

Draft: December 14, 2013
Please don't quote without permission

Rational learners and non-utilitarian rules*

Shaun Nichols

University of Arizona

Shikhar Kumar

Carnegie Mellon University

Theresa Lopez

University of Arizona

1. Introduction

“Moral distinctions are not derived from reason.” Thus does Hume begin his discussion of morals in the *Treatise*. Rather, Hume says, moral distinctions come from the sentiments. Contemporary work in moral psychology has largely followed Hume in promoting emotions

* Acknowledgements: Audiences at Royal Institute of Philosophy, Brazilian conference on analytic philosophy, Science of Morality, MPRG, ISSAS, IACAP. We'd also like to thank Mark Alfano, Mike Bruno, Colin Dawson, Jerry Gaus, Michael Gill, Tom Griffiths, Daniel Jacobson, Don Loeb, Sarah Raskoff, Josh Tenenbaum, and Jen Wright for discussion and comments on this work. Thanks to Andy Simpson, Calvin Yassi, and Hannah Robb for coding.

rather than reason as the basis for moral judgment (e.g., Blair 1995; Greene 2008; Haidt 2001; Nichols 2004; Prinz 2007). While emotions do seem vital to moral judgment, we will discuss one way in which rational processes play a critical and unnoticed role in how we make moral distinctions.

Moral dilemmas have been a key tool for uncovering the moral distinctions people make. Philosophers have recruited moral dilemmas to show that we intuitively draw distinctions that are at odds with utilitarianism (e.g. Thomson 1985). In the new millennium this theme has been reinforced by hundreds of empirical studies on moral dilemmas. The most intensively studied moral dilemmas involve trains rushing towards oblivious rail-workers. In *Switch*, an agent sees that a train is bound to kill five people on the track unless the agent throws a switch that will divert the train to a side track where it will kill one person. When given this scenario, people tend to say that it is permissible for the agent to flip the switch (e.g. Greene et al. 2001; Mikhail 2007). In the *Footbridge* dilemma, an agent is on a bridge overlooking the train tracks along with a large man; again there is a train bound to kill five people, and the agent knows that he can save the five people only by pushing the large man in front of the train. People given this scenario tend to say that it is not permissible for the agent to push the man.

The results on Footbridge provide just one example in which people make judgments that appear to contravene a simple utilitarian calculation. But there are dozens of experiments that confirm the basic pattern: people often think that an action is wrong even if it produces greater benefits than any of the alternatives (see, e.g., Cushman et al. 2007; Lopez et al. 2009; Mikhail 2011). These empirical findings have underscored a further question – *why* do people make non-utilitarian judgments? We will champion a rational learning approach to the issue, drawing on

recent work in statistical learning theory. But before we set out our own view, we need to review the prevailing accounts.

2. Background

2.1. Emotional process theory

The most prominent psychological account of the observed pattern of judgments is the dual-process theory of moral judgment, according to which non-utilitarian judgments are characteristically generated by emotional processes (e.g., Greene 2008). The proposal is that cases like Footbridge trigger emotions that subvert utilitarian cost-benefit analysis. Some claim this provides the foundation for an argument for the irrationality of non-utilitarian judgment. Joshua Greene has been the main advocate of this view. He suggests that deontological judgments, like “it’s wrong to push the guy in front of the train,” are defective because they are insensitive to rational considerations:

[T]he consequentialist weighing of harms and benefits is a weighing process and not an ‘alarm’ process. The sorts of emotions hypothesized to be involved here say, ‘Such-and-such matters this much. Factor it in.’ In contrast, the emotions hypothesized to drive deontological judgment are ... alarm signals that issue simple commands: ‘Don’t do it!’ or ‘Must do it!’ While such commands can be overridden, they are designed to dominate the decision rather than merely influence it (Greene 2008, 64-5).

Greene maintains that since our deontological judgments derive from irrational emotional responses, we should ignore them in normative theorizing (2008; see also Singer 2005, 347).¹

¹ For direct responses to this argument, see Berker (2009) and Timmons (2008).

Although there is a diverse array of evidence supporting the view that emotions play a role in judgments about Footbridge (Amit & Greene 2012; Bartels & Pizarro 2011; Koenigs et al. 2007), emotions cannot explain the basic phenomenon of non-utilitarian moral judgment. Many dilemmas for which people report very *low* in emotional content – e.g., those involving lying, stealing, and cheating – elicit non-utilitarian responses (see, e.g., Dean 2010). Indeed, the asymmetry between Footbridge and Switch is found even when the potential human victims are replaced by *teacups* (Nichols & Mallon 2006).²

The fact that people make non-utilitarian judgments in the absence of significant affect indicates that there must be some further explanation for these responses. It's plausible that these judgments depend critically on internally represented *rules* (Nichols & Mallon 2006). Presumably, there is something about the structure of the rules such that they apply in certain cases and not in others. But to appeal to some such rules to explain non-utilitarian judgment is a manifestly incomplete explanation. For we still need to characterize what their structure is and why the rules have this structure.

2.2. Nativism

The most prominent alternative to the dual-process theory is a nativist account of moral distinctions (e.g., Dwyer 2004; Harman 1999; Mikhail 2011). While we will offer an alternative, empiricist learning account, our proposal is largely inspired by the kinds of considerations that motivate moral nativism.

² In addition, recent work indicates that Switch is just as emotionally arousing as Footbridge (Horne & Powell 2013).

First, moral nativists reject the idea that non-utilitarian judgment is driven exclusively by emotions. Instead, nativists maintain that there are basic principles that underlie people's judgments. As Hauser and colleagues put it, the view "builds on non-consequential moral philosophy by exploring how the psychology of such distinctions as that between killing and letting die and intended harm and foreseen harm bears on the nature of our moral judgments" (Hauser et al., 2007, 2). The idea is that our inclination to distinguish between cases like Footbridge and Switch is a product of principles, not simply emotions. Indeed, in further contrast to the emotional process theory, nativists attempt to explain the acquisition of such moral distinctions rather than focus narrowly on online processing.

A second important point for the nativist is that these discriminations appear early in development. John Mikhail writes, "The judgments in trolley cases appear to be widely shared among demographically diverse populations including young children" (Mikhail 2007, 144). Pellizzoni and colleagues (2010) showed that 3-year-old children make the familiar distinctions on footbridge and switch. And Powell and colleagues show that 5-year-old children regard actions as worse than omissions (2012). Even young children have a facility with tracking intentions and forming judgments based on non-utilitarian rules.

The final point concerns the evidence available to the child. Although children acquire these abilities very early, nativists maintain that the evidence available to the child is scant.

Susan Dwyer and colleagues put the point well:

[A]lthough children do receive some moral instruction, it is not clear how this instruction could allow them to recover moral rules... [W]hen children are corrected, it is typically by way of post hoc evaluations... and such remarks are likely to be too specific and too

context dependent to provide a foundation for the sophisticated moral rules that we find in children's judgments about right and wrong (Dwyer et al. 2009, 6).

Nativists use these points to argue for an innate, domain-specific contribution to the acquisition of moral distinctions.

Although we offer an empiricist alternative, we think that the nativists are quite right about a number of important features of the acquisition of moral distinctions. We agree that children do acquire a facility with subtle distinctions, like that between acts and omissions. More interestingly, the nativists are right that children don't get a lot of explicit training on rules. They are certainly not told things like: *this rule applies to what agents do but not to what agents allow to happen*. Jen Wright and Karen Bartsch conducted a detailed analysis of a portion of CHILDES, a corpus of natural language conversations with several children (MacWhinney 2000). They coded child-directed speech for two children (ages 2 to 5) for moral content. Wright and Bartsch found that only a small fraction of moral conversation adverted to rules or principles (~5%). By contrast, disapproval, welfare, and punishment were frequently implicated in moral conversation (2008, 70).

The lack of explicit training on rules is compounded by the fact – stressed by nativists – that any particular instance of disapproval will carry many specific features, and the child has to learn to abstract away from those features to glean the general rule. Although there is very little reference to rules in child-directed speech, there is a lot of *no!*, *don't!*, and *stop!* But it seems as if these injunctions won't provide enough information to fix on the content of the rule, and this promises to be a pervasive problem for the young learner. To repeat a key point from Dwyer and colleagues, “such remarks are likely to be too specific and too context dependent to provide a foundation for the sophisticated moral rules that we find in children's judgments about right and

wrong” (Dwyer et al. 2009, 6). Any particular case of training will typically be open to too many different interpretations to allow for the child to draw the appropriate inferences about the relevant distinctions. The nativists are right that the evidence available to the child is scant and seems to underdetermine the content. But it is at this juncture that we think that new work in statistical learning can help explain how these distinctions may be acquired.

3. Bayesian learning

Like nativists, we think that emotional process models by themselves are bound to be inadequate to the task of explaining why our judgments diverge from utilitarian verdicts. Unlike nativists, we will explore an empiricist learning theoretic explanation for how the rules driving these judgments might be acquired (see also Lopez forthcoming). Bayesian statistical inference has emerged as a powerful theoretical approach for understanding learning across a variety of domains. Over the last decade, a wide range of learning problems have been illuminated by Bayesian learning models, including categorization (Kemp et al. 2007), the acquisition of grammar (Perfors et al. 2011a), and word learning (Xu & Tenenbaum 2007). The principles of Bayesian learning extend naturally to modeling the acquisition of rules of conduct. Unlike other approaches to learning and reasoning (e.g., Rumelhart & McClelland 1986), Bayesian approaches allow a central role for structured, symbolic representations, which can serve as hypotheses that are assigned different levels of certainty (e.g., Goodman et al. 2010; Perfors et al. 2011b). Thus, a Bayesian explanation of the acquisition of moral rules can model different candidate moral rules as structured representations, and these representations will be assigned different levels of certainty in light of available evidence. These assignments are guided by principles of rational statistical inference. In what follows, we will provide a Bayesian account of

why children acquire rules focused on what agents *do* rather than utilitarian rules focused on maximizing valued outcomes, even where both kinds of rules are consistent with the evidence.

3.1. The size principle

The model that we offer involves a simple Bayesian principle, the *size principle* (e.g., Perfors et al. 2011; Tenenbaum & Griffiths 2001). To get an intuitive sense of the principle, imagine that a friend has a box of 4 fair dice, each with a different denomination: 4, 6, 8, and 10. He pulls out one die at random and rolls it 10 times, reporting that the outcomes were 3 2 2 3 4 2 3 4 2 2. Is it likely that he's rolling the 10 sided die? Of course not. Why? Because you would have expected *some* numbers over 4 if it were the 10. If it were the 10, it would be a *suspicious coincidence* that all the observations were ≤ 4 . The size principle offers a systematic way to capture this intuitive fact. Let's call the hypothesis that the die is 4-sided h_4 , the hypothesis that the die is 6-sided h_6 , and so on. We can represent the size of the hypotheses by a nested structure (figure 1).

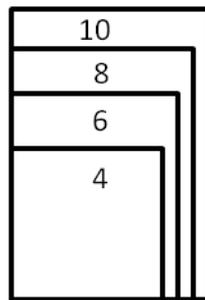


Figure 1: The numbers represent the highest denomination of the die; the rectangles represent the relative sizes of the hypotheses

Again, suppose that your friend pulls out a die at random, so the prior probability is the same for h_4 , h_6 , h_8 , and h_{10} . Suppose again the first roll comes up 3. That result is consistent with both h_4

and h_{10} , but the probability of 3 under h_4 is .25, and the probability of 3 under h_{10} is .1. The second roll is 2. That result too has probability .25 under h_4 and .1 under h_{10} ; since we now have two rolls that are consistent with both h_4 and h_{10} , we square those probabilities for the joint probability, yielding .0625 for h_4 and .01 for h_{10} . With three consistent rolls (3, 2, 2), we cube the probabilities to yield .015 as the joint probability given h_4 , and .001 for h_{10} . This process illustrates the fact that smaller hypothesis that are consistent with the data (e.g., h_4) are significantly preferred to larger hypotheses (e.g., h_{10}), and this advantage increases exponentially with each new data point.³

3.2. The size principle and word learning

Xu and Tenenbaum use the size principle to explain a striking feature of word learning in children. When learning the word “dog,” children need only a few positive examples in which different dogs are called “dog” to infer that the extension of the term is $[[\text{dog}]]$ rather than $[[\text{animal}]]$. Pointing to a Dalmatian, a terrier, and a mutt suffices. You don’t also need to point to a robin or a fish and say “that’s not a dog.” Xu and Tenenbaum explain this in terms of the size principle. Put most succinctly, the likelihood of getting those particular examples (a Dalmatian, a terrier, and a mutt) is much higher if the extension of the word is $[[\text{dog}]]$ as compared with $[[\text{animal}]]$. Consider the nested hypothesis space in which $[[\text{dog}]]$ has the smallest extension, and

³ The general principle can be expressed as follows:

$$p(d|h) = \left[\frac{1}{\text{size}(h)} \right]^n$$

The size principle is an instance of Bayesian Occam’s razor (MacKay 2003), a more general principle that emerges naturally in Bayesian inference.

this extension is contained in the larger hypothesis [[animal]], which itself is contained in the larger hypothesis [[living thing]] (Xu & Tenenbaum 2007, 248). If there are several examples in the training set, all of which are consistent with the hypothesis that the extension of the word is [[dog]], then the size principle would dictate that it's much more likely that one would get these examples if the word means [[dog]] as compared to [[animal]] or [[living thing]]. As a result, if the prior probabilities for those interpretations are roughly the same, then the statistically appropriate inference is that the extension of the word corresponds to the hypothesis with the smallest extension, i.e., [[dog]].

Xu and Tenenbaum report word learning experiments and Bayesian simulations that suggest that participants conform to this kind of inference. Our own experiment and simulation are closely modeled on Xu and Tenenbaum's work, so we will explain it in some detail. In a word learning task, adult participants were presented with a nonsense syllable, e.g., "Here is a fep," accompanied by a pictured object; the task was to generalize the application of that word to other depicted objects. In some trials, participants saw one sample application of the word. For example, they might be told "Here is a fep" and shown a picture of a Dalmatian. In other trials, they were shown three sample applications. For instance, they might be told "Here are three feps" and shown pictures of a Dalmatian, a terrier, and a mutt. In other trials, they were shown three examples of the new word drawn from the same superordinate-level category (e.g., a Dalmatian, a toucan, and a pig) (253). When given examples across the superordinate level, participants generalized to animals in general (e.g., extending the reference of the term to a cat, a seal, and a bear). However, when given examples of a Dalmatian, a terrier and a mutt, participants generalized only to other dogs.

After the word learning portion of the task, participants were presented with pairs from the learning phase (e.g., Dalmatian and terrier) and asked to indicate, for each pair, how similar they are. They were explicitly told to base their similarity ratings on the features of the objects that were important to their judgments in the word-learning phase. The similarity ratings provide natural clustering (e.g., Dalmatians cluster more with other dogs than with birds) and this is used to generate a hierarchical representation of the hypothesis space guiding subjects' word learning. Using this representation of the hypothesis space, Xu and Tenenbaum ran a Bayesian simulation of word learning and found that the Bayesian model closely approximated human performance (263).

4. Bayesian analysis of rule learning

Just as the hypotheses concerning the dice form a subset structure (figure 1), a subset structure characterizes several distinctions of interest in the normative domain, depicted in Figure 2.

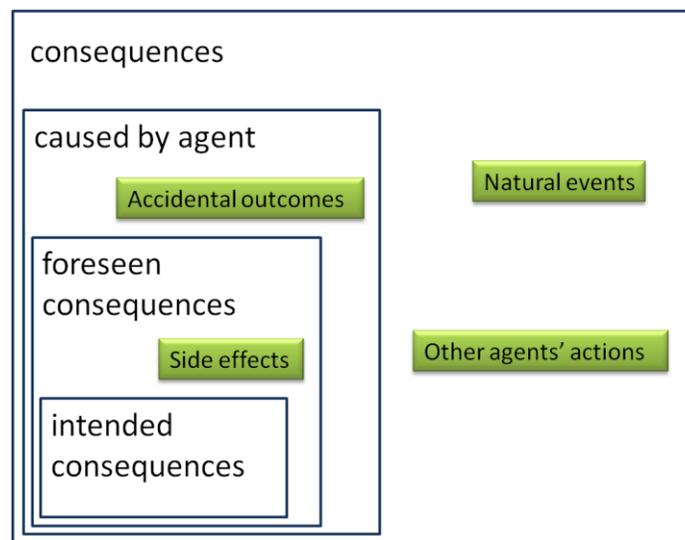


Figure 2: The scope of rules represented in a subset structure

The class of actions in which one intentionally produces an outcome (*intended consequences*) is the narrowest set. For instance, if I intentionally scratch a car, this fits into the narrow class of a consequence I intentionally produce. A wider class is formed by including cases in which my action leads to a side effect that I foresee but don't actually want to produce (*foreseen consequences*). For instance, I might open my car door wide enough to get out, knowing that this will scratch the car next to me. A wider class still includes accidental production of the consequence, like accidentally scratching a car. This wider class that includes accidents can be thought of as the set of consequences *caused by the agent*. A much wider class is created if we also include consequences that are *not* caused by the agent, for instance, outcomes that are caused by natural events or by other agents (*consequences*). Rules might be formulated at any of these "scopes." A rule at the narrowest scope might prohibit an agent from intentionally producing an outcome, e.g., intentionally scratching a car. At the broadest scope, a rule might prohibit tolerating the outcome, even if it is produced by someone or something else. For instance, there might be a rule indicating that agents must ensure that cars don't get scratched.

Familiar moral distinctions can be captured in terms of this subset structure.⁴ Consider, for instance, the act/allow distinction. In many cases, a prohibition applies to what an agent *does* but not to what the agent *allows to happen*. In the subset structure, that means that the rule is not

⁴ The precise boundaries of these distinctions is a delicate issue in philosophy (see, e.g., McNaughtan & Rawlings 1991; Parfit 1984; Schroeder 2007). For present purposes, we will not attempt to give precise renderings of these distinctions, but only note that there is considerable overlap between the familiar distinctions and the subset structure in figure 2.

extended to the widest scope. Or consider the intended/foreseen distinction. In some cases, a prohibition might apply to what an agent *intends*, but not to what an agent foresees as an unintended side effect of his action. In that case, the rule would have the narrowest scope in the subset structure. Utilitarian considerations – on which one maximizes desirable outcomes – might be represented as rules at the widest scope. This conforms to the idea that utilitarian considerations generally seek to maximize desirable outcomes independently of who produces the desired outcomes.

Given this subset structure, the size principle has the potential to explain critical features of rule learning. Imagine trying to learn a rule of conduct for a different culture. The available hypotheses are: h_n – the rule prohibits putting things on the sand, and h_w – the rule prohibits allowing things to be on the sand. Hypothesis h_n has *narrow* scope, applying to an agent's action; h_w has *wide* scope, applying to what the agent allows. Now imagine that there are several known violations of the rule, all of which are cases in which a person has intentionally put something on the sand. Following the size principle, one should assign higher probability to the narrow scope hypothesis that the rule prohibits intentionally putting things on the sand. As with the dice, it would be a statistically suspicious coincidence if h_w were the right hypothesis, given that all the evidence is consistent with h_n .

5. Evidence

The foregoing analysis suggests how Bayesian learning theory might explain how people acquire narrow-scope (and hence non-utilitarian) rules. If (i) when people are acquiring rules, they approximate Bayesian learners and (ii) the sample violations for a given rule are consistent with a narrow scope interpretation, then people should infer that the rule has narrow scope. A full

demonstration of this is obviously beyond the reach of this paper, but we will examine several critical components

5.1. The data: Evidence from child-directed speech

The first question concerns the kind of evidence available to children. We investigated this by looking at parental instruction in a large corpus of child-directed speech (CHILDES; MacWhinney 2000). Sensitivity to moral distinctions, like that between acting and allowing, has been observed in children from 3 to 5 years old (e.g., Pellizzoni et al. 2010; Powell et al. 2012). So we looked at child-directed speech from 33-36 months. In the CHILDES database, there are four children (Abe, Adam, Ross, and Sarah) for whom there are data in that age range. Naïve independent coders went through this portion of the database and identified child-directed speech relevant to rules of conduct, in particular, speech in which adults were communicating something directly relevant to how to behave. Those statements were then coded again for consistency with narrow-scope interpretations of the rules. Coders were trained on the distinction between narrow-scope and wide-scope. They were told that narrow-scope rules have the form “agent(s) shouldn’t cause outcome S”, whereas wide-scope rules have the form: “agent(s) should not cause outcome S nor allow such a state of affairs to persist.”⁵ There was very high inter-coder agreement (over 99%), and the few disagreements were settled by discussion.

⁵ The database also includes cases in which an action is required. For such positive cases coders were told that narrow scope rules have the form “agents should produce this (agent-specific) outcome”, so different outcomes should be produced by different agents (e.g., *one should brush one’s own teeth* or *one should care for one’s own children*); wide scope rules have the form “agents should maximize this sort of outcome,” so the same outcome should be sought by all

The results were clear. Over 99% of the cases of adult communication on behavior was consistent with a narrow scope interpretation. Typical examples include “don’t hit anybody with that Adam,” “don’t throw paper on the floor,” and “don’t write on that.” Of course, there were also many many cases of parents just saying “no!” to express disapproval over a child’s action. Thus, the database evidence indicates that if children learn rules by approximating Bayesian inference, for most rules, they would naturally come to believe that the rules prohibit *acting*.⁶

5.2. The Likelihood

Given the available evidence about rules, Bayesian inference would point to a narrow-scope interpretation. But it is a further question whether people approximate Bayesian learners. In particular, when people are learning rules, are they sensitive to the likelihood – the fit between the data (i.e. examples of violations) and the hypothesis (i.e. the scope of the rule).

5.2.1. Learning task 1

We investigated this first by adapting Xu and Tenenbaum’s (2007) word learning task into a rule-learning task. The subjects’ job was to figure out the meaning of a rule from a foreign

agents (e.g., *one should ensure that children are cared for* or *one should ensure that children are cared for by their own parents*).

⁶ The one clear case that was coded as inconsistent with narrow scope is itself interesting. It’s a case in which the child is told not to let his little brother fall. The case is interesting because the protection of children is one area of commonsense ethics that does tend to involve wide-scope rules. It’s not enough to refrain from intentionally hurting children; one is obligated to ensure the safety of children in one’s vicinity.

culture, given sample violations of the rule.⁷ The rules were labeled with nonsense terms, e.g. “taf byrnal” or “zib matan”. Since our interest was in rule learning, we used examples that were arbitrary and unemotional. For each rule, participants were presented with examples of violations of that rule. In some trials, all of the sample violations were consistent with a narrow scope interpretation in which an agent *acts*, e.g., “Mike puts a block onto the shelf.” In other trials, some of the sample violations were *inconsistent* with a narrow-scope interpretation, e.g., “Dan doesn’t pick up a jump-rope that he notices on the shelf.” After being exposed to the examples for a given rule, participants then had to indicate which other cases were violations of that rule.

We found that when participants were exposed only to examples that were consistent with narrow scope, participants overwhelmingly selected only cases in which the person *acted* (e.g., “Chris places a toy truck on the shelf”). However, if participants were exposed to two cases that were inconsistent with narrow scope interpretation, participants overwhelmingly selected cases in which the person either acted or *allowed* a state of affairs to persist (e.g., “Emily sees a marble on the shelf and walks past it”) (see figure 3).⁸

⁷Participants were recruited through Amazon mturk. The task itself was rather tedious and time-consuming. Since many participants completed the survey far too quickly, we calculated their cumulative time and defined a threshold based on that. Subjects lying below the 75th percentile were rejected, leaving 18 participants (11 female).

⁸ As expected, people did not confuse domains – they tended to generalize from chalkboard examples to other chalkboard examples and not to shelf cases (there was < 1% errors on domain (5 out of 1296)). When presented with two examples that are inconsistent with narrow scope, participants overwhelmingly generalize to include wide-scope cases (one sample t-test

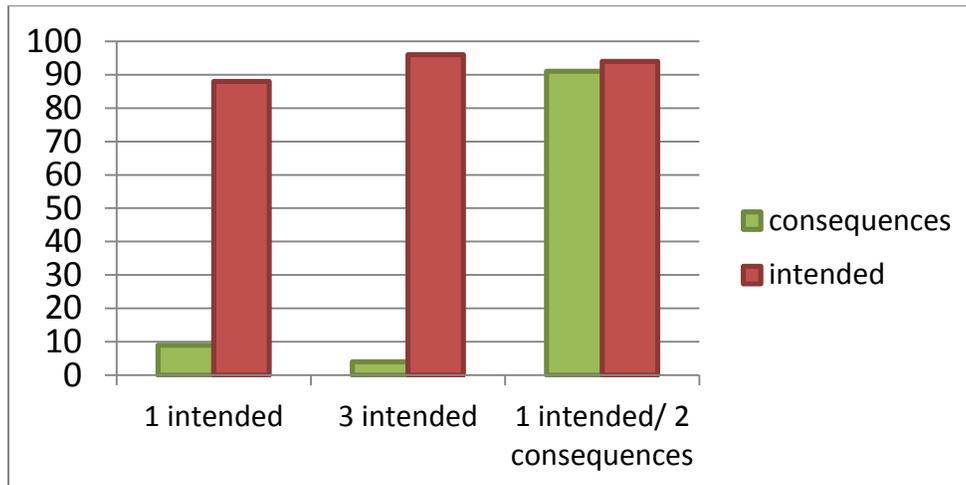


Figure 3: Generalization of scope of rules

Thus, in our learning task, people are appropriately sensitive to whether the examples are consistent or inconsistent with a narrow scope interpretation. In effect, people shift to a wide-scope interpretation when given two examples that don't fit the narrow scope interpretation.

5.2.2. Learning task 2

Although our first learning study showed that people are sensitive to information about the scope of the rule, this study did not show that people conform to the size principle in rule learning. The size principle predicts that people should be more inclined to adopt a narrow scope interpretation

$t(17)=20.4, p<.0001$); when presented only with 3 cases that are consistent with narrow scope, participants overwhelmingly refrained from generalizing to wide scope cases (one sample t-test $t(17)=7.9, p<.0001$). And of course, there were a significant difference between these two conditions ($t(17)=19.96, p<.0001$).

when given 3 examples consistent with narrow scope than when given a single such example. However, even when given a single example, people strongly favor the narrow scope interpretation.⁹ In effect, people show a strong bias in favor of the narrow scope interpretation.¹⁰

The fact that there is a strong bias for narrow scope doesn't exclude a rational learning theory. For there are natural resources for explaining the acquisition of this bias. If most of the rules that children acquire are narrow-scope rules, then this plausibly forms the basis for developing an *overhypothesis* (Goodman 1955) about the nature of rules, according to which most rules are narrow-scope. Recent work indicates that children, including infants, do form overhypotheses in learning (Dewar & Xu 2010; Smith et al. 2002). Obviously it will be important in future work to explore whether the narrow-scope bias is acquired as an overhypothesis. But for present purposes there is a more pressing concern – *do* people make judgments that conform to the size principle? To investigate this, we developed a version of the rule learning task that is more sensitive to variations in probability judgments.

As before, participants were told that they were learning foreign rules. In each block they were first given one example of a violation. The example was always consistent with a narrow-scope interpretation. For example, in one case, the rule was called “nib weighs” and this was the sample violation: “Lisa draws a circle on the chalkboard”. Participants were then presented with

⁹ Responses differed in the right direction – there were more generalization to ‘wide scope’ interpretation when participants were given a single example. However, this difference was far from significance ($t(17)=1.31, p=.20$).

¹⁰ This parallels Xu and Tenenbaum’s finding that in word learning tasks, adults show a distinct bias for basic-level categories (2007, 263).

two cases and asked whether the people in those cases were also violating the rule. In all blocks, these two cases were *inconsistent* with a narrow-scope interpretation, e.g., “Susan sees a drawing of a cat on the chalkboard and doesn’t rub it out”. Participants were first asked, in light of the one example they had seen, to assess how likely it is that the individuals in the two cases are also breaking the rule. Then, in the critical part of the study, participants were shown five more examples of violations of the rule. For some rules, two of the five examples were inconsistent with narrow scope interpretation; for the other rules, all five examples were consistent with narrow-scope interpretation.

Corroborating results from task 1, we found that when two of the new examples were inconsistent with narrow scope, participants judged that it was more likely that the “allowers” were violating the rule (one sample $t(31)=3.974, p<.001$). More importantly, we found that when all five of the new examples were consistent with narrow scope, participants change their judgment in the other direction, saying that it’s less likely that the agents are breaking the rule (one-sample $t(31) = 3.72, p<.0001$). Thus, participants’ inferences about the scope of rules do conform to the size principle.

5.3. The Prior

As theorists, we find the subset structure in figure 2 intuitive. But it is a further question whether ordinary people carve things up the same way. Following Xu and Tenenbaum (2007), we used a similarity task to assess the hypothesis space that people bring to rule learning. After participants completed the rule-learning portion of the learning task 1, they rated how similar they regarded dozens of pairs of scenarios assembled from items included on the rule-learning component. Of course, similarity judgments depend on the background task. If asked to group things by size, then a dachshund is more similar to a ferret than it is to a Rottweiler. To ensure that participants

were focusing on the relevant similarity metric, they were instructed to make these similarity judgments based on the same aspects of the scenarios that were important in making their earlier decisions about the meaning of the rules. These similarity judgments provided the data for an average linking algorithm (Duda & Hart, 1973) that we used to generate a hierarchical representation of the hypothesis space. The results are presented in figure 4. The hierarchy indicates that people are highly sensitive to scope features when making inferences about rules. In particular, *intended consequences* form a unique cluster under each domain.

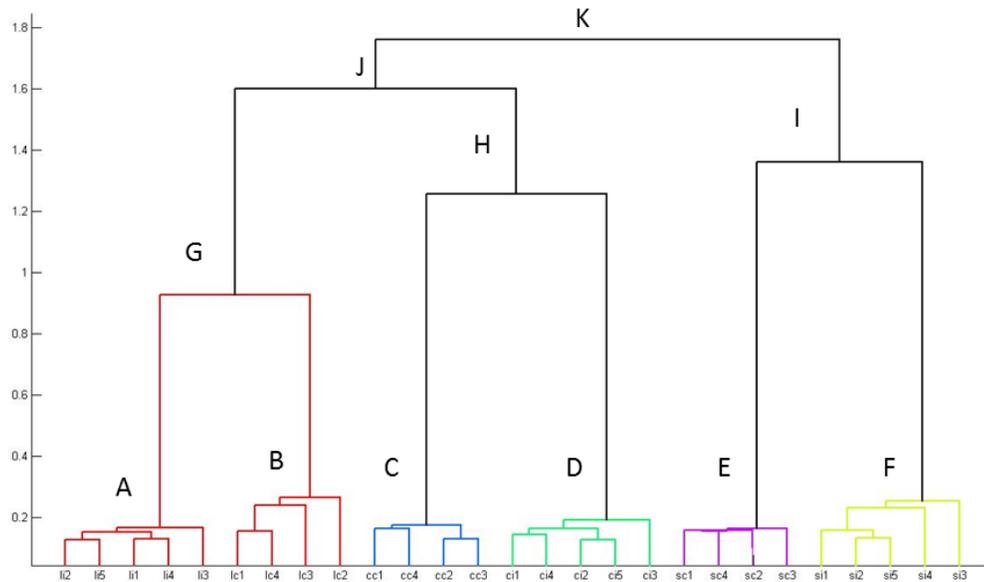


Figure 4: Hierarchical representation of hypothesis space, based on similarity judgments. Letters G, H, I represent domains (litter, chalkboard, shelf). Letters A-F represent unique clusters. All the cases in A, D, and F are cases of intended consequences. None of the cases in B, C, or E are cases of intended consequences. The *y-axis* represents the distance (height) between clusters. For example, the distance between G and A would be $Height(G) - Height(A)$.

5.4. The Bayesian model

We used this hierarchical representation of the hypothesis space to build a Bayesian model to simulate rule learning.¹¹ Formally, the problem can be defined as learning a single rule R from a set of examples I drawn from some known domain D , where $I = i_1, \dots, i_n$. As with other standard Bayesian learning models, our model assumes that the learner has access to a hypothesis space H , containing a set of candidate hypotheses for representing the rule R and a probabilistic model to relate hypotheses $h \in H$ to the evidence I . Given this information, the Bayesian framework provides a statistical measure for inferring the rule R based on the evidence I .

For the observed evidence I , the Bayesian learner computes the posterior probabilities $p(h|I)$ for different hypotheses $h \in H$, using Bayes' rule:

$$p(h|I) = \frac{p(I|h)p(h)}{\sum_{h' \in H} p(I|h')p(h')}$$

As noted in 5.3, for our model, we generated the hypothesis space using the similarity ratings of the scenarios. The hierarchical tree represents the distance between different clusters, reflecting how similar/dissimilar they are from each other. All the examples in A are intended consequences involving litter, so A is naturally interpreted as a narrow scope hypothesis; by contrast, G contains intended consequences involving litter and consequences involving litter that are not intended, so G is naturally interpreted as a wide scope hypothesis.

¹¹ Again, we are closely following the technique used by Xu and Tenenbaum (2007).

In order to understand how the Bayesian learner would choose different hypotheses from the hypothesis space based on the evidence, let's consider an example where the evidence is li1. In this case there are 2 possible hypotheses to consider: A and G. All other hypotheses are ignored because they don't contain the evidence li1. The prior for each hypothesis is computed as the difference in heights of the node and its parent:

$$p(h) = \text{height}(\text{parent}[h]) - \text{height}[h]$$

Thus, for hypothesis A, the prior is determined by $(\text{height}[G] - \text{height}[A])$ and for hypothesis G, the prior is determined by $(\text{height}[J] - \text{height}[G])$.

Similarly, the likelihoods are computed for each hypothesis. The likelihoods are computed based on the size principle:

$$p(I|h) = \left[\frac{1}{\text{height}(h) + \sigma} \right]^n$$

where n is the number of examples and σ is a small constant ($\sigma = .05$) introduced to prevent the likelihood of the lowest nodes from going to infinity. For hypothesis A, the likelihood is $[1/(\text{height}[A] + \sigma)]$ and for G, the likelihood is $[1/(\text{height}[G] + \sigma)]$. Since we are considering the case in which we have only one example (li1), $n=1$; when we have three examples, the expression is raised to the power of 3.

To run the simulation, the model was presented with examples corresponding to the sample violations presented to participants in learning task 1. For each example or set of examples presented to the model, the model used Bayes' rule to calculate the posterior probability for each hypothesis. The model responded to the evidence much like human subjects do. When given cases that are exclusively narrow (e.g., li1), the probability of the model choosing intended cases is 100% and the probability of the model choosing unintended cases is

very low; when given cases that are wide (e.g. li1, lc1, lc2), the probability of the model choosing both intended and consequence based cases is 100% (figure 5).

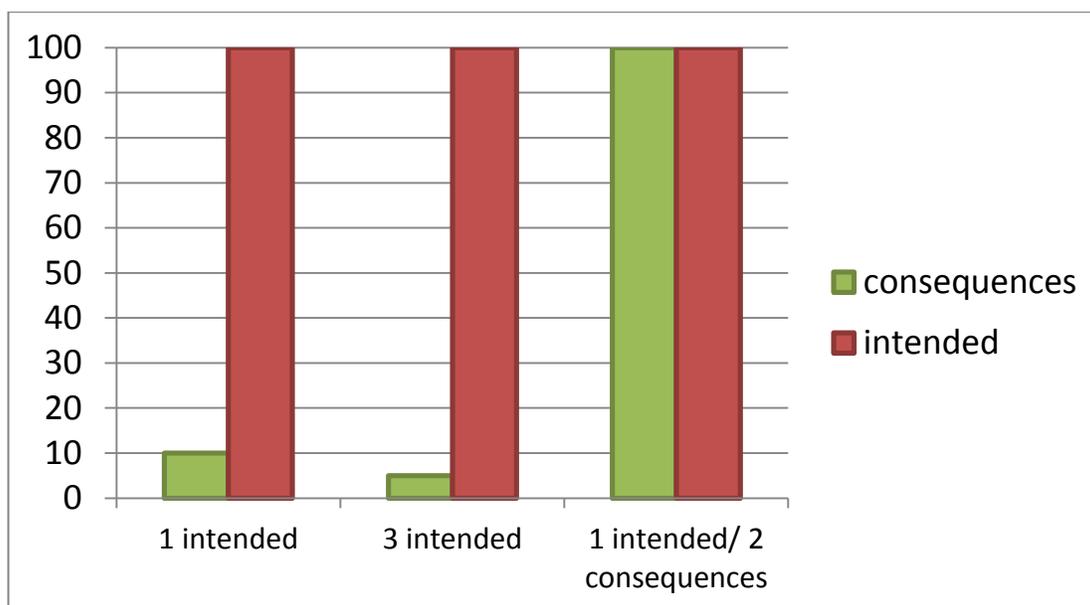


Figure 5: Predictions of the Bayesian model

6. Conclusion

When people are presented with moral dilemmas, they often respond in ways that do not conform to utilitarian principles. For instance, people tend to judge that it's worse to cause a bad outcome than to allow a bad outcome to persist. One explanation for non-utilitarian judgment is that people actually operate with non-utilitarian rules. However, identifying and explaining the structure of the rules has remained elusive. Moral judgment is sensitive to a wide range of factors, including emotions, framing, and values. This had made it extremely difficult to identify precisely which aspects of judgment derive from structured rules and which aspects of judgment derive from other factors. The difficulty here is reflected by the fact that moral philosophers have

failed to achieve anything approaching a consensus concerning the detailed character of moral rules.

We have approached this issue through the lens of statistical learning. Our hypothesis is that non-utilitarian judgment derives from learning narrow-scope rules, i.e., rules that prohibit *acting*, in a way that approximates Bayesian learning. Our learning experiment indicates that, when learning a new rule, adults are sensitive to evidence concerning the scope of transgressions. When exposed only to cases that are consistent with a narrow-scope interpretation, people overwhelmingly favor the narrow-scope interpretation. By contrast, when exposed to cases that are inconsistent with a narrow-scope interpretation, people quickly move to a wide-scope interpretation of the rule focused on maximizing consequences. The evidence thus suggests that if people were exposed to sample moral violations that were inconsistent with a narrow-scope interpretation, they would acquire a rule with wider scope. However, the evidence from CHILDES suggests that children are generally *not* exposed to this kind of evidence. The overwhelming preponderance of child-directed speech concerning conduct is consistent with a narrow-scope interpretation of many rules. While we did not conduct a learning study on children, a growing body of evidence indicates that children learn aspects of language in ways that approximate Bayesian inference (Gerken 2010; Dawson & Gerken 2009, 2011; Xu & Kushner 2013). Indeed, children learn words in ways that conform to the size principle (Xu & Tenenbaum 2007).

Our evidence supports the hypothesis that non-utilitarian rules are acquired through a process that approximates Bayesian inference. This account obviously aims to provide an alternative to nativist accounts of the acquisition of moral distinctions (see also Lopez forthcoming). In addition, the Bayesian account provides new grounds for thinking that non-

utilitarian judgment derives in part from the structure of moral rules. As noted earlier, the complexity of factors in moral judgment makes it difficult to determine which aspects of judgment are contributed by rules and which by emotions, frames, or values. Our statistical learning approach provides a new way in to this problem. For our account suggests that children would learn rules that have narrow scope built into their structure. This provides reason to think that it is part of the structure of moral rules that they are encoded as narrow scope.

For our empirical investigations, we took on the starkest distinction in the subset structure (figure 2). We looked at rules aimed at intended consequences as compared to rules directed at consequences in general. Our results suggest that statistical learning provides a plausible explanation for why people come to have rules with a narrow-scope that applies to intended consequences as opposed to consequences in general. Although our focus was limited, the basic idea might extend to explain the acquisition of other, more fine-grained distinctions, like that between outcomes that are intended and those that are unintended but foreseen.

These results also promise wider conclusions about the nature of moral judgment. Most broadly, the results suggest that the way people come to draw moral distinctions derives in a significant part from reason. Insofar as sentimentalists eschew any role for reason in the genesis of moral distinctions, they will be missing a critical element of human moral judgment. This point applies more immediately to recent work on non-utilitarian judgment. As noted above (section 2.1), one prominent proposal is that irrational features of the human mind interfere with the kind of rational cognition epitomized by utilitarian reasoning, and this provides reason to disregard those non-utilitarian judgments (Baron 1994; Greene 2008; Singer 2005; Unger 1996). On this view, people's non-utilitarian judgments are a result of rational failures that occur when we evaluate cases. Our Bayesian approach paints quite a different picture. Given the evidence

that is available to the learner, it would be statistically *irrational* to infer utilitarian rules. Of course, we might have other grounds for rejecting commonsense non-utilitarianism. But the Bayesian account undercuts wholesale attempts to cast commonsense non-utilitarianism as the product of irrationality.

References

- Amit, Elinor, & Greene, Joshua (2012). You See, the Ends Don't Justify the Means. *Psychological Science*, 23(8), 861-868.
- Baron, Jonathan (1994). Nonconsequentialist decisions. *Behavioral and Brain Sciences*, 17, 1-1.
- Bartels, Daniel, & Pizarro, David (2011). The mismeasure of morals. *Cognition*, 121, 154-161.
- Berker, Selim (2009). The normative insignificance of neuroscience. *Philosophy & Public Affairs*, 37, 293-329.
- Blair, James (1995). A Cognitive-Developmental Approach to Psychopathy. *Cognition* 57,1-29.
- Cushman, Fiery, Young, Liane, & Hauser, Marc (2006). The role of conscious reasoning and intuition in moral judgment. *Psychological science*, 17(12), 1082-1089.
- Dawson, Colin, & Gerken, LouAnn (2009). Language and music become distinct domains through experience. *Cognition*, 111(3), 378-382.
- Dawson, Colin, & Gerken, LouAnn (2011). When global structure "explains away" evidence for local grammar. *Cognition*, 120(3), 350-359.
- Dean, Richard (2010). Does neuroscience undermine deontological theory? *Neuroethics*, 3(1), 43-60.

- Dewar, Kathryn, & Xu, Fei (2010). Induction, Overhypothesis, and the Origin of Abstract Knowledge Evidence From 9-Month-Old Infants. *Psychological Science*, 21(12), 1871-1877.
- Duda, Richard, & Hart, Peter (1973). *Pattern classification and scene analysis*. New York: Wiley.
- Dwyer, Susan (2004). How good is the linguistic analogy. *The innate mind*, 2, 237-256.
- Dwyer, Susan, Huebner, Bryce, & Hauser, Marc (2009). The linguistic analogy. *Topics in cognitive science*, 2(3), 486-510.
- Gerken, LouAnn (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition*, 115(2), 362-366.
- Goodman, Nelson (1955). *Fact, fiction, and forecast*. Harvard University Press.
- Goodman, Noah, Tenenbaum, Joshua, Feldman, Jacob, & Griffiths, Tom (2010). A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science*, 32(1), 108-154.
- Greene, Joshua, Sommerville, R. Brian, Nystrom, Leigh, Darley, John, & Cohen, Jonathan (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105-2108.
- Greene, Joshua (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (ed.), *Moral Psychology*, Vol. 3, Cambridge, MA: MIT Press, 59-66.
- Haidt, Jonathan (2001). The Emotional Dog and Its Rational Tail. *Psychological Review* 108: 814-834.
- Harman, Gil (1999). Moral Philosophy and Linguistics. In K. Brinkmann (ed.), *Proceedings of the 20th World Congress of Philosophy: Volume 1: Ethics*. Bowling Green, OH:

- Philosophy Documentation Center, 107-115. Reprinted in his *Explaining Value*, Oxford: Oxford University Press, 217-226.
- Hauser, Marc; Cushman, Fiery; Young, Liane; Kang-Xing Jin, R.; & Mikhail, John (2007). A dissociation between moral judgments and justifications. *Mind & Language*, 22, 1-21.
- Horne, Zachary and Powell, Derek (2013). More than a feeling. In M. Knauf et al. (Eds.) *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Kemp, Charles; Perfors, Amy; & Tenenbaum, Joshua (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental science*, 10(3), 307-321.
- Koenigs, Michael; Young, Liane; Adolphs, Ralph; Tranel, Daniel; Cushman, Fiery; Hauser, Marc; & Damasio, Anthony (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446(7138), 908-911.
- Lopez, Theresa (forthcoming). The Bayesian Mind: Moral Rules and Moral Nativism.
- Lopez, Theresa; Zamzow, Jennifer; Gill, Michael; & Nichols, Shaun (2009). Side constraints and the structure of commonsense ethics. *Philosophical Perspectives*, 23(1), 305-319.
- McNaughton, David & Rawling, Piers (1991). Agent-relativity and the doing-happening distinction. *Philosophical Studies*, 63(2), 167-185.
- MacWhinney, Brian (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum Associates.
- Mikhail, John (2007). Universal moral grammar. *Trends in Cognitive Sciences*, 11, 143–152.
- Mikhail, John (2011). *Elements of Moral Cognition*. Cambridge: Cambridge University Press.
- Nichols, Shaun (2004). *Sentimental Rules*. New York: Oxford University Press.

- Nichols, Shaun & Mallon, Ron (2006). Moral dilemmas and moral rules. *Cognition*, 100 (3), 530-542.
- Parfit, Derek (1984). *Reasons and Persons*. Oxford University Press.
- Pellizzoni, Sandra; Siegal, Michael; & Surian, Luca (2010). The contact principle and utilitarian moral judgments in young children. *Developmental science*, 13(2), 265-270.
- Perfors, Amy, Tenenbaum, Joshua & Regier, Terry (2011a). The learnability of abstract syntactic principles. *Cognition*, 118(3), 306-338.
- Perfors, Amy; Tenenbaum, Joshua; Griffiths, Tom; & Xu, Fei (2011b). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, 120, 302-321.
- Powell, Nina; Derbyshire, Stuart; Guttentag, Robert (2012). Biases in children's and adults' moral judgments. *Journal of experimental child psychology*.
- Prinz, Jesse (2007). *The Emotional Construction of Morals*. Oxford, UK: Oxford University Press.
- Rumelhart, David & McClelland, James (1986). *Parallel distributed processing* (Vol. 1). Cambridge, MA: MIT Press.
- Schroeder, Mark (2007). Reasons and agent-neutrality. *Philosophical Studies*, 135(2), 279-306.
- Singer, Peter (2005). Ethics and Intuitions. *Journal of Ethics*, 9, 331-352.
- Smith, Linda; Jones, Susan; Landau, Barbara, Gershkoff-Stowe, Lisa; & Samuelson, Larissa (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, 13(1), 13-19.
- Tenenbaum, Joshua & Griffiths, Tom (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24, 629-641.
- Thomson, Judith (1985). Double Effect, Triple Effect and the Trolley Problem. *Yale Law Journal*, 94, 1395-1415.

Timmons, Mark (2008). Towards a Sentimentalist Deontology. In W. Sinnott-Armstrong (ed.)

Moral Psychology, Vol. 3. Cambridge, MA: MIT Press

Unger, Peter (1996). *Living High and Letting Die*. Oxford University Press.

Xu, Fei & Tenenbaum, Joshua (2007). Word learning as Bayesian inference. *Psychological review*, 114(2), 245.

Xu, Fei & Kushnir, Tamar (2013). Infants Are Rational Constructivist Learners. *Psychological Science*, 22(1), 28–32.