Review of Shaun Nichols's *Rational Rules*

Joshua May

Published in *Ethics* Vol. 133, No. 3 (2023), pp. 434-440.

Nichols, Shaun. *Rational Rules: Toward a Theory of Moral Learning*.
New York: Oxford University Press, 2021. Pp. 272. $70 (cloth).

Commonsense morality is often viewed as a set of rules. *Don't lie. Be fair. Keep promises. Love thy neighbor. Respect your elders. Don't eat pork.* And so on. Ethicists have also drawn out some common features of these rules. Moral rules can apply either universally or only to certain people; to one's actions or omissions; to what one intends or also to what one foresees. How are these moral rules and distinctions learned? And is the process rational?

These are the guiding questions of Shaun Nichols's innovative and instructive book, *Rational Rules*. Like Hume, Nichols draws on our understanding of the human mind to answer questions about rationality in ethics. Against recent nativists (like Susan Dwyer, Gilbert Harman, and John Mikhail), he argues that moral rules are not innate but largely *learned*. Against the debunkers of commonsense morality (like Peter Singer and Josh Greene), he argues that ordinary moral learning is typically *rational*. And against many sentimentalists (like Jesse Prinz, early Jonathan Haidt, and to some degree Nichols himself), he argues that moral learning involves unconscious statistical *reasoning*. Although Nichols maintains that emotions are essential for normal moral cognition in humans, in this book reason is the star of the show. Even moral motivation, on Nichols's account, can't be explained by natural emotions.

In no uncertain terms, *Rational Rules* represents a significant shift in moral psychology. There are now major strands of evidence and theorizing that highlight rational inference over emotional responses, which paints a more optimistic picture of moral cognition and its ability to improve in response to reasons. Although the book only focuses on certain aspects of moral learning, it adds an important piece to the explanatory puzzle and paves the way for future work on moral development and moral progress.

Moral Psychology

For Nichols, moral learning isn't particularly mysterious. Fancy new theories have drawn on mechanisms of reinforcement learning that we share with many other animals. Yet Nichols convincingly argues that we probably don't learn rules like "Respect your elders" or "Incest is wrong" through good or bad experiences with these act types, like a rat who learns to avoid the electrified corner of a cage. Rather, we're simply *told* as children what the rules are. Humans do learn through reinforcement, but they are also rule-following creatures with language, and "it would be a sadistic parent who opted to use reinforcement learning on their child rather than simply telling them the rule" (43). "Rats are great and all," says Nichols, but "we need to go beyond rodent psychology" (28).

A fair point, but we should be wary of embracing a false dichotomy here. Testimony and reinforcement learning probably both shape the acquisition of sophisticated moral rules. Parents tell children not to lie, cheat, and steal, but these rules have exceptions (or can be overridden by other considerations), the details of which are rarely specified. Instead, we plausibly learn the nuances of rules and conflicts among them through reinforcement learning and some degree of deliberation (more on this later). Over time, we learn that people aren't punished for white lies but also that we personally don't experience much hurt or betrayal when on the receiving end of such lies, at least some of them—it depends on the context. Nichols seems well aware of these issues and contends only that, absent rule representations, reinforcement learning *alone* "can't possibly explain the character of our moral judgments" (45).

Nichols relies heavily on the model of parental instruction in his defense of moral empiricism. Criticisms of moral nativism abound, but Nichols aims to defend a systematic alternative with sophisticated machinery that can compete with Chomskyan nativism. The result is a direct response to "poverty of the stimulus" arguments (John Mikhail, Elements of Moral Cognition [Cambridge: Cambridge University Press, 2011]). There is no such poverty, according to Nichols, for children have plenty of data from which to infer basic moral rules and their characteristic features. Let's get a sense of this machinery before considering its philosophical implications.

Start with the familiar distinction between doing and allowing. Many moral rules, such as "Keep promises" or "Don't steal," apply only to one's own *actions*; they are not prohibitions on *allowing* other people to steal or break their promises. Is this doing/allowing distinction just an innate feature of many rules, instilled in us by evolution to detect rule violators and thus to promote cooperation and coordination among group members in the Pleistocene?

Perhaps that is part of the story, but Nichols believes the distinction can be explained by a rational principle in statistical inference: the "size principle." Imagine you're in quarantine and a kind friend calls you up to play a game with dice to pass the time. A fan of tabletop role-playing games, she naturally has many dice with different numbers of sides—ranging from 6 to 10 sides. She chooses a die at random and rolls it 12 times. It lands on a number less than 6 each time. The question of the game is: Which die is she rolling? All of the dice are compatible with the rolls, yet they are not all equally likely. Roughly, according to the size principle, when the relevant hypotheses have such a nested structure, the smallest hypothesis consistent with the evidence is the most probable (57). The principle is intuitive enough—it would seem a coincidence if the 10-sided die, say, were responsible for so many rolls under 6—but it can also be proven with a little Bayesian probability. Size matters, or at least that's what Bayes says.

Nichols's insight is that the doing/allowing distinction appears to fit the same structure. In brief, children must infer the scope of moral rules by observing when elders deem events to be rule violations. Ordinary experience and a corpus of recorded conversations with children suggest that kids are often told only what *they* should or shouldn't do. If children are almost exclusively told not to do something, like break promises, that's consistent with interpreting the rule as having wide scope: to minimize all promise-breaking. After all, one's own promise-breakings are a subset of all promise-breakings. Yet people tend to infer that rules have narrow scope, unless they are told explicitly that *allowing* is forbidden. Is that inference irrational? Or even an arational innate reflex? Not necessarily. The narrow scope interpretation is the "smaller" hypothesis consistent with the evidence, so it's more likely, according to the size principle. On this account, viewing certain moral rules as applying to doings rather than allowings is both learned (albeit unconsciously) and rational (apportioning one's credence to the evidence).

Nichols provides similar accounts of how we learn other common features of moral rules. Included in his analysis are *parochial* rules that apply only to some people, such as voting rights for men but not women, prohibitions against stealing but only from members of one's own community, and rules against eating certain animals. These can be thought of as restricting the scope of moral rules, and learners tend to infer that a rule has narrow scope unless they receive evidence otherwise. A similar story is given for the distinction between bringing about a bad outcome intentionally and doing so merely as a foreseen side-effect (central to the Doctrine of Double Effect). Outcomes that are brought about intentionally, such as making a classmate cry in order to create a diversion, are a subset of the same consequences that are foreseen, such as making a classmate cry as a side-effect of setting off the fire alarm. In both cases, the size principle makes it rational to infer the narrower scope version of the moral rule.

Of course, a narrower interpretation of a moral rule is rational only if there is no evidence for a wider interpretation. If learners receive evidence that, say, merely bringing about a side-effect is impermissible, then it's rational instead to infer that the rule has wide scope. Similarly, with different evidence, learners can infer that a rule applies to allowings or to more than one's ingroup. Nichols seems happy to admit that some moral rules have wide (or wider) scope. Perhaps, for example, rules related to negligence, recklessness, and carelessness have a wide scope that includes certain omissions, allowings, and foreseeable harms. Even if compatible with Nichols's account, it's important to recognize how common these rules are, at least in relatively civil society. Many of the ways people wrong you are not through intentional actions like violence or theft but omissions and foreseeable side effects, such as unpaid debts, unanswered emails, unwashed dishes, missed apologies, and forgotten birthdays. These aren't instances of the rare or unusual case of strict liability but rather failures to show due care, consideration, or respect. A theory of moral learning should be sure to fully explain both wide a narrow scope rules.

Nichols also provides statistical learning accounts of other aspects of moral psychology. He draws on the notion of an "overhypothesis" to explain the tendency to assume that most moral rules are act-based, thus grounding a general doing/allowing distinction in commonsense morality as a Bayesian "prior" (Chapter 4). Like Mikhail, Nichols also draws on a "closure principle" to explain why people tend to assume that an action is morally permissible unless there is a rule against it. Unlike Mikhail, however, Nichols shows that such a closure principle might be learned through "pedagogical sampling," rather than a feature of an innate moral grammar (Chapter 5).

Nichols even attempts to explain some folk metaethics: why some rules are thought of as universal, others relative (Chapter 6). Compared to wearing pajamas to school, for example, stealing tends to be treated as universally wrong and regardless of whether an authority says it's OK (authority-independence being a core element of the famous moral/conventional distinction). Nichols argues that learners unconsciously use information about disagreement and consensus among people to determine whether a statement should be understood as universally true or only true relative to a particular location, time, or standard. Nichols draws a useful comparison with the seasons (116). Imagine a child living in Australia who learns that July is a winter month. She might assume this is universally true at first, but she eventually learns that it's only true relative to the southern hemisphere. One way to learn this is by explicit instruction, but another is through evidence of disagreement or lack of consensus: some people in movies say that July is a *summer* month. One way to explain the disagreement is that both sides are right relative to some standard. (The statistical principle here is, roughly, just fitting a hypothesis to the data.).

Nichols argues for a similar approach to moral statements. Some are typically regarded as universal, such as "Stealing is wrong," while others are treated as more relative, such as "Don't marry your cousin." On Nichols's account, these differences can be explained by learners using disagreement as evidence that, like “Pajamas shouldn’t be worn to school,” the statement “Don’t marry your cousin” is true relative to some standards but not others. Of course, this is a primarily psychological claim about how people view the truth conditions of these statements. However, since Nichols is providing a statistical learning explanation, the psychological process plausibly counts as rational (whether or not the resulting beliefs are accurate).

Moral Implications

Does it matter much whether moral rules, or their characteristic features, are learned rather than innate? Empiricism and nativism both seek to explain common moral rules and their features, such as ranging over doings rather than allowings. Nativists say these are largely innate or organized in advance of experience; empiricists say they’re primarily learned. But why would *these* types of rules be innate or taught? Nichols says act-based norms are better able to achieve human aims (161, 183-6), such as having rules that are easy to follow and enforce. Yet presumably this would also be the distal explanation for nativists: humans tended to adopt act-based rules because they facilitated coordination and cooperation among large groups that can outcompete others. Does it really matter whether certain moral rules are ecologically rational *now* or *were* in the Pleistocene (or some more recent episode of cultural evolution)?

Yes, says Nichols, for several reasons, but let’s focus on two. The first relates to moral epistemology. Nichols’s empiricism supports a relatively optimistic view of commonsense morality, partly because it’s a form of rationalism. Despite the historical marriage between rationalism and nativism, Nichols’s empiricism draws on statistical learning rather than emotional resonance, which implies that moral learning deploys domain-general inference. This is important because moral cognition then resembles other forms of cognition in applying domain-general mechanisms to a particular subject matter (Joshua May, “Moral Rationalism on the Brain,” Mind & Language). As Nichols draws out, the role of domain-general reasoning mechanisms paves the way for a defense of commonsense morality. Even if not perfect, common moral distinctions are insulated from certain debunking arguments because learners aren’t relying on rigid emotional heuristics driven by morally irrelevant factors like an aversion to up close and personal contact. I’d add that Nichols’s learning approach identifies specific mechanisms that could help engineers implement moral competence in artificial intelligence. Deep learning algorithms might be trained to learn not only the rules of games, like chess and Go, but of human moral systems. (Or, if it turns out that AI don’t learn moral rules under the conditions human children are in, then we might gain evidence against Nichols’s account.)

Of course, commonsense morality isn’t perfect, and Nichols realizes there is room for improvement. A second important implication of his empiricism relates to moral progress. If moral learning is statistical inference, then it’s flexible enough to update in response to new evidence. We aren’t locked into a fixed moral grammar. Consider greater inclusivity, a prominent form of moral progress (Allen Buchanan & Rachel Powell, The Evolution of Moral Progress [New York: Oxford University Press, 2018]). Societies have improved by expanding rules about voting, marriage, and freedom to include women, racial minorities, and same-sex couples. Again, Nichols allows that we can adopt wider scope rules, and these improvements could be construed as moral norms gaining wider scope by applying to more people. At one point, Nichols proposes that his account is “consistent with thinking that it’s possible to move people to more inclusive moralities by giving them different evidence” (80).

However, the model we're given is not only learning-theoretic but largely testimonial. For all the statistical principles that underwrite Nichols's account, the *evidence* is primarily what parents or society deem to be a rule violation, which isn't necessarily a good guide to ethics. The focus on how children learn moral rules is instructive for resisting theories of moral learning that draw only on evolutionary explanations or animal models, but it's important to emphasize that moral learning continues throughout the lifespan. The focus on child development becomes less informative when we turn to adults and teens who go beyond uncritical acceptance of society's rules.

Some elements of Nichols's statistical model of moral learning might apply well to the adoption of new moral rules in adulthood. When discussing universal versus relative rules, Nichols draws heavily on the notion of consensus information, but the more general phenomenon is social learning, which continues well past childhood. Consumers consult aggregated restaurant reviews as well as skilled and trusted food critics. People's opinions about the death penalty and same-sex marriage are similarly influenced by community leaders and other moral exemplars. These well-studied psychological mechanisms of social learning aren't just a matter of mindlessly following what others tell us to do. We use other people's behavior as *evidence* of what to do, even on controversial moral issues. For example, information about increasing rates of vegetarianism has been shown to reduce meat consumption (Joshua May & Victor Kumar, "Harnessing Moral Psychology to Reduce Meat Consumption," Journal of the American Philosophical Association). When other people you know and trust change their minds in response to reasons, you take those reasons more seriously. That's the power of social learning, which is responsive to consensus but also to particular influencers and trends. Although the mechanism here isn't the size principle or mere consensus, an empiricist theory like Nichols's might be expanded to accommodate other ways that social learning drives moral cognition.

Nevertheless, the explanatory can is kicked further down the street. Why, we might ask, do the moral trend setters change *their* minds? Here the mechanisms of moral learning might involve good old-fashioned reasoning that is less statistical, though no less rational (Joshua May, Regard for Reason in the Moral Mind [Oxford: Oxford University Press, 2018]).

Consider the rapid change in Americans' attitudes toward gay people over the past few decades. It doesn't look as though heterosexual Americans were simply being told by some authority that "Love thy neighbor" and "Don't be a bully" applies to one's treatment of gay people too. Nor does it seem that younger people with more progressive parents were being taught more inclusive moral rules. Polling data suggest that attitudes toward homosexuality changed across a wide range of age groups, ethnicities, and religious denominations. So generational turnover is at best a small piece of the puzzle. An arguably larger factor is that people started to love and respect their children, relatives, friends, co-workers, and church members who later came out as gay. An inconsistency in one's moral outlook ultimately surfaced: How can I love and respect my daughter's same-sex relationship yet reject other similar relationships? It seems many individuals resolved this inconsistency by accepting rather than rejecting all gay people (Victor Kumar & Richmond Campbell, A Better Ape [New York: Oxford University Press, 2022], 210-14). Although the relevant moral norms became more inclusive, notice how different this form of moral belief revision looks from statistical learning of rules. It certainly isn't just correcting a "sampling error" (190).

Conclusion

Fitting with the size principle, *Rational Rules* is a relatively small book that aims to explain a great deal. By developing a detailed theory of how some elements of moral cognition might rationally unfold in humans, Nichols moves many debates forward in moral psychology and moral epistemology. A complete theory of moral learning would emphasize much more than testimonial learning in childhood, including the myriad other forms of social learning and consistency reasoning. Nevertheless, Nichols provides a crucial counterweight to theories that view the foundations of moral judgment as innate or learned primarily through reinforcement. The book should be read widely, not just by ethicists interested in the origins and rationality of commonsense morality, but by philosophers and scientists concerned with moral development, moral education, and how moral knowledge might be implemented in AI.

JOSHUA MAY
University of Alabama at Birmingham