Universalization Reasoning Guides Moral Judgment

Sydney Levine[1,2,*], Max Kleiman-Weiner [1,2], Laura Schulz[1],
Joshua Tenenbaum[1,†] & Fiery Cushman[2,†]

[1] *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology*
[2] *Department of Psychology, Harvard University*

Abstract

To explain why an action is wrong, we sometimes say: "What if everybody did that?" In other words, even if a single person's behavior is harmless, that behavior may be wrong if it would be harmful once universalized*. We formalize the process of universalization in a computational model, test its quantitative predictions in studies of human moral judgment, and distinguish it from alternative models. We show that adults spontaneously make moral judgments consistent with the logic of universalization, and that children show a comparable pattern of judgment as early as 4 years old. We conclude that alongside other well-characterized mechanisms of moral judgment, such as outcome-based and rule-based thinking, the logic of universalizing holds an important place in our moral minds.*

INTRODUCTION

Many people feel morally obligated to vote [1], recycle [2], and contribute to the public good in general [3]. Yet, current theories of moral psychology have trouble explaining precisely why. We know that people sometimes judge actions according to the utilitarian principle of whether they help or harm others [4, 5, 6, 7, 8]. But a single person's decision to vote in a national election, for instance, almost certainly makes no difference. Other times we judge actions according to the emotions they elicit—when judging, for instance, stabbing a person [4, 5], or french kissing your sibling [9]. Voting and recycling, however, are rarely so arousing. Other times we judge actions wrong when they violate clear social norms [10] or rules [11, 12, 13, 8]. But Americans are not required to vote, and only a minority do so consistently—if anything, the norm is to often skip.

Why, then, does anybody consider it wrong to skip voting, or to withhold contributing to similar public goods? Ask, and sooner or later you'll hear something like this: "Imagine what would happen if *everybody* did that!" This logic arises in everyday assessments of different social dilemmas. Why not pick flowers for your home from the nice bushes in the public park? Why not take all the money from the change jar at the checkout counter and buy yourself a chocolate? Why not flush a paper napkin down the toilet at work? To any of these questions a person might reasonably respond:"What if everybody did that?" Our goal is to understand what they mean, whether they really mean it, and what it means for current theories of moral psychology.

Universalization

We call this mechanism for making moral judgments "universalization": People determine whether it is morally permissible for a person to perform an action by asking what would happen if (hypothetically) everybody felt free to do the same. If things would go better, the action is permissible. If things would go worse, it is not.

Universalization differs from the dominant psychological models of moral judgment. According to models of utilitarian moral judgment, people ask "What would *actually* happen if *I* did that?", not "what would *hypothetically* happen if *everyone* did?". Neither is the process of universalization grounded in automatic emotional responses, or in existing social norms and rules.

Rather, universalization generates new rules by considering their hypothetical consequences. This involves a distinct composition of elements essential to the construction of other theories of moral psychology. Specifically, universalization respects the joint constraints of utility ("What would happen..."), impartiality ("... if *everybody*...") and rules ("followed this principle?"). It is a recurrent theme in philosophical theories ranging from Kant's categorical imperative [14] to rule utilitarian theories [15, 16, 17]. It also echoes agreement-based

[*] To whom correspondence should be addressed. E-mail: smlevine@mit.edu
[†] These authors contributed equally.

methods of collective moral decision-making, such as bargaining or negotiation [18, 19], and the associated philosophical tradition of contractualism [20, 21, 22, 23]. Like successful bargaining, universalization guides us towards impartial rules ensuring mutual benefit.

We do not propose—nor do our studies suggest—that universalization is the sole or dominant method of making moral judgments. Instead, we show that many people invoke it in one particular and important kind of social dilemma, which we call "threshold problems". In the discussion we return to consider whether it may be applied even more broadly, and how it relates to other well-established methods of moral judgment.

Threshold problems

We define threshold problems by a basic structure: If only a few people defect nobody is harmed, but when many people defect (i.e., more than the "threshold" number) everyone is harmed. Because universalizing asks what happens when everyone abides by the same principles, it renders distinctive moral judgments in these cases. For instance, consider a fishery where a new and more powerful fishing hook becomes available. If only one person uses the hook, then that person is better off, nobody else is worse off, and so overall utility increases. But if lots of people use the new hook, the fishery will collapse. Thus, if everyone feels morally at liberty to use the hook, overall utility will decrease. According to the logic of universalization, it is therefore wrong for even one person to start using the hook.

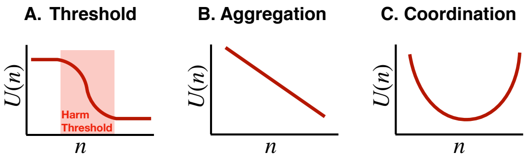

Figure 1: *Three categories of social dilemmas and their utility functions $U(n)$ as a function of the number of people doing an action. A. Threshold problems (such as over-fishing) possess a threshold structure where utility is high at small values of n, low at high values of n, and decreases exclusively within an intermediate range. B. Littering, on the other hand, is an aggregation problem, where each piece of litter adds the same negative utility to the outcome and thus there is no delimited threshold range. C. Another class of problems, coordination problems, have a high expected utility if everyone either decides to take the action or not take the action (such as driving on the left side of the road).*

Figure 1 defines the structure of a threshold problem in terms of the relationship between the number n of people who take the action and the resulting aggregate utility $U(n)$. In threshold problems there is a critical range of people acting for which $U(n)$ decreases, but outside of which it does not. Thus, utility may be unaffected (or increase) if only one person takes an action, but decrease if many people do. We call this range the *harm threshold*, because it is the threshold past which harm occurs. This contrasts with utility functions for other kinds of social dilemmas. For example, in aggregation problems $U(n)$ is strictly decreasing (perhaps linearly). An example is armed robbery: Each unique instance decreases total utility by roughly the same amount. In coordination problems $U(n)$ is maximized when everybody acts identically. Examples include driving on the right versus the left side of the road. Intuitively, universalization does not make sense in these cases: We would not explain why it is wrong to rob a person, or drive on the left side of the road, by asking, "What if everyone did that?"

Ordinary psychological theories of utilitarian moral judgment can explain why unilateral defection is impermissible in aggregation and coordination problems, because a single person taking an action makes others worse off (i.e., $U(1) < U(0)$). But they have trouble explaining why unilateral defection is impermissible in threshold problems, because a single person taking an action makes nobody worse off while bettering herself (i.e., $U(1) \geq U(0)$). For instance, if one fisherman unilaterally begins to catch more fish with a new hook, he may be able to catch more fish without changing the size of anybody else's catch. While his action does not actually reduce utility, it would hypothetically reduce total utility if universalized as an impartial rule. This is why universalization reasoning makes its most distinctive predictions in threshold problems.

The logic of universalization

Universalization involves imagining the hypothetical consequences of collective action. Two key questions immediately arise: What "action" gets universalized, and how are the "consequences" evaluated?

Although the colloquial phrase, "What if everybody did that" implies that we universalize *performing an action*, this literal interpretation leads to absurd conclusions. It would be wrong, for instance, to be a dentist. After all, what if *everyone* became a dentist? (Clean teeth, but social collapse).

We propose that people universalize not the action itself, but the sense of moral liberty or moral constraint that attaches to it [24]. If everyone felt at liberty to skip voting, for instance, the civic outcome might be bad. But if everyone felt at liberty to be a dentist (as, presumably, we do), the outcome would be just fine: Liberty or not, most people are uninterested in dentistry.

Thus, we predicted that, in threshold problems, people would make moral judgments sensitive to the number of "interested parties" n_i—all those who would perform the action if they felt morally at liberty to do so. Specifically, we predicted that people would judge an action wrong when the utility of all the interested parties acting is worse than the utility of nobody acting, i.e. $U(n_i) < U(0)$. This will happen, of course, if the number of interested parties exceeds the harm threshold. For example, suppose a fishery can sustain up to seven people using a new more effective hook. If only three are interested, than the principle "use it if you want!" can be universalized with no harm. On the other hand, if 10 fishermen are interested, then universalizing the principle would cause harm, and therefore universalization predicts that it is impermissible for any of them to use it. Pursuing this logic, we test sensitivity to the number of interested parties in Study 1, and its interaction with the critical utility threshold in Study 2.

A second key question is precisely what "utility" people are concerned with. Perhaps people are concerned with personal utility of the actor. Or, perhaps they are concerned with the social utility of all interested parties. We test these and other possibilities in Studies 3 and 4.

We are not the first to propose universalization as a mechanism for moral judgment. In fact, this was the hallmark idea of the philosopher Immanuel Kant [14], which he called the "Categorical Imperative".[3] Similar ideas have been proposed in the normative theories of R.M. Hare [16], Marcus Singer [15] and others [26]. Lawrence Kohlberg [27] suggested universalization as a psychological mechanism for making moral judgments. However, Kohlberg argued that this sophisticated form of reasoning emerged only in adults, and typically only after explicit philosophical training [28]. We explore the developmental emergence of universalization in Study 5.

Our work makes three main contributions. First, we state a formal model of universalization with sufficient precision to generate distinctive qualitative and quantitative predictions. Second, we show that many adults spontaneously use universalization to make moral judgments in threshold problems. Third, we show that universalization reasoning is present in children as young as four years old.

FORMALIZING UNIVERSALIZATION AND ITS ALTERNATIVES

We begin by defining an idealized model of universalization as applied to threshold problems, along with several alternative models of moral judgment. Although these models are quite simple, formalizing them allows us to clearly organize their competing qualitative predictions, and also to test their quantitative fit to experimental data.

Universalization

When universalizing, people ask which of two hypothetical worlds would yield greater utility: One in which $n = 0$ because everybody feels morally constrained, or one in which $n = n_i$, the number of "interested parties"—i.e., those who would perform the action if they felt morally unconstrained. Following a common approach in models of choice [29, 30], we model moral judgment as a stochastic relationship of difference in utility between these hypothetical worlds, $U(0) - U(n_i)$, as given by the logistic (or "softmax") function:

$$P_{\text{Univ}}(\text{Acceptable}) = \frac{1}{1 + e^{\tau(U(0)-U(n_i))+\beta}} \qquad (1)$$

where the "temperature" τ governs the strength of the effect of utility maximization on moral judgment and the "bias" β governs whether people err on the side of acceptability or unacceptability judgments when the relevant utilities are approximately equal (see Fig. 2, panel A). This model makes the unique prediction that moral acceptability is a function of both (1) the number of interested parties and (2) whether it exceeds the harm threshold given by the utility function $U(n)$.

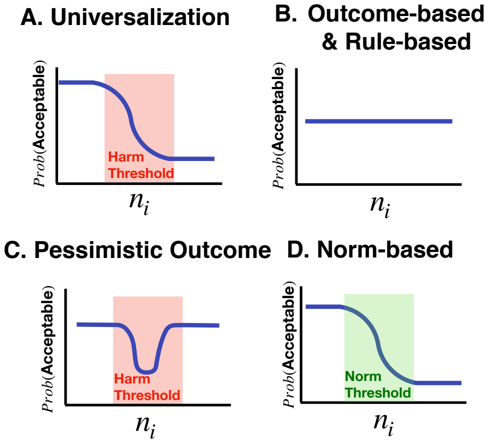

Figure 2: *Predictions of 5 different models of moral judgment. The graphs show the relationship between the number of interested parties n_i and the probability of judging one person's action morally acceptable, assuming none of the interested parties will actually act. (A) Universalization is sensitive to the number of interested parties and the harm threshold. (B) Simple outcome-based and rule-based models are not sensitive to the number of interested parties. (C) The Pessimistic Outcome model assumes that interested parties will act, and thus considers the action of one person to be least acceptable in the range of the harm threshold; here their action makes the greatest difference to outcomes. (D) A norm-based model is sensitive to the proportion of people who endorse a prohibitive norm. This may be approximated by the number of interested parties in some cases, as discussed in Study 2 but this model is not sensitive to harm thresholds.*

[3] Kant is often associated with the philosophical view deontology, but he also had a robust social contract theory with the categorical imperative at the foundation [25].

Outcome-based

Whereas the universalization model predicts that moral acceptability will be a function of the number of interested parties, n_i, standard outcome- or rule-based methods of moral judgment would not. A standard outcome-based model applied to our cases would depend on the difference in utility between the current actual number of people performing the action n_a and the utility obtained if one more person did, i.e., $U(n_a) - U(n_a + 1)$. Again we model moral judgment by applying a logistic function to this comparison:

$$P_{\text{Outcome}}(\text{Acceptable}) = \frac{1}{1 + e^{\tau(U(n_a) - U(n_a+1)) + \beta}} \tag{2}$$

where τ and β have the same interpretation as above. Although this decision rule is sensitive to utility and makes use of the utility function U, it is not directly sensitive to the number of interested parties n_i (see Fig. 2, panel B).

Pessimistic Outcome

We also consider a modified version of an outcome-based model that pessimistically assumes that anybody who is currently *interested* in performing some action will eventually *actually* do it. Thus, substituting $n_a = n_i$ into equation 2 yields

$$P_{\text{Pess. Out.}}(\text{Acceptable}) = \frac{1}{1 + e^{\tau(U(n_i) - U(n_i+1)) + \beta}}. \tag{3}$$

Like universalization, this model is sensitive to the number of interested parties and to the harm threshold. But, it predicts a very different pattern of moral judgments (Fig. 2, panel C). An action is judged wrong only when the number of interested parties falls within the range of the "harm threshold"; only here could adding one more actor plausibly influence outcomes. In other words, this model asks, "How likely am I to be the pivotal straw that breaks the camel's back?" Below the harm threshold, adding an actor carries little risk of harm; above this zone, utility is doomed anyway and additional actors make no difference. Moreover, in all the experiments we conduct below, we make it clear that interested parties will *not* actually perform the action, violating the pessimistic assumption of this model. We test whether participants accept this premise, and we exclude those who do not.

Rule-based

Standard rule-based models would be sensitive to neither the number of interested parties nor the harm threshold. Rather, it predicts acceptability judgments as a function of the presence or absence of a rule:

$$P_{\text{Rule}}(\text{Acceptable}) = \begin{cases} p & \text{if no rule} \\ 1 - p & \text{if rule} \end{cases} \tag{4}$$

where p governs the influence of rules on moral judgment (see Fig. 2, panel B).

Norm-based

A third family of models [31] proposes that people judge an action wrong by considering how many other people judge it wrong—i.e., whether there is a norm against it. Thus, it is not directly sensitive to the utility function U, but rather to the proportion of people subscribing to a prohibitive norm n_p/n:

$$P_{\text{Norm}}(\text{Acceptable}) = \frac{1}{1 + e^{\tau(n_p/n) + \theta}} \tag{5}$$

where the "temperature" τ governs the influence of descriptive norms on moral judgement and the "threshold" $0 < \theta < 1$ governs the threshold proportion of the population that must exhibit a norm in order for an agent to be more likely than not to also exhibit it (see Fig. 2, panel D). This model does not directly predict sensitivity to the number of interested parties, but in Study 2 we consider the possibility of an indirect relation where n_p is approximated by n_i. (Roughly, because interested parties who don't act might be inferred to adhere to a moral norm.)

Experimental Approach

Studies 1-3 rule out the alternative models based on the distinctive qualitative predictions each model makes. Study 1 rules out the Rule, Outcome, and Pessimistic Outcome models. Study 2 rules out the Norms model. Study 3 differentiates between two versions of the universalization model. Study 4 compares idealized utility functions with subjective utility functions to establish a quantitative fit of the Universalization model to the data. Study 5 extends these methods to children 4-11 years old.

Study 1: Sensitivity to "interested parties"

We begin by testing a distinctive and central feature of universalization: It is sensitive to the number of "interested parties". These are people who would hypothetically choose to perform an action if they felt morally at liberty (i.e., under the universalized principle).

Participants read about a threshold problem arising in a small vacation town located on a lake. Twenty vacationers currently fish in a sustainable way, but then a new fishing hook becomes available that allows each vacationer to catch many more fish. If fewer than 3 vacationers start using the hook there will be no negative consequences; if more than 7 vacationers start using the hook then there is guaranteed to be a total collapse of the fish population by summer's end. Thus, the harm threshold occurs at 3-7 interested parties. The protagonist of this vignette, John, is interested in using the new hook. Participants are asked if doing so would be morally acceptable.

Our critical manipulation is the number of interested parties, i.e., people who are actually interested in catching more fish. John knows this number because he speaks to each one of the other vacationers individually. In the Low Interest Condition, none of the other vacationers are interested in the new hook (i.e., $n_i = 0$, below the harm threshold): They feel like they have enough fish, and enjoy the traditional fishing method. Thus, moral permission to use the hook can be universalized without harm. In the High Interest Condition, all the other vacationers would be interested in using the new hook and catching more fish ($n_i = 19$, above the harm threshold), and yet each of them has personally decided against it because they are worried about sustainability. Thus, moral permission to use the hook would be harmful if universalized. For this reason, universalization predicts that participants will judge John's action less acceptable in the High Interest condition than in the Low Interest condition.

Crucially, however, in both versions of the vignette John knows that none of the other vacationers will use the new hook. He can therefore safely start using the hook without any actual negative consequences, because one person using the hook is below the harm threshold. An outcome-based model therefore predicts that John's action should be acceptable. John also knows that there is no "rule" in his community against using the hook. We checked whether participants agreed with these premises about rules and outcomes; nearly all did, and our results hold whether or not we include the remainder.

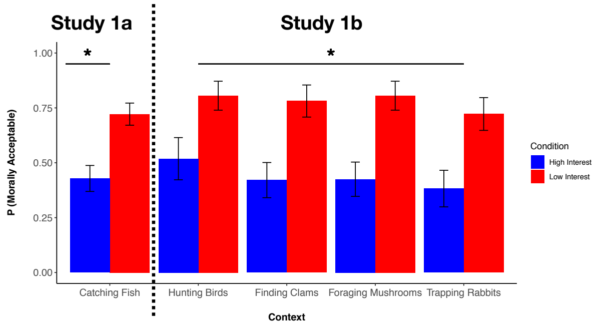

Figure 3: *Study 1 results. There is a significant difference between the High Interest and Low Interest Conditions, as predicted if participants use universalization to make moral judgments in this case. Error bars are standard error of the mean. * indicates* $p < 0.001$

As predicted, subjects judged John's action to be significantly more morally acceptable in the Low Interest Condition than in the High Interest Condition (Low Interest: 72.1%, High Interest: 42.9%, $\chi^2(1) = 13.11, p < .001$, two-tailed, $V_{Cramer} = .30, CI_{95\%}[.16, .46], n = 149$, see Fig. 3.)

Manipulation checks confirmed that participants understood the scenarios as we intended. All but one subject responded that there was no rule forbidding John from fishing. On the outcome measure, most subjects (87% in the High Interest condition; 81% in the Low Interest condition) reported that John's action wouldn't make a difference to the fish population. The remaining subjects said that John's action would decrease the health of the fish population (13% in the High Interest condition and 19% in the Low Interest condition) and no one said that John's action would increase the health of the fish population. There was no significant difference across the conditions on answers to the outcome measure ($\chi^2(1) = 1.03, p = .310$, two-tailed, $V_{Cramer} = .08, CI_{95\%}[-.05, .20], n = 149$). Moreover, our main results hold even if we exclude participants who said that John's action would make a difference to the fish population; see SI for details.

To ensure that our finding was robust, we replicated Study 1 using four additional contexts. Rather than fishing in a lake, the contexts involved stories where a group of people were foraging for mushrooms, hunting birds, trapping rabbits, or gathering clams. As Figure 3 shows we found similar results across all scenarios; see SI for details.

In summary, we find that more people consider John's action morally acceptable when other relevant parties are disinterested. This pattern of judgment is predicted by the logic of universalization, but not by standard theories of outcome- or rule-based moral judgment.

Whether it can be explained by a norm-based model is more ambiguous. We attempted to write our scenarios so that there was no overt expression of moral norms correlated with the number of interested parties. Nevertheless, in the context of our experiments, "interested parties" who say that they would *like* to perform an action but have *chosen not to* (perhaps because they consider it morally wrong) may establish a relevant descriptive norm. Thus, a norm-based mechanism of moral judgment could explain the sensitivity to the number of interested parties, but without any role for universalization reasoning. Study 2 was designed specifically to provide a strong test of the norm versus universalization models.

Finally, while our data are consistent with some participants employing the logic of universalizing, they clearly show that not all participants do so. For instance, even in the "high interest" case, 43% of participants judged that it was permissible for John to act. The judgments of these participants are most consistent with an outcome- or rule-based mechanism of moral judgment. A straightforward estimate of the proportion people who employed universalization in this case, based on our sample, is the difference in proportions between conditions: about 29% of participants.

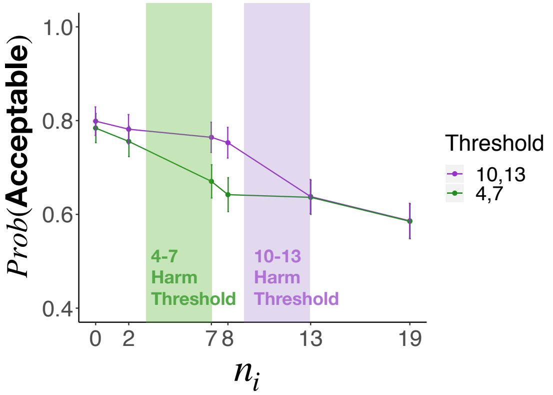

Figure 4: *Moral acceptability judgments for Study 2. The probability of a subject judging the action morally acceptable as a function of the number of people interested in using the new hook. The location of the threshold (indicated by shaded areas) impacts moral permissibility judgments, suggesting that the way that the utility aggregates impacts moral permissibility. This distinguishes universalization from a norm-based account of moral judgment.*

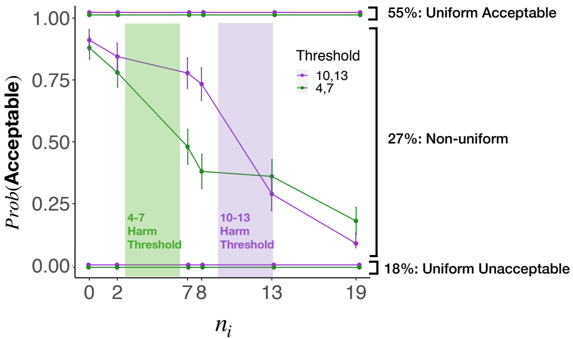

Figure 5: *Moral acceptability judgments for Study 2, broken down into three response patterns: subjects who uniformly considered the action acceptable, those who uniformly considered the action unacceptable, and those with non-uniform response patterns. This latter group exhibits the pattern predicted by universalization.*

Study 2: Sensitivity to the harm threshold

Universalization reasoning is sensitive not just to the number of interested parties, but to whether that number exceeds the "harm threshold," the number of actors sufficient to decrease utility. In the context of our fisheries case, the question is whether the number of interested parties n_i is sufficient to cause a collapse of the fish population if everyone felt morally at liberty to use the new, more powerful hook. In other words, n_i matters because it affects the output of the utility function $U(n_i)$.

This feature is not shared by the norm-based model. It is only sensitive to the proportion of people who endorse a moral prohibition on action, n_p/n. In our studies, people may approximate this by assuming that interested-non-actors endorse a moral prohibition: $n_p/n \approx n_i/n$. Crucially, however, this model has no dependency on the structure of the utility function $U(n)$.

We therefore designed Study 2 to test whether moral acceptability judgments are sensitive to the structure of the utility function. Specifically, we modified the design of Study 1 to vary both the number of interested parties $n_i \in [0, 2, 7, 8, 13, 19]$ and the harm threshold—i.e., the value of n at which the fish population will go extinct. In the "low threshold" condition this threshold occurred between 4 (harm) - 7 (collapse), while in the "high threshold" condition it occurred between 10 (harm) - 14 (collapse).

In both cases, subjects are asked to make a moral judgment about a scenario where just one person decides to use the new hook, given that a certain number of other people are interested in using the new hook. As in Study 1, it was made clear that this is the only person who is going to use the hook (and control questions verified that subjects agreed with this premise). This design allows us to test whether moral acceptability is a function both of the number of interested parties and the form of the utility function that depends on it.

Our results show a clear effect of the harm threshold on moral judgment (Figure 4). We analyzed moral acceptability judgments with a linear mixed effects model including fixed effects for n_i, condition (low threshold vs. high threshold) and their interaction. We included participants as a random effect and specified maximal random slopes following [32]. We compared this with a model excluding both condition and its interaction with n_i. The full model is significantly preferred ($\chi^2 = 8.64, p = .013$).

As a further test of the possibility that people infer moral rules from the behavior of interested-non-actors, we asked subjects whether any rule prohibited using the new hooks. We analyzed their rule judgments with a logistic regression with n_i as the predictor. Although there was a marginal effect of the number of interested parties on rule perception ($z = 1.798; p = 0.072$), when rule judgments and n_i were entered into a logistic regression to predict moral judgments, rule judgments were not a significant predictor of moral judgments ($z = 0.037; p = 0.97$), nor was their interaction with the number of interested parties ($z = -0.038; p = 0.97$).

Individual Differences

We next used the rich within-subjects design from Study 2 to estimate the proportion of participants in our studies whose data conform to different models of moral judgment. Outcome- and rule-based models of moral judgments make the distinctive prediction that participants will render a uniform pattern of judgment across all six values of n_i that we tested in Study 2 (see Fig. 2, Panel B). Consistent with these predictions, we found that 55% of participants judged John's action uniformly morally acceptable, while 18% judged it uniformly unacceptable. Figure 5 plots the moral acceptability judgments of the remaining 27% of participants whose judgments were non-uniform. These participants clearly exhibit the pattern associated with universalization, with the predominant decrease in moral acceptability occurring at the harm threshold specific to each condition. We provide a more precise quantitative fit of our model to these participants' judgments in Study 4. We elaborate on the implications of these findings for our theory in the Discussion.

Study 3: Whose utilities matter?

Universalization asks what the outcome would be if everyone felt at liberty to act in a certain way. But whose utilities count?

One possibility is that people are concerned with the welfare of everyone involved—in our fisheries case, for instance, all the fishermen. This wide scope of concern is a hallmark of utilitarian moral theories (including rule consequentialism, e.g. [16]), which tend to be concerned with impartial maximization of aggregate utility. A wide scope is also predicted by contractualist theories that project the outcome of ideal bargaining and negotiation. On such theories the welfare of everyone involved will influence the bargain (not through simple maximization of aggregate utility, but through taking everyone's perspective into account, e.g. [22, 21, 20]).

Alternatively, people may be narrowly concerned with their own personal utility. For instance, if the fish population collapsed, John would personally be worse off. A variety of this possibility, closely associated with Kant [14], is that people reject actions that undermine their own purpose once universalized. For instance, John wants to use the new, powerful fishing hook as a means of catching extra fish, but if the fish population disappears (as would happen if everyone used the new hook), using the new hook will no longer enable John to achieve his goal.

Thus, in Study 3 we designed a case that sharply dissociates John's utility from the other fishermen's. We modified the case used in Study 1 so that John operates a motorized tour boat and can increase his profits by using a new kind of motor oil. Using this motor oil can also increase the profits of the fishermen. John's utility U_j is unaffected by how many *other* people (i.e., fishermen) use the motor oil. Thus, $U_j(20) > U_j(0)$. By contrast, the fisherman's utility U_f decreases if too many people use the new motor oil because it will destroy the fish population. Thus, $U_f(20) < U_f(0)$.

If participants are concerned with everyone's utility, they should judge this new "Tour Boat" case identically to the original case used in Studies 1 and 2. Although John would not personally be harmed if everybody started using the new motor oil, the other fishermen would. On the other hand, if participants are narrowly concerned with John's utility only, they should judge this new tour boat case quite differently. Since John suffers no harm from adoption of the new motor oil, it does not matter whether there is high or low interest among fishermen in adopting it. Thus, they should judge John's action permissible whether there is high or low interest among fishermen. (Similarly, by Kant's measure, John's act can be universalized without undermining its own purpose.)

We asked participants to make moral judgments of the high- and low-interest conditions of the original "Fisherman" case and the modified "Tour Boat" case in a 2 × 2 between-subjects design. As Figure 6 (top panel) shows, we observed a large effect of condition (high interest vs. low interest) in both cases. We analyzed this data with a logistic regression, which shows a significant effect of condition on subjects' judgments ($p < .0001$), a significant main effect of context ($p < .001$), but no condition-by-case interaction ($p = .59$). These results suggest that people did not attend selectively to John's utility, but instead broadly to the utility of both John and the fishermen.

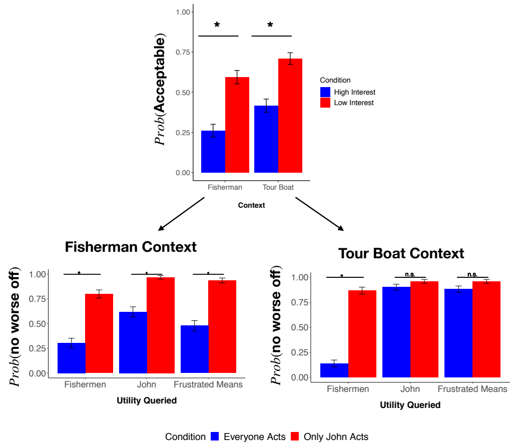

Figure 6: *Top panel shows moral acceptability judgments for the Fisherman and Tour Boat contexts. Each context exhibits a similar pattern of moral acceptability: the action is more permissible in the Low Interest Condition compared to the High Interest Condition. The Bottom panel shows utility measures for each context. In the Fisherman context, all three measures (Everyone's Utility, John's Utility, and Frustrated Means) show a similar pattern. This context therefore does not differentiate between the measures. In the Tour Boat Context, the measures exhibit very different patterns. In this context, it becomes clear that considering everyone's utility explains the moral acceptability data better than considering John's utility in isolation or whether John's action is a means to his end.*

In order to confirm that participants perceived the utility functions for John and for the fishermen as we intended, a separate group read the same scenarios but provided ratings on the utility consequences for various parties. Specifically, for each case they rated change in utility for John $U_j(20) - U_j(0)$ (subjects responded "better off", "worse off" or "the same"), the change in utility for the rest of the fishermen $U_f(20) - U_f(0)$ (again, as "better off", "worse off", or "the same"), and the likelihood that John would be able to bring about his goal ("more likely than before", "less likely than before", "the same as before"; testing Kant's concept of a frustrated means). While subjects chose between three options, in Figure 6 we report the more illustrative metric of the percentage of subjects that reported that the relevant parties would be no worse off (i.e., $U(20) - U(0) > 0$; we determine this by collapsing the answers of "better off" and "the same"). See SI for expanded data.

As expected, in the context where John is a fisherman, there are significant and large difference across the conditions both when subjects are asked about the utility of the fishermen (Everyone Acts: M=.30, SE=.048; John Acts: M=.80, SE= .041; $\chi^2(1)$ Yates' Corrected = 44.5, $\phi = .49$, $p < .0001$), John's utility (Everyone Acts: M =.48, SE=.052; John Acts: M=.94, SE=.025; $\chi^2(1)$ Yates' Corrected = 45.7, $\phi = .49$, $p < .0001$), and whether John's action would no longer be a means to his goal (Everyone Acts: M =.62, SE=.051; John Acts: M=.97, SE=.018; $\chi^2(1)$ Yates' Corrected = 33.0, $\phi = .42$, $p < .0001$). Specifically, when everyone uses the new hook, most subjects judged that John and the fishermen would be worse off. In contrast, when only John uses the new hook, most participants judged that John and the fishermen would be no worse off 6. In contrast, in the tour boat context, when we ask subjects about John's utility, the difference between the conditions is only marginally significant with a small effect size (Everyone Acts: M=.88, SE=.032; Only John Acts: M=.96, SE=.020; $\chi^2(1)$ Yates' Corrected = 3.11, $p = .078$; $\phi = .12$). We see a similar pattern of findings when we ask whether John's action would no longer be a means to his goal (Everyone Acts: M=.90, SE=.030; Only John Acts: M=.96, SE=.020; $\chi^2(1)$ Yates' Corrected = 1.76, $p = .18$; $\phi = .094$). In contrast, when subjects are asked about everyone's utility, the difference between conditions is large and significant (Everyone Acts: M=.14, SE=.034; Only John Acts: M=.87, SE=.034;$\chi^2(1)$ Yates' Corrected = 102.8, $p < .0001$, $\phi = .72$). Notably, the 8% of subjects who felt that John would be worse off if everyone adopted the new oil is not sufficient to explain the 29% of subjects that treat the cases differently when asked about moral acceptability judgments, because the upper bound on the difference in the conditions for John's EU (95% CI upper bound = .13) is smaller than the lower bound on the difference across the conditions for for moral judgment (95% CI lower bound = .216; High Interest: M=.417, SE=.041, Low Interest: M=.710, SE=.036). This same logic extends to the 6% difference in the conditions when subjects are asked about John's action being a means to his goal.

In summary, when applying universalization reasoning, people consider a wide scope of utilities. In our cases, they are concerned not just with the utility of John, or with Kant's conception of self-undermining action, but with the utility of all those who use the lake. In the discussion we return to consider what this implies about the nature and function of universalization.

STUDY 4: EMPIRICAL UTILITY FUNCTIONS

So far we have assumed that participants represent the utility function U precisely as it is described in our scenarios (Figure 1a). On this assumption, utility is flat[4] until the harm threshold, drops precipitously through the threshold range, and then is low and flat beyond it. Presumably, however, participants' subjective impressions of the utility function deviate somewhat from this idealization. For example, if subjects know that utility will precipitously fall at n = 4, they may be uncertain about whether it actually remains constant between n = 3 and 4, and thus impute a negative slope to the utility function before the threshold. In Study 4, we empirically estimate the utility function imputed by participants. Our experiment focuses on the regions before and after the threshold range, about which our stimuli were most explicit. We then ask whether substituting this empirical estimate for the idealized function improves the fit of our universalizing model (as it should if participants are universalizing with respect to their subjective utility functions) [33].

Measuring the empirical utility function

We asked participants to read our stimuli and describe the utility outcomes at various settings of n. Specifically, participants read the stories from Study 2, either for high-threshold or low-threshold conditions in a between-subjects design. We asked them how much better or worse things would go for all the fishermen if various numbers of people ($n \in [1, 3, 8, 9, 14, 20]$) started to use the new fishing hook, as compared to the status quo in which nobody is using it (i.e., $U(n) - U(0)$). They responded on a scale from -50 ("a lot worse off") to 50 ("a lot better off"), where 0 indicates no change.[5]

As Figure 7 shows, participants generally felt that things would go slightly better if a below-threshold number of

[4] Or perhaps slightly increasing, given that the people who act are slightly better off while nobody else is harmed.

[5] This experiment also contained two additional conditions (John's Expected Utility and Frustrated Means), which queried the change in utility in different ways. These ways of measuring utility produce results very similar to the ones we report here, so the results from the remaining two conditions are reported in the SI.

people began to use the hook. This makes sense: Things should go better for that minority of hook-users, and go no worse for the non-users. And, generally, participants felt that things would go much worse if an above-threshold number of people began to use the hook. This also makes sense: Above the threshold the fish population collapses and things go worse for everyone. Put another way, the empirical utility function seems to represent "the average effect on those affected" rather than, for instance, the total or average utility for all the fishermen (which would increase before the threshold rather than staying flat or dropping). Notably, however, the empirical utility function does not conform precisely to any idealized utility function, insofar as it shows slightly negative slopes both above and below the critical threshold.

Applying the empirical utility function

We next assessed whether the universalization model achieves a superior fit to moral acceptability data when we substitute the empirical utility function obtained above in place of the idealized utility function assumed so far. In order to do this, we fit participants' moral judgment data from Study 2 to our formal model of universalization (i.e., Equation 1) by optimizing the values of the parameters τ and β. We only included data from the 27% of participants who gave a non-uniform pattern of judgments across values of the number of interested parties n_i (see Figure 8), since these are the only subjects who show evidence of applying universalization to these cases. In order to generate an empirically-estimated value of $U(n_i)$ we substituted the mean of the data described in the preceding section (i.e., the values plotted in Figure 7). In order to generate an idealized value of $U(n_i)$ we set $U(n) = 0$ for all n_i less than the threshold and $U(n) = 1$ for all n_i at or above the threshold. The critical feature of this idealized model is that the values before and after the threshold are constant (not, for instance, that they be symmetric around a utility mid-point; see SI for further details).

Figure 8 shows the predicted moral acceptability judgments of each model overlaid on the data obtained from Study 2. Overall we find a strong correspondence between the model predictions and the data. It is especially striking that we find a strong correspondence between the model and the data across our two different harm threshold conditions, given that we fit identical parameter values to both. In other words, the difference between these conditions is not explicit in the model, but rather derives exclusively from the shape of the utility functions: participants perceptions of the $U(n_i)$ for different harm thresholds. We also find that the model fit when using empirical utility functions ($AIC = 609$) is substantially better than the model fit when using an idealized utility function ($AIC = 621$). This finding adds quantitative support to our argument that subjects use universalization as a method of moral decision-making by showing that subjects' own subjective understanding of the utility functions predict their moral judgments.

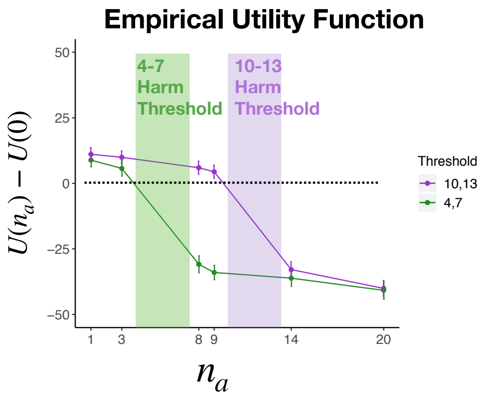

Figure 7: *Subjective utility functions produced by the subjects in Study 4 for both the 4,7 and 10,13 threshold conditions. These utility functions deviate from our idealized function in that utility slopes downward rather than being completely flat before and after the threshold regions.*

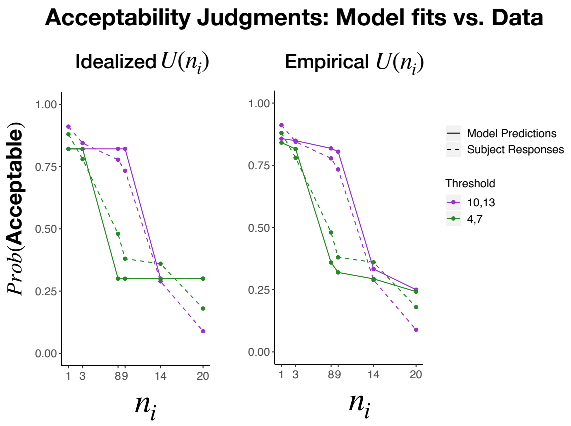

Figure 8: *Moral judgments predicted by the universalization model (solid lines) overlaid against the observed moral acceptability data (dotted lines) for those participants who exhibit not-uniform patterns of judgment. Left panel shows the model based on the ideal utility curve. Right panel shows the model based on the empirical utility curve.*

Having obtained an empirical utility function that differs slightly from our idealized assumed utility function, it is important to ask whether it can improve the predictive power of any of our alternative models of moral judgment. It obviously will not improve the predictive power of rule- or norm-based models, since these do not involve the utility function at all. But what about outcome-based models, which do?

Applying the outcome-based model described in Equation 2 to the threshold problems we are investigating here, the key question is how the utility of one person acting compares with the utility of nobody acting, i.e. whether $U(1) \geq U(0)$. The slope of our idealized utility function is flat in this range and so, according to Equation 2, it is permissible for John to begin using the hook. But, our empirically-derived utility function shows that some participants (14.6%) actually believe that one person using the hook will result in a negative utility. According to our outcome-based model, such participants would therefore be predicted to judge it impermissible for John to start using the hook. This is consistent with the observation in Study 2 that 18% of participants judged John's action impermissible uniformly, for any number of interested parties—a pattern of judgment inconsistent with universalization, but consistent with an outcome-based model applied to the empirical utility function. (Of course, this outcome-based model is insensitive to the number of interested parties, and thus cannot explain the substantial proportion of participants whose moral acceptability judgments depend on it.)

Study 5: Universalization across development

Kohlberg famously described a series of six stages that we naturally progress through as our ability to reason about morally-charged actions develops [27]. On his view, the 6th and highest stage of moral development is the ability to provide moral justifications that appeal to the abstract concept of universalization. Kohlberg [28] argued that "Without formal moral theory men naturally attain to a'stage 5' [the penultimate moral stage]." While we agree that universalization receives its fullest treatment in formal philosophical theories, in this study, we investigate whether the seeds of this sophisticated reasoning process are present in our minds from a very young age.

Kohlberg diagnosed his subjects' moral stage by classifying the justifications they provided for their moral judgments—their ability to explicitly introspect on their moral cognition—rather than by attending to patterns of judgments themselves. Since Kohlberg, a rich body of work has shown that young children and even pre-verbal infants exhibit a sophisticated moral sense that is not necessarily dependent on their ability to linguistically communicate their reasoning [34, 35, 36, 37]. Yet, while we have learned much about the development of moral cognition by studying the judgment patterns of young children, we know far less about whether children use universalization to make moral judgments.

We presented children aged 4-11, and a small group of adults, with two stories that were similar in structure to the fisheries cases used in Studies 1-4. In one story, for instance, there is a path made of rocks that the kids in the story like to walk on to get home. Jimmy has a rock collection and would like to take one of the rocks for his collection. In the Low Interest Condition, Jimmy is the only one who wants to take a rock and therefore moral licence to take a rock can be universalized without harm. In the High Interest condition, all the kids have rock collections and would like to take a rock. However, they have all decided not to take the rocks because they want to path to stay intact. In this case, universalizing permission to take the rocks would cause the path to disappear.

In these stories, the actor is completely alone when he acts to ameliorate the worry that our subjects might imagine that other kids will see Jimmy act and thereby start a chain reaction of everyone acting this way. We also make it clear that no other kids are going to do the action, confirming comprehension with a series of control questions (see SI for details).

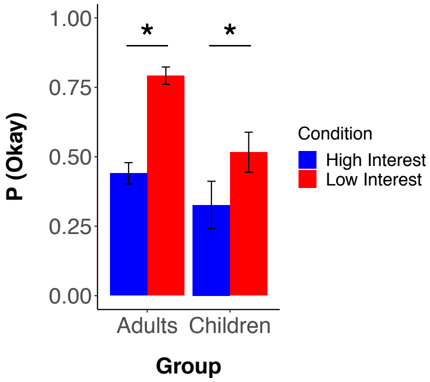

Figure 9: *Probability children and adults judging the story "Okay" in Study 5. There is a significant difference between the High Interest and Low Interest cases for both adults and children. Error bars are standard error of the mean. * indicates p<.01.*

Results

The results from the adult sample replicate the findings in Study 1: the High Interest condition is judged to be significantly more permissible than the Low Interest condition (($\chi^2(1) = 42.3, p =< .0001$). See Fig. 9, left panel.

Each subject received two stories in counterbalanced order, one Low Interest and one High Interest. Five children opted to hear only one story. For those who answered both, 143 (79.4%) subjects gave the same answer to both stories. 74 subjects (41.1%) said not OK to both stories and 69 subjects (38.3%) said OK to both stories. Of those who gave different answers to the two stories, 26 (14.4%) gave the expected pattern of answers (permissible in Low Interest, impermissible in High Interest) and only 11 (6.1%) gave the opposite pattern (McNemar's test of changes $p = .02$). Because so few subjects switched their answers, however, the remainder of the analysis was conducted on answers to the first story only.

The central finding (see Fig. 9) is that children were significantly more likely to judge the Low Interest case permissible compared to the High Interest case ($\chi^2(1) = 6.85, p = .009$, two-tailed, $V_{Cramer} = .19, CI_{95\%}[.06, .34], n = 185$, BF=5.58 in favor of H_1. Moreover, this pattern holds for the entire age range we tested (4-11 years old). To assess potential developmental trends, we compared three models of the data: one including condition only, one including the main effect of age, and one including an age × condition. In the latter two models, there is no significant effect of age or the age × condition interaction and the data is best explained by the model that includes only condition (see SI for details).

Discussion

Across five studies, we show that both adults and children sometimes make moral judgments well described by the logic of universalization, and not by standard outcome, rule or norm-based models of moral judgment. We model participants' judgment of the moral acceptability of an action as proportional to the change in expected utility in the hypothetical world where all interested parties feel free to do the action. This model accounts for the ways in which moral judgment is sensitive to the number of parties hypothetically interested in an action, the threshold at which harmful outcomes occur, and their interaction. By incorporating data on participants' subjectively perceived utility functions we can predict their moral judgments of threshold problems with quantitative precision, further validating our proposed computational model. These data suggest a new and intriguing correspondence between ordinary people's moral judgments and common features of diverse philosophical theories which all draw on versions of universalization [14, 16, 15, 20, 24].

Our findings contribute to current theories of moral psychology in several ways. First, many current theories emphasize the role of automatic, intuitive processes [4, 38, 39], often to the exclusion of reasoning [9]. Our results suggest a highly structured and complex cognitive architecture that contributes to moral judgment. Second, a dominant view in moral development holds that young children are incapable of universalization reasoning [28]. In contrast, we find its fingerprints in the patterns of judgments by children as young as four. Third, many current theories of moral psychology focus on the twin contributions of outcome- and rule-based moral judgment [4, 11, 12]. We offer universalization as a key element of the moral mind that composes elements of outcome- and rule-based theories in a distinctive way.

Our work raises several important questions about the cognitive mechanisms supporting universalization. The philosophical literature includes several related, yet distinct, versions of universalization, and each of these corresponds to a viable cognitive model. For instance, people might ask, "Which rule would be best for us all?" (more akin to rule utilitarian theories, e.g. [16]), or instead "Which rule would everyone agree to?" (more akin to contractualist theories, e.g. [22]), and so on. Distinguishing these possibilities is an important direction for future research.

Another important question is how people infer the relevant general principle to universalize from the observation of a single person's single act. For instance, if on July 17th John uses the new fishing hook, is the rule to be universalized "Everybody may use the hook", or "Everybody named John may use the hook", or "Everybody may use the hook on July 17th", etc.? This echoes a general problem in the sciences of *program induction*: Given the output of a policy or rule, how can we infer its generative content [40]?

Other important questions concern the scope of universalization: how widely is it employed?. Although many participants in our studies employed universalization when faced with a threshold problem, many used different strategies for moral judgment in these cases. These other participants may instead have relied on norm-, outcome- or rule-based strategies, or others yet to be described. If the choice of strategy is a stable individual difference, what demographic or cultural factors can predict it? Alternately, it is possible that a single subject may sometimes use universalization and sometimes a different strategy to make moral judgments for threshold problems? If this is the case, subjects are then faced with a "strategy selection problem" [41]. How do they know which strategy for moral judgment to use when? Can we manipulate the cases to encourage participants to adopt one strategy over another? There are some cases where it seems quite clear that universalization is the wrong strategy to apply. It would be strange to say that it is wrong to punch a person because "What if everybody did that?" Likewise with betraying a friend, french kissing a sibling, or compensating an employee unfairly. Following most contemporary theories of moral psychology, we assume that human moral judgments are generated by multiple complementary mechanisms. We have demonstrated a key place for universalization among this set.

Finally, what is the relationship between universalization and these other mechanisms of moral judgment? They might be entirely independent. In contrast to this possibility, however, we suggest that universalization is intimately related to both outcome- and rule-based mechanisms. Specifically, universalizing identifies rules that would be good candidates for everyone to agree on: When impartially applied, they tend to improve utility. It is, in short, meta-moral rule: it endorses rules that bring about good outcomes.

In this respect, universalization is a cognitive mechanism that achieves outcomes similar to social processes that generate moral norms—ones such as negotiation, bargaining, and cultural or biological evolution. For instance, when negotiating, people will typically agree upon arrangements that are both fair and ensure mutual benefit [18, 42, 23, 22, 43]. As we show in Study 3, universalization is sensitive to the welfare outcomes for all interested parties, and not just the welfare of the actor. This may reflect a simple aggregation of social welfare (as in utilitarian theories), but it might instead reflect a more complex bargaining solution (as in contractualist theories). In other words, when universalizing, people may simulate a virtual bargaining process [19, 44] to determine which moral liberties and constraints everybody would agree to. Our study cannot speak directly to this possibility, but it stands out as an important direction for future research. Similarly, the logic of universalization mirrors some key concepts in evolutionary game theory. In general, biologically or culturally evolved moral norms should be stable equilibria, in the sense that nobody can improve their position by unilateral defection [45]. But when there are several such viable equilibria, which should we expect to observe? All else being equal, evolutionary dynamics favor those yielding greater aggregate payoffs [46]. Thus, biologically or culturally evolved moral norms tend towards payoff-maximizing rules that everyone would agree to [47]. In sum, universalization can generate moral norms similar to social processes such as negotiation and bargaining, or cultural and biological evolution. Unlike these processes, however, universalization can be quickly and efficiently implemented within a single person's mind.

As these connections illustrate, when we judge an action by universalizing we receive not just a judgment of the action, but also a new rule. This suggests that universalization might be a strategy deployed in cases where no clear moral rule exists. Once the rule is derived, it may be re-used in similar cases [48] to enable more computationally efficient moral decision-making. For instance, having decided that John shouldn't use his more powerful fishing hooks in this instance, we have established the rule "don't use the hooks" to govern the whole lake population across time. Rules established by universalization reasoning may be reused across an individual's life and transmitted across generations as a cultural inheritance. In this manner, the logic of universalization holds the power to transform individuals' outcome-based preferences into impartial and binding moral rules on the community.

Methods

Data and analysis for all studies are available at github.com/sydneylevine/universalization.

Study 1

This study was preregistered[6]. 202 subjects participated in this study, recruited from Amazon MTURK through turkprime and were paid a small amount for their participation. (We intended to recruit 200 subjects as we indicated in our preregistration, but 2 additional subjects participated due to an error in the turkprime recruitment service.) 53 subjects were excluded for failing control questions.

Study 1a was a replication of Study 1 using multiple story contexts. This study was preregistered[7]. The dependent variables, exclusion criteria, and study design were the same as in Study 1. The only difference was the context of the story. Rather than fishing in a lake, the contexts involved stories where a group of people were foraging for mushrooms, hunting birds, trapping rabbits, or gathering clams. See Supplemental Materials for full story text. 400 subjects completed the experiment. 121 subjects were excluded for failing control questions.

Study 2

This study was preregistered[8]. 700 subjects participated in this study, recruited from Amazon MTURK through turkprime and were paid a small amount for their participation. 350 subjects were excluded for failing control questions.

Subjects were randomly assigned to 1 of 2 conditions. 4,7 Condition: Up to 4 people can use the new hook with no effect on the fish population; once 7 people use the new hook the fish population will go extinct. 10,13 Condition: Up to 10 people can use the new hook with no effect on the fish population; once 13 people use the new hook the fish population will go extinct.

Each subject was told that N people are interested in using the new hook. Subjects answered a series of questions about the story. Subjects then read the same story, the only change being that a new value of N was given. N was chosen at random without replacement from the following values until all values of N were been seen by each subject: 0,2,7,8,13,19. See SI for further details.

Study 3

1242 subjects participated in this study, recruited from Amazon MTURK through turkprime and were paid a small amount for their participation. Subjects were divided into two groups: the Moral Judgment Group (n=840) and the Expected Utility Group (n=402; 2 additional subjects were accidentally allowed to take the experiment after our 400 subject cap). 284 subjects were excluded from the study for failing one or more control questions in the Moral Judgment Group. 16 subjects were excluded in the Expected Utility Group. Subjects in both groups were randomly assigned to one context (Fisherman or

[6] See: http://aspredicted.org/blind.php?x=2qb4nc)

[7] See: https://aspredicted.org/blind.php?x=ag83iq

[8] See: https://aspredicted.org/blind.php?x=c44jr2

[9] See: http://aspredicted.org/blind.php?x=at7cs8

Tour Boat) and one condition (High Interest or Low Interest). Subjects in the Moral Judgment Group answered different questions about the scenarios than did subjects in the Expected Utility Group. See SI for further details.

Study 4

This study was pre-registered[9]. 300 subjects participated in this study, recruited from Amazon MTURK through turkprime and were paid a small amount for their participation. 18 subjects were excluded from the study for failing one or more control questions. Subjects were randomly assigned to one of two conditions (4,7 Condition and 10,13 Condition, as described in Study 2) and one of three questions (yielding a 2x3 design). The three questions were Everyone's Expected Utility, John's Expected Utility, and Frustrated Means (see SI). Only the data from the Everyone's Expected Utility Condition are reported in the main text. Results from the remaining two conditions are reported in the SI.

Each subject read the story and was asked what would happen if N subjects used the new hook (exact wording varied depending on the utility curve, see SI). N was chosen at random without replacement from the following values until all values of N were seen by each subject: 0,2,7,8,13,19. [This list is for John's EU and Frustrated Means. For Everyone's EU, subjects see N+1.]

Study 5

Children (ages 4-11) were recruited in the Boston Common. We planned to analyze our data using a Bayesian analysis to avoid having to plan for a specific stopping rule due to our uncertainty about the effect size for this study and the difficulty of recruiting subjects [49, 50, 51, 52]. In the results section, we report the Bayes Factor as the main item of analysis, though we also include p-values to conform with current standards for data reporting. Ultimately, 191 subjects were included in the analysis (mean age = 7.5 years). 28 additional children were recruited but excluded from the analysis for failing the screening or control questions.

Children were first told simple stories accompanied by pictures to verify their competence with English and to ensure that they could use "OK" and "not OK" to make simple moral judgments. Subjects who did not answer these questions correctly were excluded from the analysis. Following the screening, children heard two test stories accompanied by pictures, counterbalancing condition and context. See Supplemental Materials for complete stimuli.

201 adult subjects received the same stimuli as children. Adults were recruited from Amazon MTURK through turkprime and were paid a small amount for their participation. See SI for details.

References

[1] André Blais and Carol Galais. Measuring the civic duty to vote: A proposal. *Electoral Studies*, 41:60–69, 2016.

[2] Jean-Daniel M Saphores, Oladele A Ogunseitan, and Andrew A Shapiro. Willingness to engage in a pro-environmental behavior: An analysis of e-waste recycling based on a national survey of us households. *Resources, Conservation and Recycling*, 60:49–63, 2012.

[3] John O Ledyard. Public goods: A survey of experimental research. In J. Kagel and A. Roth, editors, *Handbook of Experimental Economics*, chapter 2. Princeton University Press, NJ, 1995.

[4] Joshua David Greene. *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin, 2014.

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

[6] Max Kleiman-Weiner, Tobias Gerstenberg, Sydney Levine, and Joshua B Tenenbaum. Inference of intention and permissibility in moral decision making. In *CogSci*. Citeseer, 2015.

[7] Sandra Pellizzoni, Michael Siegal, and Luca Surian. The contact principle and utilitarian moral judgments in young children. *Developmental science*, 13(2):265–270, 2010.

[8] Sydney Levine and Alan Leslie. Preschoolers use the means-ends structure of intention to make moral judgments, under review.

[9] Jonathan Haidt. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):814, 2001.

[10] P Wesley Schultz, Jessica M Nolan, Robert B Cialdini, Noah J Goldstein, and Vladas Griskevicius. The constructive, destructive, and reconstructive power of social norms. *Psychological science*, 18(5):429–434, 2007.

[11] Shaun Nichols and Ron Mallon. Moral dilemmas and moral rules. *Cognition*, 100(3):530–542, 2006.

[12] John Mikhail. *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press, 2011.

[13] Elliot Turiel. *The development of social knowledge: Morality and convention*. Cambridge University Press, 1983.

[14] Immanuel Kant. *Groundwork for the Metaphysics of Morals*. 1785.

[15] Marcus G Singer. Generalization in ethics. *Mind*, 64(255):361–375, 1955.

[16] Richard Mervyn Hare, Richard Mervyn Hare, Richard Mervyn Hare Hare, and Richard M Hare. *Moral thinking: Its levels, method, and point*. Oxford: Clarendon Press; New York: Oxford University Press, 1981.

[17] Richard B Brandt. *Ethical Theory the Problems of Normative and Critical Ethics*. 1959.

[18] Nicolas Baumard, Jean-Baptiste André, and Dan Sperber. A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, 36(1):59–78, 2013.

[19] Jennifer B Misyak, Tigran Melkonyan, Hossam Zeitoun, and Nick Chater. Unwritten rules: virtual bargaining underpins social interaction, culture, and society. *Trends in cognitive sciences*, 18(10):512–519, 2014.

[20] Derek Parfit. *On what matters: volume one*, volume 1. Oxford University Press, 2011.

[21] Thomas Scanlon. *What we owe to each other*. Harvard University Press, 1998.

[22] John Rawls. *A theory of justice*. Harvard university press, 1971.

[23] Jürgen Habermas. *Moral consciousness and communicative action*. MIT press, 1990.

[24] Thomas Pogge. The categorical imperative. *Kant's Groundwork of the Metaphysics of Morals: Critical Essays (Lanham, MD: Rowman & Littlefield)*, pages 189–213, 1998.

[25] Onora O'Neill. Kant and the social contract tradition. *Kant's Political Theory: Interpretations and Applications, E. Ellis (ed.), The Pennsylvania Press, Pennysylvania*, pages 25–41, 2012.

[26] Nelson T Potter and Mark Timmons. *Morality and universality: Essays on ethical universalizability*. D. Reidel Publishing Company, 1985.

[27] Lawrence Kohlberg. *Stage and sequence: The cognitive-developmental approach to socialization*. Rand McNally, 1969.

[28] Lawrence Kohlberg. The claim to moral adequacy of a highest stage of moral judgment. *The journal of philosophy*, 70(18):630–646, 1974.

[29] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[30] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*. Institute of Urban and Regional Development, University of California, 1973.

[31] Cristina Bicchieri. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, 2005.

[32] Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278, 2013.

[33] Matt Jones and Bradley C Love. Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, 34(4):169, 2011.

[34] Julia W Van de Vondervoort and J Kiley Hamlin. The early emergence of sociomoral evaluation: infants prefer prosocial others. *Current Opinion in Psychology*, 20:77–81, 2018.

[35] Melanie Killen and Judith G Smetana. Origins and development of morality. *Handbook of child psychology and developmental science*, pages 1–49, 2015.

[36] Katherine McAuliffe, Peter R Blake, Nikolaus Steinbeis, and Felix Warneken. The developmental foundations of human fairness. *Nature Human Behaviour*, 1(2):42, 2017.

[37] F Ting, MB Dawkins, M Stavans, and R Baillargeon. Principles and concepts in early moral cognition, 2019.

[38] David G Rand, Alexander Peysakhovich, Gordon T Kraft-Todd, George E Newman, Owen Wurzbacher, Martin A Nowak, and Joshua D Greene. Social heuristics shape intuitive cooperation. *Nature communications*, 5:3677, 2014.

[39] Jim AC Everett, David A Pizarro, and Molly J Crockett. Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, 145(6):772, 2016.

[40] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[41] Falk Lieder and Thomas L Griffiths. Strategy selection as rational metareasoning. *Psychological Review*, 124(6):762, 2017.

[42] John C Harsanyi, Reinhard Selten, et al. A general theory of equilibrium selection in games. *MIT Press Books*, 1, 1988.

[43] David Charny. Hypothetical bargains: The normative structure of contract interpretation. *Michigan Law Review*, 89(7):1815–1879, 1991.

[44] Sydney Levine, Max Kleiman-Weiner, Nick Chater, Fiery Cushman, and Joshua B Tenenbaum. The cognitive mechanisms of contractualist moral decision-making. In *CogSci*. Citeseer, 2018.

[45] John Maynard Smith. *Evolution and the Theory of Games*. Cambridge university press, 1982.

[46] Robert Boyd and Peter J Richerson. Group selection among alternative evolutionarily stable strategies. *Journal of Theoretical Biology*, 145(3):331–342, 1990.

[47] Joseph Henrich. Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization*, 53(1):3–35, 2004.

[48] Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.

[49] Ward Edwards, Harold Lindman, and Leonard J Savage. Bayesian statistical inference for psychological research. *Psychological review*, 70(3):193, 1963.

[50] Eric-Jan Wagenmakers, Michael Lee, Tom Lodewyckx, and Geoffrey J Iverson. Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses*, pages 181–207. Springer, 2008.

[51] Eric-Jan Wagenmakers. A practical solution to the pervasive problems ofp values. *Psychonomic bulletin & review*, 14(5):779–804, 2007.

[52] Jeffrey N Rouder. Optional stopping: No problem for bayesians. *Psychonomic bulletin & review*, 21(2):301–308, 2014.