Flexibility in moral cognition: When is it okay to break the rules?

Joe Kwon, Joshua Tenenbaum, Sydney Levine
Correspondence to: smlevine@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
43 Vassar St, Cambridge, MA 02139 USA

Abstract

Rules undoubtedly guide our moral lives. Simple moral rules prohibit lying, cheating, and stealing, for instance. But the moral mind is more flexible than a theory based only on rule-adherence can account for. In this paper, we look at one particular kind of flexibility: the ability to figure out when it is okay to break a moral rule. We elicit judgments of the moral acceptability of breaking a simple rule: it's wrong to cut in line. We created a video game environment in which agents attempt to gather water – and sometimes must stand in line behind others to do so. Subjects watch clips of the game being played and make judgments of the moral acceptability of cutting in line across a wide range of spatio-temporally varied and dynamic scenarios. Our data suggests that subjects make judgments by using a generative understanding of the *underlying function of the rule* about waiting in line. We further show that our data cannot be accounted for by either 1) simple rule adherence or 2) utility maximization.

Keywords: moral judgment; moral psychology; contractualism

Introduction

How do we make moral judgments? Two main ideas dominate the moral psychology literature: outcomes and rules. On outcome-based views, people make moral judgments by considering the good or bad consequences (positive or negative utility) of an action (Greene, 2014; Cushman, 2013; Harsanyi, 1978; Crockett, 2013). Simply adding up these utilities tells you whether an action is morally acceptable or not. On rule-based views, pre-compiled rules dictate whether an action is morally acceptable (Mikhail, 2011; Nichols & Mallon, 2006). Rules are generally thought of as simple, articulable, general directives on behavior that restrict wide classes of actions ("don't lie", "don't cheat", "don't steal"). Interestingly, even theories of moral psychology that have a significant role for outcomes, also have *some* role for rules. In fact, rules appear in nearly every contemporary theory of moral psychology (see also Baumard (2016); Kleiman-Weiner, Gerstenberg, Levine, and Tenenbaum (2015)).

Why do rules play such a prominent role in our moral lives and our theories? There are many reasons that rules are useful. Rules enable coordinated action, allow for consistent social judgment, are easily communicable, act as commitment devices, ensure reliable planning, and guard against the impulse to treat yourself as an exception (Hare, 1981). One of the most important reasons – and the one we focus on here – is that moral cases can often be complex, so relying on pre-established rules can be an efficient way to come to an answer that is pretty good most of the time. Of course, this strategy isn't limited to moral judgment and decision-making; we see rules used effectively across a wide range of decision-making contexts. Relying on heuristics when time, information and cognitive processing power are tight can be a good strategy (Simon, 1955; Chater & Oaksford, 1999; Anderson, 1990; Gigerenzer & Gaissmaier, 2011). This idea of "bounded rationality", in many ways a foundational one to cognitive science, has received increasing recent attention (Lieder & Griffiths, 2020; Lewis, Howes, & Singh, 2014; Gershman, Horvitz, & Tenenbaum, 2015). In this paper, we draw on the notion of bounded rationality developed in other decision-making contexts to shed light on mysteries of the moral mind.

Moral Rules are Flexible

Despite the centrality of rules to our theories of moral cognition, our current picture of moral rules remains in many ways dissatisfying. We are left with a series of mysteries about the nature of rules and how they work. Where do they come from? What do we do when there is no rule? How do we make new ones? How do we know when the rules apply and don't apply? How do we know when it's OK to break rules? Put another way, while we tend to think of rules as rigid, rules are actually quite *flexible* – and our current theories fail to capture that flexibility. The central goal of this paper is to explain and describe the flexibility of the moral mind.

In this paper, we focus on one particular kind of flexibility – the capacity to know when a simple rule should be broken. A moment's reflection reveals that we can think of exceptions (often *many exceptions*) to the seemingly-simple rules that we all know. It's wrong to steal, but it's probably OK to duck into a cafe to "steal" a napkin if you really need one to stop a bloody nose – but we also know that you can't take the whole stack of napkins to refill your supply at home. It's wrong to trespass, but it's probably OK to rest my foot on your doorstep to tie my shoe. It's wrong to lie, but white lies are sometimes recommended. How do we know when it is OK to break the rules?

The thesis we will argue for is this: People not only know the simple versions of rules that everyone can articulate, but they also know the *generative principles* that produced those rules in the first place. One of those generative principles is an *understanding of the function* of a rule. The idea (often present in "contractualist" theories of moral philosophy, e.g. Scanlon (1998)) is that, most of the time, we figure out what

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

is morally acceptable by adhering to the simple, articulable version of a rule (e.g., "don't steal"). But other times – in cases we've never seen before, or in unusual edge cases – we consider what the purpose of the rule is. If the action under consideration instantiates the function of the rule (even if it violates the simple version of the rule) it may be permitted. Inversely, if the action seems permitted by the simple version, it may be disallowed when the function of the rule is considered.

Possessing a *functional understanding* of rules that can generate novel moral content may be a resource-rational strategy. Deploying a simple rule most of the time is fast and efficient, while having a generative understanding of the rule allows for a more resource-intensive mechanism that can be used in unusual or novel contexts to make more fine-tuned judgments.

Our Test Case: Waiting in Line

Our test cases is the rule surrounding standing in line (or: "queuing"). It seems like one simple rule governs the process of waiting in line: no cutting. But, when confronted with individual cases, it becomes apparent that we can flexibly evaluate all kinds of exceptions to the rule. For example, if you're in a deli and drop your spoon on the floor, it seems okay to cut to the front of the line to ask for a new one.

Our proposal is that participants in our experiments will make moral judgments about waiting in line by using their understanding of the *function* of the rule about waiting in line, namely, to treat each person's claim to the resource as equivalent (Adrian, Seyfried, & Sieben, 2020; Bose, 2013; Sundarapandian, 2009).[1] We operationalize this idea by asking whether the action taken by any given agent could be *universalized* – that is, if all the other agents could feel free to do the same without things going badly for everyone (Levine, Kleiman-Weiner, Schulz, Tenenbaum, & Cushman, 2020). If not, that's a sign that one agent is taking an advantage for themselves that can't also be taken by others. When everyone's claim to the resource is equivalent, this ensures a "fair" distribution.[2]

We predict that it will be judged impermissible to leave a line and head directly for the resource ("cut the line") if doing so is not universalizable. We rule out the possibility that our data can be explained by simple rule-following or utility maximization processes.

Is this even moral? There is neither scholarly (Stich, 2018) nor lay (Levine et al., 2021) consensus about what sorts of norm violations count as moral (as opposed to *conventional*). Rules about lines blur this boundary further because they are designed to ensure fairness (a topic often associated with the moral domain, Haidt and Joseph (2004)), though they are also established by individual societies and understood not to apply universally (features often associated with conventional norms, Turiel (1983)). We consider line rules to be an appropriate test case for our purposes because our broad interest is in understanding norms that help navigate the problem of *interdependent rational choice*, the struggle to achieve mutual benefit when agents have some compatible and some conflicting interests (Braithwaite, 1955; Gauthier, 1987). As shorthand, we call these rules "moral," though this theoretically-driven definition will necessarily exclude some actions that seem to some people to be moral and include some actions that seem not moral to others.

Experimental Strategy

In a series of experiments we ask participants to make moral judgments about agents who either stay in line to get a resource (specifically, a bucketful of water) or get out of line to try to get the resource more quickly. These scenarios play out in a video game environment; participants watch videos of the game being played. The game environment allowed us to create a wide range of spatio-temporally manipulated dynamic scenes, which enabled subjects to express their implicit functional understanding of line rules. Moreover, The novelty of these cases ensures that subjects' responses cannot be explained by their having seen the cases in real life and memorized the answers.

The game involves eight agents who are tasked with getting water from a water source (wells or streams). A well can only be accessed by one agent at a time, whereas a stream can often be accessed by many agents simultaneously. Each agent's goal is to get a bucketful of water and then bring the water to a set of water storage barrels. Agents get a higher reward the faster they get their bucket of water to the barrels. Each scene begins with a set of eight agents standing in a line in front of one of the water sources. Subjects are asked to make a moral judgment of a target agent who gets out of line and heads straight to the water source without waiting. The scene ends when all the agents have gotten a bucket of water and deposited it in the barrels. Then a new scene begins with a different arrangement of the game environment and agents. Two main parameters of the game are manipulated: 1) arrangement of water sources and 2) number of agents leaving the line.

Arrangement of water source(s) Some game maps have just one water source while others have multiple sources (e.g. multiple wells, a series of wells and streams, multiple stream access points, etc.). The water sources are arranged such that for some game maps, the rule about cutting in line *should apply* – waiting in line ensures that everyone's claim to the resource is treated as equivalent ("Line Necessary" cases). For instance, if a map contains just one well or one stream access point (e.g. Fig 1, panels a, c and d), then cutting in line may speed things up for an individual agent (and not actually slow anyone else down), but this action is not universalizable. If everyone tried to cut, chaos would ensue, which doesn't reli-

[1] The function might also involve ensuring a predictable, efficient, and orderly distribution of resources. We sideline these other elements of the function here.

[2] We use the word "fair" as shorthand for the concept we lay out above, but see McAuliffe, Blake, Steinbeis, and Warneken (2017) for a range of definitions.

ably benefit anyone (as demonstrated on panel d). So, when a single person cuts in line, they are prioritizing their own claim to the resource. This violates the function of the rule about cutting in line – fairness is undermined – so the action should be judged impermissible.

In contrast, there are some game maps where the rule about waiting in line *shouldn't apply*, because waiting in line actually isn't necessary to ensure that everyone's claim to the resource is equivalent ("No Line Necessary" cases, e.g. Fig 1, panels b, e, and f). For instance, if there is a stream that everyone can access simultaneously (panel b), standing in line needlessly slows down the people in the back; leaving the line and heading straight for the water is universalizable (can be done by everyone without negative consequences) and should therefore be treated as a permissible override of the cutting rule.

Number of agents getting out of line In some scenes, the target agent is the only one who gets out of line and goes directly to the water source (e.g. Fig 1, panel c). In other scenes, some (3-4), or all (8) agents leave the line and head directly towards the water source (e.g. Fig 1, panels d, e and f). This manipulation allows us to ask how the permissibility of one person's action is impacted by how many other people also decide to do it. One possibility is that there will be no impact of the number of line-leavers on permissibility (a "two wrongs don't make a right" effect). Another possibility is that as the number of line-leavers increases, leaving the line is more acceptable (possibly because continuing to stand in line is less beneficial for the individual and for the social good). Our experiment was designed to differentiate between these two broad possibilities.

Study 1

Methods

Subjects read instructions and were shown a video of one agent moving around a game map, which familiarized them to the actions available to the agents (including moving, picking up water, and depositing water into barrels). Subjects answered questions about the instructions and were excluded for wrong answers. Videos were categorized into two conditions: those where a line ensures fairness ("Line Necessary" cases, 4 maps with different arrangements of water sources) or those where a line was not necessary for fairness ("No line Necessary" cases, 3 maps). In each video, either one, some (3 or 4), or all 8 people left the line (number of "line-leavers"). Subjects saw all 21 videos in randomized order. Subjects were asked to focus on a target agent and were allowed to watch each video as many times as they wanted and then decided if the agent's action was morally acceptable or not (binary response).

Subjects. Subjects were recruited from MTURK through CloudResearch and paid for participating. 60 subjects finished the task and 2 were excluded for failing control questions. 41 reported demographic data: 24.4% female, 75.6% male. Mean age: 37.6 years, SD: 9.3, min: 23, max: 65

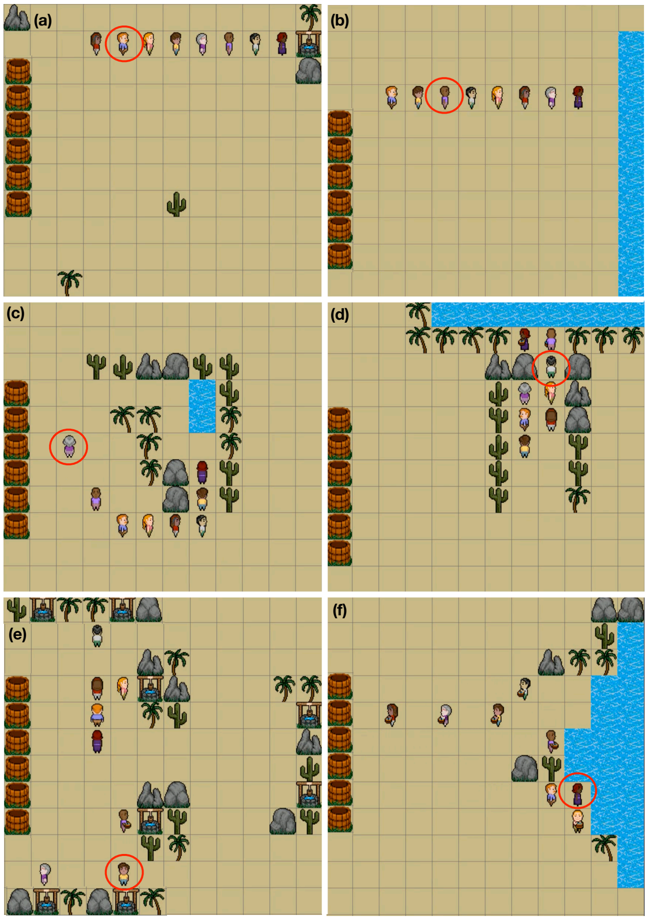

Figure 1: Examples of stimuli used in Studies 1, 2, and 3. Target agent leaving the line is circled in red. (a) Example of a Line Necessary Case. (b) Example of a No Line Necessary Case. (c) Line Necessary case with one line-leaver, who is about to enter through the "exit". (d) Line Necessary case with eight line-leavers. A blockage has formed causing substantial delays. (e) No Line Necessary case with four line-leavers. Due to the number of available wells, it is fair for everyone to leave the line and go right to an available well. (f) No Line Necessary case with eight line-leavers. There are enough access points on the stream that everyone leaving the line is more fair than everyone waiting.

Results

Collapsing across the "No Line Necessary" Cases: when there was one line-leaver, 89.1% of trials were judged morally acceptable. When there were some line-leavers, 96.6% were judged acceptable. When they were all line-leavers, 95.4% were judged acceptable. Collapsing across the "Line Necessary" Cases: when there was one line-leaver, 28% of trials were judged morally acceptable. When there were some line-leavers, 57.3% were judged acceptable. When they were all line-leavers, 72% were judged acceptable. See Fig 2 (left panel).

Fig 2 (left panel) reveals that moral judgments of agents leaving the line in “Line Necessary” cases were harsher than of agents leaving the line in“No Line Necessary” cases. It is also apparent that as the number of people leaving the line increases, acceptability of leaving the line increases. Participant response was predicted by a logistic mixed effects model that included Condition (Line Necessary/No Line Necessary), Line Leavers (One/Some/All), and their interaction as fixed effects. Subject and map were included as random effects. For each level of line-leavers (One/Some/All) the contrast between the Line Necessary and No Line Necessary Conditions was significant ($OddsRatios > 10, p < .0001$). We also compared a model with Condition as the only fixed effect to a model with Condition and Line-Leavers as fixed effects. The full model is significantly preferred ($\chi^2 = 17.306, p < 0.001$). Adding the interaction to the model was not significant ($\chi^2 = 2.99, p = 0.22$).

Ruling out simple rule-following. It is immediately apparent that subjects can’t simply be using the articulable rule “don’t cut in line” to make their judgments. In fact, there is not a single subject in our sample who judged that it was always unacceptable to leave the line. There was one subject who judged that it was always unacceptable to leave the line if the target agent was the sole line-leaver. All other subjects judged that it was at least sometimes acceptable to leave the line.

Ruling out outcome-based reasoning. Is it possible that subjects’ judgments are a simple reflection of how much delay is caused by agents getting out of line? To investigate this hypothesis, we calculated the total delay time that was created by the line leavers in each video. We calculated the amount of time (in video frames) it would have taken everyone in the scene to get water had no one left the line. We compared this baseline time to the amount of time it took for everyone to actually get water. Each video received a “delay score”, where delay = actual - baseline. Delay values for each video were entered into a logistic mixed effects model as a fixed effect and compared to a model with Delay and Condition (Line Necessary/No Line Necessary) as fixed effects. (Both models treated subject as random effects.) Even when including Delay in the model, the full model is still significantly preferred ($\chi^2 = 383.19, p < 0.0001$), indicating that delay cannot entirely account for the effect of condition on subjects’ judgments.

Discussion

The central finding of Study 1 is that subjects treat getting out of line in the Line Necessary Cases to be less morally acceptable than getting out of line in the No Line Necessary Cases. If subjects were simply applying the articulable rule “don’t cut in line” to make moral judgments, then there would be no difference in acceptability across the conditions. Instead, subjects spontaneously know when the rule about waiting in line can be broken, and their judgments follow the pattern we would expect if they are responsive to the function of the rule about waiting in line, namely, to treat everyone’s claim to the resource as equivalent.

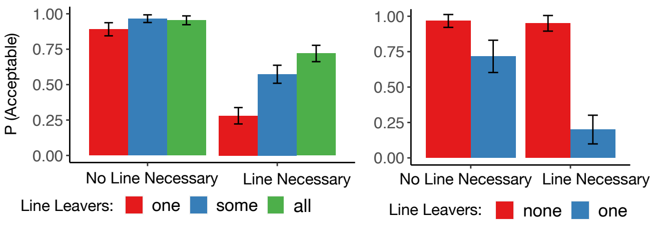

Figure 2: Study 1 (left) and Study 2 (right) results. Study 1: There is a significant effect of Condition (whether or not a line is needed to ensure fairness) on the acceptability of getting out of line. In addition, as the number of line-leavers increases, acceptability of getting out of line increases. Study 2: Even when game players “lock in” their strategies, it is judged worse to leave the line in Line Necessary Cases. This suggests that the difference in judgments across Line Necessary/No Line Necessary cases cannot be due to the worry that one person getting out of line will cause others to follow along, thereby leading to overall delays. Error bars are 95% CI.

What should we make of the fact that acceptability goes up as more people leave the line? It is possible that this is evidence that moral judgment for socially-constructed norms (like standing in line) is (partially) impacted by how strong adherence to the norm is; if fewer people follow the norm it is judged less morally problematic to break it.

On the other hand, judgments in the Line Necessary Cases when everyone is leaving the line do not reach ceiling. This indicates that there is some hesitancy on the part of subjects to say that it is acceptable to break a norm even if *literally everyone else is also doing it.* This is particularly striking in the cases of standing in line, because there actually is no way to stand in line if no one else is. One possibility is that subjects are expressing the fact that *no one* should have left the line, so even though they are making a judgment about the target agent only, the responsibility for a collective action falls on every individual contributing to it. However, waiting in line is rather unique in this respect. In many cases where everyone is doing the wrong thing, it is still meaningful (that is: conceptually coherent) to do the right thing. When everyone is engaging in activities that emit greenhouse gases, one person’s decision to follow suit may still be the wrong thing to do. Perhaps our subjects’ judgments reflect this intuition. Future work is needed to differentiate between these possibilities.

Study 2

Study 2 was designed to rule out a possible alternate explanation to our findings in Study 1. In Study 1, we argued (via calculating delay scores for each scenario) that our data cannot be explained by appeal to simple utility maximization. However, it is possible that subjects were taking into account the possibility that one person (or a few people) getting out of line and going directly for the water source could start a chain-reaction where the other people standing in line would also go directly for the water source, leading to massively more chaos than was actually observed and much lower utility. Even though, in our videos, those who leave the line do so simultaneously, it is still possible that subjects are imposing their experience of line-cutting onto these scenarios and making moral judgments based on the inference that people getting out of line often cause others to do so as well. Study 2 was designed to respond to this critique.

Methods

The experimental setup was similar to that of Study 1, except subjects were asked to imagine that all players (except one) had to "lock in" their choice of behavior before the game begins. The same explanation and instructions were given regarding the game mechanics and the goal of the players in the game. However, subjects were told that before each scene began, the players had to make independent decisions about how they would behave in the scene.

The instructions explained that the players would see what the game map looks like and where the line forms, but not which position in the line they would occupy. Then each player would decide whether they would definitely stay in line, or take opportunities to leave the line and head directly for a water source. It was made clear that the players would lock-in their choices and the game would not let them change their strategy afterwards. Uniquely, the target agent (whose actions the participants would judge), would not be required to decide their strategy ahead of time. They would be able to decide once the game starts. Subjects' judgments, therefore, could not be based on the possibility that the cutting agent might start a chain reaction.

This experiment used only two game environments: one in which the only source of water was a well (Line Necessary Condition, Fig 1, panel a) and one in which the only source of water was a stream that spanned the entire length of the board (No Line Necessary Condition, Fig 1, panel b). Either zero or one agent left the line on each trial. The total amount of time that it took for all 8 agents in the scene to get water was held constant across all four scenes. Subjects were asked to judge whether the action of the target agent was morally acceptable. The target agent is standing in line in Zero Line-Leavers trials and is leaving the line in the One Line-Leavers trials.

Subjects. Subjects were recruited from MTURK through CloudResearch and paid a small amount for participating. 61 subjects finished the task and 1 was included for failing control questions. 59 reported demographic data. 23.7% female, 76.3% male. Mean age: 37.1, SD: 9.8, min: 20, max: 70.

Results

As is apparent in Fig 2 (right panel), in both the Line Necessary and No Line Necessary Conditions it is less acceptable to leave the line than to stay in line. Importantly, the effect is significantly larger in the Line Necessary Condition (Line-Leavers 0: 95.1% acceptable, Line-Leavers 1: 19.7%) than in the No Line Necessary Condition (Line-Leavers 0: 96.7%, Line-Leavers 1: x=72.1%). Condition (Line Necessary/No Line Necessary) and number of line-leavers (0/1) were entered in a mixed-effects logistic regression with subject and stimulus as random effects. This was compared to a model with the same fixed effects as well as their interaction. As expected, the full model was preferred ($\chi^2 = 4.9, p = 0.027$).

Discussion

The main finding in Study 2 is that, even when participants are told that game players "lock in" their play strategies before the scene begins, it is less morally acceptable to cut in Line Necessary Cases compared to No Line Necessary Cases. We demonstrated this in scenes where there is no difference in the total time it takes all the agents to gather water, whether one person cuts, no one cuts, or whether it is in a Line Necessary Case or No Line Necessary Case. This suggests that subjects cannot be making moral judgments based on 1) actual delay times or 2) concerns about potential downstream consequences ("band wagon effects") of getting out of line. Instead, knowledge of the function of the rule about waiting in line better explains the pattern of judgments we report here and in Study 1.

Study 3

Study 1 looked for broad differences between two conditions (Line Necessary vs No Line Necessary) and simply asked whether the hypothesized differences in conditions were found. In this study, we directly test our proposed mechanism of rule-breaking by creating stimuli designed to elicit graded responses that our model can predict with quantitative precision. Our model predicts that subjects figure out when it is acceptable to get out of line by asking if the line is necessary to treat everyone's claim to the resource as equivalent. Or put another way, is getting out of line "universalizable"? To test this hypothesis, we created game maps where uncertainty exists around the universalizable nature of getting out of line. If we are right that universalizability is a critical step of the mechanism of making moral judgments in these cases, then there should be a direct relationship between the probability that an action is universalizable and the probability that it is morally acceptable.

Methods

The experimental setup is similar to that of Study 1, except that each scene has only one line-leaver. Moreover, videos were designed to exhibit a range of universalizability and moral acceptability and were thus designed to fall into one of three categories: Line Necessary, No Line Necessary, and Maybe Line Necessary. The Line Necessary and No Line Necessary maps were identical to the ones used in Study 1, with 6 additional maps created for Maybe Line Necessary. Each video stops a few time steps after the target agent leaves the line. Participants are thus able to see the action and intention of the target agent, but unable to see the exact outcome of the scene (thus maintaining the uncertainty about whether the

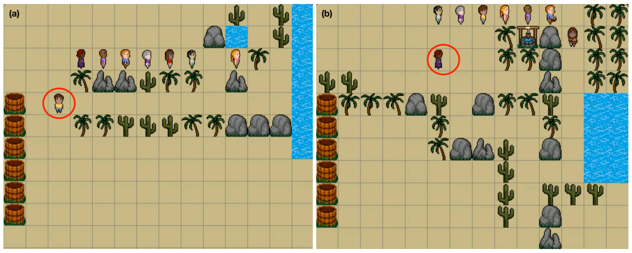

Figure 3: Example of two different kinds of uncertainty generated in the Maybe Line Necessary Cases. Panel (a): Uncertainty about whether the target agent will go down the narrow path thereby blocking others and slowing things down or will continue down towards the lowest part of the stream. Panel (b): Uncertainty about whether the target agent can get to the stream and out again before delaying others.

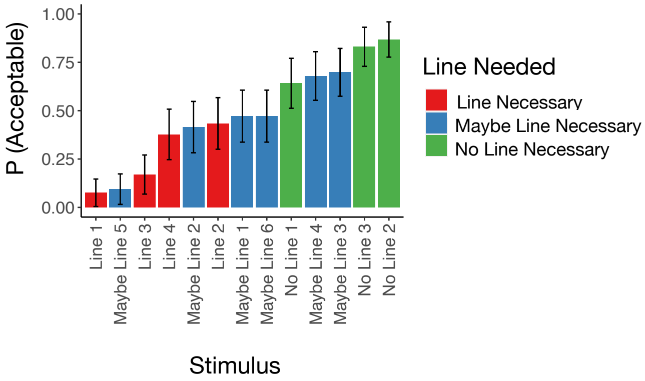

Figure 4: Study 3 results. The maps used in this experiment elicited graded responses from subjects, roughly corresponding to the hypothesized categories (Line Necessary, No Line Necessary, Maybe Line Necessary). Error bars are 95% CI.

action is universalizable or not). Subjects were asked to judge whether the action of the target agent was morally acceptable.

A second group of subjects saw the same stimuli but were asked to indicate how much better or worse off everyone would be if everyone felt at liberty to leave the line to try to get the water, rather than everyone staying in line. Subjects responded with a continuous slider scale anchored at -50 (much worse off), 0 (the same), and 50 (much better off).

Subjects. Subjects in both groups were recruited from MTURK. For the judgment task, 53 subjects finished, 6 were excluded for failing control questions. 42 reported demographic data. 40.5% female, 59.5% male. Mean age: 38.5, SD: 9.6, min: 23, max: 62. For the universalization task, 57 subjects finished, 3 were excluded for failing control questions. 48 reported demographic data. 31.3% female, 66.6% male, 2.1% other. Mean age: 37.8, SD: 8.2, min: 23, max: 56.

Results

Fig 4 shows that, as intended, the cases that we constructed have a wide range of smoothly graded permissibility judgments – from highly unacceptable to highly acceptable to many levels of permissibility in between.

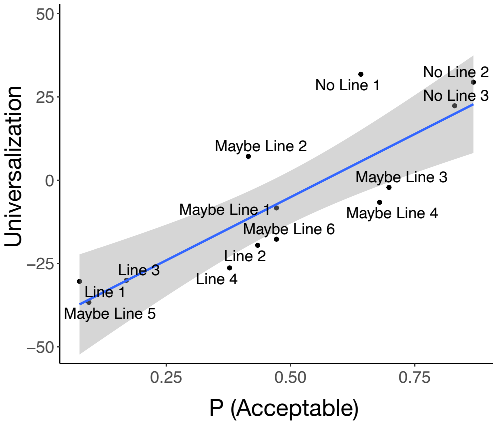

Figure 5: Study 3 results. Inferences about universalizability are strongly correlated with moral acceptability judgments across a range of line cutting cases.

Do universalization inferences capture the graded nature of these moral acceptability judgments? As an initial test, we find a strong correlation between mean universalization judgments for a case and acceptability judgments for that case (R=.84, see Fig 5). A logistic mixed-effects model predicting acceptability judgments with subject and stimulus as random effects was compared to a model with mean universalization judgments as a fixed effect. The full model is significantly preferred ($\chi^2 = 15.7 p =< .0001$).

Discussion

The main finding of Study 3 is that subject inferences about the universalizability of an action of line cutting is a strong predictor of the moral permissibility of that action in completely novel cases. This suggests that subjects use their generative understanding of the rule about waiting in line to make moral judgments, which are predictable by understanding the function of the rule.

Conclusion

If people use their understanding of rule function to know when it is permissible to break a rule, how do they figure out the function in the first place? This question is of particular interest in the case of line rules because they are entirely socially constructed. Environmental input must explain people's competence, though the function of the rule about waiting in line is rarely (if ever) discussed. Building on the ideas of moral philosophers in the contractualist ("agreement-based") tradition (e.g. Scanlon (1998); Rawls (1971)), we suggest that people can infer the function of a rule by thinking about what everyone governed by the rule could agree to – the functions that lead to mutual benefit. Future work will seek to answer this question by studying the developmental trajectory of functional rule understanding.

References

Adrian, J., Seyfried, A., & Sieben, A. (2020). Crowds in front of bottlenecks at entrances from the perspective of physics and social psychology. *Journal of the Royal Society Interface*, *17*(165), 20190871.

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Baumard, N. (2016). *The origins of fairness: How evolution explains our moral nature*. Oxford University Press.

Bose, S. K. (2013). *An introduction to queueing systems*. Springer Science & Business Media.

Braithwaite, R. B. (1955). Theory of games as a tool for the moral philosopher.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in cognitive sciences*, *3*(2), 57–65.

Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, *17*(8), 363–366.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review*, *17*(3), 273–292.

Gauthier, D. (1987). *Morals by agreement*. clarendon Press.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, *62*, 451–482.

Greene, J. D. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55–66.

Hare, R. M. (1981). *Moral thinking: Its levels, method, and point*. Oxford: Clarendon Press; New York: Oxford University Press.

Harsanyi, J. C. (1978). Bayesian decision theory and utilitarian ethics. *The American Economic Review*, *68*(2), 223–228.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In *Cogsci.*

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*.

Levine, S., Rottman, J., Davis, T., O'Neill, E., Stich, S., & Machery, E. (2021). Religious affiliation and conceptions of the moral domain. *Social Cognition*, *39*(1), 139–165.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science*, *6*(2), 279–311.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

McAuliffe, K., Blake, P. R., Steinbeis, N., & Warneken, F. (2017). The developmental foundations of human fairness. *Nature Human Behaviour*, *1*(2), 1–9.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press.

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

Rawls, J. (1971). *A theory of justice*. Harvard university press.

Scanlon, T. (1998). *What we owe to each other*. Harvard University Press.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, *69*(1), 99–118.

Stich, S. (2018). The quest for the boundaries of morality. *The Routledge handbook of moral epistemology. Taylor and Francis Group, New York*.

Sundarapandian, V. (2009). 7. queueing theory. *Probability, statistics and queueing theory. PHI Learning*.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge University Press.