support to take his organs?
NormAct.ConsAct (Y)	NormAct.ConsOmit (Z)
You are a surgeon in a small hospital. One day, your hospital receives five badly hurt patients from a car accident. These patients all need organ transplants or they will have serious health problems for the rest of their lives. You have no spare organs, but there is a patient who has been in a coma for several weeks and it seems unlikely that he will wake up again. One of your co-workers plans to terminate his life support and take his organs for the five accident victims, **so that they won't suffer from health problems**. You could stop your co-worker by informing the director of the hospital. Is it acceptable in this case to stop your co-worker from terminating the patient's life support to take his organs?	You are a surgeon in a small hospital. One day, your hospital receives five badly hurt patients from a car accident. The patients all need organ transplants or they will die. You have no spare organs, but there is a patient who has been in a coma for several weeks and it seems unlikely that he will wake up again. One of your co-workers plans to terminate his life support and take his organs for the five accident victims, **so that their lives will be saved**. You could stop your co-worker by informing the director of the hospital. Is it acceptable in this case to stop your co-worker from terminating the patient's life support to take his organs?

3. Torture

NormOmit.ConsAct (W)

You are a member of a special police department who is trained to obtain information in particularly difficult cases. You are dealing with a case involving a male adult who is accused of having abducted several children. You don't know where he is hiding the children, and he refuses to tell you where they are. **The children will likely die from dehydration if they are not found** within the next 24 hours. You have tried every legal interrogation technique, but none of them were successful. To get information on where the children are, you consider the use of illegal techniques that are deemed torture. Is it acceptable in this case to use illegal interrogation techniques?

NormOmit.ConsOmit (X)

You are a member of a special police department who is trained to obtain information in particularly difficult cases. You are dealing with a case involving a male adult who is accused of having stolen several paintings. You don't know where he is hiding the paintings, and he refuses to tell you where they are. **The paintings will likely be shipped to a different country if they are not found** within the next 24 hours. You have tried every legal interrogation technique, but none of them were successful. To get information on where the paintings are, you consider the use of illegal techniques that are deemed torture. Is it acceptable in this case to use illegal interrogation techniques?

NormAct.ConsAct (Y)

You are a member of a special police department who is trained to obtain information in particularly difficult cases. You are dealing with a case involving a male adult who is accused of having stolen several paintings. You don't know where he is hiding the paintings, and he refuses to tell you where they are. **The paintings will likely be shipped to a different country if they are not found** within the next 24 hours. You have tried every legal interrogation technique, but none of them were successful. To get information on where the paintings are, your partner starts using illegal techniques that are deemed torture, and you consider stopping him by reporting him to your supervisor. Is it acceptable in this case to stop your partner from using illegal interrogation techniques?

NormAct.ConsOmit (Z)

You are a member of a special police department who is trained to obtain information in particularly difficult cases. You are dealing with a case involving a male adult who is accused of having abducted several children. You don't know where he is hiding the children, and he refuses to tell you where they are. **The children will likely die from dehydration if they are not found** within the next 24 hours. You have tried every legal interrogation technique, but none of them were successful. To get information on where the paintings are, your partner starts using illegal techniques that are deemed torture, and you consider stopping him by reporting him to your supervisor. Is it acceptable in this case to stop your partner from using illegal interrogation techniques?

4. Assisted suicide

NormOmit.ConsAct (W)

You are a doctor and are treating a seriously ill, suffering patient. Even the strongest drugs do not relieve him from his pain anymore. **He feels terrible agony and you know that his condition will never improve.** For days, he has been waiting for his death. Since he no longer wants to endure his pain, he repeatedly asks you to end his life. You could give him a drug causing his death, which will release him from his pain. Is it acceptable in this case to provide your patient with this drug?

NormOmit.ConsOmit (X)

You are a doctor and are treating a seriously ill, suffering patient. Even the strongest drugs do not relieve him from his pain anymore. **He feels terrible agony, but you know that he will get better again soon.** For days, he has been waiting to see improvements. Since he no longer wants to endure his pain, he repeatedly asks you to end his life. You could give him a drug causing his death, which will release him from his pain. Is it acceptable in this case to provide your patient with this drug?

NormAct.ConsAct (Y)

You are a doctor and are treating a seriously ill, suffering patient. Even the strongest drugs do not relieve him from his pain anymore. **He feels terrible agony, but you know that he will get better again soon.** For days, he has been waiting to see improvements. Since he no longer wants to endure his pain, he repeatedly asks you to end his life. Suddenly, he has a severe heart attack. You could give him a drug to save him from dying. Is it acceptable in this case to provide your patient with this drug?

NormAct.ConsOmit (Z)

You are a doctor and are treating a seriously ill, suffering patient. Even the strongest drugs do not relieve him from his pain anymore. **He feels terrible agony and you know that his condition will never improve.** For days, he has been waiting for his death. Since he no longer wants to endure his pain, he repeatedly asks you to end his life. Suddenly, he has a severe heart attack. You could give him a drug to save him from dying. Is it acceptable in this case to provide your patient with this drug?

5. Immune deficiency

NormOmit.ConsAct (W)

You are the director of a hospital in a developing country. A foreign student who is volunteering in the country got infected with a rare virus. **The virus is highly contagious and deadly to seniors and children.** The only medication that can effectively stop the virus from spreading has severe side-effects. Although the virus will not kill her, the student suffers from a chronic immune deficiency that will make her die from these side-effects. Is it acceptable in this case to give the student the medication?

NormOmit.ConsOmit (X)

You are the director of a hospital in a developing country. A foreign student who is volunteering in the country got infected with a rare virus. **The virus is highly contagious and can cause severe stomach cramps.** The only medication that can effectively stop the virus from spreading has severe side-effects. Although the virus will not kill her, the student suffers from a chronic immune deficiency that will make her die from these side-effects. Is it acceptable in this case to give the student the medication?

NormAct.ConsAct (Y)

You are the director of a hospital in a developing country. A foreign student who is volunteering in the country got infected with a rare virus. **The virus is highly contagious and can cause severe stomach cramps.** The student suffers from a chronic immune deficiency that will make her die from the virus if she is not returned to her home country for special treatment. However, taking her out of quarantine involves a considerable risk that the virus will spread. Is it acceptable in this case to take the student out of quarantine to return her to her home country for treatment?

NormAct.ConsOmit (Z)

You are the director of a hospital in a developing country. A foreign student who is volunteering in the country got infected with a rare virus. **The virus is highly contagious and deadly to seniors and children.** The student suffers from a chronic immune deficiency that will make her die from the virus if she is not returned to her home country for special treatment. However, taking her out of quarantine involves a considerable risk that the virus will spread. Is it acceptable in this case to take the student out of quarantine to return her to her home country for treatment?

6. Vaccine

NormOmit.ConsAct (W)

You are a doctor in an area that suffers from an outbreak of a highly contagious disease. Preliminary tests have shown the success of a new vaccine that is not approved by the health department of your country, because of its severe side-effects. **The side-effects of the vaccine will likely cause the death of dozens of people who are not infected, but the vaccine will save hundreds of lives by preventing spread of the virus**. Is it acceptable in this case to use the vaccine?

NormOmit.ConsOmit (X)

You are a doctor in an area that suffers from an outbreak of a highly contagious disease. Preliminary tests have shown the success of a new vaccine that is not approved by the health department of your country, because of its severe side-effects. **The side-effects of the vaccine will likely cause the death of dozens of people who are not infected, but the vaccine will save about the same number of lives by preventing spread of the virus**. Is it acceptable in this case to use the vaccine?

NormAct.ConsAct (Y)

You are a doctor in an area that suffers from an outbreak of a highly contagious disease. Preliminary tests have shown the success of a new vaccine that is not approved by the health department of your country, because of its severe side-effects. **The side-effects of the vaccine will likely cause the death of dozens of people who are not infected, but the vaccine will save about the same number of lives by preventing spread of the virus**. One of your colleagues plans to use the vaccine, but you could stop him by reporting his plans to the health department. Is it acceptable in this case to report your colleague to the health department?

NormAct.ConsOmit (Z)

You are a doctor in an area that suffers from an outbreak of a highly contagious disease. Preliminary tests have shown the success of a new vaccine that is not approved by the health department of your country, because of its severe side-effects. **The side-effects of the vaccine will likely cause the death of dozens of people who are not infected, but the vaccine will save hundreds of lives by preventing spread of the virus**. One of your colleagues plans to use the vaccine, but you could stop him by reporting his plans to the health department. Is it acceptable in this case to report your colleague to the health department?

Appendix B

We used a CNI design in which the consequences were numbers of lives saved by medical interventions and the norm was equal treatment of two groups.

Method

We presented subjects with two sets of items, Screening and Vaccine, with four items each. The two basic item types, NormOmit.ConsAct, were:

> **[Screening:]** A health provider has a limited budget for screening tests. It plans to offer **all its members** a simple screening test suitable for a population at low risk of colon cancer. The test would save **1000 lives** in the next 5 years.
>
> A new, more expensive, test has just become available. It can be offered to **half of the members, selected at random**. The two tests differ in effectiveness. The new test would save **1100 lives** in the next 5 years.
>
> **[Vaccine:]** A developing country has a limited budget for COVID-19 vaccines. It can afford to offer **all adults** a vaccine that would reduce the total number of new cases by **30%**.
>
> A new vaccine has become available. The country could afford to offer it to **half of the adults, selected at random**. It would reduce the total number of new cases by **35%**.

We derived three other cases from each basic case by switching the two percents, and/or whether the vaccine was given to all or half of the population. For example, using the vaccine case:

> **[NormAct.ConsOmit:]** A developing country has a limited budget for COVID-19 vaccines. It can afford to offer **half of the adults, selected at random** a vaccine that would reduce the total number of new cases by **35%**.
>
> A new vaccine has become available. The country could afford to offer it to **all adults**. It would reduce the total number of new cases by **30%**.
>
> **[NormAct.ConsAct:]** A developing country has a limited budget for COVID-19 vaccines. It can afford to offer **half of the adults, selected at random** a vaccine that would reduce the total number of new cases by **30%**.
>
> A new vaccine has become available. The country could afford to offer it to **all adults**. It would reduce the total number of new cases by **35%**.
>
> **[NormOmit.ConsOmit:]** A developing country has a limited budget for COVID-19 vaccines. It can afford to offer **all adults** a vaccine that would reduce the total number of new cases by **35%**.

> A new vaccine has become available. The country could afford to offer it to **half of the adults, selected at random**. It would reduce the total number of new cases by **30%**.

The 8 items were presented in an order randomized for each subject. Subjects were 118 members of the same panel who did the studies described in Baron and Goodwin (2020). The median age was 48 (range, 22–84), and 65% were women.

After each case, the subject answered three questions:

> Which one should be chosen?
> Clearly the old one On balance, the old one On balance, the new one Clearly the new one.
>
> Which option is more effective on the whole?
> [same options]
>
> Which option is more fair on the whole?
> [same options]

We call these Choice, Effective, and Fair.

Results

TABLE 2: Responses to the 8 items.

	NormOmit ConsAct	NormAct ConsOmit	NormOmit ConsOmit	NormAct ConsAct
Scoring	−1.5	−0.5	0.5	1.5
Screening test				
NormOmit.ConsAct	28	42	30	18
NormAct.ConsOmit	18	34	43	23
NormOmit.ConsOmit	100	16	2	0
NormAct.ConsAct	1	3	16	98
Vaccine				
NormOmit.ConsAct	26	32	44	16
NormAct.ConsOmit	21	36	32	29
NormOmit.ConsOmit	97	17	1	3
NormAct.ConsAct	2	2	10	104

Table 2 shows all the responses. The perverse response categories are those with the small numbers (going down the rows): 2, 0, 1, 3, 1, 3, 2, 2. This is less than 3% of the total number of resonses to these items. Other evidence suggests that these responses were careless errors.

The number of perverse responses per subject was correlated −.30 (p = .001) with a meaure of response time (the log of the mean total time per page of the fastest 4 of the 8 pages). The three fastest subjects were among the 10 subjects who made any perverse responses.

We derived two estimates of action/inaction bias from the incongruent responses, without using the perverse cases. This is simply the sum of NormOmit.ConsAct and NormAct.ConsOmit for each of the two sets (Screening and Vaccine). These two measures were correlated across subjects (r = .31, p = .001). Thus, we can measure this bias without using the congruent cases.

An analogous measure for the congruent cases alone showed no apparent correlation (r = .07) and the combined measure for the congruent cases likewise did not seem to correlate with the measure for the incongruent cases (r = .05). These results are consistent with the conclusion that the perverse responses were not sensitive to act/omission biases (contrary to the assumption made by the CNI model).

The incongruent cases are also sensitive to the norm/consequence trade-off, the relative sensitivity to norms and consequences. The difference between ratings for NormOmit.ConsAct and NormAct.ConsOmit was correlated for the two sets (r = .65, p = .000).

In sum, when we eliminate most of the perverse responses by removing possible ambiguity of interpretation, we can indeed get two different measures from switching the assignment of action between favoring norms or consequences, as the CNI model tries to do. One measure is this trade-off between norms and consequences, the original intent of sacrificial dilemmas. The other is act/omission bias, which may indeed affect the results of the basic case.

This conclusion preserves one basic insight of the CNI model, which is that group (or experimental-condition) differences in apparent sensitivity to norms or consequences could result from differences in action/inaction bias.

Note that the example works because the norm in question is not specific to acts or omissions, unlike the norms used in sacrificial dilemmas, where conformity to a norm such as “Do not kill people actively” is confounded with a bias toward inaction in the basic case.