Predictions from Uncertain Moral Character

Samuel G. B. Johnson[1], Gregory L. Murphy[2], Max Rodrigues[3] & Frank C. Keil[3]
(sgbjohnson@gmail.com, gregory.murphy@nyu.edu, mrod1791@gmail.com, frank.keil@yale.edu)
[1]School of Management, University of Bath, Bath, BA2 7AY UK
[2]Department of Psychology, New York University, New York, NY 10003 USA
[3]Department of Psychology, Yale University, New Haven, CT 06520 USA

Abstract

People assess others' moral characters to predict what they will do. Here, we study the computational mechanisms used to predict behavior from *uncertain* evidence about character. Whereas previous work has found that people often ignore hypotheses with low probabilities, we find that people often account for the possibility of poor moral character even when that possibility is relatively unlikely. There was no evidence that comparable inferences from uncertain non-moralized traits integrate across multiple possibilities. These results contribute to our understanding of moral judgment, probability reasoning, and theory of mind.

Keywords: Moral psychology; theory of mind; prediction; causal reasoning; categorization

Introduction

People intensely scrutinize others' moral characters. Is Hillary Clinton a bastion of moral sanity or a devious opportunist? Is Donald Trump a man of the people or a corrupt plutocrat? Is your neighbor Todd a good person because he donates 20% of his income to charity, or a bad person because he received a citation for reckless driving?

This obsession with moral character makes good evolutionary sense: People track reputation to assess who will cooperate (Sperber & Baumard, 2012). For this reason, some have argued that our moral judgments about individual acts are primarily determined by what those acts reveal about the actors' character, rather than the intrinsic properties of those acts (Goodwin et al., 2014; Uhlmann et al., 2015). This explains our interest in intentions when judging an act's wrongness (Cushman, 2008). More broadly, knowing another's character allows us to predict what he or she will do, just as knowing a thing's category tells you about its properties. A bird is likely to fly; a snake is likely to be venomous. A good person may lend a helping hand; a bad person may stab you in the back.

But often we do not know someone's moral character with any certainty. Todd gives money to charity, but might the charity be a money-laundering operation? He was driving at a reckless speed, but what if he may have been doing so because he needed to perform an emergency surgery? How do we predict Todd's actions when we cannot be sure of his intentions?

In this paper, we ask what computational principles govern our predictions of others' actions from uncertain beliefs about moral character. This work falls in a research tradition studying predictions from uncertain categories and uncertain beliefs about causation. For example, if you have uncertain evidence leading you to identify a bird as a heron with 65% probability and a crane with 35% probability, then when you predict the bird's behavior, you may assume it is definitely a heron, without hedging for the possibility it is a crane (Malt et al., 1995; Murphy & Ross, 1994). If you think there's a 75% chance that the Fed chair's statement implies a tightening of the money supply but a 25% chance it does not, you will act as though the Fed is certainly tightening the money supply when you are predicting the stock market (Johnson & Hill, 2017; Johnson et al., 2018). People often ignore uncertainty because it is computationally difficult to consider two possibilities simultaneously, considering the implications of each and integrating across those two possible worlds.

But perhaps people *would* integrate across possibilities when reasoning about moral character. First, we may have encapsulated, module-like mechanisms for aspects of mental-state understanding (Leslie, 1995) and moral judgment (Mikhail, 2007). Perhaps such domain-specific mechanisms perform more efficiently than domain-general mechanisms (Cosmides, 1989). Second, people *do* seem to integrate across possibilities for categories when one of the categories is dangerous (a shark) rather than neutral (a school of fish) (Zhu & Murphy, 2013).

By analogy, if you think an act is probably caused by a morally neutral motive, and then encounter evidence that the motive may actually have been immoral, you might take account of that motive when making further predictions about the person's future behavior. But if you instead encounter evidence that the motive may have been some *other* morally neutral motive, you may very well ignore that possibility when predicting behavior.

The Current Studies

Participants read scenarios about various actions taken by characters. For example, in one item, a driver struck a bicyclist while heading the wrong way on a one-way street. For each action, there were three possible explanations. One explanation was neutral (e.g., the driver did not know that the street was one-way; hypothesis H_{Neut}), one implied poor moral character (e.g., the driver hit the bicyclist deliberately to teach him a lesson; hypothesis H_{Imm}), and one implied that the person had some other, non-moralized trait, such as forgetfulness, risk-aversion, or poor eyesight (e.g., the driver had forgotten to turn her headlights on; hypothesis H_{Other}).

We developed a set of actions predicted by H_{Imm} or

H_{Other}. If H_{Imm} were true (the driver hit the bicyclist on purpose), then her immoral character would suggest other immoral actions (driving with expired registration; prediction Z_{Imm}). If H_{Other} were true (she forgot to turn her lights on), then her trait (forgetfulness) would suggest other related actions (leaving her windows open before a rainstorm; prediction Z_{Other}). In Pretest B, we obtained judgments of how likely each prediction was given each hypothesis. For example, we measured $P(Z_{Imm}|H_{Imm})$, the probability the driver would drive with an expired registration given that she had hit the bicyclist on purpose. We also measured $P(Z_{Imm}|H_{Neut})$, $P(Z_{Imm}|H_{Other})$, $P(Z_{Other}|H_{Neut})$, $P(Z_{Other}|H_{Imm})$, and $P(Z_{Other}|H_{Other})$.

For the Main Study, we were interested in predictions about these actions (Z_{Imm} and Z_{Other}) when participants had evidence rendering her motives (H_{Neut}, H_{Imm}, H_{Other}) uncertain, and how these predictions from uncertain motives would compare to the predictions from certain motives from Pretest B. We constructed two versions of each scenario. In one version (uncertain evidence U_{Imm}), the neutral explanation H_{Neut} was presented as most likely, but the immoral explanation H_{Imm} was also introduced as possible.[1] For example, the driver probably didn't know the street was one-way, but possibly hit the bicyclist on purpose. In Pretest A, we ensured that participants viewed the neutral intention (H_{Neut}) as likelier than the immoral intention (H_{Imm}) given the uncertain evidence—that she really *was* likelier to have forgotten about the one-way street—but also that H_{Imm} was still reasonably likely.

In the other-trait version (uncertain evidence U_{Other}), H_{Neut} was presented as most likely, but the other-trait explanation H_{Other} was also introduced as possible. For example, the driver probably didn't know the street was one-way, but possibly forgot to turn on her lights. Pretest A ensured that people viewed the neutral explanation as likelier than the other-trait explanation. Thus, Pretest A overall elicited judgments of $P(H_{Imm}|U_{Imm})$, $P(H_{Neut}|U_{Imm})$, $P(H_{Other}|U_{Other})$, and $P(H_{Neut}|U_{Other})$.

Our Main Study then tested whether people account for uncertainty about the actor's character given uncertain evidence (U_{Imm} and U_{Other}) when they are making predictions, measuring $P(Z_{Imm}|U_{Imm})$ and $P(Z_{Other}|U_{Imm})$. Would participants think the driver is likelier to perform an immoral act like driving with an expired registration (Z_{Imm}) when she possibly hit the bicyclist on purpose (U_{Imm}) than when she definitely did not (H_{Neut})? If people focus on the most likely hypothesis (i.e., ignore uncertainty about character), then they should view these immoral acts as equally likely regardless of whether there is a chance the driver behaved immorally. Moreover, they should view Z_{Imm} as much less likely if it is merely possible that the driver has poor moral character (U_{Imm}) compared to knowing this for sure (H_{Imm}). That is:

$$P(Z_{Imm}|H_{Neut}) = P(Z_{Imm}|U_{Imm}) < P(Z_{Imm}|H_{Imm})$$

Likewise, if people ignore uncertainty about non-moralized traits, the driver should be seen as equally likely to do other forgetful things regardless of whether it is possible that she forgot to turn on her lights:

$$P(Z_{Other}|H_{Neut}) = P(Z_{Other}|U_{Other}) < P(Z_{Other}|H_{Other})$$

But as mentioned earlier, people might attend to the lower-probability trait when it is moralized, but not when it is non-moralized. If so, people would think the driver likelier to commit other immoral acts even if it is merely possible that she hit the bicyclist on purpose. But people would not consider the driver likelier to commit other forgetful acts if it is merely possible that she forgot to turn on her lights:

$$P(Z_{Imm}|H_{Neut}) < P(Z_{Imm}|U_{Imm}) < P(Z_{Imm}|H_{Imm})$$
$$P(Z_{Other}|H_{Neut}) = P(Z_{Other}|U_{Other}) < P(Z_{Other}|H_{Other})$$

To test these hypotheses, Pretest A normed judgments of $P(H_i|U_k)$—inferences of character from uncertain evidence of intentions—and Pretest B normed judgments of $P(Z_i|H_j)$—predictions of future actions from certain knowledge of character. In the Main Study, we tested judgments of $P(Z_i|U_k)$—predictions of future actions from uncertain evidence of intentions, using the pretest norms to generate normative predictions for comparison.

Pretest A: Intentions from Uncertain Evidence

We first sought to norm the values of $P(H_{Neut})$ and $P(H_{Imm})$ given evidence U_{Imm} and the values of $P(H_{Neut})$ and $P(H_{Other})$ given evidence U_{Other}. That is, we evaluated the scenarios we constructed to be sure that readers interpreted them as intended. This serves two purposes. First, for an item to be included, we need the neutral explanation to be deemed likelier than the moral or non-moral trait explanations—that is, $P(H_{Neut}|U_{Imm}) > P(H_{Imm}|U_{Imm})$ and $P(H_{Neut}|U_{Other}) > P(H_{Other}|U_{Other})$. Second, these estimates are needed to compute normative responses in the Main Study.

Method

We recruited 100 participants from Amazon Mechanical Turk (50% female, M_{age} = 36.9). Participants were excluded if they incorrectly answered more than 30% of a set of 10 check questions (N = 9).

Each participant read eight items, with each item in one of two versions. In one version (U_{Imm}), the evidence suggested two possible explanations, H_{Neut} and H_{Imm}, with H_{Neut} designed to be more plausible. For example:

> Navigation through Tabbsboro is complicated by a set of one-way streets, which were put into place because the streets are too narrow to allow parking and traffic in both directions. The police recently reported on an accident that happened in one of them. One late afternoon, Cindy Harlan

[1] That is, U_X refers to a case in which the neutral explanation of the person's behavior is offered as likely, but X is mentioned as a less likely explanation. H_X refers to cases in which only explanation X is offered.

struck a bicyclist who was riding towards her car. The bicyclist was in her way and injured his hip in the accident. He was taken away in an ambulance.

[H_{Neut}] The police questioned Cindy, and she denied knowing that it was a one-way street. This was the first time she had driven on this road. There was no sign near the driveway, and she had not noticed that it was one-way when she arrived there.

[H_{Imm}] The reporting officer noted that the bicyclist, who was a teenager, had seen Cindy earlier that day, and that she seemed irritated when he and his friends didn't get out of the street fast enough when Cindy was driving to her acquaintance's home. The officer asked if Cindy went the wrong way down the road because she saw the bicyclist playing in the street again and wanted to teach him a lesson. Cindy denied this and said that she simply didn't know it was one-way.

Participants then estimated the probabilities of H_{Neut} ("Cindy hit the bicyclist because she didn't know she was driving the wrong way down a one-way street") and H_{Imm} ("Cindy hit the bicyclist because she was trying to teach the teenager a lesson"). These judgments were entered in separate text boxes on scales from 0 to 100. Since the hypotheses were not strictly exhaustive, the judgments did not have to sum to 100 (though most did).

In the other-trait version of each item (U_{Other}), the evidence suggested two possibilities, H_{Neut} and H_{Other}, with H_{Neut} again more plausible. For the item above, the last paragraph of the U_{Imm} version was replaced with:

[H_{Other}] The reporting officer noted that Cindy's lights were not on. He asked if she might not have seen the bicyclist because she had forgotten to turn her lights on. Cindy pointed out that the accident had happened almost an hour ago, when it was light out.

Participants then judged H_{Neut} and H_{Other} ("Cindy hit the bicyclist because she didn't have her lights on").

Participants saw one version of each item, with half of the items in version U_{Imm} and half in version U_{Other}, counterbalanced across participants. The order of the probability judgments was randomized for each item.

Results

All eight items met our desired conditions (see Appendix). Mean judgments of $P(H_{Neut}|U_{Imm})$ and $P(H_{Neut}|U_{Other})$ ranged from 65% to 82% across items (Ms = 75.1% and 74.3%, respectively), and judgments of $P(H_{Imm}|U_{Imm})$ and $P(H_{Other}|U_{Other})$ ranged from 17% to 32% (Ms = 24.2% and 25.3%, respectively).

Pretest B: Predictions from Certain Intentions

Next, we normed the values of the predictions, $P(Z_{Imm})$ and $P(Z_{Other})$, given certain intentions H_{Neut}, H_{Imm}, and H_{Other}. Once again this has two purposes. First, an inclusion criterion: We want the immoral prediction Z_{Imm} to be more plausible given the immoral than the neutral intention—that is, $P(Z_{Imm}|H_{Imm}) > P(Z_{Imm}|H_{Neut})$—and likewise for the prediction Z_{Other} to be more plausible given the other-trait than the neutral intention—that is, $P(Z_{Other}|H_{Other}) > P(Z_{Other}|H_{Neut})$. For example, since the immoral prediction Z_{Imm} in our example was driving with an expired registration, we needed to ensure that people agree that someone who hits a bicyclist on purpose (H_{Imm}) is more likely to drive with an expired registration (Z_{Imm}) than someone who hit the bicyclist accidentally (H_{Neut}). Second, these values—predictions given certain intentions—are needed for comparison with the Main Study, which measured predictions given uncertain intentions.

Method

We recruited 149 participants from Mechanical Turk (29% female, M_{age} = 33.9). Participants were excluded using the same criterion as Pretest A (N = 25).

Each participant read eight items, with each item in one of three versions. In one version, H_{Neut} was true. For the Tabbsboro example, the first paragraph was the same as in Pretest A, and the remainder of the item read:

Cindy didn't realize that it was a one-way road. This was the first time she had driven on this road. There was no sign near the driveway, and she had not noticed that it was one-way when she arrived there.

In a second version, H_{Imm} was true:

Cindy had pulled out of an acquaintance's driveway and turned left, even though that was the wrong way for this one-way street. She went the wrong way because she saw several kids who had irritated her earlier in the day for not getting out of the street fast enough, so when she saw them again, she wanted to drive close to them to teach them a lesson.

Finally, in a third version, H_{Other} was true:

Cindy had pulled out of an acquaintance's driveway and turned left, but she didn't see the bicyclist because she had forgotten to turn her lights on.

Participants then estimated five probabilities for each item. We included two versions each of Z_{Imm} ("What is the probability that Cindy would drive her car with an expired vehicle registration?") and Z_{Other} ("What is the probability that Cindy would forget to shut her window the night before a thunderstorm?"). For the Main Study, we chose the best version for each item to maximize the chance we could use a given item. We also included a filler item ("What is the probability that the city will install a clearer sign in the next week?") which would not necessarily vary based on Cindy's intention. These judgments were made using the same procedure as Pretest A.

Participants saw one version of each item, with the eight items distributed about evenly across the three versions, counterbalanced across participants. The order of the probability judgments was randomized for each item.

Results

The probability of Z_{Imm} was consistently judged higher given H_{Imm} than given H_{Neut}; that is, $P(Z_{Imm}|H_{Imm}) > P(Z_{Imm}|H_{Neut})$ for all items (Ms = 65.6% and 26.4%, respectively), with the difference between these conditional probabilities ranging from 26.7% to 53.4% across items (see Appendix). $P(Z_{Other}|H_{Other})$ was higher than $P(Z_{Other}|H_{Neut})$ for all items but one (Ms = 60.7% and 49.8%, respectively), with the difference between these probabilities varying from –1.5% to 23.2% across items.

Since all items satisfied the desired conditions for Z_{Imm}, we did not exclude any items for the Main Study. However, these results suggest two caveats. First, it is difficult to compare participants' inferences about moral versus other kinds of traits, since the morally laden predictions (Z_{Imm}) were much more responsive to knowledge of intentions compared to the non-morally laden predictions (Z_{Other}—see later discussion). Given this limitation, any conclusions about moralized versus non-moralized character traits must be provisional. Second, because some of these items were not very robust for the non-morally laden predictions, we repeat key analyses on individual items.

Main Study: Predictions from Uncertain Evidence

The Main Study tested inferences about people's future actions based on uncertain knowledge of their intentions. Participants saw the evidence normed in Pretest A, making predictions about the actions normed in Pretest B.

Method

We recruited 99 participants from Mechanical Turk (54% female, M_{age} = 37.5). Participants were excluded using the same criterion as in the pretests (N = 1).

Participants read the eight vignettes used in Pretest A, each in one of the two versions (either U_{Imm} or U_{Other}). For each item, participants were asked questions across two pages (with the vignette text displayed on the screen for both). On the first page, participants made predictions of Z_{Imm} and Z_{Other}, using the phrasing normed in Pretest B. On the second page, participants indicated which intention they thought was likelier. For the U_{Imm} version of the item, participants chose between H_{Neut} ("Cindy hit the bicyclist because she didn't know she was driving the wrong way down a one-way street") and H_{Imm} ("Cindy hit the bicyclist because she was trying to teach the teenager a lesson"); for the U_{Other} version, participants chose between H_{Neut} and H_{Other} ("Cindy hit the bicyclist because she didn't have her lights on").

Results

Overall, participants tended to place positive weight on the immoral explanation H_{Imm} when making predictions, even when they acknowledged that the neutral explanation H_{Neut} was likelier. This is a departure from most previous studies of predictions from uncertain beliefs. On the other hand, there was little evidence that participants placed any weight on the other-trait explanation H_{Other} when making predictions, which raises the possibility that people might attend selectively to evidence of immoral intentions. Finally, there was modest evidence that people underweighted H_{Imm} relative to normative standards, and considerable evidence for underweighting H_{Other}.

We tested reliance on H_{Imm} in two ways. First, we conducted an item-level analysis, averaging probability judgments across all participants (see Appendix). Unsurprisingly, mean judgments of $P(Z_{Imm}|U_{Imm})$ (31.9%) were lower than $P(Z_{Imm}|H_{Imm})$ in Pretest B (65.6%), $t(7) = 8.80$, $p < .001$, $d = 3.11$, reflecting the fact that H_{Imm} had a low prior probability given evidence U_{Imm}. More interestingly, mean judgments of $P(Z_{Imm}|U_{Imm})$ (31.9%) in this study were higher than $P(Z_{Imm}|H_{Neut})$ in Pretest B (26.4%), $t(7) = 3.37$, $p = .012$, $d = 1.19$. That is, when the evidence is uncertain between H_{Neut} and H_{Imm}, predictions of Z_{Imm} fall between predictions made when either H_{Neut} or H_{Imm} is certain. This shows that people take both H_{Imm} and H_{Neut} into account when predicting Z_{Imm}. (In English: When there are two possibilities, the induction will take both into account and therefore lie in between the predictions given either possibility alone.)

We can also use the data from Pretests A and B to calculate normative values of $P(Z_{Imm}|U_{Imm})$:

$$P(Z_{Imm}|H_{Neut})P(H_{Neut}|U_{Imm}) + P(Z_{Imm}|H_{Imm})P(H_{Imm}|U_{Imm})$$

These normative responses are given in the Appendix for each item (M = 35.7%). Participants' actual judgments (M = 31.9%) were marginally more conservative, $t(7) = 1.99$, $p = .087$, $d = 0.82$, compared to the normative responses, suggesting that participants underweighted H_{Imm}. Although statistically not very robust, this would be consistent with previous studies, finding that people underweight unlikely hypotheses, even when they do not ignore them entirely (Johnson, Merchant, & Keil, 2018).

This item analysis, however, can be criticized because it lumps together participants who agreed that H_{Neut} was likelier than H_{Imm} (which should be the dominant belief, based on Pretest A), with those who believed the converse. In fact, about 19% of responses disagreed with our assumption that H_{Neut} was likelier. Thus, the analysis above could be lumping together two populations: Those who believed H_{Neut} was likelier and assigned no weight to H_{Imm}, and those who believed H_{Imm} was likelier and assigned no weight to H_{Neut}. The item means would look like both hypotheses are being considered, but this is an illusion due to mixing two populations (Malt et al., 1995).

Thus, our second approach analyzed the data at the level of individual participants, including only participants for each item who agreed that H_{Neut} was the likelier than H_{Imm}. Using this approach, participants rated $P(Z_{Imm}|U_{Imm})$ numerically higher than the average pretest ratings of $P(Z_{Imm}|H_{Neut})$ for 6 of the 8 items, significantly so for three of the items (items 2, 5, and 6; ps < .02); one

item was significant in the opposite direction (item 7; p = .026). Overall, this evidence is moderately consistent with the idea that people often place weight on H_{Imm} even when they view H_{Neut} as likelier.

We also used this approach to test whether people would *underweight* H_{Imm} even when they assigned positive weight to it. For this analysis, we included all participants (even those indicating that H_{Imm} was likelier than H_{Neut}) because the estimates from Pretest A, used to calculate normative values, average across both kinds of participants. Participants rated $P(Z_{Imm}|U_{Imm})$ numerically lower than its normative value for 5 of the 8 items, but significantly for only one item (item 7); no items were significant in the opposite direction.

The above analyses focused on the U_{Imm} condition. What about the U_{Other} condition? The item analysis found that judgments of $P(Z_{Other}|U_{Other})$ (M = 45.2%) were lower than $P(Z_{Other}|H_{Other})$ from Pretest B (M = 60.7%), $t(7)$ = 2.85, p = .025, d = 1.01. But unlike the U_{Imm} condition, there was no evidence that people took account of H_{Other}, since $P(Z_{Other}|U_{Other})$ judgments (45.2%) did not differ from $P(Z_{Other}|H_{Neut})$ judgments from Pretest B (49.8%), and indeed were in the wrong direction on average, $t(7)$ = –0.86, p = .42, d = –0.31. That said, these judgments also did not differ significantly from their normative values (52.6%), $t(7)$ = 1.47, p = .18, d = 0.50 (although the normative values themselves did differ significantly from $P(Z_{Other}|H_{Neut})$; p = .023). These inconclusive results are probably due to the poor diagnosticity of intentions for predicting Z_{Other}, shown in Pretest B.

Since the diagnosticity differed across items, it is useful to conduct subject-level analyses for each item, as we did for the U_{Imm} condition. Looking at just those who agreed that H_{Neut} was likelier than H_{Other}, ratings of $P(Z_{Other}|U_{Other})$ were higher than the average pretest ratings of $P(Z_{Other}|H_{Neut})$ for only 3 out of the 8 items, with only one item reaching significance (item 1; p = .037), and three items reaching significance in the opposite direction (items 6, 7, and 8; ps < .001). Conversely, looking at all participants, ratings of $P(Z_{Other}|U_{Other})$ were lower than the normative scores for 5 out of the 8 items, with 4 items reaching significance (items 2, 6, 7, and 8; ps < .02). These results cast doubt on the idea that people place positive weight on the other-trait hypotheses when making predictions, suggesting that people underweight such hypotheses. However, this conclusion must be provisional given the poor diagnosticity of some of the non-moralized traits.

Discussion

Judgments of moral character are central to social life. They guide our decisions about who we interact with, inform our beliefs about what others are thinking, and help us to predict what others are going to do. But moral character is often ambiguous, since we often cannot know others' intentions with certainty. How do we predict others' behavior when their character is uncertain?

First, in contrast to other studies of predictions from uncertain beliefs (Johnson et al., 2018; Malt et al., 1995), we find that people have at least some ability to account for the possibility of immoral character, even when it is relatively unlikely. In vignettes where characters were assigned a 25% probability of a nefarious motive, people took this motive into account when predicting other immoral behaviors. Although this result was not consistent across all of our items, it was statistically robust for some of them and was significant overall.

Second, it is possible that this ability to account for uncertain traits is specific to moral traits. There was little evidence that participants weighted uncertain non-moral traits (e.g., poor eyesight) in predictions. This result is limited by the relatively poorer quality of our non-moral than of our moral items, suggesting the need for future research with more directly comparable items.

This problem, however, may reflect a real issue in making predictions about human behavior. When someone makes a mistake of some kind (as all these examples are), there are many factors that could be involved, and it may be hard to rule any out. If someone makes a wrong turn while driving, the person could well have not been paying attention, the sign might not have been very clear, the traffic might have been distracting, and so on. The presence of one of these explanations does not greatly reduce the possibility that one of the others also applied. Thus, such explanations based on non-moral character traits may not be very diagnostic about future actions. Morality, in contrast, may be of special interest to people precisely because it is thought to be diagnostic.

Third, we compared judgments to normative benchmarks. There was a trend toward underweighting the less-likely hypothesis. For the moral traits, this trend reached only marginal significance overall because participants' judgments were actually quite close to the normative benchmarks; only one individual item revealed significant evidence of underweighting. For the non-moralized traits, there was less room for reliable differences to emerge between actual and normative judgments overall, given the poor quality of some of the non-moral items. But there was considerable evidence for underweighting for half of the individual items. Thus, there seems to be more robust underweighting of unlikely non-moral traits than of unlikely moral defects.

These results contribute to several conversations in cognitive and social psychology. First, they add to our understanding of when people account (or fail to account) for uncertainty in probabilistic reasoning (e.g., Johnson et al., 2018; Zhu & Murphy, 2013). Second, they help to elucidate the mechanisms by which we compute others' mental states (Jara-Ettinger et al., 2016; Leslie, 1995). Finally, they sharpen our understanding of the interplay between domain-specific and domain-general computational principles in moral judgment (Cosmides, 1989; Mikhail, 2007). Moral reasoning may be more than just a special case of general-purpose thought.

References

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*, 187–276.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*, 353–380.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*, 148–168.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*, 589–604.

Johnson, S.G.B., & Hill, F. (2017). Belief digitization in economic prediction. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2314–2319). Austin, TX: Cognitive Science Society.

Johnson, S.G.B., Merchant, T., & Keil, F.C. (2018). *Belief digitization: Do we treat uncertainty as probabilities or as bits?* Available at SSRN.

Leslie, A. M. (1995). A theory of agency. In *Causal cognition: A multidisciplinary debate* (pp. 121–149). New York: Oxford University Press.

Malt, B.C., Ross, B.H., & Murphy, G.L. (1995). Predicting features for members of natural categories when categorization is uncertain. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 646–661.

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

Murphy, G.L., & Ross, B.H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*, 148–193.

Sperber, D., & Baumard, N. (2012). Moral reputation: An evolutionary and cognitive perspective. *Mind & Language*, *27*, 495–518.

Uhlmann, E.L., Pizarro, D.A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*, 72–81.

Zhu, J., & Murphy, G.L. (2013). Influence of emotionally charged information on category-based induction. *PLoS ONE*, *8*, e54286.

Appendix

		Pretest A			
		$P(H_{Neut}\|U_{Imm})$	$P(H_{Imm}\|U_{Imm})$	$P(H_{Neut}\|U_{Other})$	$P(H_{Other}\|U_{Other})$
Item 1	Hitting bicyclist	65.4	31.5	73.4	29.4
Item 2	Taking someone's umbrella	78.7	21.4	72.5	23.6
Item 3	Assigning jobs	75.7	24.8	67.5	32.1
Item 4	Staring at student	69.4	29.8	73.1	26.2
Item 5	Writing essay	81.0	17.9	74.2	25.4
Item 6	Hitting sports opponent	81.4	18.2	75.1	25.0
Item 7	Child's eye injury	73.3	25.8	82.0	16.9
Item 8	Medical recommendation	76.0	24.3	76.8	23.5
Mean		75.1	24.2	74.3	25.3
		Pretest B			
		$P(Z_{Imm}\|H_{Imm})$	$P(Z_{Imm}\|H_{Neut})$	$P(Z_{Other}\|H_{Other})$	$P(Z_{Other}\|H_{Neut})$
Item 1	Hitting bicyclist	55.3	28.6	45.9	23.3
Item 2	Taking someone's umbrella	71.4	35.2	66.5	43.3
Item 3	Assigning jobs	84.3	31.0	56.2	54.4
Item 4	Staring at student	72.1	36.6	67.2	64.8
Item 5	Writing essay	60.2	9.0	72.9	57.8
Item 6	Hitting sports opponent	63.1	18.2	35.9	14.9
Item 7	Child's eye injury	72.0	33.3	71.1	67.9
Item 8	Medical recommendation	46.3	19.5	70.3	71.8
Mean		65.6	26.4	60.7	49.8
		Main Study			
		Actual		**Normative**	
		$P(Z_{Imm}\|U_{Imm})$	$P(Z_{Other}\|U_{Other})$	$P(Z_{Imm}\|U_{Imm})$	$P(Z_{Other}\|U_{Other})$
Item 1	Hitting bicyclist	37.6	35.1	37.3	29.7
Item 2	Taking someone's umbrella	43.0	39.5	42.9	49.0
Item 3	Assigning jobs	37.3	61.6	44.1	55.0
Item 4	Staring at student	41.7	62.7	47.3	65.4
Item 5	Writing essay	16.2	27.2	18.2	61.6
Item 6	Hitting sports opponent	27.1	25.7	26.4	20.2
Item 7	Child's eye injury	28.1	55.6	43.4	68.4
Item 8	Medical recommendation	24.2	54.7	26.0	71.4
Mean		31.9	45.2	35.7	52.6

Note. Entries are the mean probability judgments for each item (expressed as percentages), averaged across participants.