Rev.Phil.Psych. (2012) 3:351–368
DOI 10.1007/s13164-012-0106-3

Explaining the Abstract/Concrete Paradoxes in Moral Psychology: The NBAR Hypothesis

Eric Mandelbaum · David Ripley

Published online: 31 August 2012

Abstract For some reason, participants hold agents more responsible for their actions when a situation is described concretely than when the situation is described abstractly. We present examples of this phenomenon, and survey some attempts to explain it. We divide these attempts into two classes: affective theories and cognitive theories. After criticizing both types of theories we advance our novel hypothesis: that people believe that whenever a norm is violated, someone is responsible for it. This belief, along with the familiar workings of cognitive dissonance theory, is enough to not only explain all of the abstract/concrete paradoxes, but also explains seemingly unrelated effects, like the anthropomorphization of malfunctioning inanimate objects.

> In the minds of many peoples of the world, death would not take place without the push of human agency (Shweder et al 1997, p.130)

Recent research has uncovered a number of paradoxes in judgment created by pairs of stimuli that can be characterized as 'abstract' and 'concrete.' In this paper, we give a brief tour of the explanations that have been given for this phenomenon, followed by a new explanation of the data, one which not only explains these abstract/concrete paradoxes, but also sees the paradoxes as part of a broader phenomenon running through some strands of recent social psychological and experimental philosophy research.

E. Mandelbaum (✉)
Department of Philosophy, Emerson Hall, Harvard University, 25 Quincy St., Cambridge, MA 02138, USA
e-mail: eric.mandelbaum@gmail.com

D. Ripley
University of Connecticut, Storrs, CT, USA
e-mail: davewripley@gmail.com

D. Ripley
University of Melbourne, Melbourne, VIC, Australia

1 Abstract/Concrete Paradoxes

What we, following Sinnott-Armstrong (2008a), term 'abstract/concrete paradoxes' are pairs of vignettes that produce strikingly different responses from each other, where the crucial difference between the vignettes is in their 'concreteness'. What concreteness amounts to varies from case to case: sometimes it amounts to describing in more detail an action that had previously been left undescribed; sometimes asking about a particular case instead of a general case; sometimes asking about an action happening in our world vs. one in an alternate universe; sometimes something else. Nonetheless, the effects seem to hang together; we expect, with Sinnott-Armstrong, that a good explanation for one will be a good explanation for all.

The abstract/concrete paradoxes are *paradoxes* because they uncover inconsistencies in people's judgments; for example, a majority of participants think that an agent cannot be responsible for any action of a certain sort, but that the agent is responsible for a *particular* action of that very sort (for further discussion see Sinnott-Armstrong 2008a).

1.1 Nichols and Knobe

Nichols and Knobe (2007) gave participants vignettes that describe a deterministic universe, and then asked about the moral responsibility of agents in that universe. When participants are asked (abstract condition) whether it's "possible for someone to be fully morally responsible" in that universe, 86 % of them answer no. This condition is the abstract condition because participants are not given a specific act committed by the hypothetical agent. But when presented with a particular agent in that universe who kills his wife and children to be with his secretary (concrete condition), 72 % of them say that the agent is "fully morally responsible." This condition is concrete because the agent's action is specified. Somehow, participants' judgments of responsibility for agents in a deterministic universe are swayed by the difference between these conditions.

1.2 Nichols and Roskies

Nichols and Roskies (2008) gave participants vignettes that described a deterministic universe. Half the participants received vignettes where the deterministic universe is an alternate universe and half the participants received vignettes where the deterministic universe is our actual world. The participants were then asked how morally responsible a person in the described universe is for their actions. Participants held the alternate world's inhabitants significantly less morally responsible than our world's inhabitants. Thus, the more abstract condition—the alternate universe—caused significantly different responses from the more concrete condition—our world.

1.3 Small and Loewenstein

Another example of the abstract/concrete phenomenon is provided by Small and Loewenstein (2005). Their participants played a trust game in groups of ten. Each participant received a small sum of money and then chose either to keep the money, or to give it to the group. If a participant decided to give their money to the group, that

participant got none of it, but the other nine group members each received some extra money from the experimenters. Thus, a participant would get the most money if they chose to keep their own money while everyone else gives their money up, and the least money if they gave their money up while everyone else keeps theirs; such a participant would get nothing at all.

After the game is played, each participant who chose to give up their money was presented with an opportunity to (pay to) punish a single randomly-selected defector. In the abstract condition, the random selection happened after the punishment was chosen; in the concrete condition, it happened before. In neither condition did any participant learn who they were punishing—the random selection was simply of a slip of paper with a number on it. Participants in the concrete condition chose significantly harsher punishments than did participants in the abstract condition. Note the minimal difference between the two conditions here; somehow the bit of concreteness provided by the specific number was enough to affect participants' choices.

1.4 De Brigard, Mandelbaum, and Ripley

In De Brigard et al. (2009) participants were randomly assigned to one of two conditions, the abstract condition or the concrete condition. In the abstract condition participants saw the following prompt:

> Dennis has recently found out from his doctor that he has a neurological condition that has caused him to behave in certain ways. Were someone else to have this neurological condition then that person would have had to behave in the same ways as Dennis.
>
> On a scale of 1–7, 1 being not responsible, 7 being very responsible, how morally responsible is Dennis for the behaviors that are caused by his neurological condition?

This condition was abstract because the action that the agent committed was not specified. In contrast, the concrete condition contained the following sister prompt where the action is specified:

> Dennis has recently found out from his doctor that he has a neurological condition that has, in the past, caused him to rape women. Were someone else to have this neurological condition then that person would have had to behave in the same ways as Dennis.
>
> On a scale of 1–7, 1 being not responsible, 7 being very responsible, how morally responsible is Dennis for raping women?

The results showed the expected abstract/concrete effect. In the abstract condition, participants held Dennis significantly less responsible than in the concrete condition.

1.5 Nahmias, Coates, and Kvaran

In Nahmias et al. (2007), participants were confronted with vignettes, set in a deterministic world called "Erta," recounting various actions. Participants were asked

a number of different questions; among them questions about responsibility. There were three conditions that are relevant for our purposes. In one, the Abstract condition, no particular action was mentioned; it was simply pointed out that Ertans' actions are determined, and participants were asked to respond, on a six-point rating scale, to the sentences "Ertans should be held morally responsible for their decisions." In the two remaining conditions, an Ertan named Smit makes a particular decision: in the Good condition, Smit donates a large sum of money to an orphanage in his community, while in the Bad condition, Smit kills his wife so he can marry his lover. In the Good condition, participants were asked to respond to sentences about Smit such as "Smit should be held morally responsible for his decision to donate money to the orphanage;" the same, mutatis mutandis, for the Bad condition.

Participants' responsibility judgments were significantly higher in the Bad condition than in either the Abstract or Good condition. Although judgments were higher in the Good condition than the Abstract condition, this difference was not significant. Nahmias et al. conclude that good concrete actions seem to drive up responsibility judgments somewhat, but to a lesser extent than do bad concrete actions. (Note that they conclude this despite the absence of any significant difference between the Abstract and Good conditions on questions of moral responsibility. Presumably they think that a larger study sample would produce a significant difference here.)

In sum, there is a clearly identifiable change in participants' responses between abstract and concrete stimuli. We are not the first to notice the pattern, and several competing hypotheses have been offered in an attempt to explain it. In the next section, we turn to an important class of hypotheses: what we'll call the 'affective theories'.

2 Affective Theories

Nadelhoffer (2004), Nichols and Knobe (2007), and Prinz (2007) try to explain this type of data by appealing to participants' emotions. These theorists differ among themselves as to the details of emotion's involvement, and in particular as to whether emotion sways participants from a cooler, more reasonable responsibility judgment or whether emotions are intimately involved in responsibility judgments in the first place. Despite these differences, we won't distinguish between these theories; for our purposes, what's important is that they all want to explain the abstract/concrete phenomenon purely via affect.

According to these theorists, what accounts for the raised responsibility judgments in the concrete cases is the heightened affect the concrete cases provoke. It seems reasonable to assume that the words 'rape women' produce higher affect in participants than do the words 'behave in certain ways' (Brigard et al. 2009). What's more, there's evidence that heightened emotion can increase judgments of responsibility, even when those judgments are about cases unrelated to what raised participants' emotion in the first place (Lerner et al. 1998). Given this, we'd expect that the increased judgments of responsibility in concrete cases are at least partially due to emotional factors.

Let's look more closely at one of Nichols and Knobe's studies. They presented participants with a deterministic universe, and then asked one of the following two questions:

> High Affect Condition
>
> As he has done many times in the past, Bill stalks and rapes a stranger. Is it possible that Bill is fully morally responsible for raping the stranger?
>
> Low Affect Condition
>
> As he has done many times in the past, Mark arranges to cheat on his taxes. Is it possible that Mark is fully morally responsible for cheating on his taxes?

There were notable differences in participants' responses across these conditions: 64 % thought it possible that Bill was fully morally responsible, while only 23 % thought it possible that Mark was. Since Nichols and Knobe take the main difference between their High Affect and Low Affect conditions to be the amount of affect evoked, they attribute the difference in responses to the difference in affect. They hypothesize that the differences observed in abstract/concrete cases are also due to differences in affect.

As Sinnott-Armstrong (2008b) points out, however, it remains open that affect is not exclusively responsible for the observed effects. It is also possible that there are cognitive effects of the cases that influence participants' responses; we will proceed to explore this hypothesis. These cognitive differences might have their effects alone, in addition to affective factors, or in interaction with affective factors. Moreover, there are reasons to suspect that affect alone cannot be responsible for the range of results covered above. In particular, new data from Cova et al. (2012) paint a troubling picture for pure affective accounts. Cova et al. found patients that suffer from extreme emotional blunting: patients suffering from a behavioural variant of frontotemporal dementia (bvFTD). Patients with bvFTD are dissimilar from healthy adults in a variety of ways; for our purposes it is enough to note that bvFTD patients lack affect in situations that normally trigger affective responses in other people. If an affective theory of the above effects is sufficient, then we should see a different pattern of judgments by bvFTD patients; in particular, the difference between the abstract and concrete cases should narrow as emotional blunting rises.

However, this is not what Cova et al. discovered. They administered the Nichols and Knobe (2007) prompts to their participants and found no differences between the ways that healthy individuals and bvFTD patients responded to these prompts. This result is difficult for an affective theory to explain. Though we do not believe that there is ever such a thing as a 'crucial test', Cova et al.'s data poses a serious challenge to the idea that affect plays a major role in explaining the abstract/concrete paradoxes. Thus, in searching for an explanation, we should look beyond just affective theories.

In such a search, it is natural to look for consilience with other effects. We should see whether we can assimilate the abstract/concrete effects to other effects we already know a bit about. Eventually, we will argue that a theory that takes both affective and cognitive factors into account can offer a natural account of these abstract/concrete

effects: an account that sees them as one piece of a larger pattern. In order to get to this hybrid theory, we first turn our attention to purely cognitive theories.

3 Cognitive Theories

For our purposes, a cognitive theory will be any theory where the critical causal variable is a non-affective one, be it a cognitive state (like a belief) or a cognitive process (like memory).

The first sort of cognitive theory we will focus on, before we present our own, is one that hypothesizes that the difference in judgments is created by our tendency to think about abstract situations differently than concrete situations. Although this suggestion may be on to something, it's also seriously underspecified, and the devil is (as usual) in the pudding. A way to start filling this in is to suppose that concrete stimuli are processed differently *because they are concrete* and not because (e.g.) concrete stimuli happen to coincide with emotionally charged stimuli; that is, that we have a certain mental mechanism whose processing is dedicated to concrete inputs. It is surely possible that we have some mental mechanism which parses information differently based on its abstractness or concreteness, sending abstract information to one mental process and concrete information to a different mental process.

This pregnant suggestion is endorsed in Sinnott-Armstrong (2008b). In what follows we will focus on Sinnott-Armstrong's view because it is, to our knowledge, the only cognitive theory detailed enough to be fit for detailed evaluation. Sinnott-Armstrong's view is a cognitive theory that we will characterize as the 'Separate Capacities hypothesis.' On the subject of explaining the abstract/concrete paradoxes he writes, "… one intuition reflects an abstract way of thinking, whereas the other reflects a concrete way of thinking" (p 214). The Separate Capacities hypothesis attempts to explain the difference between intuitions in abstract and concrete cases by pointing to different encoding systems. Sinnott-Armstrong is inspired by the hypothesis that we have two systems of memory: semantic memory and episodic memory, and he postulates semantic and episodic processes in play in the abstract/concrete paradoxes. The semantic processes are brought to bear in the abstract cases; these are processes that handle information in a sentence-like propositional manner. On the other hand, the episodic system encodes information not in propositional form, but as phenomenological episodes; it has its effect in the concrete cases. The Separate Capacities hypothesis posits that the difference between these two systems is the source of the abstract/concrete phenomena.

We can see how this might work by considering a rule like "if an action is determined, then the agent is not responsible for the action" (hereafter Rule R) (Sinnott-Armstrong 2008b). It might be a good idea to believe Rule R, because for most cases we are likely to stumble on, the rule accords with our intuitive judgments. If you're reaching for a cup of coffee and someone bumps into you causing you to spill your coffee, intuitively *you* should not be held responsible for the ensuing mess, even if that scalding mess is in our laps. At the very least, you would be less responsible than if you had poured the coffee on us because you thought it would be fun to watch us squirm. Rules like Rule R, Sinnott-Armstrong posits, are brought to bear by the semantic system, in judging the abstract cases.

However, though a rule like Rule R is useful, it has the usual downfall of heuristic rules: it sometimes over-generalizes to situations where it doesn't seem to work as well. If you were to spill scalding coffee on our laps for fun and then claim that your behavior was determined by laws of nature and the initial conditions of the universe, we probably would not be very understanding. This type of situation is at the heart of the Separate Capacities hypothesis: paradoxical judgments arise when meshing abstract situations with concrete situations because we represent some information abstractly, as in semantic memory, and some information concretely, as in episodic memory. In short, the Separate Capacities hypothesis entails that abstract situations get coded by a different representational system than concrete representations. Importantly, whether something is encoded as abstract or concrete does not rely on the affective nature of the stimulus, for Sinnott-Armstrong; thus, the Separate Capacities hypothesis is not an affective theory.

The Separate Capacities hypothesis has something important going for it: it recognizes that abstract/concrete paradoxes can arise even in situations that aren't affectively loaded. However, we are skeptical. First, the mere abstractness or concreteness of the stimuli doesn't appear, prima facie, to be itself the source of the puzzling differences. For example, all of the abstract/concrete effects have had the consequence that people judge the concrete actions more harshly than the abstract actions. But consider a pair of cases where the abstract case describes X as a murderer, whereas the concrete case vividly and concretely describes an overwhelmingly mundane act (X butters her bread, in the hallway, at midday, with a paring knife, slowly, deliberately, while slightly salivating, etc.). Without more detail, the Separate Capacities hypothesis seems to predict that concrete cases will produce higher levels of responsibility judgments in virtue of their mere concreteness (although see fn. 4), but this prediction must be wrong. Responsibility judgments, we predict, are tied to the breaking of a norm rather than the concreteness of a situation.

Additionally, the Separate Capacities hypothesis suffers from a gap, between the stimuli that cause representations and the representations themselves. The theory tells us that we represent some information abstractly (as in semantic memory) and some information concretely (as in episodic memory). But we want to know why abstract *stimuli* cause different responses from concrete *stimuli*. It remains to say why the different stimuli (abstract or concrete) trigger the processing of different systems. It's not enough to say that abstract stimuli are processed by a different system in virtue of their abstractness. The Separate Capacities hypothesis, to be complete, needs to explain how and why abstract situations get encoded by abstract representations, while concrete situations get represented by concrete representations.[1]

In fact, Sinnott-Armstrong gives us little reason to believe that there is such a mapping between situation and representation. Red stimuli needn't be represented by

[1] It might be that more 'vivid' stimuli somehow are more effective at creating a mental image (or a proxytype, (Prinz 2004) or some such) which leads to episodic-style rather than semantic-style encoding. But until we have some grip on what vividness is and how it does this work, as well as how mental imagery (or whatever) leads to episodic encoding, our charge of underspecification stands.

red representations (they needn't even be represented *as red*); similarly, abstract stimuli needn't be represented abstractly (and mutatis mutandis for concrete stimuli/representations). So even if the Separate Capacities hypothesis theory is correct, it is incomplete: we still need a story about why abstract stimuli are represented abstractly and why concrete stimuli are represented concretely. The Separate Capacities hypothesis doesn't on its own have the resources to explain the paradoxes we've seen.

Moreover, even if we grant that this gap can be filled, the Separate Capacities hypothesis still encounters some trouble. Remember, we want to explain how responsibility judgments arise. Once we have a story about how information from the external world gets subsumed, we still need a story about how these representational systems end up making certain judgments. Specifically, there must be some explanation for the fact that abstract processes lead to lower levels of responsibility judgments while concrete processes lead to higher levels of responsibility judgments. Though Sinnott-Armstrong does not hypothesize as to how the processes and judgments are linked, we will speculate on his behalf, in order to point to a dilemma.

We see two stories that can cover the whole range of data to be explained, and neither is congenial to his view. (If there is some third story, then perhaps the Separate Capacities hypothesis can dodge this dilemma.) The first: concrete stimuli cause harsher responsibility judgments because of the heightened affect that comes along with concrete situations. This would explain why the concrete stimuli correlate with higher responsibility judgments: heightened affect creates a heightened tension, and this tension causes harsher judgments.[2] However, if Sinnott-Armstrong chooses this explanation, then his view simply is an affective view. It's clear that he takes himself to offer an alternative to affective views. Moreover, a theory filled out in this way would have the same difficulty as other affective theories in explaining the data from Cova et al. (2012).

The other way we see to fill in the Separate Capacities hypothesis is to postulate a difference in belief states that corresponds to differences between abstract and concrete stimuli or representations. This is the type of view we have to offer.[3] In the next section we'll specify how such a view would work. We think that a belief-based view is a legitimate competitor to affective views. However, we

[2] We find this very intuitive fact somewhat mystifying. The connection between anger and harsher judgments is not in doubt, but it would be nice if we could *explain* why this connection arises rather than just note that it does.

[3] Sinnott-Armstrong's theory is not the only cognitive theory available besides ours; see, for example, Nado (2008) or Uttich and Lombrozo (2010). One could also propose a modular account to explain these effects. It's not hard to see extending a view like Mikhail's (e.g., 2007) to a modular system whose input is concrete social situations and which gives responsibility judgments as output. A modular system would still need to pick up on some difference in the stimuli to 'direct' the stimuli to different modular processes. If one takes the relevant difference in the stimuli to be an affective difference (i.e. concrete stimuli are affect-laden, abstract are not) then the modular view collapses into an affective view. If one takes the difference in processing to just be the difference between abstract and concrete situations, one would need to explain how the modular processor picks up on this difference (perhaps concrete situations set off agency detection in a way abstract stimuli don't?). We see no in-principle reason why such a modular view couldn't work, and perhaps this would be a fruitful avenue for future research. However, since no such modular views have been proffered, we will not pay the potential view any more attention here.

also think that such a belief-based view has no need to appeal to different representational systems to explain the data; we will show this below. So if Sinnott-Armstrong were to adopt a belief-based answer to this dilemma, his main theory would be obviated.[4]

4 NBAR

We think there is a more fruitful way of understanding the data. Our hypothesis belongs to the family of theories involving an interaction between affective and cognitive effects. It explains the abstract/concrete paradoxes without either appealing to affect as the main cause or by giving a cognitive architectural account like the Separate Capacities hypothesis. On our theory, unlike on the Separate Capacities hypothesis, both abstract and concrete stimuli are processed in similar ways.

The data seems to indicate that whenever a norm is broken, people think that an agent has to be responsible.[5] Our theory explains this by positing a certain unconscious belief and drawing on the resources of cognitive dissonance theory.[6] We propose that all of the abstract/concrete data can be explained by appealing to this single unconscious belief, made salient in concrete cases, but not in the abstract cases.

Our specific hypothesis is that participants have an unconscious belief that whenever a norm is broken, an agent is responsible for the breaking of the norm. We will refer to this belief as NBAR, for 'Norm Broken, Agent Responsible'. Bad happenings are situations in which norms have been broken (since being bad is one way to break a norm). Thus, in situations where a bad thing happens, people will conclude that an agent is responsible for the bad occurrence. It should be stressed that we, qua theorists, are not endorsing NBAR; we are endorsing the hypothesis that abstract/

[4] An anonymous reviewer astutely pointed out a third option, one which appears quite tantalizing on first glance. The reviewer suggested that Sinnott-Armstrong could flesh out his view by claiming that abstractly represented cases are judged against previously held heuristics (such as Rule R) while concrete cases are judged by comparison to prototypes of the moral violation that are recruited from memory. This suggestion does mesh well with Sinnott-Armstrong's overall motivations, and would certainly block our 'buttering the bread' objection above. However, the suggestion is of no use to Sinnott-Armstrong, for if he took it on he would be unable to explain some of the cases that he takes to be paradigmatic abstract/concrete effects, such as the Roskies and Nichols case. In that case both the abstract and concrete cases should recruit the same prototypes, for the differences in the cases don't consist in the level of description, but instead consist in the world under consideration. Additionally, this explanation, we think, would struggle to explain Small and Loewenstein's results, although Sinnott-Armstrong does not discuss their work.

[5] By 'agent' we just mean some being that can act on the basis of their intentional or phenomenological states; that is, some agent that is capable of some minimal forms of practical rationality.

[6] An anonymous referee points out that a number of the benefits of our approach do not strictly depend on supposing that it is a *belief* involved, and claims that something like a quasi-belief or disposition to believe might do as well (while also bringing along fewer metaphysical and epistemological commitments). As far as we can see, this is indeed possible, depending on just how these other notions are understood. But putting our hypothesis in terms of belief keeps it simple and clear, and helps us derive predictions from it via the familiar mechanisms of cognitive dissonance. Of course, we're open to finding evidence that might weigh on this issue.

concrete effects are to be explained by appealing to *participants*' tacit endorsement of NBAR.

We use 'norms' quite broadly: The norms we are talking about form a broader class than merely moral, or even deontic, norms. We are not offering anything like a definition of norms, but we think that the class of norms we are discussing constitutes an interesting psychological kind. For our purposes, norms are something like expectations. For example, one doesn't wear a clown suit to a business meeting, one shouldn't eat steak with one's feet, objects don't just appear out of nowhere, tables don't talk, etc.[7] Roughly speaking, each of the beliefs one has about the way the world ought to be (in the broadest possible sense of 'ought') counts as a norm on our view.

Unsurprisingly, there are many situations that would threaten NBAR: e.g. a storm-blown tree falling on one's firstborn. In a case like this something bad happens without any agent in particular being responsible for the bad event.[8] Cognitive dissonance theory hypothesizes that having inconsistent beliefs puts the believer in a negative motivational state (see, e.g., Elliot and Devine 1994). The theory dictates that when one's belief is threatened, one will feel a basic drive to relieve this threat, either by giving up the belief or finding some way to see it as not really threatened (ignoring evidence, engaging in rationalization, adding more consonant information, etc.). Beliefs can be threatened (and thus dissonance can be created) in many ways; sometimes the dissonance is created because a belief conflicts with the world, sometimes beliefs are threatened because the belief conflicts with one's behavior, and sometimes beliefs are threatened because one holds contradictory beliefs. This account allows for interplay between cognition and affect. On the one hand, cognition drives affect when conflicting beliefs lead to an unpleasant motivational state; on the other hand, affect drives cognition when the unpleasant motivational state leads to, inter alia, changes in beliefs.

Most germane are the cases where multiple beliefs contradict one another. Take the aforementioned case where a storm-blown tree falls on one's firstborn. This type of situation contradicts NBAR; it seems to be an example of something bad happening (thus an example of a broken norm) without anyone being responsible. The familiar phenomenon of cognitive dissonance should lead us to either i) decide that what's happening isn't bad after all, ii) give up NBAR, or take this case to provide an exception to it, or iii) decide that someone is in fact responsible for the bad thing in question.

Since people seem to deeply care about their firstborn, option i) seems difficult at best in these cases; option ii) might be possible, but if NBAR is relatively well-entrenched, as we hypothesize it is, giving it up will be difficult; and option iii) is also possible, if only somebody could be found who is responsible for the storm. (Witches, Satan, &c. don't seem like bad candidates here, if one is inclined to believe

[7] An anonymous referee objects: "The problem here is that "expectation" is ambiguous between prediction (I expect the sun to rise soon) and evaluation ('I expect you to clean your room')." But we don't think this is an ambiguity; we think the single notion of expectation is broad enough to cover both of these cases; it is this broad notion encompassing both prediction and evaluation we're aiming at.

[8] The 'bad' here and below, is broader than morally bad. For example, stubbing your toe is bad, but it isn't morally bad. Badness is sufficient for norm violation.

in witches or Satan or the like. But if one isn't, candidates for responsibility seem harder to find.[9])

If this is right, we can now explain the observed abstract/concrete results. In most of the studies cited above (all but one), it's clearer that a norm is being violated in the concrete conditions than in the abstract conditions.[10] That is, we take the perception of norm violation to be a confounding variable, and one that explains the observed effects: abstractness and concreteness are neither here nor there, except as they have an effect on the perception of norm violation. For example, Nichols and Knobe's cases compare family-killing to an unspecified action. Assume (pace Nahmias et al. 2007) that people are intuitive incompatibilists. Thus, when they are told that a situation is in a deterministic universe, they will conclude that people in that situation do not have moral responsibility. In Nichols and Knobe's abstract case (the unspecified action case) participants infer that the character is not responsible for his behavior. However, in Nichols and Knobe's concrete case the situation is different. Participants follow the same chain of reasoning, but then they end up with an inconsistency: they think that the character is not responsible (because his behavior is determined) and they think that some agent must be responsible for the norm violation (the family-killing). This inconsistency will cause the participants dissonance, which can be alleviated by overriding their intuitive incompatibilism and holding that the character must be free and thus responsible. The intuitive incompatibilism will be jettisoned as opposed to NBAR because NBAR, we suppose, is a well-entrenched belief, while the folk's pre-reflective belief in incompatibilism isn't as well-entrenched. Ceteris paribus, any time a norm is broken, NBAR will be invoked and thus strengthened, whereas the incompatibilist belief will be a fairly rarely used belief for the folk. Thus, the NBAR belief can explain the relevant abstract/concrete asymmetry.[11]

[9] It would be interesting to examine the role that ideas of "luck" or "coincidence" play in situations like this, to see if they are being subtly agentified. See also Gray and Wegner (2010).

[10] An anonymous reviewer asks, "What if the case description says only that the act violates a norm without specifying which norm? This description makes it an abstract case but also makes it clear that a norm is violated." What would happen depends on the specifics of the structural description of NBAR, along with the mechanisms that interact with it. (See also footnote 6.) If the antecedent can be triggered by mere mention of a norm violation, then we'd expect agency detection and attributions of agency to be enhanced in general (because there'd be no specific action to use as the attribution base for the responsibility of the agent). However, it's an open question exactly what the cognitive instantiation of NBAR looks like—one could instead imagine that what happens is that norm violations are detected by a separate mechanism (e.g., some sort of change detection mechanism) and it's the detection of the norm violation that immediately sets off the 'agent responsible' belief; in this case, NBAR would be more of a process and not a belief, strictly speaking).

[11] What happens when more than one norm is violated, which happens in both conditions of Nichols and Knobe's follow-up experiment, where the tax cheat is pitted against the rapist? Since both cheating on one's taxes and raping violate norms, we expect NBAR to come under threat when these actions are done deterministically, and we expect responsibility to be attributed to the agent in both cases, as Nichols and Knobe observed. Further, since rape is worse than cheating on your taxes we expect the dissonance created to be higher, and since rape is more emotionally charged than tax cheating we expect the overall affect to be higher. Moreover, since affect and dissonance interact in predictable ways (with higher negative affect causing higher dissonance, see Cooper 2007) we expect the dissonance created and the overall level in the rape case to be greater, producing a greater response (i.e., the rapist is held more responsible than the tax cheat). This is also what Nichols and Knobe observe.

The explanation above can also be applied to the mentally-ill-agent cases in Brigard et al. People seem to intuitively think that the mentally ill aren't responsible for their behaviors (presumably because they think that the behaviors are caused in some aberrant way). Thus, in the abstract case, people let this general judgment come out and they judge that the mentally ill aren't responsible for their behaviors. However, in the concrete case the participants are faced with a mentally ill character who commits a bad action. In this case NBAR is threatened and dissonance created. In order to appease the dissonance participants discard their much less well-entrenched belief about how responsible the mentally ill are for their actions in favor of their very well entrenched belief, NBAR.[12] Thus, they end up holding the mentally ill character responsible for his actions in the concrete, but not abstract, case.

NBAR can also explain the data found by Nahmias et al. (2007). The difference between their Bad condition and their Abstract condition is clear, and can be explained in the same way as the Nichols and Knobe study. Suppose in addition that Nahmias et al. are right in their conjecture that their Good condition produces higher judgments of responsibility than their Abstract condition. NBAR predicts this as well—after all, donating a large sum of money to a local orphanage is certainly a violation of a norm. Obviously, it's not a violation of a *moral* norm, but it's nonetheless a deviation from the ordinary course of events. The Bad condition, on the other hand, features not only a violation of this ordinariness norm, but also a violation of the moral norm against killing. We thus predict that concrete neutral actions, which violate neither moral norms nor ordinariness norms, will result in lower responsibility judgments than either out-of-the-ordinary good or bad actions.

The Small and Loewenstein cases are a bit more involved than the other cases cited above. Small and Loewenstein (2005) hypothesize that the key difference between their conditions is the salience of the group of the defectors; in the abstract condition, where the random defector isn't selected yet when the punishment is chosen, the full group of defectors is still salient (even though the punisher doesn't know how large this group is), but in the concrete condition, where the random defector has already been selected, the full group of defectors is less salient, and the punisher can focus on one specific defector (even though only as a number). We think this fits well with NBAR: a salient group of defectors can reduce the impression that any individual defector has broken a norm, and thus reduce the pressure to hold that individual defector responsible. In particular, since the game participants played in this study was unusual to them, they may not have had clear ideas about what behavior would and would not be acceptable; that is, they may be unclear as to what norms were in play. The salience of a group of defectors might have reduced the impression that defection is a norm violation at all.

[12] One could interpret the situation as one where the participants 'agentify' the mentally ill. Perhaps what is happening is that participants don't think that the mentally ill (nor people in deterministic universes) are actually agents. Then when they come to case of a norm violation by the mentally ill (or by a person in a deterministic universe) they reverse their stance and turn the person into an agent. We think that something like this, perhaps best (paradoxically) called the anthropomorphization of people, occurs in normal everyday cases.

While this fits with the part of Small and Loewenstein's explanation mentioned above, it differs from their full story, which sees the difference between their abstract and concrete cases as one of identifiability: they think participants punish identifiable defectors more harshly than unidentifiable ones. (NB: The notion of 'identifiability' in play here must be loose enough to count seeing a number on a slip of paper as making someone identifiable.) Their explanation would make a different prediction from ours in a case where a single defector is identifiable but a group of defectors is kept salient. Small and Loewenstein should predict that in such a case participants will choose to punish the defector in line with their concrete condition, since the defector is identifiable, while we predict that participants will choose to punish in line with Small and Loewenstein's abstract case, since the group of defectors remains salient. In sum, since a norm is more clearly violated in the concrete condition, we expect more dissonance to arise in the concrete condition, and this dissonance can be resolved by holding the character in question more responsible.

The Roskies and Nichols data, on the other hand, does not fit the above mold. In both their abstract and concrete cases it is equally clear that a norm is being broken. In their abstract cases dissonance is not as aroused because of the appeal to an alternate universe, but the action in question remain unspecified in both cases. Nonetheless, the NBAR hypothesis predicts this outcome as well. We should expect participants to be more (implicitly) confident that NBAR holds of our world than that it holds in other worlds. After all, take some general principle you believe: say, that when a pen is dropped it falls to the ground. Estimate your level of confidence in this belief. Now, how confident are you that when a pen is dropped *in an alternate universe* it drops to the ground? Presumably less so. We're just less sure how things work in alternate universes than we are for our own, at least when it comes to everyday generalizations like these. However, in the concrete case, NBAR is fully in play. The participants then get primed to think about a norm violation (the crime that Roskies and Nichols' vignettes refer to), cease to hold their incompatibilism, and hold the characters responsible for their actions.[13]

NBAR can also avoid the problems that arise from Cova et al.'s work. The bvFTD patients suffer from emotional blunting, which is why the data is so problematic for affective accounts. However, emotional blunting should make no difference for the detection of norm violations. Since NBAR works off the detection of norm violations, NBAR alone is perfectly suited for dealing with Cova et al.'s findings.[14]

[13] A referee objects: "I imagine I am equally *implicitly* confident gravity holds on other planets, and I can't ascertain this by assessing my level of confidence in the explicit belief." But this seems quite odd to us. Connolly et al. (2007) seems to show that participants are less confident that at least some of the general beliefs they hold apply *in Peru.* Alternate universes, it seems natural to think, would do even more to undermine confidence. (Of course, this is a testable hypothesis.) We also want to note that the implicit/explicit distinction may be beside the point here. Since we suppose NBAR to be an unconscious belief, we suppose its effects to be implicit; but we don't think there has to be anything different in kind between implicit and explicit effects at work here. We expect people to be less confident about unfamiliar situations both implicitly and explicitly. Finally, we don't mean to imply that there aren't other factors active in the Nichols and Roskies case (e.g., perhaps participants are also affected by in-group/out-group statuses).

[14] One might object that the negative motivational state created by dissonance is a type of affect; after all, we said as much above. But if it is a type of affect, it is a very different sort than the basic emotions that get blunted in bvFTD patients. For one thing, there is no evidence that bvFTD patients are less susceptible to dissonance (in which case characterizing the negative motivational state as 'affect' may be a bit misleading.)

What's more, the NBAR hypothesis has greater explanatory power than the other extant accounts, affective or cognitive, because of its ability to mesh with and explain data beyond the abstract/concrete paradoxes. For example, there's evidence that people are more apt to anthropomorphize inanimate objects when those objects aren't working as they are supposed to than when they are working as designed (Epley et al. 2007). This phenomenon is ubiquitous in everyday life. Suppose you had a 2-year-old Honda Accord that always starts whenever you want it to. How much personality would you say your car has? Presumably not very much. Contrarily, imagine that you have a 1985 Iroc-Z Camaro, which only starts every other Tuesday. Here you'd be more apt to say that the Iroc *chooses* to start when it wants, that it doesn't start on cold days because it's *crazy* or *mad* at you. Likewise, when your computer keeps crashing as you're trying to type an important document you might say that it's possessed by the devil or that it hates you. However, we never say that our computer is being guided by angels when it turns on just fine.[15]

Our account has the resources to explain this puzzling data along with the abstract/concrete data. Take the anthropomorphization of cars for example. Imagine you are trying to start your car on a frigid day when you are already running a bit late to an important meeting. If the car works when you start it, we can describe why the car works by taking the design stance (Dennett 1987). We say that the car starts because it was designed to start—after all, it's supposed to work that way. However, when the car doesn't work, why do we grant the car agency? Perhaps it's because we're left with the feeling that something abnormal or bad has happened. Since when something bad happens we assume that an agent must have done it, we search for an agent. If there is an agent in sight, then we can blame that agent (if, for example, you see a hooded man laughing at you as your car doesn't start, he can get the blame).[16] If not, we take the car to be an agent.[17]

Additionally, our theory has the resources to explain the puzzling 'side-effect effect' (e.g. Knobe 2003). In the side-effect effect cases a CEO chooses a certain plan of action that will make him the most profit. In one condition he knows that this plan of action will have the side effect of hurting the environment; in the other condition the plan will have the side effect of helping the environment. Participants tend to think that the CEO intentionally harms the environment, but does not intentionally help the environment. Many explanations have been offered for this

[15] Yet the point in the text doesn't necessarily speak in favor of the norm violations needing to have a negative valence per se. Contra Morewedge (2009), we do not think that there is a negativity bias in attributions of external agency. Rather, NBAR predicts that attributions of agency will arise even when a norm is broken in a supererogatory way. Just because people wouldn't say that their car is protected by angels when it starts just fine (when it has always done so in the past), doesn't mean that they wouldn't invoke angels to explain when the car breaks a norm in a positive manner. Imagine that as you are driving down the interstate you fall asleep and wake up 45 minutes later to see that your car has just come to gently rest on the shoulder of the interstate, out of harm's way. Here, it seems quite natural to anthropomorphize the car's behavior ('What an angelic car!') or to invoke mysterious unobservable agents ('An angel must be watching over my Hyundai!').

[16] Of course if one were to blame, say, the mechanic for the car trouble, then one would not be inclined to anthropomorphize the car. The operative point is that the harder it is to find an agent the more likely people are to create one (assuming a norm has been broken).

[17] One may see parallels between NBAR and the anthropological work of Shweder and colleagues (e.g., Shweder et al. 1997), where he finds that many non-Western cultures appear to engage in similar agentic attributions in non-normative situations.

effect, with Knobe's preferred explanation being that it is the moral difference between the situations that cause the difference in participants' judgments. We'd like to suggest that it is not the *moral* status per se, but rather the fact that it is a norm violation. In the case where the CEO's actions harm the environment, we end up with a bad action that is no one's main intention, in which case we have a bad action which no one was particularly responsible for. This situation creates dissonance by contradicting NBAR, dissonance which can be appeased by doing the closest thing Knobe's cases (or more specifically, the questions asked in Knobe's cases) allow for: claiming that the character in the vignette acted intentionally (e.g. harmed the environment). There is some evidence that this effect occurs even in the case of nonmoral norm violations, supporting this explanation.[18]

In sum, the NBAR hypothesis gives us not only a way to understand abstract/concrete paradoxes, but also a way to predict when people will agentify[19] non-agent-like things. It thus allows for unification of what appeared to be disparate phenomena.

What's more, it allows us to make a number of predictions that would help evaluate the NBAR hypothesis. Some we've already encountered, but we'll take the opportunity to list a few more here.

First, we fully expect the NBAR hypothesis to hold even of people who insist that it doesn't hold of them; that is, people who insist that norms are often broken without an agent being responsible. In fact, we think that most people probably would assert that norms can be broken without responsible agents. But the NBAR effect, like most psychological effects, is causally active outside the realm of consciousness. Thus, we hypothesize that were we to bring the nay-sayers into the laboratory, we would find that they do hold the NBAR belief. For example, we could run a lexical decision task on people who claim that they don't hold NBAR. We predict that after they witness a norm broken without an agent present, they will be faster to respond to whether 'God' is a word or non-word, or more likely to call a mouse (or some other borderline case) an agent than a non-agent, than they would if they were faced with the same lexical decision without having been exposed to the agentless norm-breaking.[20]

Here's another example: an experiment that would distinguish our account from the Separate Capacities hypothesis. If our theory is correct, then, ceteris paribus, it should take people less time to respond to abstract cases than to concrete cases. This is because abstract cases do not engender dissonance in participants while concrete cases do (and becasue the dissonance resolution process takes time). The Separate Capacities Hypothesis itself does not seem to predict any difference in reaction times

[18] See Machery (2008), Mandelbaum and Ripley (2010), and Guglielmo and Malle (2010).

[19] We say 'agentify' as opposed to 'personify' because it is a specific type of personification that happens in these cases: the participants make the objects specifically into *agents*.

[20] This isn't to say that there aren't individual differences between people; nor is it to say that there aren't performance constraints that could swamp the NBAR effect (for example, if the abstract/concrete studies were run within subjects as opposed to between subjects, we bet that the effect would dissipate because of people's desire to appear consistent). It is just to say that if we couldn't find NBAR evidence across the majority of people, then we'd take the theory to be defeated, even if there were some gerrymandered subset of people who held NBAR.

to the abstract and concrete cases. In fact, if the Separate Capacities Hypothesis is right we'd suspect that the response times for abstract cases should take longer, since presumably dealing with abstract stimuli is more difficult than dealing with concrete stimuli.[21] Here NBAR theory makes a very specific and counterintuitive prediction.

Additionally, NBAR makes predictions about the types of data we should find when it comes to cases where a norm is or is not broken. For example, NBAR predicts that participants will hold characters more responsible (or freer) when the character is doing some good action than when the character is partaking in a neutral (i.e. non-norm-breaking) action.[22] That this might be right is suggested by the Nahmias et al. studies discussed above.

5 Sinnott-Armstrong's Epistemological Abstract/Concrete Study

Sinnott-Armstrong (2008a) reports a pilot abstract/concrete study that might at first seem to be outside of NBAR's purview, since it asks about knowledge rather than responsibility. In a between subjects design, he gave each participant one of the following vignettes. The abstract vignette read, "People sometimes believe things for no good reason. For example, people sometimes believe what a politician says about the economy when they have no good reason to trust what the politician says. Our question is about knowledge: If a person cannot give any good reason to believe a claim, is it possible that the person *knows* that the claim is correct?" (p. 220–221). The concrete vignette had the same set-up with a different question at the end of it: "If you cannot give any good reason to believe that the person whom you believe to be your mother really is your mother, is it possible that you *know* that she is your mother?" (p. 221). Significantly more respondents answered 'yes' in the concrete condition than in the abstract.

Sinnott-Armstrong goes onto conclude that there are abstract/concrete effects that cannot be explained by affective accounts. Although our theory is not purely affective, it seems natural to suppose that he would make the same claim about our account. However, we don't think that his pilot study is good evidence for this claim. First, his two cases differ not just in their levels of abstraction but also in a few other important ways. In the latter case, but not in the former, Sinnott-Armstrong asks if it is possible that *you* know that the person is your mother, whereas in the former case the participants is just asked whether any arbitrary person can know any arbitrary thing. The participant in the concrete case is much more inclined to be engaged in the vignette; after all, it is the participant whose knowledge is in question. It is reasonable to suppose that this engagement is enough to drive up affect on its own. Moreover,

[21] Though the Separate Capacities hypothesis should predict that people will respond faster to abstract than concrete cases, Sinnott-Armstrong in particular would not make such a prediction, for unrelated reasons. Sinnott-Armstrong also holds a 'moral heuristics' view, one where simple heuristic rules are at play in abstract reasoning. Sinnott-Armstrong posits that in concrete cases people take specific information into account and reason more, thus presumably increasing their response time. (Sinnott-Armstrong et al. 2010). So he, too, makes this counterintuitive prediction.

[22] To reiterate an earlier point made in footnote 15, norm violations can be either positively or negatively valenced; supererogatory acts may be just as unexpected, thus agency enhancing (though maybe not as attention grabbing) as bad actions.

not only is the participant in the concrete case being asked about her own epistemic status, but she is being asked a Very Loaded Question: whether she knows that her mother is really her mother. Merely considering the question is apt to drive up people's affect. Just being asked this question may cause participants to think quite unpleasant thoughts: perhaps they were switched at birth, perhaps their mother has been deceiving them a long while, perhaps they will end up on the Maury Povich show, etc. These thoughts will cause emotional stress and presumably make the participants more defensive, thus causing them to claim that of course they know that their mother is their mother. We hypothesize that if Sinnott-Armstrong took out the pronoun asymmetry between the cases and changed the concrete question from an affectively-loaded question about one's mother to, say, an affectively-neutral question (e.g., whether someone can know if something is a chair), then the asymmetry would dissipate.

6 Conclusion

We suspect that the initial abstract/concrete phenomena look like a grab bag of effects because they are only a loosely related group. In the end, we don't think that abstract/concrete effects are any sort of natural kind. Rather, we think that what makes the class an interesting group is that all of the disparate effects share a similar etiology: they are all caused by the workings of NBAR. NBAR theory opens a new window on the abstract/concrete paradox. We do not need to suppose that the effect is caused entirely by affect, nor that two different cognitive systems are involved. It is enough to combine general principles of dissonance theory with a single hypothesized belief: that whenever a norm is broken, someone is responsible. This not only would explain the observed abstract/concrete effects, but would see them as part of a larger pattern in people's agentification behavior. If our account is correct, then a number of seemingly disparate, puzzling phenomena can be reduced to one very widely held belief: NBAR.

References

Brigard, F., E. Mandelbaum, and D. Ripley. 2009. Responsibility and the brain sciences. *Ethical Theory and Moral Practice* 5(4): 511–524.

Connolly, A., J. Fodor, L. Gleitman, and H. Gleitman. 2007. Why stereotypes don't even make good defaults. *Cognition* 103: 1–22.

Cooper, J. 2007. *Cognitive Dissonance: Fifty Years of a Classic Theory*. London: Sage Publications Ltd.

Cova, F., Bertoux, M., Bourgeois-Gironde, S., and Dubois, B. 2012. Judgments about moral responsibility and determinism in patients with behavioral variant of frontotemporal dementia: Still compatibilists. *Consciousness and Cognition* 21(2): 851–864.

Dennett, D. 1987. *The Intentional Stance*. Cambridge: MIT Press.

Elliot, A., and P. Devine. 1994. On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology* 67(3): 382–394.

Epley, N., A. Waytz, and J. Cacioppo. 2007. On seeing human: A three factor theory of anthropomorphization. *Psychological Review* 114(4): 864–886.

Gray, K., and D.M. Wegner. 2010. Blaming God for our pain: Human suffering and the divine mind. *Personality and Social Psychology Review* 14: 7–16.

Guglielmo, S., and B. Malle. 2010. Can unintended side–effects be intentional? Solving a puzzle in people's judgments of morality and intentionality. *Personality and Social Psychology Bulletin* 36 (12): 1635–1647.

Knobe, J. 2003. Intentional action and side effects in ordinary language. *Analysis* 63: 190–193.

Lerner, J., J. Goldberg, and P. Tetlock. 1998. Sober second thought: The effects of accountability, anger, and authoritarianism on attributions of responsibility. *Personality and Social Psychology Bulletin* 24: 563–574.

Machery, E. 2008. The folk concept of intentional action: Philosophical and experimental issues. *Mind and Language* 23(2): 165–189.

Mandelbaum, E., and D. Ripley. 2010. Expectations and morality: A dilemma. *The Behavioral and Brain Sciences* 33(4): 346.

Mikhail, J. 2007. Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences* 11(4): 143–152.

Morewedge, C. 2009. Negativity bias in attribution of external agency. *Journal of Experimental Psychology. General* 138(4): 535–545.

Nadelhoffer, T. 2004. Blame, badness, and intentional action: A reply to Knobe and Mendlow. *The Journal of Theoretical and Philosophical Psychology* 24: 259–269.

Nado, J. 2008. Effects of moral cognition on judgments of intentionality. *The British Journal for the Philosophy of Science* 59(4): 709–731.

Nahmias, E., D. Coates, and T. Kvaran. 2007. Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest Studies in Philosophy* 31: 214–242.

Nichols, S., and J. Knobe. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous* 41: 663–685.

Nichols, S., and A. Roskies. 2008. Bringing moral responsibility down to earth. *Journal of Philosophy* 105 (7): 371–388.

Prinz, J. 2004. *Furnishing the Mind*. Cambridge: MIT Press.

Prinz, J. 2007. *The Emotional Construction of Morals*. New York: Oxford University Press.

Shweder, R., Much, N., Manamohan, M., and Park, L. 1997. The 'Big Three' of Morality (Autonomy, Community, Divinity) and the 'Big Three' Explanations of Suffering. In *Morality and health*, eds. A. Brandt and P. Rozin. New York: Routledge.

Sinnott-Armstrong, W. 2008a. Abstract + Concrete = Paradox. In *Experimental Philosophy*, ed. J. Knobe and S. Nichols. New York: Oxford University Press.

Sinnott-Armstrong, W. 2008b. 'Framing moral intuitions' in Moral Psychology, Volume 2. In *The Cognitive Science of Morality*, ed. W. Sinnott-Armstrong, 47–76. Cambridge: MIT Press.

Sinnott-Armstrong, W., L. Young, and F. Cushman. 2010. Moral heuristics. In *The moral psychology handbook*, ed. J. Doris, G. Harman, S. Nichols, J. Prinz, W. Sinnott-Armstrong, and S. Stich. Oxford: Oxford University Press.

Small, D., and G. Loewenstein. 2005. The devil you know: The effects of identifiability on punitiveness. *Journal of Behavioral Decision Making* 18(5): 311–318.

Uttich, K., and T. Lombrozo. 2010. Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition* 116: 87–100.