Forthcoming in *Mind & Language*

Rational learners and moral rules*

Shaun Nichols

University of Arizona

Shikhar Kumar

University of Arizona

---

* Acknowledgements: We'd like to thank audiences at the Royal Institute of Philosophy, the Brazilian Conference on Analytic Philosophy, the International Association for Computing and Philosophy, UC Irvine, Princeton University, Oxford University, and Cambridge University. We'd also like to thank Mark Alfano, Mike Bruno, Colin Dawson, Jerry Gaus, Michael Gill, Tom Griffiths, Steven Gross, Daniel Jacobson, Don Loeb, Edouard Machery, Sarah Raskoff, Josh Tenenbaum, Jen Wright, and an anonymous referee for discussion and comments on this work. Thanks to Andy Simpson, Calvin Yassi, and Hannah Robb for coding. This research was supported by Office of Naval Research grant #11492159 to SN.

Address for correspondence: Shaun Nichols, Department of Philosophy, University of Arizona, Tucson, AZ 85721, USA.

Email: sbn@email.arizona.edu

Theresa Lopez

Hamilton College

Alisabeth Ayars

University of Arizona

Hoi-yee Chan

University of Arizona

Abstract

People draw subtle distinctions in the normative domain. But it remains unclear exactly what gives rise to such distinctions. On one prominent approach, emotion systems trigger non-utilitarian judgments. The main alternative, inspired by Chomskyan linguistics, suggests that moral distinctions derive from an innate moral grammar. In this paper, we draw on Bayesian learning theory to develop a rational learning account. We argue that the 'size principle', which is implicated in word learning (Xu & Tenenbaum 2007), can also explain how children would use scant and equivocal evidence to interpret candidate rules as applying more narrowly than utilitarian rules.

---

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 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 these kinds of 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 briefly review prevailing accounts.

2. Background

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 particular kinds of emotions that subvert utilitarian cost-benefit analysis.

An alternative view is that many non-utilitarian judgments depend critically on internally represented *rules* (Mikhail 2011, Nichols & Mallon 2006). On this account, 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 how they come to have this structure.

Moral nativists maintain that the structure of the rules has an innate foundation.

(e.g., Dwyer 2004; Harman 1999; Mikhail 2011). We will offer an alternative, empiricist account of how aspects of structured rules are acquired. But our proposal is largely inspired by two considerations that motivate moral nativism.

First, as nativists observe, moral 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). For instance, Pellizzoni and colleagues (2010) showed that 3-year-old children make the familiar distinctions on footbridge. Even young children have a facility with tracking intentions and forming judgments based on non-utilitarian rules.

Second, 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 that our capacity for moral judgment, and specifically our acquisition of general moral rules, requires innate morality-specific constraints**.**

Here we offer an alternative account of the acquisition of moral rules that does not ground their acquisition in innate, morality-specific constraints. However, 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 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

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.

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).

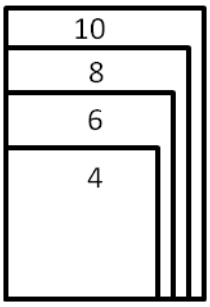

Figure 1: The numbers represent the highest denomination of the die; the rectangles represent the relative sizes of the hypotheses

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 .0156 as the joint probability given h_4, and .001 for h_{10}. This process illustrates the fact that smaller hypotheses 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.[1]

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 ⟦dog⟧ rather than ⟦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 ⟦dog⟧ as compared with ⟦animal⟧.

Xu and Tenenbaum report experiments and Bayesian simulations that suggest that participants conform to this kind of inference. Our own experiments 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

---

[1] 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.

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 3 Dalmatians. When shown three examples of Dalmatians, participants were more likely to generalize only to Dalmatians than when given a single example of a Dalmation, suggesting that they are sensitive to the number of samples – as they get more examples consistent with the narrowest hypothesis, they are more likely to restrict their generalizations. In addition, when given three examples of the new word drawn from the basic-level category (e.g., a Dalmatian, a terrier, and a mutt), participants generalized to other dogs, but not to items that were not dogs (253).

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.

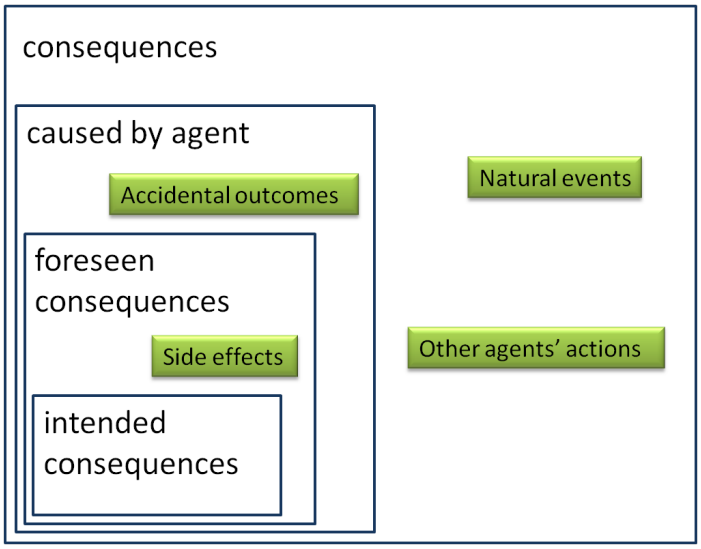

Figure 2: Potential scopes of rules represented in a subset structure

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 aim 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*).

The subset structure represents different ways in which consequences can be categorized (cf. Mikhail 2011, 134). A further issue concerns the characterization of the rules or principles (cf. Mikhail 2011, 150). Rules might be formulated at any of these 'scopes.' A rule at the narrowest scope might prohibit agents from intentionally producing an outcome, e.g., intentionally scratching a car. At the broadest scope, a rule might prohibit agents from 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.[2]

Moral distinctions familiar to ordinary thought and philosophical theory can be captured in terms of this subset structure.[3] Consider, for instance, the doing/allowing distinction. In many

---

[2] We are supposing that such a rule would involve multiple concepts, including AGENT, CAR, and SCRATCH. However, each of these concepts might be associated with a prototype structure that influences the application of the rule.

[3] The precise boundaries of these distinctions is a delicate issue. For present purposes, we will not attempt to give precise renderings of these distinctions, but note that while our distinction maps onto the doing/happening distinction (see, e.g., McNaughton & Rawlings 1991; Nagel 1986; Parfit 1984), it does not map onto the act/omission distinction. For it's plausible that I can *do* things by omission. To take a famous example, if the lead actor deliberately skips a

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 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. Act-utilitarian considerations – on which one is to maximize desirable outcomes – might be represented with 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.

---

performance, he thereby does something with his omission – he spoils the performances (Foot 1967).

In addition, it seems possible for there to be intentional violations that are not foreseen. For instance, an evil nephew who attempts to shoot his uncle from a great distance might not expect to succeed, but if the bullet lands, he has intentionally shot his uncle, despite not foreseeing it. Thus, our representation of the relation between intended outcomes and foreseen outcomes might be an oversimplification. However, it's possible that these cases are too peripheral to play a significant role in learning rules of conduct. More interestingly, it might be that the union of intended-consequences and foreseen-consequences form a psychologically natural class in rule learning, and that intended-consequences is a proper subset of that broader class.

Given this subset structure, the size principle has the potential to explain critical features of rule learning. We will begin with the most extreme distinction – that between *intended consequences* and *consequences in general*, and we will refer to rules applying only to intended consequences as having 'narrow scope', while rules applying to consequences overall will be said to have 'wide scope'. Now 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.

The foregoing analysis illustrates how Bayesian learning theory might explain how people acquire rules that generate something like the act/allow distinction. 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 (iii) people should infer that the rule has narrow scope, i.e., that it applies to what an agent intentionally does. A full demonstration of this is obviously beyond the reach of this paper, but we will examine several critical components.

5. 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 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.'[4] There was very high inter-coder agreement (over 99%), and the few disagreements were settled by discussion.

---

[4] 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

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.'[5] 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]

6. Study 1: The Likelihood: A learning study

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

---

'agents should maximize this sort of outcome,' so the same outcome should be sought by all agents (e.g., *one should ensure that children are cared for* or *one should ensure that children are cared for by their own parents*).

[5] Typical examples of positive cases coded as narrow are: 'eat over your plate', 'finish your juice', 'that's his give it back to him', and 'tell him you're sorry'.

[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.

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)? We investigated this first by adapting Xu and Tenenbaum's (2007) word learning task into a rule-learning task.

6.1. Study 1

Participants and procedures

Twenty four adult participants were recruited from an online panel (Amazon's Mechanical Turk) to complete a lengthy survey in return for a nominal cash payment. The task itself was rather tedious and time-consuming, and many participants completed the survey far too quickly. As a result, we calculated their average time and defined a threshold based on that. Subjects lying below the 75th percentile of completion rate were excluded from the analyses, leaving 18 participants (11 female).[7]

The subjects' job was to figure out the meaning of a rule from a foreign culture, given sample violations of the rule. 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. Three 'domains' were used: some rules concerned a chalkboard, some concerned a shelf, and some concerned litter. For each rule, participants were presented with examples of

[7] All statistically significant results remain significant (and the non-significant ones remain non-significant) if all participants are included.

violations of that rule. There were three kinds of trials. In *one intended sample* trials, participants received a single example which was consistent with a narrow scope interpretation in which an agent *acts*, e.g., 'Mike puts a block onto the shelf.' In *three intended samples* trials, participants received three examples that were consistent with a narrow scope interpretation. In *three mixed samples* trials, one of the examples was consistent with a narrow scope interpretation and two were *in*consistent with a narrow-scope interpretation, e.g., 'Dan doesn't pick up a jump-rope that he notices on the shelf.' For each trial, after being exposed to the examples for a given rule, participants then had to indicate which other cases were violations of that rule. The test cases included two examples of intended consequences and two examples of overall (not intended) consequences for each domain.

Results and discussion

As expected, participants 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)). We found that when participants were exposed to three 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') (one sample t-test $t(17)=7.9$, $p<.0001$). By contrast, when presented with two examples that are inconsistent with narrow scope, participants overwhelmingly generalize to include wide-scope cases 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') (one sample t-test $t(17)=20.4$, $p<.0001$) (see figure 3). And of course, there were a significant difference between these types of trials ($t(17)=19.96$, $p<.0001$). There was

not, however, a significant difference between one narrow sample trials and three narrow sample trials ($t(17)=1.31$, $p=.20$). (We will explore this issue in study 2.)

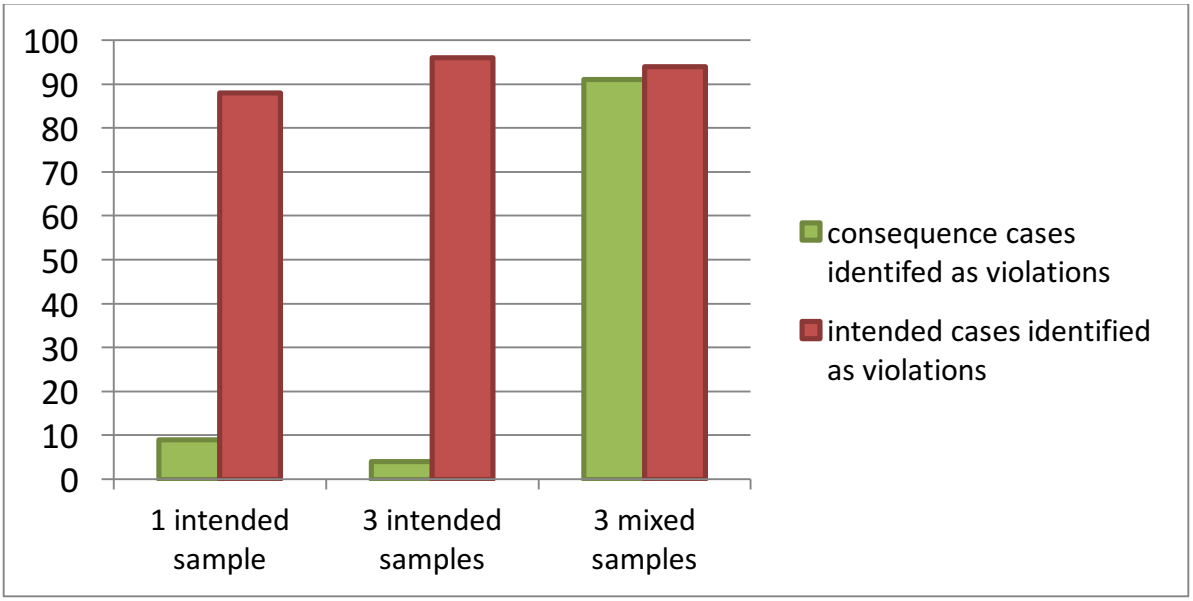

Figure 3: Generalization of scope of rules

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.

6.2. 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 (2007b), 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, 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 (cf. Xu & Tenenbaum 2007b, 254). 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.

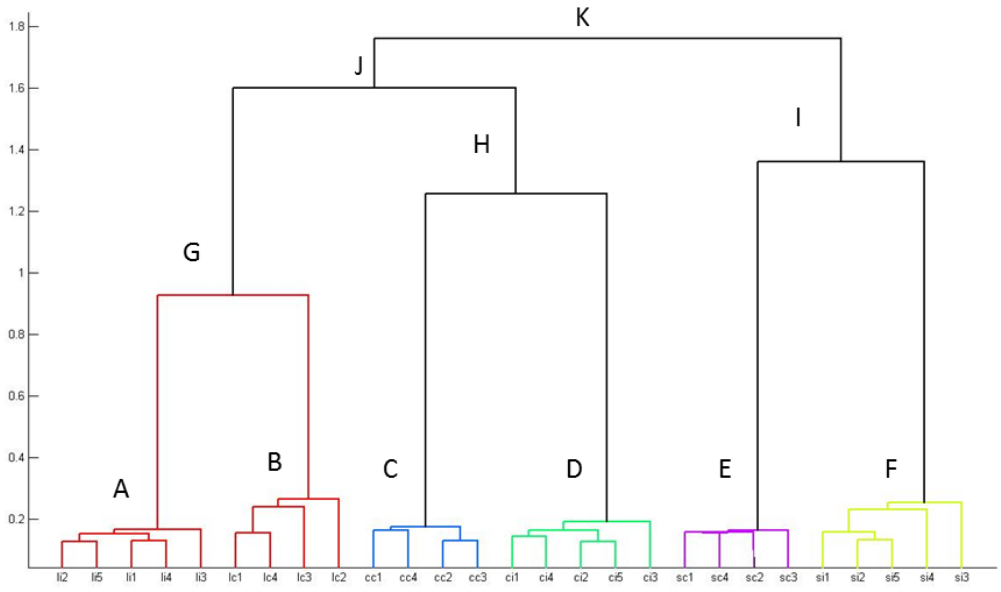

Figure 4: Hierarchical representation of hypothesis space, based on similarity judgments. Letters G, H, I represent domains (litter, chalkboard, shelf). Letters A-F represent unique clusters. All the cases in A, D, and F are cases of intended consequences. None of the cases in B, C, or E are cases of 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)*.

6.3. The Bayesian model

We used this hierarchical representation of the hypothesis space to build a Bayesian model to simulate rule learning.[8] Formally, the problem can be defined as learning a single rule *R* from a

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

set of examples *I* drawn from some known domain *D*, where $I = i_1, \ldots, i_n$. 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 nprovides 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, 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 cluster 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.

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:

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

Thus, for hypothesis A, the prior is determined by (*height* [G] – *height* [A]) and for hypothesis G, the prior is determined by (*height* [J] – *height* [G]).

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

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

where *n* is the number of examples and σ is a small constant (σ = .05) introduced to prevent the likelihood of the lowest nodes from going to infinity. For hypothesis A, the likelihood is [1/ (*height* [A]) + σ] and for G, the likelihood is [(1/(*height* [G] + σ)]. 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 the learning task. 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 (figure 5).

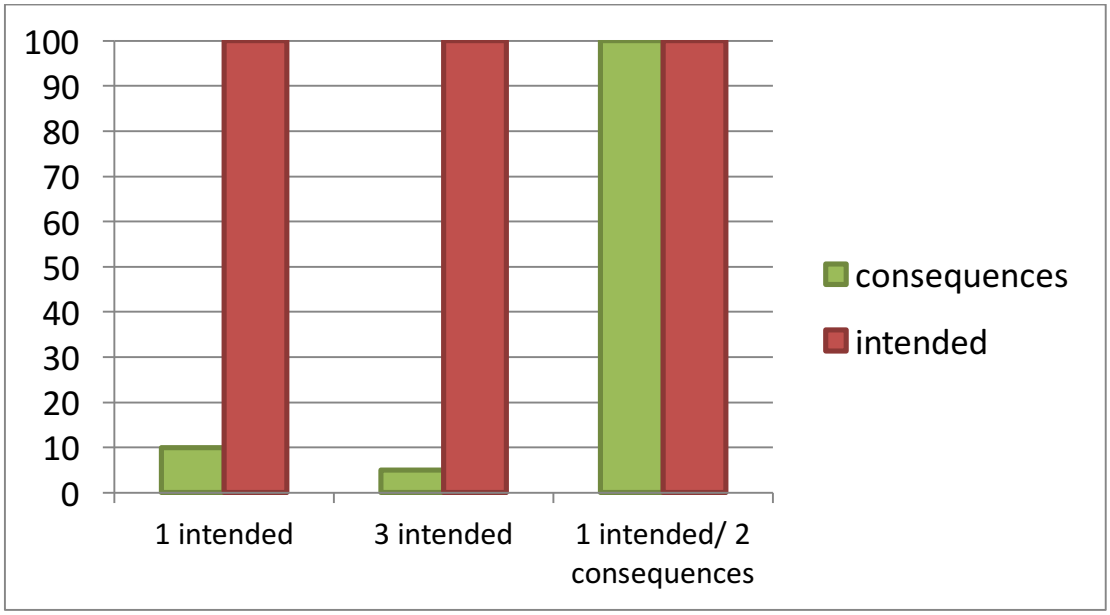

Figure 5: Predictions of the Bayesian model

7. Study 2

Our first study showed that participants learning novel rules are quite sensitive to scope information. The presence of examples that are inconsistent with a narrow-scope interpretation leads participants to infer that the rule is wide scope. In addition, our first study showed that people have a hypothesis space that conforms to philosophical distinctions of interest. In particular, when trying to learn a novel rule, people in our studies naturally distinguished whether an agent was producing a state of affairs from whether the agent was allowing a state of affairs to persist. That is, they clustered 'intended consequences' and 'allowed consequences' separately. However, in our first study, there was no significant difference between one sample trials and three sample trials when all cases were consistent with a narrow-scope interpretation. As a result, although this learning study shows that people are sensitive to information about the

scope of the rule, it does not show that people conform to the size principle in rule learning. An alternative learning account that is consistent with these results is the 'subset principle' (Berwick 1986). Like the size principle, the subset principle predicts that the learner will prefer the smallest subset consistent with the data; but unlike the size principle, the subset principle does not predict that the number of samples should change estimations (cf. Xu & Tenenbaum 2007b, 265). If the subset principle provides the best account of the results, that still provides a new account of the acquisition of moral rules. However, we wanted to explore the size principle more directly in a new set of studies.

One explanation for why we didn't get a size principle effect in study 1 is that the bias for narrow-scope interpretations drives responses to the floor (in the 1 sample cases only 9% of wide-scope cases were registered as violations), and this makes it difficult for genuine differences to be revealed. In addition, the measures used in study 1 were binary decisions – participants had to indicate that a case either was or wasn't violation of a rule. Accordingly, for study 2 we strengthened the design in several ways. First, we picked a context in which it's more likely that there are wide-scope rules – working in a restaurant.[9] Second, we focused closely on the distinction between wide and narrow scope. Study 1 already shows that people are sensitive to domain differences (shelf, litter, chalkboard) and can accommodate slight differences in the instances of action types (e.g., whether a book or a truck on the shelf). To maximize the

[9] Thanks to Tori Morris for this suggestion.

strength of the current studies, we only used a single domain and uniform instances of the action type (putting a napkin on the windowsill) and allowing type (seeing a napkin on the windowsill and leaving it there). Third, we used a 1-7 scale for how likely it was that the person in the scenario was violating the rule. Finally, we used two comprehension checks to exclude participants who did not read instructions carefully.

7.1. Study 2a

Participants and procedures

Participants were recruited from an online panel (Amazon's Mechanical Turk). 56 completed the study for a nominal cash payment. 8 were excluded due to failing comprehension checks, leaving 48 participants (15 female).

As in study 1, participants were told that they were learning a rule in a foreign language. Participants were told that a foreign person was helping them to learn the rule by showing them brief film clips of people. Participants were told that the clips would be described for them and that the teacher would select among the clips to help the participant identify the rule. They were then told, 'After you get the examples from the teacher, you will have to determine whether the employee in one of the other clips is violating the rule.' Two comprehension checks were included. Participants who failed either of these were excluded. Participants were then presented with the list of descriptions of the 11 clips (e.g., Mike puts a napkin on the windowsill, Sarah puts a napkin on the windowsill, Amy sees a napkin on the windowsill and leaves it there…) and asked to summarize them.

Following this setup material, participants were told that the teacher was teaching them the rule *yag survist* and the teacher was allowed to select either 1 clip (1-sample condition) or 3 clips (3-sample condition). The list of 11-clips was presented with the teacher selected clip(s) highlighted (cf. Xu & Tenenbaum 2007a). Then participants were asked about a case that was *inconsistent* with a narrow-scope interpretation: 'Matthew sees a napkin on the windowsill and leaves it there' and asked to rate 'HOW LIKELY you think it is that Matthew is violating the rule *yag survist* in this clip' on a scale from 1 (Not at all likely that Matthew is violating the rule) to 7 (Extremely likely that Matthew is violating the rule). Next they were asked about a case that was consistent with a narrow-scope interpretation: 'Amanda puts a napkin on the windowsill.' and asked to rate how likely they thought it was that Amanda was breaking the rule.

Results and discussion

Recall the prediction of the size-principle. In both the 1 sample and 3 sample conditions, the examples are consistent with a narrow-scope interpretation – in each instance the person is putting a napkin on the windowsill. However, when there are three such examples one should think the wide-scope interpretation is less likely than when there is only one such example. It's a more suspicious coincidence that all the examples are consistent with a narrow-scope interpretation when there are 3 samples as compared to 1. We found that people's judgments conformed to this prediction of the size principle. The case of interest is the one that is inconsistent with the narrow-scope interpretation (Matthew leaving a napkin on the windowsill). People were less likely to say that this was a violation in the 3-sample condition (M=2.26) than in the 1-sample condition (M=3.56) ($t(46)=2.2$, $p<.05$). For the example that was consistent with

the narrow-scope interpretation (Amanda putting a napkin on the windowsill), there was no significant difference between conditions. In the 1-sample condition, M=5; in the 3-sample condition, M=5.95 ($t(46)=1.49$, $p=.142$, n.s.).

7.2. Study 2b

To reinforce the findings of study 2a, we ran a within-subjects version of the study.

Participants and procedures

Participants were recruited from an online panel (Amazon's Mechanical Turk). 31 completed the study for a nominal cash payment. 7 were excluded due to failing comprehension checks, leaving 24 participants (12 female).

The setup for this study was the same as 2a except for the following changes. All participants were first given the 1-sample condition and asked to make the same probability judgment as in study 2a. Participants were then told that the teacher was allowed to give 2 more samples, which effectively yielded the 3-sample condition from study 2a. After answering the questions for the 3-sample condition, participants were asked for an explanation: 'For the question about Matthew, if you gave the same answer as the first time, please explain why you did; if you gave a different answer to that question this time, please explain why you changed in the direction you did (more likely vs. less likely).'

Results and discussion

Once again we found a size-principle effect. People rated the case that was inconsistent with a narrow-scope interpretation (Matthew) as less likely a violation in the 3 sample condition (M=3) than in the 1 sample condition (M=3.625) ($t(23)=2.22$, $p<.05$). There was also a significant difference for the case that was consistent with a narrow-scope interpretation (Amanda). People rated that case as more likely a violation (M=6.625) in the 3-sample condition than in the 1-sample condition (M=6.166) ($t(23)=2.11$, $p<.05$).

In addition to the statistical results indicating an effect of the size principle, several subjects gave explanations that indicated some fairly explicit reasoning in accordance with the size principle in explaining why they responded differently in the 3-sample condition than the 1-sample condition. Here are three such examples:

> 'I said it was less likely because the teacher never used that action as an example, even when they could give more examples'
>
> 'It is now less likely, because the teacher would want to provide the most diverse set of examples as possible. I now believe it is only against the rules to put a napkin on the windowsill.'
>
> 'All the examples of the specific rules involve putting the napkin on the windowsill.'

8. Study 3

Studies 1 and 2 focus on the starkest distinction – between producing an outcome and allowing an outcome to persist. But much work in moral psychology has focused on a finer distinction – between intending to produce an outcome and producing an outcome that is foreseen but not

intended. *Switch* is, of course, just such a case. In that case, a person intends to save the 5 people on the main track, but knows that the diverted train will kill a different person on the side track. For a final experiment, we wanted to explore this subtle distinction in the context of a rule-learning framework.

The first thing to note is that despite the apparent simplicity of a case like *Switch,* in order for the cases to work intuitively, a complex set of conditions has to be met. For instance, if I intend to divert a train to save an ant, knowing that the train will then kill a person, the fact that I don't *intend* this side effect does not absolve me from blame. The complexity of the conditions is evident when we consider the philosophical attempts to articulate a normative principle that conforms to the judgments. According to a standard characterization of the doctrine of double effect, an action that has a foreseen effect that would be wrong to intend is permissible only if:

1. the intended action is permissible
2. the foreseen bad outcome is not intended
3. there is no way to produce the good outcome without also producing the bad outcome
4. the bad outcome is not disproportionate to the good outcome (see, e.g. Uniacke 1998, 120).

Cases like *switch* are both complex and rare. Typically, if we foresee a bad side effect, there is plenty of time to develop a new plan that won't produce the side effect. This is not to discount the philosophical importance of the cases. But it does mean that a properly constructed case will be rather complex.

Participants and procedures

Participants were recruited from an online panel (Amazon's Mechanical Turk). 221 people completed the study for a nominal cash payment (75 female; 2 participants didn't report gender).

As in the previous studies, participants were given a rule learning task. As before, it was a novel rule ('*nib weigns*') in a foreign culture. And as before, it was an utterly boring case. For this study we used a 3x2 design. On the training dimension, one third of the participants were assigned to the *intended* condition, a third to the *accident* condition, and a third to the *unintended foreseen* condition; on the case dimension half were assigned to the *intended* condition and half to the *unintended but foreseen* condition. In the *intended training condition*, the training cases consisted of three examples that were consistent with a narrow-scope interpretation of the rule. Here were the actual training examples:

- John moved a piece of construction paper from the desk to the shelf.
- Bill took a letter from the desk and put it in a folder on another desk.
- Mary took a piece of paper off the desk and put it on the table.

In the *accident training condition*, the cases consisted of one example that was consistent with a narrow-scope interpretation and two examples in which the agent broke the rule without anticipating the outcome. They were given these examples:

- John moved a piece of construction paper from the desk to the shelf.
- Mary coughed because her throat felt scratchy and this caused a gum wrapper to fall off the back of the desk.
- Bill put down the garage door and the vibration caused a letter to fall off of the desk.

In the *foreseen training condition,* the first case was again consistent with a narrow-scope interpretation, but the other two cases involved an unintended but foreseen outcome[10]:

- John moved a piece of construction paper from the desk to the shelf.
- Mary closed the garage door, even though, as she expected, the vibrations caused a letter to fall off the desk.
- Bill used a book to scoot a bug off the desk even though, as he expected, this also caused a paper to slip from the desk onto a table.

Within each of these conditions, half of the participants were given a case in which the outcome was *intended* – Ed intends to move the paper off the desk; the other half were given a case in which the outcome was *unintended but foreseen*. The full case in the *intended* condition was as follows:

> Ed notices that wind is coming up through a vent under the desk. It's clear to him that the wind is about to blow all the papers off the desk that aren't secured. There are 6 pieces of construction paper on the desk. To stop the 5 pieces from blowing off the desk, Ed has only one option, and he takes that option: Ed grabs one of them and puts it over

---

[10] It's important that these unintended-but-foreseen cases don't suggest negligence. For negligence can be coded as failing to have the right intentions. To avoid this, we developed cases where one's immediate interests can only be met by producing a (potentially unwanted) side effect.

> the vent to stop the wind from blowing off the other pieces of paper. Once the piece of paper is put over the vent, the wind is stopped and the other papers remain on the desk.

The full case in the *unintended but foreseen* condition was:

> Ed notices that wind is coming up through a vent under the desk. It's clear to him that the wind is about to blow all the papers off the desk that aren't secured. One piece of blue construction paper happens to be under a large box; 5 pieces of construction paper are scattered next to the box. To stop 5 pieces from blowing off the desk, Ed has only one option, and he takes that option: Ed picks up the box and sets it down over the 5 pieces of paper, knowing that the blue piece will be blown off the desk. The blue piece is blown off the desk, but the other papers remain on the desk.

All participants were then asked 'To what extent do you think the person is violating the rule *nib weigns*?' and responded on a scale from 1 (definitely not breaking the rule) to 7 (definitely breaking the rule).

Results and discussion

In the *intended* training condition, the transgressing agent intends and foresees the outcome (removing a piece of paper from the desk); in the *foreseen* training condition, the transgressing agent foresees but doesn't intend the outcome; in the *accident* training condition the transgressing agents neither intend nor foresee the outcome. In the dependent measure, for the *intended* case, the agent intended and foresaw the outcome; in the *unintended but foreseen* case, the agent foresaw but didn't intend the outcome. We predicted an interaction between the training conditions and the cases such that participants would show a difference between the

cases in the *intention* training condition but not in the *accident* or *foreseen* training conditions. The idea is that if the training set includes examples of both intended and unintended outcomes, then people will expect that an agent is transgressing when an outcome is not intended; however, if the training set includes only examples of intended outcomes, then people will think that an agent is not transgressing when the outcome is unintended. Our prediction was borne out (see figure 6).

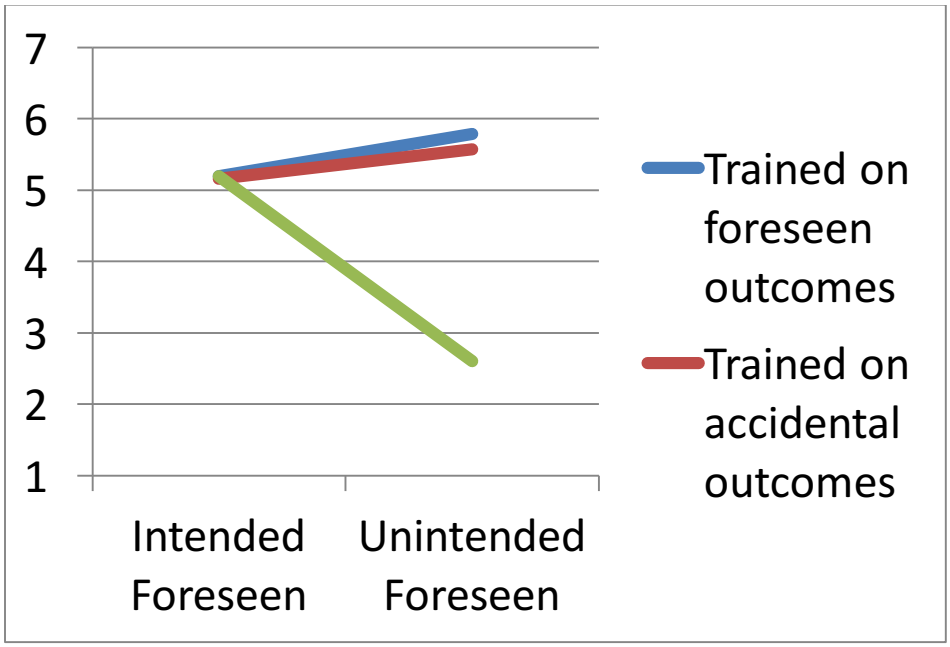

Figure 6: Judgments of how likely the agent was violating the rule

A univariate analysis of variance (ANOVA) was conducted to determine whether there was an influence of training condition and case on participants' judgments of whether the person violated the rule. There was a main effect for both variables. The type of training cases participants received (act, accident, or foreseen) influenced the extent to which participants thought the person violated the rule, $F(2,215)=15.33$, $p < .001$. The type of test case participants

received (intended or unintended but foreseen) also influenced their judgments, $F(1,215)=4.74$, $p=.03$. As expected, there was a significant interaction between training condition and case, $F(2,215)=16.90$, $p<.001$). Because this interaction was significant, we followed up with a test of the significant main effect of the case variable.

A post hoc least significant difference (LSD) test was conducted to assess the simple main effect of the case variable. The test revealed that the case variable only had an effect for participants in the act training condition. Participants in the act training condition were much more likely to say that Ed violated the rule in the intended case ($M=5.25$) than in the unintended but foreseen case ($M=2.6$), $F(1,215)=34.26$, $p <.001$. There was no such difference in the accident training condition, where participants judged the unintended but foreseen case just as clearly in violation ($M=5.57$) as they did the intended case ($M=5.16$), $F(1,215)=.86$, $p=.36$, n.s. Similarly, in the foreseen training condition, participants judged the unintended but foreseen case just as clearly in violation ($M=5.78$) as they did the intended case ($M=5.19$) $F(1,215)=2.08$, $p=.15$, n.s.

9. 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 produce 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 *intentionally producing an outcome*, in a way that approximates Bayesian learning. Our first 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. In accordance with the size principle, our second set of studies showed that participants' judgments about the probability that a person is violating the rule is sensitive to the number of examples. When given 3 examples consistent with a narrow-scope interpretation (rather than 1 such example), participants judged it less likely that a person who failed to intervene was breaking the rule. Experiments 1 and 2 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. Our final experiment looked at the finest distinction in the subset structure, that between intended consequences and unintended foreseen consequences. In

keeping with our earlier studies, we found that when learning a new rule, if participants were trained exclusively on the narrow-scope intention-based cases, they tended to think it very unlikely that a person who generated an unintended but foreseen side effect was breaking the rule. By contrast, if participants were trained on cases that were not exclusively narrow-scope, they thought it very likely that the person who produced the unintended but foreseen side effect was indeed breaking the rule.

In our studies we examined how adults learned rules in light of evidence of violations. 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 2013). However, there are two significant issues concerning nativism that remain unanswered by the current work. First, we offer no explanation for how learners arrive at the hypothesis space itself (figure 2). A nativist account of how the hypothesis space is formed might well be correct. Our

project attempts to show how, *given this hypothesis space,* a rational learner would come to infer narrow-scope rules from the available evidence.[11] The second outstanding issue concerns the fact that we found a strong bias for narrow-scope rules (study 1). When given just a single intended example, participants tended to interpret the rule as narrow scope. One explanation is that people have an innate bias to think that rules are intention-based. Our current work does not exclude this possibility. However, there are natural resources in rational learning theory for an empiricist explanation of 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. With such an overhypothesis in place, the learner will tend to expect that a new rule will also be 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.

In addition to its relevance to issues of nativism, the Bayesian account we have offered provides new grounds for thinking that non-utilitarian judgment derives in part from the structure

---

[11] This parallels Xu & Tenenbaum's work on word learning (see, e.g., 2007b, p. 251). Xu & Tenenbaum start from the assumption that people have various categories of Dalmatians, dogs, animals, etc. that stack into a subset structure. Xu & Tenenbaum ask, *given* that people have these categories, how do they learn how to map new words onto these categories.

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 of addressing 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 indeed part of the structure of moral rules that they are encoded as narrow scope.

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. It is a further question, of course, how the rules emerged in the first place, and we have not attempted to address that question. However, 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. 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 *ir*rational 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.

Department of Philosophy

University of Arizona

School of Information

University of Arizona

Department of Philosophy

Hamilton College

Department of Psychology

University of Arizona

Department of Philosophy

University of Arizona

References

Amit, E. and Greene, J. 2012: You See, the Ends Don't Justify the Means. *Psychological Science*, *23*(8), 861-868.

Baron, J. 1994: Nonconsequentialist decisions. *Behavioral and Brain Sciences*, *17*, 1-1.

Bartels, D. and Pizarro, D. 2011: The mismeasure of morals. *Cognition*, 121, 154-161.

Berker, S. 2009: The normative insignificance of neuroscience. *Philosophy & Public Affairs*, 37, 293-329.

Berwick, R. C. 1986: Learning from positive-only examples: The subset principle and three case studies. In R. S. Michalski, J. C. Carbonell & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2, pp. 625–645). Los Altos, CA: Morgan Kaufmann.

Blair, J. 1995: A cognitive-developmental approach to psychopathy. *Cognition* 57,1-29.

Cushman, F., Young, L., and Hauser, M. 2006: The role of conscious reasoning and intuition in moral judgment. *Psychological science*, 17(12), 1082-1089.

Dawson, C., and Gerken, L. 2009: Language and music become distinct domains through experience. *Cognition*, 111(3), 378-382.

Dawson, C., and Gerken, L. 2011: When global structure 'explains away' evidence for local grammar. *Cognition*, 120(3), 350-359.

Dean, R. 2010: Does neuroscience undermine deontological theory? *Neuroethics*, *3*(1), 43-60.

Dewar, K., and Xu, F. 2010: Induction, overhypothesis, and the origin of abstract knowledge evidence from 9-month-old infants. *Psychological Science*, *21*(12), 1871-1877.

Duda, R., and Hart, P. 1973: *Pattern classification and scene analysis*. New York: Wiley.

Dwyer, S. 2004: How good is the linguistic analogy. *The innate mind*, *2*, 237-256.

Dwyer, S., Huebner, B., and Hauser, M. 2009: The linguistic analogy. *Topics in cognitive science*, *2*(3), 486-510.

Foot, P. 1967: The problem of abortion and the doctrine of double effect. *Oxford Review* 5:5-15.

Gerken, L. 2010: Infants use rational decision criteria for choosing among models of their input. *Cognition*, 115(2), 362-366.

Goodman, N. 1955: *Fact, fiction, and forecast*. Harvard University Press.

Goodman, N., Tenenbaum, J., Feldman, J., and Griffiths, T. 2010: A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108-154.

Greene, J., Sommerville, R. B., Nystrom, L., Darley, J., and Cohen, J. 2001: An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105-2108.

Greene, J. 2008: The secret joke of Kant's soul. In W. Sinnott-Armstrong (ed.), *Moral Psychology*, Vol. 3, Cambridge, MA: MIT Press, 59-66.

Haidt, J. 2001: The emotional dog and its rational tail. *Psychological Review* 108: 814-834.

Harman, G. 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, M.; Cushman, F.; Young, L.; Kang-Xing Jin, R.; and Mikhail, J. 2007: A dissociation between moral judgments and justifications. *Mind & Language*, *22*, 1-21.

Horne, Z. and Powell, D. 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, C., Perfors, A., and Tenenbaum, J. 2007: Learning overhypotheses with hierarchical Bayesian models. *Developmental science*, *10*(3), 307-321.

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

Lopez, T. 2013: *The Moral Mind: Emotion, Evolution, and the Case for Skepticism*. PhD Thesis, University of Arizona.

Lopez, T.; Zamzow, J.; Gill, M.; and Nichols, S. 2009: Side constraints and the structure of commonsense ethics. *Philosophical Perspectives*, *23*(1), 305-319.

MacKay, D. 2003: *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.

McNaughton, D. and Rawling, Piers 1991: Agent-relativity and the doing-happening distinction. *Philosophical Studies*, *63*(2), 167-185.

MacWhinney, B. 2000: *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum Associates.

Mikhail, J. 2007: Universal moral grammar. *Trends in Cognitive Sciences*, 11, 143–152.

Mikhail, J. 2011: *Elements of Moral Cognition.* Cambridge: Cambridge University Press.

Navarro, D. J., Dry, M. J., and Lee, M. D. 2012: Sampling assumptions in inductive generalization. Cognitive Science, 36, 187–223.

Nichols, S. 2004: *Sentimental Rules*. New York: Oxford University Press.

Nichols, S. and Mallon, R. 2006: Moral dilemmas and moral rules. *Cognition*, *100* (3), 530-542.

Parfit, D. 1984: *Reasons and Persons.* Oxford University Press.

Pellizzoni, S., Siegal, M., and Surian, L. 2010: The contact principle and utilitarian moral judgments in young children. *Developmental science*, *13*(2), 265-270.

Perfors, A., Tenenbaum, J. and Regier, T. 2011a: The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306-338.

Perfors, A., Tenenbaum, J., Griffiths, T., and Xu, F. 2011b: A tutorial introduction to Bayesian models of cognitive development. *Cognition*, 120, 302-321.

Powell, N., Derbyshire, S. Guttentag, R. 2012: Biases in children's and adults' moral judgments. *Journal of experimental child psychology*.

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

Rumelhart, D. and McClelland, J. 1986: *Parallel distributed processing* (Vol. 1). Cambridge, MA: MIT Press.

Samuels, R. 2002: Nativism in cognitive science. *Mind & Language, 17(3),* 233-265.

Schroeder, M. 2007: Reasons and agent-neutrality. *Philosophical Studies*,*135*(2), 279-306.

Singer, P. 2005: Ethics and Intuitions. *Journal of Ethics,* 9, 331-352.

Smith, L., Jones, S., Landau, B., Gershkoff-Stowe, L., and Samuelson, L. 2002: Object name learning provides on-the-job training for attention. *Psychological Science,* 13(1), 13-19.

Tenenbaum, J. and Griffiths, T. 2001: Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24, 629–641.

Thomson, J. 1985: Double effect, triple effect and the trolley problem. *Yale Law Journal*, 94, 1395-1415.

Timmons, M. 2008: Towards a sentimentalist deontology. In W. Sinnott-Armstrong (ed.) *Moral Psychology*, Vol. 3. Cambridge, MA: MIT Press

Unger, P. 1996: *Living High and Letting Die.* Oxford University Press.

Xu, F., and Tenenbaum, J. B. 2007a: Sensitivity to sampling in Bayesian word learning. *Developmental science*, 10(3), 288-297.

Xu, F. and Tenenbaum, J. 2007b: Word learning as Bayesian inference. *Psychological review*, *114*(2), 245.

Xu, F. and Kushnir, T. 2013: Infants are rational constructivist learners. *Psychological Science, 22*(1), 28–32.