Moral empiricism and the bias for act-based rules

Alisabeth Ayars[a], Shaun Nichols[b*]

[a] Princeton University, Department of Philosophy

[b] University of Arizona, Department of Philosophy

[*] Corresponding author.

E-mail address: sbn@email.arizona.edu (S. Nichols)

ABSTRACT

Previous studies on rule learning show a bias in favor of act-based rules, which prohibit intentionally producing an outcome but not merely allowing the outcome. Nichols et al. (forthcoming) found that exposure to a single sample violation in which an agent intentionally causes the outcome was sufficient for participants to infer that the rule was act-based. One explanation is that people have an innate bias to think rules are act-based. We suggest an alternative empiricist account: since most rules that people learn are act-based, people form an *overhypothesis* (Goodman 1955) that rules are typically act-based.

We report three studies that indicate that people can use information about violations to form overhypotheses about rules. In study 1, participants learned either three "consequence-based" rules that prohibited allowing an outcome or three "act-based" rules that prohibiting producing the outcome; in a subsequent learning task, we found that participants who had learned three consequence-based rules were more likely to think that the new rule prohibited allowing an outcome. In study 2, we presented participants with either 1 consequence-based rule or 3 consequence-based rules, and we found that those exposed to 3 such rules were more likely to think that a new rule was also consequence based. Thus, in both studies, it seems that learning 3 consequence-based rules generates an overhypothesis to expect new rules to be consequence-based. In a final study, we used a more subtle manipulation. We exposed participants to examples act-based or accident-based (strict liability) laws and then had them learn a novel rule. We found that participants who were exposed to the accident-based laws were more likely to think a new rule was accident-based. The fact that participants' bias for act-based rules can be shaped by evidence from other rules supports the idea that the bias for act-based rules might be acquired as an overhypothesis from the preponderance of act-based rules in society.

Keywords: Act/allow distinction; Bayesian learning; overhypotheses; moral rules

1. Introduction

Moral judgment is complex and nuanced. People judge actively harming someone to be worse than allowing the same harm to occur. In some cases, judgments seem to conform to utilitarianism: people judge that it is okay to direct a trolley onto a track on which it will kill fewer people. In other cases, people appear to abide by deontological constraints: people deem it impermissible to launching someone in front of the trolley to prevent it from killing many more people. What explains these complex judgments? One answer is that moral intuitions are responses to an internally represented "moral grammar," characterized as a set sophisticated, internally represented rules or principles.

The existence of a moral grammar is supported by a number of considerations drawing on an analogy with linguistics. As with linguistic expressions, people have the capacity to generate moral judgments about an unbounded number of actions and arrangements. Chomskians have long stressed that the fact that people can recognize an infinite number of grammatically proper sentences entails that people do not store individual expressions, for this would be incompatible with the finite storage capacity of memory. Rather, people must store "instructions" or rules for building expressions—a grammar. Proponents of moral grammar point out that moral judgment appears to be unbounded in a similar way: agents are "prepared to make a potentially unlimited number and variety of moral judgments about the moral properties of various acts, agents, and institutional arrangements, including in entirely new situations" (Mikhail 2011, 46). Call this the capacity for *moral projection*. Given the fact of moral projection, the moral system cannot merely be the storage of particular judgments, says the proponent of moral grammar. It must also contain an internally represented system of generative rules and principles.

Once one accepts the existence of a moral grammar, a puzzle arises: how is such a sophisticated grammar acquired? Moral nativists maintain that, at least to some extent, the principles people use to arrive at moral judgments are innately specified (e.g., Dwyer 2004; Harman 2000; Mikhail 2011).

According to this view, moral judgment is the product of innate principles (a "Universal Moral Grammar") and environmental input that solidifies the particular culturally-specific manifestation of the principles.

Moral nativists draw on a variety of evidence. Children acquire moral distinctions at a young age. For instance, 3-year old children judge intended harm to be worse than harm that is merely foreseen or allowed (Mikhail 2007; Pellizzoni et al., 2010). But the evidence available to children at this age regarding moral rules is arguably minimal. Susan Dwyer and colleagues note:

> [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).

This point is corroborated by a detailed analysis conducted by Jen Wright and Karen Bartsch of CHILDES, a corpus of natural language conversations with several children (MacWhinney 2000). They found that only 5% of moral conversation referenced rules or principles (Wright & Bartsch 2008). Given the meagre data available to the child, it's perhaps surprising that children apparently have knowledge of these moral distinctions. This has led several theorists to promote a kind of poverty of the stimulus argument for moral nativism. Gil Harman writes, "An ordinary person was never taught the principle of double effect, and it is unclear how such a principle might have been acquired by the examples available to the ordinary person. This suggests that [it] is built into . . . morality ahead of time" (Harman, 2000, p. 225).

Nativists have also emphasized that conscious reasoning plays only a minimal role in moral judgment: people are often unaware of the principles they use to make moral evaluations. For instance, according to Hauser et al. (2008), although trolley intuitions conform to the *doctrine of double effect*, people never articulate this principle as justification for their moral judgments. In addition, people often confabulate reasons for their moral judgments (Haidt, 2005). Thus, implicit moral knowledge seems to exceed what we can consciously formulate. This point was already made by Rawls, who maintained that an adequate characterization of our moral faculty must "involve principles and theoretical constructions which go much beyond the norms and standards cited in everyday life" (Rawls, 1971; 47). If many of the moral principles we use to arrive at judgments are never articulated, it is harder to see how they could be learned. Finally, nativists emphasize that there is a high degree of agreement in moral intuitions across cultures in, e.g., intuitions about trolley cases. This, nativists suggest, is plausibly explained by the hypothesis that some moral principles are innate.

Moral empiricists are skeptical of the nativist's supposition of an innate moral faculty, or that this faculty is as rich as the Nativist assumes. Empiricists instead appeal to general purpose learning mechanisms (e.g., statistical inference, association, prototype-building) to explain the acquisition of key elements of morality. We will advocate a moderate empiricism. We will not argue that all moral responses are the product of learning. Indeed, it's plausible that there are innate contributions to our moral capacity (e.g., certain emotions that drive moral responses), but we articulate a moderate moral empiricism on which domain general learning mechanisms can explain the acquisition of important and subtle features of moral rules.

Previous critiques of moral nativism have challenged the linguistic analogy in several ways. First, while moral rules are complex, they are not nearly as complex as linguistic rules. Second, while much of moral knowledge may be tacit, conscious moral evaluations are not epiphenomenal to moral judgment. People change their moral views in response to arguments. People become utilitarians, egalitarians, or libertarians after exposure to relevant information. Many moral responses therefore seem to be the product of moral knowledge acquired over the lifetime, rather than innate reactions. In addition, the fact that people are unable to articulate their reasons for moral judgments is little evidence that the principles applied are innate. Many forms of acquired expertise are constituted by intuitive judgments for which the justification is not entirely transparent to the expert—e.g., the rapid identification of a musical style by a music critic (Sterelny, 2008).

While these replies are promising rejoinders to the linguistic analogy, the empiricist position about moral grammar currently lacks a systematic positive defense. The nativist and empiricist disagree about the role of general-purpose learning mechanisms in explaining moral judgment. A positive empiricist program would take the form of demonstrating ways in which general-purpose learning mechanisms can be applied to acquire these refined moral rules. To our knowledge, there has been little systematic defense of the application of specific learning mechanisms to the acquisition of moral rules. We have aimed to jump-start such a project. It remains to be seen whether a complete defense is possible, but the only way to examine the question is to begin to explore how various learning mechanisms can account for key elements of moral judgment.[1]

In a previous paper, we and our co-authors argued that recent developments in learning theory provide a new way to defend an empiricist explanation of how moral rules are acquired (Nichols, Kumar, Lopez, Ayars, & Chan, forthcoming; see also Lopez 2013). We argued that the "size principle"

[1] It is important to clarify what we mean by "empiricist defense." We do not pretend to show that the learning mechanisms we propose are *in fact* the explanation for elements of morality; this would require evidence beyond that provided in this paper. Rather, our defense consists in an elaboration of how such mechanisms *could* be used to acquire key moral knowledge. "How possible" defenses are common in evolutionary theory and certain areas cognitive science. They are apt when the opposing side challenges the very possibility or conceivability of a certain explanation. Because we take moral nativism to have this form, we think possibility defenses are appropriate for the moral empiricist program.

can be used to learn the scope of moral rules (e.g., whether the rule applies only to actions or also to things that are allowed) from only a few examples.

In this paper we argue that learned *overhypotheses* can help account for crucial elements of moral projection. Overhypotheses are second-order generalizations about first-order categories. For instance, children exhibit a "shape bias" in category learning such that when they encounter a novel object category they expect members of the same category to share the same shape but not the same color. We maintain that overhypotheses about rules can help explain some aspects of moral projection.

1.1. Statistical learning and rules

The problem of projection with respect to a domain is, generally, how it is that knowledge in the domain is applied to new contexts (e.g., grammatical judgments about unfamiliar expressions). With respect to the moral domain, "projection" is the ability to morally evaluate arrangements and situations of which the evaluator is unfamiliar. Consider the ability of students to morally evaluate the justness of various distributive schemas (e.g., egalitarianism or libertarianism) and political arrangements which they may have never considered prior to a course. Philosophers often introduce bizarre hypothetical thought experiments which people have no trouble assessing (e.g., trolley cases), even though they have never learned any specific rules about the proper response to runaway trains.

The fact of projection entails, according to proponents of moral grammar, that moral knowledge includes abstract principles that act as the "building blocks" for generating more specific judgments in particular contexts (e.g., Mikhail 2011, 46-47). The moral empiricist must therefore not only explain how specific rules may be acquired but also how people come to store the abstract principles and distinctions that allow for moral projection and guide moral learning in new situations.

In a previous project (Nichols et al. forthcoming), we drew on the "size principle" to explain how the scope of rules might be acquired with few examples. According to the "size principle", the "smallest" hypothesis (i.e., the most restrictive hypothesis consistent with all of the observations) should be preferred over "larger" hypotheses when the observations are consistent with both (Xu & Tenenbaum 2007). Consider the following problem. Imagine that a friend has 4 fair dice, each with a different number of sides: 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. How likely is it that he is rolling the 10-sided die? Intuitively, this is very unlikely. Why? Because you would have expected some numbers greater than 4 if it were the 10 sided die—the wealth of observations less than or equal to 4 is a highly suspicious coincidence. In other words, this data is improbable on the hypothesis that the friend is rolling a 10-sided die.

Turning to the moral domain, some core moral distinctions are also nested. Consider consequences done by the agent and consequences allowed by the agent. All consequences done by the agent are also "allowed" (in the sense that they are not prevented by the agent), but not all actions that are allowed by the agent are brought about by the agent. The nesting of this distinction is depicted in figure 1.

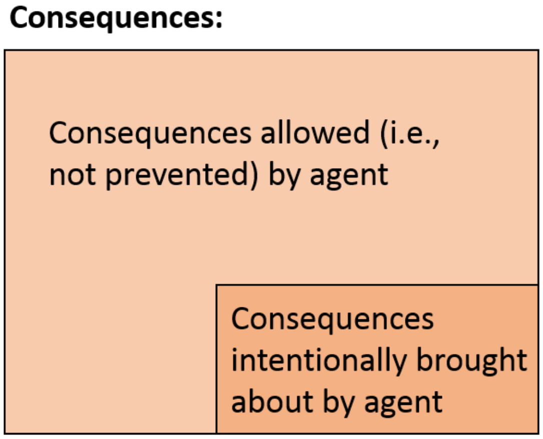

Consequences:
Consequences allowed (i.e., not prevented) by agent
Consequences intentionally brought about by agent

Figure 1: Nesting of consequence types.

A learner could learn that a rule only prohibits doing an action, rather than allowing it, if all the examples of violations of the rule she receives are consistent with the "narrow" hypothesis—i.e., the hypothesis that the rule only prohibits an agent from bringing about the consequences. For instance, suppose you're told that the following are all violations of a certain rule: John puts a ball on the shelf; Mary puts a cup on the shelf; Fred puts a pencil on the shelf. Is it more plausible that the rule simply prohibits putting things on the shelf, or also prohibits allowing things to remain on the shelf? Just as in the case of adjudicating what die a person is rolling, the more examples consistent with the narrowest hypothesis you receive, the more likely it is the rule is cast at the narrowest scope.

In general, we found that people's inferences about the scope of new, unfamiliar rules conformed to the predictions of the size principle: when the example violations people received were all consistent with the narrow scope hypothesis (i.e., involved people bringing about the consequence, people thought the rule was narrow scope. On the other hand, if they received a single example of a wide-scope violation (i.e., involving a person allowing the consequence) they thought the rule was wide-scope. Furthermore, we found (studies 2a and 2b) that participants were more inclined to think the rule was narrow-scope when given several sample violations rather than a single sample. Overall, the evidence indicates that the scope of moral rules can be acquired 1) with only a few examples and 2) without negative evidence (i.e., evidence about what the rule does NOT prohibit). In addition, an analyses of the CHILDES database revealed that over 99% of the cases of adult communication on behavior was consistent with a narrow scope interpretation of the rules (Nichols et al., forthcoming). This suggests that when children are learning a rule, the evidence typically favors the hypothesis that the rule is narrow scope.

However, the foregoing account does not have the resources to explain key elements of moral projection. We argued that statistical learning can explain the acquisition of single domain-specific rules (e.g., "do not put things on the shelf"). Of course, in a limited sense, this involves the acquisition of projectable knowledge: once a rule is learned, it can be applied to make judgments about many different individuals in many contexts. However, there is another kind of projection, at a higher level of abstraction, that plausibly operates in moral learning. Our acquisition of rules can depend on knowledge that *intention* is especially important to moral assessment, or that *acting* and *allowing* are not morally equivalent. This abstract knowledge about morality goes knowledge of specific rules. While the size-principle provides a kind of "bottom-up" data-driven mechanism, we will argue that abstract knowledge of morally important categories can act as "top-down" constraints.

1.2 Learned Overhypotheses and Moral Rules

Nelson Goodman introduced the notion of an *overhypothesis* as a hypothesis containing predicates at a higher-level than relevant first-order hypotheses. For instance, one might have a first-order hypothesis about the diversity of colors in one marble bag or an overhypothesis about diversity of color in all marble bags.

Overhypotheses are thought to inform inductive inference in many domains. For instance, in a phenomenon known as the "shape bias," children tend to think object categories are constrained by object shape (rather than, e.g., material or color) (Heibeck & Markman, 1987). Such categorization biases are plausibly also operative with respect to inferences about word meaning. In what has come to be known as the "basic level bias," both children and adults tend to think unfamiliar nouns refer to basic-level taxonomic categories (e.g., *dog*) (e.g., Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Xu & Tenenbaum (2007) observed that when given a single example of a Dalmatian and told, "this is a fep", adults tended to think that "fep" referred to *dog,* rather than to *dalmatians* or to *animal*. Remarkably, children infer a basic-level extension of some words after being exposed to a single labeled exemplar (e.g., Carey & Bartlett, 1978).

Distinctions like the act/allow distinction and the intention/foreseen distinction are plausibly thought of as informing overhypotheses about moral rules. For instance, there might be an overhypothesis of the form: rules tend to prohibit acting but not allowing. Call this an "act overhypothesis." Of course, overhypotheses can be overridden. If a learner receives information that a new rule prohibits acting or allowing an outcome, she will reject the overhypothesis, just like a child who is told "fep" actually refers to Dalmations will override the overhypothesis that it refers to the basic-level category of "dog."

In Nichols et al. (forthcoming), we unexpectedly uncovered evidence suggesting that people have an act overhypothesis. In learning a new rule, even after receiving only a single sample violation of a person producing a consequence, participants were strongly inclined to think that the rule applies only to producing the consequence rather than allowing the consequence (study 1).

Subjects' task was to learn what a foreign rule (with a foreign name, like "taf byrnal" or "zib matan") forbade based on randomly selected sample violations of the rule. In *one sample* trials, participants received a single example which was consistent with a narrow scope interpretation of the rule in which an agent *brings about* a consequence, e.g., "Mike puts a napkin on the windowsill." In *three "act" samples* trials, participants received *three* examples that were consistent with a narrow scope interpretation. In the *three mixed sample* trials, two sample violations were inconsistent with a narrow scope interpretation (i.e., a person violated the rule by allowing the outcome to occur) and one was consistent. For each trial, after being exposed to the examples for a given rule, participants then indicated whether they thought the rule applied to "acting" (narrow scope) or also "allowing" (wide scope).

Participants thought the rule was wide scope in the *three mixed samples* condition, as expected (since two of the example violations were a person allowing the outcome). However, regardless of whether participants received 1 or 3 examples consistent with the narrow scope, they overwhelmingly assumed the rule was narrow scope—i.e., that it forbade "acting" but not "allowing." This is indicated by the fact that hardly any cases of people allowing an outcome were considered violations (green bar) in either the *one intended sample* and *three intended sample* conditions (see figure 2). While we did find a size principle effect in other studies (Nichols et al. forthcoming, 2a and 2b), the fact that participants are willing to infer a narrow-scope interpretation after a single example is suggestive of a constraint resembling the basic-level bias in word learning.

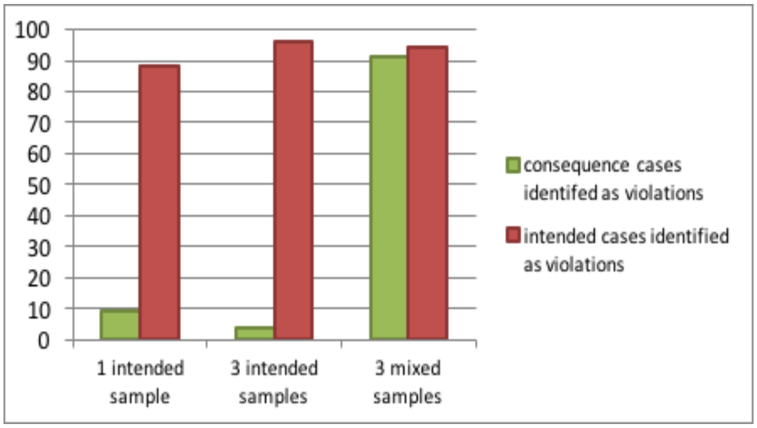


Figure 2: Percentage of cases chosen as violating the new rule (Nichols et al. forthcoming, study 1)

In recent years, several researchers have shown how overhypotheses might be learned (e.g., Smith et al. 2002; Dewar & Xu 2010; Kemp, Perfors, & Tenenbaum 2007). In the next section, we'll elaborate how overhypotheses can be acquired, and suggest that learned overhypotheses might also explain the bias for act-based rules.

2. Learned Overhypotheses and Moral Rules

Goodman elucidated how overhypotheses could be learned using an example involving marbles (Goodman 1955). Imagine you are trying to predict the color of the next marble you are to draw randomly from a bag. You first draw a green marble. What will the color of the next marble be? Without any additional information about the color distribution of the marbles in the bag, there is little evidence to predict the color of the next marble. But now suppose that before drawing from the bag, you draw marbles from 3 other bags first. In the first bag, all the marbles are blue. In the second, all the marbles are pink. In the third, all the marbles are yellow. Finally, you draw a single marble from the fourth bag and find that it is green. What will the color of the next marble from that fourth bag be? In this case you can reasonably infer that the next marble will also be green. The fact that all bags encountered have been uniform in marble color suggests an overhypotheses that within bags, marbles are uniform in color.

Turning to the normative domain, might the bias for act-based rules result from a learned overhypothesis about the scope of rules? One can think of overhypotheses about rules as analogous to the overhypothesis about the uniformity of marble color within bags: the overhypothesis is a generalization about rules themselves (e.g., that with respect to any given prohibition rule, it's more likely that the rule applies to producing an outcome rather than allowing the outcome). This overhypothesis could constrain data-driven inferences about moral requirements drawn from example violations, just like the overhypothesis about marble color uniformity constrains inferences about the color of marbles in a particular bag. Such an overhypothesis could be acquired if the majority of rules the learner is exposed to are act-based, i.e., the rules that prohibit bringing about some outcome but do not prohibit allowing such an outcome to occur.

Consider the basic-level bias regarding word meanings on which common nouns reference basic-level taxonomic classes, but not superordinate classes. This overhypothesis about the scope of nouns is plausibly learned by exposure to many common nouns that refer to basic-level categories of objects. If that's right, it means that abstract overhypotheses about the typical scope of can be acquired. If information about word scope can be encoded into second-order generalizations that inform inferences about new words, then perhaps scope information about rules can also be encoded into an overhypothesis about the typical scope of rules that informs inferences about new rules. In other words, if people are exposed to many examples of act-based rules, they might represent this and come to expect that a new rule will be act-based as well. Indeed, there are numerous prohibitions against intentionally causing an outcome, but less common are prohibitions against allowing an outcome to occur. Actions like littering, making messes, or lying are prohibited, but it is generally not forbidden to allow litter to be left on the ground, to leave messes that others have made, or to fail to prevent someone from lying. Of course, there are exceptions, just as there are exceptions to the generalization that common nouns refer to basic-level taxonomic classes (e.g., the word "animal," which refers to a superordinate class). But insofar as the *typical* rule is act-based, this information may be encoded into a second-order overhypothesis about rules of conduct.

There is considerable evidence that adults can acquire overhypotheses from data in their environment (e.g., Nisbett et al. 1983). Indeed, a growing body of work shows children too can form second-order generalizations about categories. For instance, Smith and colleagues found that children trained in the laboratory acquired the shape bias much earlier than normal (Smith et al. 2002). 17-month children were repeatedly provided with names of object categories along with exemplars of objects in each category, which all shared a common shape. Weeks after this training, the children were more likely to think that a new category was defined by shape.[2] Dewar & Xu (2010) showed that even infants can form a second-order generalization about the nature of categories after being exposed to a number of examples in each category. The experimenters pulled objects out of four boxes. If the objects within each box were uniform in shape (e.g., four balls; four cubes; four stars), infants were surprised when a second item drawn from a fourth, unknown box did not match the shape of the first item drawn from the box. The infants learned that objects within boxes are similar in shape. Clearly, the ability to form generalizations about categories is an early emerging feature of human cognition, rendering it a viable mechanism for the acquisition of moral knowledge.

One might object to the idea that overhypotheses about moral rules can be acquired on a few

[2] Kemp and colleagues (2007) showed that such an overhypothesis can be learned by a hierarchical Bayesian model.

grounds. First, one might favor a nativist position that the bias for act-based rules is a fixed constraint. Second, one could object that *rules* are not the sort of thing amenable to acquired overhypotheses. Previous studies on overhypotheses have focused on simple and seemingly fundamental categories (e.g., the extensions of nouns; categories of objects), rather than rules that prohibit classes of violations. It is not clear that "categories" of actions defined by rules are the right kind of categories for which people are likely to acquire overhypotheses. Finally, even if people do form overhypotheses about the kind of consequences that are prohibited by rules, it is not obvious that people will form overhypotheses about the specific properties of prohibited consequences relevant to contrasts like the act/allow distinction or the intended/foreseen distinction.

In the remainder of the paper, we will present three experiments indicating that people can form overhypotheses about rules. The experiments are meant to constitute a first inroad to the inquiry on whether overhypotheses relevant to important distinctions like acting and allowing can be acquired, by addressing (1) whether generalizations about rules are malleable based on the kind of rules people are exposed to and (2) Whether people are sensitive to the scope of rules (e.g., whether a rule applies to acting or also allowing) in making generalizations about rules. Because these experiments are conducted in adults, they do not address the key question of whether overhypotheses explain the acquisition of moral knowledge in children—an essential component of an empiricist defense of moral grammar. Nevertheless, they provide an important basis for thinking that rules are the sort of "categories" about which people can learn overhypotheses.

The general paradigm we used in the experiments involved exposing participants with examples of rules at a certain scope (e.g., that forbade acting but not allowing), and then asking them to learn a new rule; this allowed us to examine whether exposure to the previous rules influenced their inferences about the new rule. In studies 1 and 2, we focused on the scope distinction between acting and allowing. In study 1, participants either learned 3 action-based or 3 consequence-based ("allowing") rules; the rules they learned influenced their beliefs about the type of actions prohibited by the new rule. In study 2, we presented participants with either 1 consequence-based rule or 3 consequence-based rules, and we found that those exposed to 3 such rules were more likely to think that a new rule was also consequence based. In study 3, we used a subtler manipulation that did not involve learning rules in the first phase, but merely being exposed to descriptions of laws.

3. Study 1

In study 1, we examined whether participants' inferences about the scope of a new rule would be influenced by the scope of recently-learned rules, consistent with the formation of an overhypothesis about the scope of rules. The main purpose of the experiment was to assess whether (1) participants form generalizations about the scope of new, unfamiliar rules, and (2) whether these generalizations influence inferences about the scope of a new rule. In order to test this, we had participants learn three rules before making a judgment about a fourth rule.

Recall Goodman's example involving marbles. After sampling from 3 bags and finding the marbles in each bag to be uniform in color, the learner forms an overhypothesis that marbles within bags are uniform in color, including marbles in bags that have not yet been sampled from. Now, imagine that rather than sampling marbles, the learner is exposed to sample rule violations. If the learner finds that three previous "bags" of sample violations have turned out to be act violations and therefore indicative of an act-based rule, the learner might expect that a new "bag" of violations will also turn out to be violations of an act-based rule even if the sample from the new bag is quite limited. On the other hand, if participants are exposed to "bags" of violations that indicate allowing-based rules (because some of the violations involve allowing an outcome), they will be more likely to think a fourth rule also prohibits allowing an outcome.

Participants learned either 3 narrow scope rules (i.e., rules that only prohibit bringing about an outcome) or wide scope rules (i.e., rules also prohibit allowing an outcome to occur). They learned the rules by receiving sample violations of the rule (e.g., "Julie draws a dog on the chalkboard). This learning phase mimics the learning phases of other experiments designed to examine overhypotheses, such as Dewar & Xu (2010) who drew objects from boxes that shared a feature like shape to facilitate a second-order generalization about the boxes. We used mundane and unfamiliar rules so that participants would not use prior beliefs about the content of moral rules to inform their judgments.

Then, participants predicted the scope of a new rule after being given 2 examples of violations of the new rule by indicating what other actions counted as violations, out of a list of 16. The scope of the new rule was ambiguous based on the examples of violations provided, which all involved a person bringing about an outcome (i.e., "acting"). (Bringing about the outcome would count as a violation regardless of whether the rule was wide or narrow scope). We predicted that participants who had previously learned three rules at the allowing scope would be more likely to think that the ambiguous rule was wide scope (i.e., include cases of allowing as violations) than participants who learned three narrow scope rules.

Participants and procedures

Participants were recruited from an online panel (Amazon's Mechanical Turk). N=61 (28 female) completed the study for a nominal cash payment. None were excluded.

Participants were asked to imagine they were helping someone to figure out what several different rules from foreign cultures mean, such as the rule *taf byrnal*. They were then informed they would learn each rule by being given one or more examples in which a person has violated the rule, and that they would be asked to decide in which other cases the person violated the rule. Participants received the same list of 16 possible violations after learning each rule and asked to check off which counted as violations of the rule they just learned.

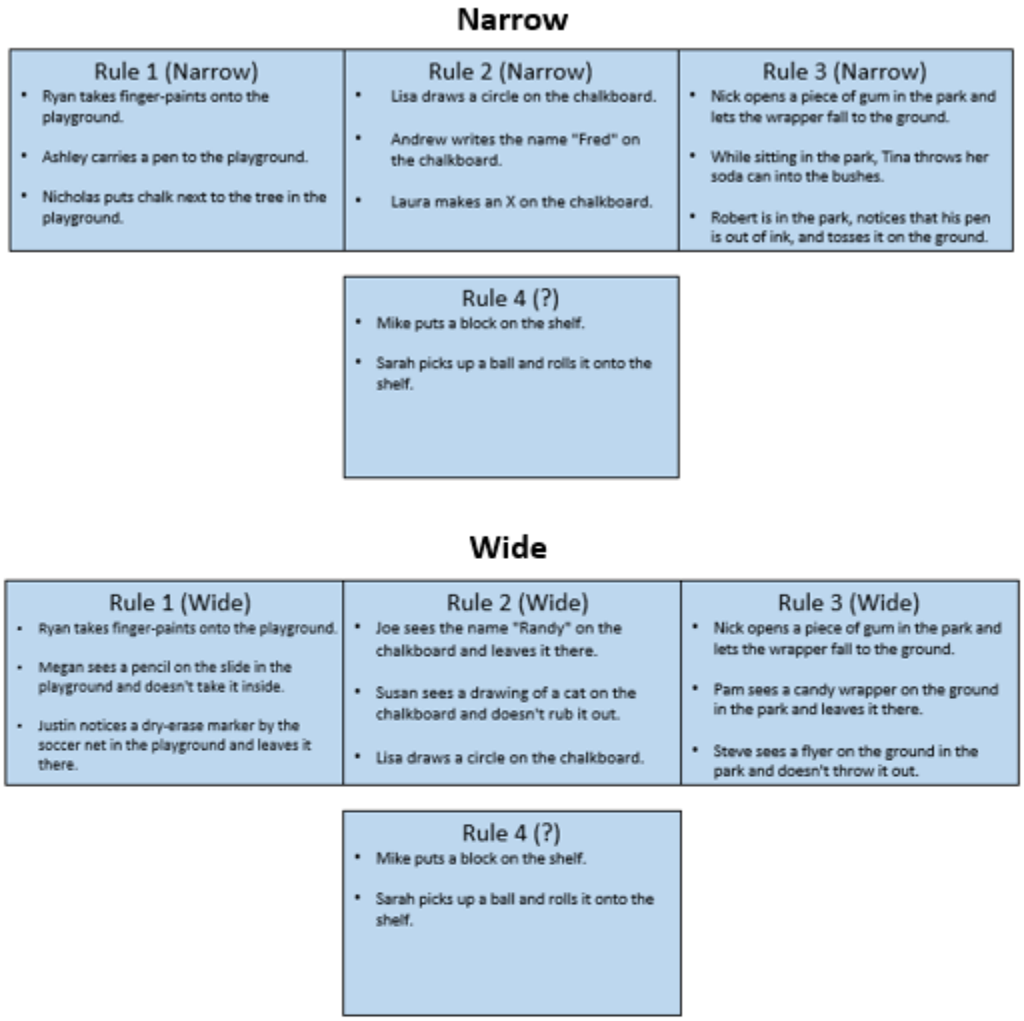
First, participants were familiarized with the 16 possible violations they would be asked to evaluate for each rule they learned. These included instances of causing an outcome such as “Julie draws a dog on the chalkboard,” and allowing an outcome, such as “John sees a face drawn on the chalkboard and doesn't wipe it away.” Participants simply read these possible violations.

Then, participants learned the first rule (“*yag survist*”). In the Narrow Scope condition, participants were given 3 examples consistent with a narrow scope interpretation of the rule *yag survist*. The examples were “Ryan takes fingerpaints onto the playground,” “Ashley carries a pen to the playground,” and “Nicholas puts chalk next to a tree in the playground.” The rule is, of course, “Do not take things to the playground.” In the Wide Scope condition, participants were given 1 example consistent with the narrow scope and two examples inconsistent with the narrow scope—that is, that involved a person allowing an outcome. The inconsistent examples were, “Megan sees a pencil on the slide in the playground and doesn't take it inside” and “Justin notices a dry-erase marker by the soccer net in the playground and leaves it there.” Clearly, the rule in this case also prohibits leaving things on the playground.

Participants then were provided with the list of 16 possible violations they saw earlier, and were asked to check off in which cases the person was violating the rule *yag survist*. The 16 possible violations concerned different domains, so participants were expected only to check off instances of violations within the domain of the examples (i.e., a playground). We knew, from our previous experiments that participants in the Narrow Scope condition would tend to infer a narrow scope rule and only check off instances of putting things in the playground, whereas participants in the Wide Scope condition would infer a wide scope rule and check off instances of both intentionally putting things in the playground as well as cases of allowing things to be left in the playground.

Then, participants in the Narrow Scope condition learned 2 more narrow scope rules using this same paradigm, and participants in the Wide Scope condition learned 2 more wide scope rules. The rules concerned drawing/leaving things on a chalkboard and putting/leaving things on the ground, respectively.

After this learning phase, participants made an inference about the scope of a fourth rule, which concerned putting/leaving things on a shelf. (This “test” phase was not presented as a distinct phase—participants were led to believe that this was simply the fourth rule of the four they were to learn for the experiment). Only two examples of violations were provided for the fourth rule, so that the scope would be more ambiguous. Participants in both conditions got the same two examples, and the examples were both consistent with the narrow scope (e.g., “Mike puts a block on the shelf.”) (see figure 3). Considering results from our previous study (Nichols et al. forthcoming, study 1) in which participants overwhelmingly thought a new rule was narrow scope after receiving only 1 example violation consistent with this scope, receiving 2 examples consistent with a narrow scope normally would cause participants to think the rule was narrow scope—that is, only forbade bringing about the outcome, not allowing it. However, we predicted that participants in the Wide Scope condition, who learned 3 wide-scope rules, would acquire a different overhypothesis about the typical scope of the rules in the experiment and be less likely than participants in the Narrow Scope condition to think the rule applied only to bringing about the outcome.



Narrow

Rule 1 (Narrow)

- Ryan takes finger-paints onto the playground.
- Ashley carries a pen to the playground.
- Nicholas puts chalk next to the tree in the playground.

Rule 2 (Narrow)

- Lisa draws a circle on the chalkboard.
- Andrew writes the name "Fred" on the chalkboard.
- Laura makes an X on the chalkboard.

Rule 3 (Narrow)

- Nick opens a piece of gum in the park and lets the wrapper fall to the ground.
- While sitting in the park, Tina throws her soda can into the bushes.
- Robert is in the park, notices that his pen is out of ink, and tosses it on the ground.

Rule 4 (?)

- Mike puts a block on the shelf.
- Sarah picks up a ball and rolls it onto the shelf.

Wide

Rule 1 (Wide)

- Ryan takes finger-paints onto the playground.
- Megan sees a pencil on the slide in the playground and doesn't take it inside.
- Justin notices a dry-erase marker by the soccer net in the playground and leaves it there.

Rule 2 (Wide)

- Joe sees the name "Randy" on the chalkboard and leaves it there.
- Susan sees a drawing of a cat on the chalkboard and doesn't rub it out.
- Lisa draws a circle on the chalkboard.

Rule 3 (Wide)

- Nick opens a piece of gum in the park and lets the wrapper fall to the ground.
- Pam sees a candy wrapper on the ground in the park and leaves it there.
- Steve sees a flyer on the ground in the park and doesn't throw it out.

Rule 4 (?)

- Mike puts a block on the shelf.
- Sarah picks up a ball and rolls it onto the shelf.

Figure 3

There were two cases in the 16-case dependent measure concerning leaving something on the shelf: “David enters the room, sees a puzzle on the shelf and leaves it there” and “Emily sees a marble on the shelf and walks past it.” We expected participants in the Wide Scope condition to be more likely to check these as violations than participants in the Narrow Scope condition.

Results and discussion

As expected, participants who learned 3 wide scope rules were more likely to think that the new rule was wide scope than participants who learned 3 narrow scope rules. The cases of interest in the 16-case dependent measure were the 2 instances in which a person leaves something on the shelf (e.g., "David enters the room, sees a puzzle on the shelf and leaves it there"), which, if checked as violations, indicate a wide-scope understanding of the rule. Participants in the Wide Scope condition were significantly more likely to think that these were violations (green bar) than participants in the Narrow Scope condition, χ^2 (2, N=61) = 7.69, *p*<.05, *V*=.355.

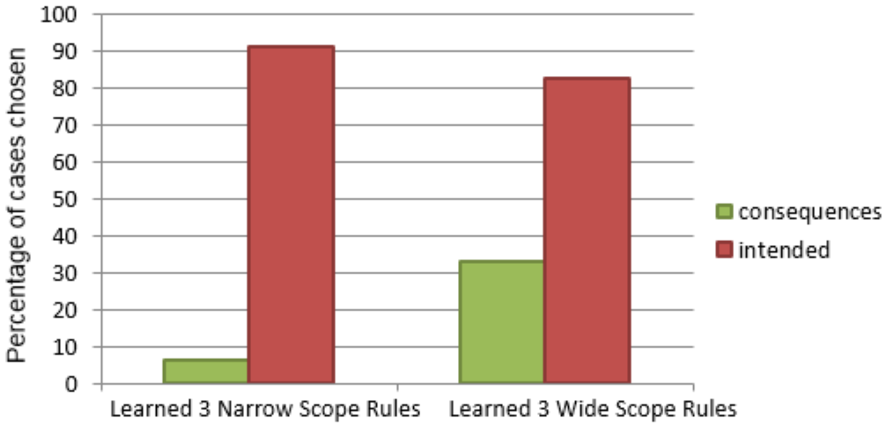


Figure 4: Percentage of cases chosen as violating the new rule (study 1)

Participants in the Wide Scope condition checked off instances of a person leaving something on the shelf as violations about a third of the time. In contrast, participants who learned 3 narrow scope rules checked off these cases as violations less than 5% of the time. This rate (5%) is about the same as the rate at which participants judge the rule to be wide scope in previous studies (Nichols et al. forthcoming). Participants' responses in the Narrow Scope condition are therefore no different than what we'd expect even if they had not been exposed to 3 narrow-scope rules first. It therefore appears that exposure to 3 wide-scope rules caused participants to be more likely to think that the rule was wide-scope than they would normally.

Participants in both conditions thought that a person who intentionally put something on the shelf violated the rule. This was expected, since the sample violations themselves involve intentionally putting things on the shelf. There was no difference between these conditions, $\chi^2(2, 61) = 2.8594$, p=.24, n.s.

Discussion

The results of Study 1 show that the bias for act-based rules observed in our previous experiment—i.e., the tendency to assume a rule is act-based even after receiving only a single example of a person bringing about the outcome—can be altered by learning 3 rules that apply to allowing. This result shows that people are sensitive to the scope of rules in making generalizations about the rules. Even though the fourth rule differed from the others in terms of the specific action that was prohibited (i.e., putting/leaving things on the shelf vs. putting/leaving things on the playground or the park), participants applied a generalization about the previous rules to the fourth rule.

It is important to note that only about a third of participants thought the fourth rule prohibited allowing the outcome in the Wide Scope condition, despite learning three allowing rules in the first part of the experiment. Most participants (2/3) in this condition did *not* infer the fourth rule to be Wide Scope. We think this is unsurprising considering participants received two sample violations consistent with the narrow scope interpretation of the rule, and none suggesting a wide-scope interpretation. (I.e., both sample violations involved a person bringing about the outcome). The “data” (in the form of the sample violations) about the rule therefore *competed* with the second-order overhypothesis in the Wide Scope condition, by design. It is therefore not surprising that many participants did not apply a generalization about the first three rules to the fourth, given that they sample violations they received for that rule suggested a narrow-scope interpreation of the rule. We take the fact that a third of participants believed the rule was wide-scope evidence of the *influence* of a learned overhypothesis in the Wide Scope condition, even if the overhypothesis did not override the impact of the sample violations for most participants in informing their inferences about scope.

4. Study 2

If people are using evidence concerning prior rules to form overhypotheses about rules, then the number of rules that they see should affect their inferences. Consider again Goodman's bags of marbles. Imagine that in one case you're shown 9 marbles drawn from 1 bag. Three of the marbles are large, one of these is red, another green and another yellow; 6 of the marbles are small, two of thesee ae red, two green and two yellow. In the other case, you're shown 3 bags, with 3 marbles drawn from each; in each bag, there is 1 large marble and 2 similarly colored small marbles. You are trying to determine whether a new bag will have both large and small marbles in it. In the 3 bag case, you have stronger evidence in favor of the overhypothesis that the next bag will have both large and small marbles in it. Similarly, if you are given 3 rules, each with 3 mixed violations (some act, some allow), that gives more evidence in favor of an allow-overhypothesis than if you are given 1 rule with 9 mixed violations. We exploited this feature of overhypothesis formation for study 2. We exposed participants to nine examples of act- and allow- violations that were said to either violate three separate rules (the 3-rule condition) or a single rule (the 1-rule condition). Then, they were asked to judge the scope of a new rule. Our prediction was that participants in the 3-rule condition would be more inclined to infer that a new rule forbade allowing the outcome.

Participants and Procedure

Participants were recruited from an online panel (Amazon's Mechanical Turk). Sixty participants (30 female) completed the study for a nominal cash payment. None were excluded.

As in Study 1, participants were asked to imagine they were learning rules from foreign cultures based on the example violations provided. In the 3-rule condition, they were told they would be learning four rules: *yag survist*, *nib weigns*, *pim storna*, and *taf byrnal.* (*Taf byrnal* was the test case with which we assess participants' tendency to generalize to the wide scope). In the 1-rule condition, they were told they would learn 2 rules: *yag survist* and *taf byrnal.*

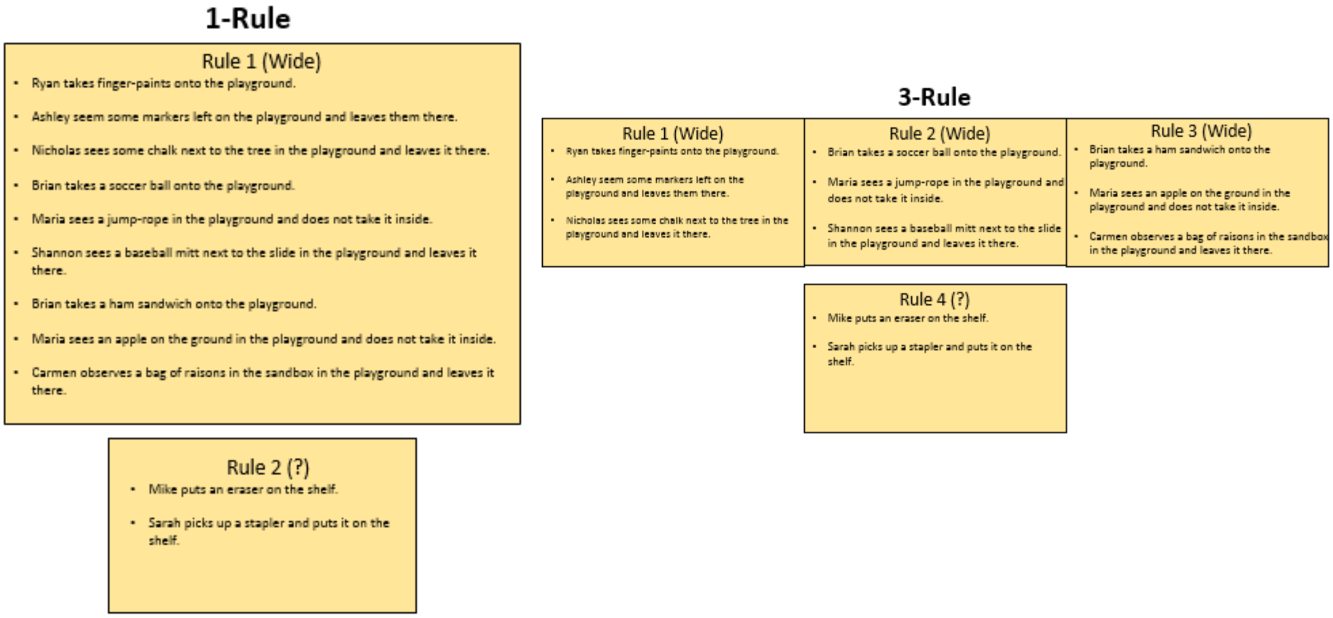
Participants in both conditions were exposed to the same nine example violations which involved putting or leaving things on the playground. Of the violations, three were "act" violations (involving someone putting something on the playground) and six were "allow" violations (involving someone leaving something on the playground) (see Figure 1). In the 3-rule condition, the violations were sorted into three distinct rules: one concerning art supplies, another involving sports equipment, and another concerning food. (These distinctions were not explicitly provided to participants; the violations were only distinguished by the name of the rule that subsumed them). Each rule contained one example of an allowing violations, meaning all the rules were wide-scope.

In the 1-rule condition, the violations comprised a single rule involving a general prohibition against leaving things on the playground. (This was also a wide scope rule as six of the example violations involved allowing the outcome). Participants were exposed to the violations both sequentially and then in summary form at the end of the first phase of the experiment. In the sequential phase, they clicked through a series of screens each of which described a violation of a rule (e.g., "the person in the following scenario is violating the rule *pim storna*: Brian takes a ham sandwich onto the playground). In the 1-rule condition the violations were all noted to be violations of the same rule (*yag survist*) while in the 3-rule condition they were described as either violations of *yag survist*, *nib weigns*, or *pim storna*. The order of violations was constant between the conditions. Participants were then shown a summary screen with the violations of each rule listed under the rule (in the 1-rule condition, they were all listed under a single rule).

Because we told participants they would learn the rules, we then asked them what they thought each rule meant using a fill-in-the-blank question. For instance, for *yag survist*, they were asked to complete the following: "According to the rule *yag survist*, it is wrong for a person to…" Participants in the 1-rule condition only completed one fill-in-the-blank since they were exposed to only a single rule, while participants in the 3-rule condition completed three fill-in-the-blanks, one for each rule. They could freely observe the violations of each rule as they guessed its meaning.

Then participants were told they would learn one more rule: *taf byrnal*. They were shown two examples violations of the rule. The violations were both consistent with a narrow-scope interpretation of the rule—i.e., that it prohibited bringing about the outcome but not allowing it. The violations involved people putting office supplies (an eraser and a stapler) on a shelf.

As the dependent measure, participants were asked whether a person who allowed something to remain on the shelf was also breaking the rule. The case was, “David enters the room, sees a tape dispenser on the shelf and leaves it there.” The greater participants’ confidence that this is a violation, the more confident they are that the rule is wide-scope. Participants made ratings on a 6-point scale ranging from “Definitely not breaking the rule” to “Definitely breaking the rule.” Participants were also asked about a case involving a person bringing about the outcome: “Amy moves a pencil case to the shelf.”



1-Rule

Rule 1 (Wide)

- Ryan takes finger-paints onto the playground.
- Ashley seem some markers left on the playground and leaves them there.
- Nicholas sees some chalk next to the tree in the playground and leaves it there.
- Brian takes a soccer ball onto the playground.
- Maria sees a jump-rope in the playground and does not take it inside.
- Shannon sees a baseball mitt next to the slide in the playground and leaves it there.
- Brian takes a ham sandwich onto the playground.
- Maria sees an apple on the ground in the playground and does not take it inside.
- Carmen observes a bag of raisons in the sandbox in the playground and leaves it there.

Rule 2 (?)

- Mike puts an eraser on the shelf.
- Sarah picks up a stapler and puts it on the shelf.

3-Rule

Rule 1 (Wide)

- Ryan takes finger-paints onto the playground.
- Ashley seem some markers left on the playground and leaves them there.
- Nicholas sees some chalk next to the tree in the playground and leaves it there.

Rule 2 (Wide)

- Brian takes a soccer ball onto the playground.
- Maria sees a jump-rope in the playground and does not take it inside.
- Shannon sees a baseball mitt next to the slide in the playground and leaves it there.

Rule 3 (Wide)

- Brian takes a ham sandwich onto the playground.
- Maria sees an apple on the ground in the playground and does not take it inside.
- Carmen observes a bag of raisons in the sandbox in the playground and leaves it there.

Rule 4 (?)

- Mike puts an eraser on the shelf.
- Sarah picks up a stapler and puts it on the shelf.

Figure 5: Division of violations into three rules (3-rule condition) or one rule (1-rule condition). The test rule, *taf byrnal*, is denoted as "Rule 2" in the 1-rule condition and "Rule 4" in the 3-rule condition.

Results

We predicted that participants in the 3-rule condition would be more likely to think that the new rule was wide-scope in the 3-rule condition than the 1-rule condition, since there is more information about the category in the 3-rule condition than the 1-rule condition. In accordance with this hypothesis, participants in the 3-rule condition were more likely to think David violated the rule by leaving the tape dispenser on the shelf (M = 4.45) than participants in the 1-rule condition (M = 3.48), $t(58) = 2.2$, $p < .05$, Cohen's $d = .57$. As expected, Amy was thought to violate the rule by intentionally putting a pencil case on the shelf in both the 3-rule condition (M = 5.45) and in the 1-rule condition (M = 5.59), $t(58) = .55$, n.s (see figure 6).

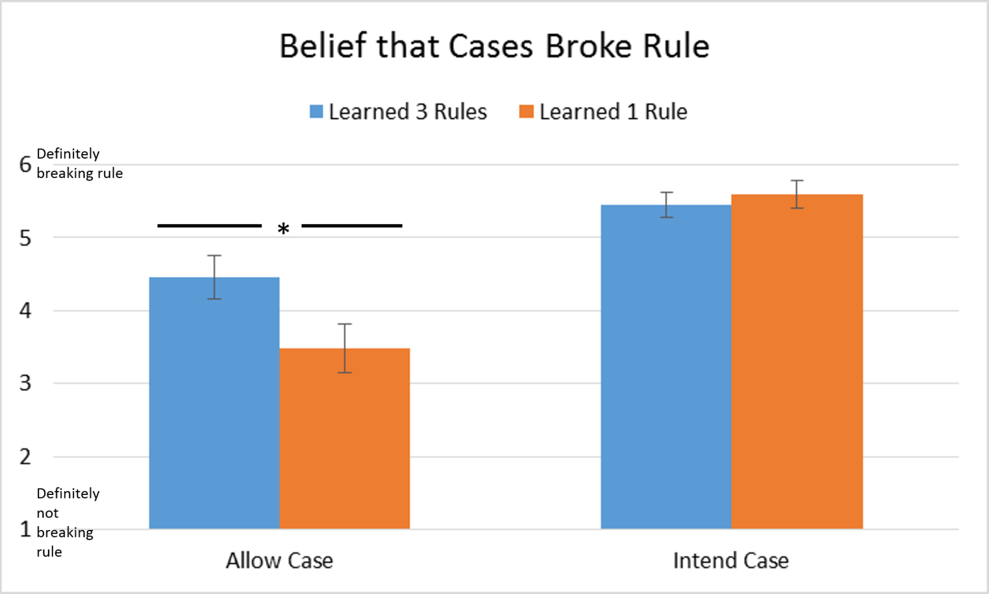


Figure 6: Extent to which an intended or allowed outcome was judged to break the new rule (study 2). (Error bars indicate standard error of the mean)

Discussion

Study 2 again suggests that people form overhypotheses about rules that inform first-order inductive inferences about new rules. Sampling 3 violations from 3 rules is more informative about rules than sampling 9 violations from a single rule. As a result, we predicted that participants in the 3 rules condition would be more likely to generate an overhypothesis that rules forbid allowing outcomes, and that this would make them more inclined to say that a new rule forbid allowing the outcome. This is exactly what we found.[3]

[3] This result also provides evidence against a deflationary explanation of study 1. While the two conditions of that study differed in terms of the scope of the rules participants were exposed to, they also differed in the type of violations participants encountered. In the Wide Scope condition, participants were given examples of violations involving a person allowing an outcome. It is therefore possible that participants in the Wide Scope condition were more likely to think that the allowing case was a violation simply because they formed a first-order hypothesis about violations (e.g., "instances of

5. Study 3

In Studies 1 & 2 we manipulated participants' overhypotheses about rules by exposing them to

a person allowing the outcome tend to count as violations"), rather than about rules (i.e., "most rules tend to prohibit allowing the outcome"). For comparison, consider again the marbles. If one is exposed to 3 bags of marbles and samples 3 different colored marbles from each bag, one might notice the variation in color of the marbles without attending to the fact that each sample was drawn from a different bag. In that case, one might come to a general view about marbles (that they vary in color), not about bags of marbles. That is, one might expect future marbles to exhibit different colors, without expecting that marbles will be diversely colored within a bag. Similarly, in study 1, it's conceivable that participants exposed to wide scope violations expect that many future violations will also involve allowing; however, on this proposal, this inference isn't about expecting that violations will include acting and allowing within a rule, just that violations in general will be diverse with respect to whether allowing an outcome is a violation.

We don't think this explanation is plausible, since we expect that people will pay attention both to bags and to rules. In the case of marbles for instance, even if one has seen a wide range of diverse colors pulled from 3 bags, if for the fourth bag, 3 red marbles are pulled, one is likely to think that the next marble will be red. That is, one attends to the fact that this is a separate bag. Similarly, we suspect that even if participants are exposed to several sample violations that are allowings, if for the new rule, exactly 4 act-based examples are pulled, participants are likely to think that the next violation will also be act-based. They won't simply rely on the total number of cases of allowing-based violations to make their predictions. But in any case, study 2 provides further reason to doubt the deflationary interpretation. For in that study, we see that having the exact same instances of violations leads to different inferences depending on whether the instances are assigned to a single rule or to three rules.

examples of either wide or narrow scope rules. In our third experiment, we decided to use a more subtle manipulation. Rather than require participants to learn rules, we merely exposed them to examples of legal violations purportedly as part of a separate study on judgments about laws. This manipulation better resembles how participants might encounter actual rules in the environment. We then had them learn a new rule. In addition, in study 3, we focused on a new distinction – between deliberate actions and accidents (i.e., an outcome unintended and unforeseen by the agent).

Recall the nested structure of consequences along the act-allow dimension, depicted in figure 1. Intended and accidental consequences also have a nested structure, as depicted in figure 7. With respect to the class of actions brought about by agents, not all are intended—some are accidental. Accidental consequences fall outside the narrow box containing intended consequences. Rules may be formulated at either the “narrow” scope (i.e., prohibit only intentionally bringing about an outcome) or at the “wide” scope (i.e., prohibit even accidentally causing the outcome). For the purposes of study 3, the term “narrow scope” will refer to rules that prohibit intentionally causing the outcome, and the term “wide scope” will refer to rules that prohibit intentionally or accidentally causing the outcome.

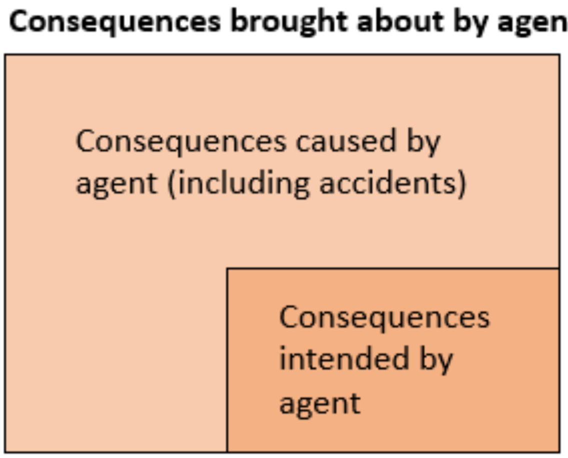

Consequences brought about by agents:
Consequences caused by agent (including accidents)
Consequences intended by agent

Figure 7: Nesting of consequence types. Considering the subset of consequences that are brought about by the agent, not all consequences brought about by the agent are intended, but all intended consequences brought about by the agent are brought about by the agent.

We take it that rules prohibiting accidents are much less common than rules prohibiting intentional acts. A rule that prohibits accidentally causing an outcome could not be action-guiding in the same way as intention-based rules are, because by definition when one accidentally causes an outcome, she fails to foresee that her action will bring about the prohibited outcome. Therefore, one cannot refrain from acting on the basis that acting will violate the rule. The only method by which one can ensure compliance with the rule is to exercise extreme caution with respect to acting in ways that might bring about the outcome—e.g., refusing to sell alcohol to anyone who looks under 35 years of age to avoid accidentally selling alcohol to a person under 21.

Although accident-inclusive rules are rare, there are indeed such rules to be found. For example, strict liability laws specify that a person is legally responsible for damages associated with an outcome even if the person did not foresee the outcome and could not reasonably be expected to. Strict liability laws are designed to encourage precisely the kind of extreme caution that would be required to avoid accidental violations. Consider the strict liability law that one may not allow one's pet to harm another person. This law specifies that if an animal causes harm to a person, the owner of the animal is responsible, even if proper measures were taken to ensure the animal's captivity (e.g., a well-constructed cage) thus making the harm unforseeable. This law encourages taking strong precautions against the possibility of one's animal(s) causing harm.

The question we addressed in Study 3 was: Does exposing participants to examples of intention-based or accident-inclusive (Strict Liability) laws influence their inferences about the scope (i.e., intention-based or accident-inclusive) of a new rule?

Participants and procedures

Participants were recruited from an online panel (Amazon's Mechanical Turk). Sixty-five participants (32 female) completed the study for a nominal cash payment. None were excluded. The description of the study noted that participation entailed completing two short studies.

Participants were told the first study concerned opinions about laws. The instructions were as follows: “For this study we are interested in peoples' opinions about laws. We will give you examples of some laws in our culture and you will be asked some questions about them.”

Participants were first exposed to either 4 wide scope (i.e., Strict Liability) or 4 narrow scope laws. Participants were led to believe that this part of the experiment constituted a separate study, on “opinions about laws.”. They were told they would be given examples of current laws, and asked some questions about the laws.

One example of a Strict Liability law in the Wide Scope condition was the following:

- If a person takes wood from someone else's property, this is a violation of the law against timber trespass. It doesn't matter whether the person knows that they were on someone else's property. For example, if person A thinks he is taking wood from his own property but is actually on the property of person B, person A has broken the law and is legally responsible.

In the Narrow Scope condition, the four rules were matched in domain but only referenced intentionally causing an outcome. For instance, the corresponding narrow version of timber trespass was:

- If a person takes wood from someone else's property, this is a violation of the law against timber trespass. For example, if person A knows that he is on person B's property and takes wood from that property, person A has broken the law and is legally responsible.

After reading each rule, participants were asked several questions about the rule that accorded with the cover story. The questions were, “To what extent do you agree with this law?”, “How important do you think the law is?", and "How upset would you be at someone who did not follow the law?" Responses were made on a 5-point scale.

Next participants learned a new rule. At the start of this part of the experiment, participants were told that they were participating in a second study regarding how people learn new rules. For this part of the study, we returned to the use of arbitrary rules so as to minimize the resemblance between the rules participants were exposed to in the first phase of the experiment and the learned rule. Participants were told they would learn a rule called *lep depros* which involved littering at a park. They were given two examples of violations of the rule *lep depros*, both of which were consistent with a narrow scope interpretation of the rule—i.e., a rule that forbade only intentionally littering. The examples were, "Nick tosses an empty gum wrapper on the ground when he is walking in a park" and "While she is in a park, Pam puts an empty can of soda on the ground and leaves it there." Next, participants judged whether two new cases were also violations of the rule. One case was an instance of a person intentionally littering:

> Jenna tosses a bottle cap on the ground while she is walking in the park.

The case of interest was a case in which the person accidentally litters:

> Mark is taking a walk with his friend in a park. There is a small hole in Mark's backpack. A bottle cap falls out without Mark noticing.

Responses were made on a 6 point scale ranging from "Definitely not breaking the rule" to "Definitely breaking the rule." If participants thought the rule applied to all outcomes caused by the agent, they should think that Mark's action was a violation even though it was entirely accidental.

Results and discussion

We predicted that participants primed with strict liability laws (the Wide Scope condition) would be more likely to think that the new rule applied to accidental outcomes. In accordance with the prediction, participants in the Wide Scope condition were significantly more likely to think that Mark violated the rule (*M* = 3.60) than participants in the Narrow Scope condition (*M* = 2.46), *t*(63) = 3.04, *p*=.003, Cohen's *d* = .751. In contrast, with respect to the case of Jenna intentionally littering, participants in the Wide Scope condition were no more likely to think that she violated the rule (*M* = 5.67) than participants in the Narrow Scope condition (*M* = 5.71), *t*(63) = .23, *p* = .82, n.s. This is as expected, since Jenna would violate the rule by intentionally littering regardless of whether the rule was Wide or Narrow scope (see figure 8).

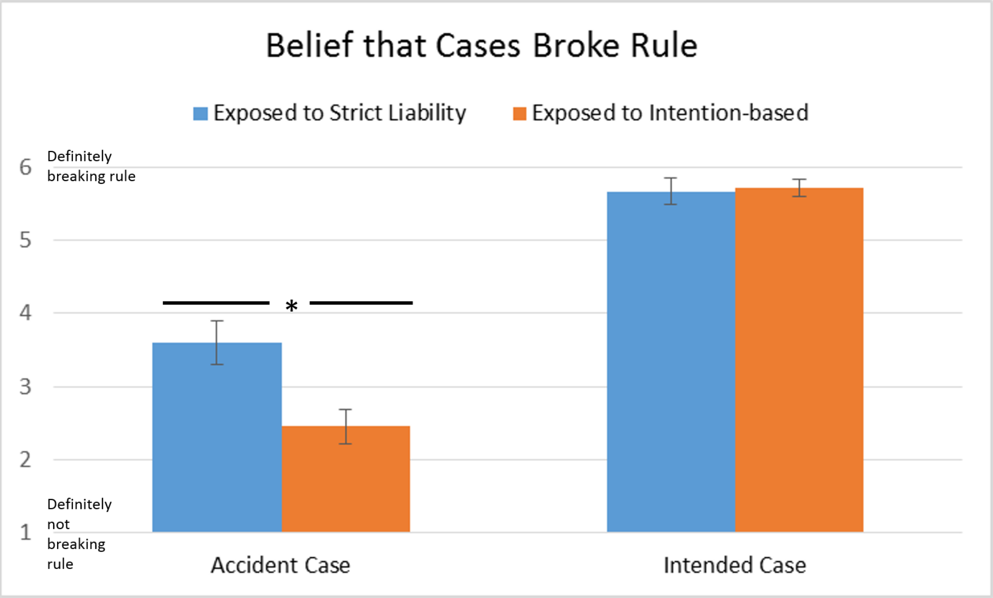


Figure 8: Extent to which an intended or accidental outcome was judged to break the new rule (study 3). (Error bars indicate standard error of the mean.)

Study 3 thus indicates that peoples' inferences about the scope of a new rule can be influenced by exposure to rules (laws) at a particular scope. Furthermore, the study suggests that directing participants' attention to the scope of the rules in the first phase of the experiment is not necessary for this information to influence judgments about a new rule. Study 3 also demonstrates that the effect extends to inferences about a different scope distinction—that between intentionally causing an outcome and accidentally causing it.

It is possible that the results on study 3 are the product, not of statistically appropriate application of an overhypothesis, but because the manipulation simply made rules of a certain scope more salient. Studies 1 and 2 provide better evidence for the role of overhypotheses in learning new rules. In the context of those results, it's plausible that participants are also reasoning with an overhypothesis in study 3. However, we think that the real significance of study 3 is that it shows that people naturally abstract scope information from examples that they are exposed to, and this abstracted scope information guides the inferences they make on a learning task.

6. General Discussion

When people are presented with moral dilemmas, they often respond in sophisticated yet predictable ways. For instance, people tend to judge that it's worse to produce a bad outcome than to allow a bad outcome to persist. This pattern holds for judgments about many different kinds of scenarios, including counter-factual scenarios that people have not before considered. Moreover, when people learn new rules, they seem to be biased in favor of expecting the rule to prohibit acting, but not allowing. One explanation for this tendency is that it derives from abstract principles of an innate moral grammar. We explore an alternative possibility on which abstract principles do play a role in the acquisition of moral rules, but the abstract principles themselves are acquired from exposure to many different specific rules. Exposure to specific rules could lead to the formation of an overhypothesis about the typical scope of rules (e.g., that they typically applying to *acting* but not *allowing*), which then informs and constrains the learning of new rules.

In three experiments, we provided preliminary evidence for the key postulates that (1) People attend to and retain information about the scope of rules they learn/encounter and (2) The scope information they retain influences judgments about the scope of unfamiliar rules and the particular actions that are prohibited by these rules.

In Study 1 we applied a paradigm for examining overhypotheses about categories to rule learning. We found that after learning three allow-based rules, participants were more likely to think that a new rule prohibited allowing the outcome to occur, even though both examples of violations of the rule involved a person intentionally causing the outcome. This suggests that people can form overhypotheses about rules—in particular, about the scope of the rules.

In Study 2, we explored whether the number of learned rules mattered. We exposed all participants to 9 violations, 6 of which involved allowing an outcome to persist. In one condition, participants were told that these were all violations of one rule; in the other condition, the 9 sample violations were divided into three groups, each of which corresponded to a different rule. We found that when participants learned 3 rules, they were more likely to judge that a new rule applied to allowing an outcome to persist.

In our third experiment, we explored whether overhypotheses about rules might be manipulated merely exposing participants to the rules rather than requiring them to learn the rules. Participants who were exposed to 4 strict-liability laws as part of a supposed separate study were more likely to think that a new rule also forbid accidentally causing an outcome.

Our current work complements our earlier results and helps to fill in an empiricist defense of moral rule learning based in rational inference. In Nichols et al. (forthcoming), we showed how one resource in rational learning theory—the size principle—provides a "bottom-up" explanation of how moral rules can be acquired even without negative evidence. In our present work, we show how an additional learning-theoretic resource—acquired overhypotheses—can augment the empiricist account by explaining "top-down" constraints on moral rule learning.

Of course, our defense of moral empiricism is limited in important ways. Our experiments are on adults. To demonstrate that learned overhypotheses in fact explain moral projection would obviously require developmental research. Our goal here is to lay the foundation for an empiricist defense of these phenomena by providing evidence of certain key elements, such as the claim that people treat rules as "categories" to which overhypotheses are applicable and that generalizations about rules influence first-order inductive mechanisms.

Also note that even if the bias for act-based rules is explained by learned overhypotheses about rules, this would not mean that no innate knowledge was operative in this process. The space of possible overhypotheses that a learner considers may be innately specified. Learners may not consider the hypothesis that, e.g., a rule will prohibit allowing the outcome when the agent is female but not when the agent is male; that a rule will prohibit allowing but *not* bringing about the outcome; that a rule prohibits allowing an outcome when one is 50% certain it will occur but not when it is 100%; and so on. The logical space of possible rules is clearly much larger than the number of rules people consider. Nevertheless, even if the hypothesis space is to a large degree innately specified, the selection of hypotheses within this space might be explained by general purpose learning mechanisms like the formation of overhypotheses.

The present work is incomplete in other important ways. We have not presented formal treatments or computational models of our hypotheses or results. We hope that future work will address these limitations. In addition, it will be important to integrate an account of moral rules into a much fuller account of moral decision making. Recent work that applies Bayesian learning models to expected utility theory provides a promising framework (see, e.g., Baker et al. 2009, Kleiman-Weiner & Tenenbaum, this issue).

Finally, though we have emphasized specific learning mechanisms like the size principle and overhypotheses, there are undoubtedly other kinds of learning mechanisms involved in the acquisition of moral rules. The learning processes we describe involve making inferences from a sample assumed to be randomly-selected. However, often children will be provided with non-randomly sampled data, and when data is known to be non-randomly sampled, this affects learning in demonstrable ways. To take just one example, in "pedagogical sampling" learners are sensitive to the pedagogical intentions of teachers (see e.g., Shafto et al., 2014; see also Gweon et al. 2010). It is likely that participants' inferences about rules would change under the assumption that the examples of rule violations were pedagogically sampled.[4]

The results presented in this paper suggest that the moral empiricist has important resources for explaining certain features of moral cognition, such as the capacity for moral projection. The empiricist defense we have presented contrasts with both moral nativism and more familiar emotion-based accounts, which hold that many moral judgments are derived from emotive responses rather than systematic representations (Baron 1994; Greene 2008; Singer 2005; Unger). While both of these alternative accounts have much support, we think a moral empiricism specifying that important moral principles are learned remains a possibility. We hope to have established some optimism about the

[4] In recent work, we explored the role of pedagogical sampling in learning "closure rules" for norms system (Gaus & Nichols forthcoming). Many act-types are never mentioned in a norm system, and so a learner must make inferences about whether those unmentioned act-types are permitted or forbidden (Mikhail 2011, 133). Our hypothesis was that an efficient teacher would be expected to mention prohibition rules if the remainder of the act-types are permitted. In keeping with this hypothesis, we found that when a teacher provided only prohibition rules, participants tended to think that unmentioned act-types were permitted; and when a teacher provided only permission rules, participants thought the unmentioned act-types were prohibited. In addition, many participants explicitly referred to considerations of pedagogical efficiency in explaining their responses.

theoretical import of well-established features of human cognition—such as the formation of second-order generalizations about categories—to human morality.

Acknowledgments

We would like to thank Victor Kumar, Fiery Cushman, and three anonymous referees for extensive comments on an earlier version of this paper. Research for this paper was supported by Office of Naval Research grant #11492159 to SN.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, 113(3), 329–349.

Baron, Jonathan (1994). Nonconsequentialist decisions. *Behavioral and Brain Sciences*, *17*.

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.

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.

Gaus, J. & Nichols, S. forthcoming. "Moral Learning in the Open Society: The Theory and Practice of Natural Liberty." *Social Philosophy & Policy.*

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, 107(20), 9066-9071.

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

Harman, Gil (2000). Moral Philosophy and Linguistics. In *Explaining Value*, Oxford: Oxford University Press (2000), 217-226.

Heibeck, T. H., & Markman, E. M. (1987). Word learning in children: An examination of fast mapping. *Child development*, 1021-1034.

Kemp, Charles; Perfors, Amy; & Tenenbaum, Joshua (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental science*, *10*(3), 307-321.

Kleiman-Weiner, M. & Tenenbaum, J. (this issue). Learning a Commonsense Theory of Morality.

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

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, S., Kumar, S., Lopez, T., Ayars, A, & Chan, H. forthcoming. "Rational Learners and Moral Rules." *Mind & Language*.

Nichols, Shaun & Mallon, Ron (2006). Moral dilemmas and moral rules. *Cognition*, *100* (3), 530-542.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*,*90*(4), 339.

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.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382-439.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, 71, 55-89.

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.

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

Wright, J. C., & Bartsch, K. (2008). Portraits of early moral sensibility in two children's everyday conversations. *Merrill-Palmer Quarterly*, 54(1), 56-85.

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

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