

When it is not out of line to get out of line: The role of universalization and outcome-based reasoning in rule-breaking judgments

Joe Kwon¹, Tan Zhi-Xuan¹, Joshua Tenenbaum¹, Sydney Levine^{1,2}

Correspondence to sydneyl@allenai.org

¹Department of Brain and Cognitive Sciences, MIT

²Allen Institute for AI

Abstract

How do we know when it's OK to break moral rules? We propose that — alongside well-studied outcome-based measures of welfare and harm — people sometimes use universalization, asking “What if everyone felt at liberty to ignore the rule?” We develop a virtual environment where agents stand in line to gather water. Subjects judge agents who get out of line to try to get water more quickly. If subjects use universalization, they would need to imagine all agents getting out of line and going straight for the water in each environment. To test this prediction, we model an action's universalizability by simulating what would happen if every agent tried to follow a path directly to the water, then evaluating the effects. We also investigate the role of several outcome-based measures, including welfare aggregation and harm-based measures. We find that universalizability plays an important role in rule-breaking judgments alongside outcome-based concerns.

Keywords: moral judgment; moral psychology; universalization; rule-breaking; cooperation

Introduction

Rules are a crucial part of making moral judgments—nearly every contemporary theory of moral psychology has some role for rules (Crockett, 2013; Cushman, 2013; Greene, 2014; Mikhail, 2011; Nichols & Mallon, 2006). However, nearly all psychological theories treat rules as rigid, while, upon reflection, it is clear that rules are actually quite flexible. After all, there are myriads of exceptions to seemingly simple rules. It is wrong to lie, but white lies are sometimes recommended. It is wrong to steal, but it might be okay to “steal” a napkin from a cafe to stop a bloody nose. How do we know when it is okay to break the rules?

While this question has seldom been raised by psychologists (for a recent exception see Levine, Chater, Tenenbaum, & Cushman, 2023), a compelling answer has been raised by a series of moral philosophers who suggest that we can navigate exceptions to rules by using the tool of *universalization* (Kant, 1785; Harsanyi, 1977; Gert, 1998; Hare, 1981; Scanlon, 1998; Anderson, 2001; Roemer, 2015). When a possible exception to a rule arises, we can ask: what would happen if everyone felt at liberty to break the rule in cases like this? This philosophical approach can be contrasted with those that instead consider *outcomes* when deciding if an instance of rule-breaking is morally permissible. For instance, an act of lying might be considered permissible if the outcome leaves everyone better off overall, in accordance with act consequentialist welfare maximization (Bentham, 1789; Hare, 1981), or

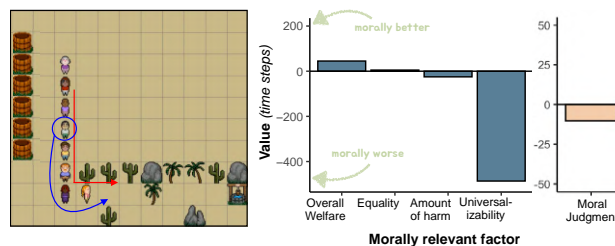


Figure 1: Players in this game are trying to gather water as quickly as they can and deposit it in the buckets on the far left side of the screen. In this particular map, the players start off standing in line next to the entrance to the narrow path that leads to the map's only water source, a well on the far right side of the screen. The players enter two at a time and when they get to the well, each player can fill their bucket with water simultaneously before reversing down the path and heading to deposit their water in the buckets. However, one person gets out of line, goes to the front, and sneaks into the narrow path when two people have already entered. Our participants judge this action to be morally wrong (graph on the right), though the action actually increases overall welfare and equality while increasing harm only marginally (see graph on the left). Participants' judgments are largely driven by universalization in this case: if everyone felt at liberty to get out of line and head straight for the water source, things would go badly for everyone. See main text for precise descriptions of the metrics reported in these graphs.

if nobody is harmed by the specific instance of lying, in accordance with the harm principle (Mill, 1859; Parfit, 2011).

In this paper, we use these competing ideas from moral philosophy as a starting point for studying the moral cognition of rule-breaking. In addition to outcome-based assessments of welfare or harm (Cushman, Young, & Hauser, 2006; Nichols & Mallon, 2006; Crockett, 2013), we study how the universalizability of a potential rule violation may affect people's judgments of its permissibility. As our test case, we look at the rule governing standing in line, often simply stated as “no cutting.” We aim to predict when people will think that it is acceptable to violate this rule.¹ To do this, we develop a vir-

¹There is no consensus about what sorts of norm violations count as moral violations, as opposed to conventional, religious or prag-

tual environment where eight agents try to collect water from wells, ponds, and streams as quickly as they can and then store the collected water in a bucket in another part of the map. The agents start each scene standing in line in front of a water source with rocks and trees sometimes blocking their path (see Fig 2). Only a single agent can occupy a given space in the environment, so collisions between agents are a natural complication. In each scene, one agent gets out of line and heads directly for a water source. After presenting a scene to human subjects, we ask them to judge if what the agent did was morally acceptable.

To predict these moral acceptability judgments, we introduce a computational model that produces quantitative metrics for the universalizability of an agent's action, as well as the outcomes of an agent's action along several morally-relevant dimensions (aggregate welfare, harm, etc.). We then evaluate whether and when each of these metrics plays a role in moral judgements. If the direct benefits or harms of rule-breaking are factors, then moral judgments should vary with the *actual* outcome of the agent getting out of line. If universalizability is a factor, then moral judgements should track *hypothetical* outcomes, namely, the outcomes that arise from everyone feeling at liberty to get out of line (Fig 1). Of course, multiple of these factors might impact moral judgments. In that case, the contribution of a factor like universalizability will be sharpest when other factors are close to zero.

Background: Universalization as a psychological process

Universalization asks us to imagine a *hypothetical world* in which everyone acts a certain way — or, more precisely, when everyone feels *at liberty* to act a certain way.² The outcomes in that hypothetical world determine the moral permissibility of the action, regardless of the effect of the action in the real world. For instance, if one person abstains from voting in a national election, it will generally have no effect on the outcome of the election. However, if everyone felt at liberty to abstain, then presumably a majority of people *would* abstain since voting takes time and effort, and democracy would collapse.

Despite its prominent role in theories of moral philosophy, universalization has been relatively neglected in theories of

matic ones, for instance (Levine et al., 2021; Stich, 2018). Rules about lines blur this boundary further because they are clearly socially constructed, though often thought of as a system designed to ensure some measure of fairness (a metric often associated with paradigmatically moral issues, (Haidt, 2003)). We call the rule associated with standing in line “moral” (and hence judgments about breaking that rule “moral judgments”) because we are primarily interested in understanding rules and judgments that deal with the pervasive problem of *interdependent rational choice*, the struggle to achieve mutual benefit when agents have some compatible and some conflicting interests (Gauthier, 1987; Braithwaite, 1969; Levine et al., 2023).

²As explained in Levine, Kleiman-Weiner, Schulz, Tenenbaum, and Cushman (2020), the “at liberty” qualification is necessary to avoid implausible conclusions. Without the qualification, dentistry would be impermissible simply because things would go poorly if everyone became a dentist.

moral cognition (though Kohlberg (1969) is an important exception). Recent research, however, has shown that universalization can sometimes explain people's moral intuitions in collective action problems (Levine et al., 2020; Kwon, Tenenbaum, & Levine, 2022), and that it plays an important role in a broader contractualist understanding of moral cognition (Levine et al., 2023). Building upon this work, we ask if this mechanism is operative in rule-breaking judgments as well.

In addition, our current study asks whether universalization reasoning can explain moral judgments even when such reasoning is not explicitly prompted. This is in contrast to previous work, where participants were explicitly informed what would happen if an action were universalized. For instance, in Levine et al. (2020), participants were told exactly what would happen if certain numbers of people in a fishing village used a new method of fishing (e.g. the fish population would collapse by the time seven people used the new fishing hook). Such work leaves open the critical question of whether people can (and do) spontaneously and accurately simulate the hypothetical world where an action is universalized and draw on that simulation to make a moral judgment. In the complex world we live in, these mental simulations could involve quite complex computations. To account for this, we study universalization in a naturalistic setting of moral judgment — line cutting — which requires people to imagine and predict, at least to some degree of approximation, the hypothetical consequences of many agents' actions and interactions unfolding over space and time.

Finally, previous work tested cases where universalization could only be detected in the absence of the use of other moral mechanisms, and furthermore used hypothetical aggregate utility as the evaluation metric for a universalized action (Levine et al., 2020). Here we test whether universalization can be used on its own or in combination with multiple other (outcome-based) approaches to moral judgment, in a setting where the distinct predictions of different judgment mechanisms can be isolated but can also synergistically combine. We also investigate the fact that universalization-based reasoning can vary in its *test conditions* or *criteria of evaluation* (Forschler, 2017): Even when imagining the hypothetical world where everyone feels at liberty to act a certain way, people might differ in how they evaluate that world, perhaps by considering hypothetical aggregate welfare, as in rule utilitarianism (Harsanyi, 1977), the acceptability of the hypothetical world to each individual, as in Scanlon's formulation of contractualism (Scanlon, 1998), or other principles of logical consistency and symmetry (Kant, 1785; Anderson, 2001; Roemer, 2015).

Modeling Universalization and Outcome-Based Reasoning

To model how people make moral judgements of potential rule violations, we used a virtual environment (examples of which are shown in Figure 2) as a simulator for rule-following and rule-breaking behavior. For universalization

reasoning, the environment was used to simulate *hypothetical* line-leaving behavior, whereas in outcome-based reasoning, the environment was simply used to compare the direct effects of the agent getting out of line versus staying in line. We describe each of these models in greater detail below.

Universalization Reasoning

Consider watching someone cut in line in order to reach their own goals more quickly, and imagining “What if everybody felt at liberty to act this way?” Modeling this hypothetical would require specifying what agents would be motivated to do, given that liberty, then simulating how they act upon those motivations. In our case, rule-unbound agents would presumably be motivated to leave the line out of self-interest — as long as they can reasonably anticipate that doing so would benefit them — and act efficiently from there on to achieve their individual goals.

Line-Leaving as Self-Interested Planning To formalize this intuition, we model our hypothetical line-leaving agents as (naively) rational self-interested actors: Each agent plans to follow the shortest path to achieve its next sub-goal (e.g. collecting water, or storing the water in a bucket), under the assumption that there are no other agents; the other agents are treated as obstacles that the agent cannot pass through. This process happens repeatedly, with each agent planning to follow a new path at every timestep, thereby accounting for the actual movement of other agents. In order to allow for some amount of stochasticity (as might be expected of real humans), and also avoid deadlocks due to the determinism, the precise model we use is a Boltzmann-rational policy:

$$\pi_i(a|s) = \frac{\exp(-C_i(s,a)/T)}{\sum_{a'} \exp(-C_i(s,a')/T)} \quad (1)$$

where $\pi_i(a|s)$ is the probability of an agent i taking the action a in state s , which increases with decreasing cost $C_i(s,a)$ of the shortest path to the agent’s goal that starts with state s and action a . In other words, agents will tend to follow shorter paths to their goal, with their efficiency in doing so controlled by a noise parameter or “temperature” T , where $T = 0$ means that agents *always* follow the shortest path (under the assumption that all other agents do not move). To efficiently compute the value of $C_i(s,a)$, we use A* search as a shortest-path planning algorithm, which was implemented by encoding our virtual environment in the Planning Domain Definition Language (Aeronautiques et al., 1998; Zhi-Xuan, 2022) and running the planner with an admissible heuristic (Hart, Nilsson, & Raphael, 1968).

Avoiding Collisions through Ordered Execution In addition to specifying how agents act at each state, our model needs to specify how agents *interact* with each other: Since agents cannot occupy the same physical location, they may collide with each other if they try to enter the same grid cell. To address this, we make the assumption that agents’ actions are *ordered*: The first agent in the original line formation executes their action, followed by the second agent, the third

agent, and so on. This prevents collisions, reflecting the fact that even line-cutting agents are somewhat able to predict the future actions of other agents, and hence are mostly successfully in avoiding gross miscoordination. It also accounts for the possibility that agents will naturally follow the line out of self-interest, especially in tight corridors where moving forward in line is the only reasonable option.

Simulation Parameters We ran simulations using Boltzmann policies while varying the noise parameter $T \in \{0, 0.0001, 0.001, 0.01, 0.1, 1\}$. We found that a value of $T = 0.0001$ best captured relatively optimal self-interested behavior (as measured by average task completion times across all maps) while avoiding the deadlocks that occurred when simulating deterministic behavior ($T = 0$). As such, we standardized our modeling results to use $T = 0.0001$.

Universalizability Metrics As noted earlier, universalization reasoning requires not only simulating a hypothetical world where everybody feels at liberty to act in a certain way, but also evaluating that hypothetical world according to some metric or criterion M (Forschler, 2017). We thus computed different metrics of an action’s universalizability Univ_M by applying each of the outcome-based metrics M described in the next section (aggregate welfare, etc.) to our *hypothetical* simulation outcomes, then taking the expectation across different simulated world histories w given an initial environment state s_0 :

$$\text{Univ}_M(s_0) = \mathbb{E}_{w \sim P(w|s_0)} [M(w)] \quad (2)$$

where $P(w|s_0)$ is the distribution over world histories w , defined by modeling how agents would act if everyone felt at liberty to leave the line. To estimate this expectation, we averaged the value of $M(w)$ across 5 simulations for each map.

Outcome-Based Reasoning

In contrast to universalization reasoning, outcome-based reasoning simply considers the outcomes of an agent taking a potentially rule-violating action, in comparison to everyone having followed the (potential) rule. Outcomes, however, can be measured in terms of many metrics. We thus evaluated our stimuli in terms of the metrics defined below.

Aggregate Welfare We measure the (change in) aggregate welfare associated with line-leaving by summing the reduction in costs incurred by each agent, such that aggregate welfare is higher whenever the line-leaving action makes everyone better off in aggregate. Let $C_i(w)$ be the cost for the i -th agent to complete the task in the actual world w where someone leaves the line, $C_i(w_L)$ be the cost for the i -th agent in the world w_L where everyone stays in line, and $C_{\text{last}}(w_L)$ be the cost incurred by the last agent when everyone stays in line. Then our aggregate welfare metric is defined as:

$$\Delta W(w) = -\frac{\sum_{i=1}^n C_i(w) - \sum_{i=1}^n C_i(w_L)}{C_{\text{last}}(w_L)} \quad (3)$$

We divide by $C_{\text{last}}(w_L)$ to standardize the metric across maps, since $C_{\text{last}}(w_L)$ measures the intrinsic difficulty of a map.

Ordinal Harm This accounts for “positional harms” that may be incurred by each agent due to someone leaving the line, as measured by how it increases the *order* in which they finish their task. Let $O_i(w_L)$ be the ordinal position of the i -th agent when staying in line, and $O_i(w)$ be the ordinal position of the i -th agent in the actual world w , as measured among those agents who collect water from the same source location. We define the net ordinal harm as:

$$OH(w) = \sum_{i=1}^n \begin{cases} O_i(w) - O_i(w_L) & \text{if } O_i(w) > O_i(w_L), \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Since we define $O_i(w)$ as the ordering of agents who aim for the same water source, this means that if the line-leaver ends up going to a different water source, then $O_i(w)$ can only stay the same or improve. This captures the intuitive notion that an agent i is not harmed if someone in front of them leaves the line to collect a different resource.

Cardinal Harm This is similar to ordinal harm, except that it considers the excess *costs* incurred by each agent as the result of someone leaving the line, where cost is measured by task completion time. As in aggregate welfare, $C_i(w)$ is the cost incurred by the i -th agent in the actual world, and $C_i(w_L)$ is the cost incurred by the i -th agent if everyone had stayed in line. Let $C_{\text{last}}(w_L)$ be the cost incurred by the last agent when everyone stays in line. Cardinal harm is then defined as:

$$CH(w) = \frac{1}{C_{\text{last}}(w_L)} \sum_{i=1}^n \begin{cases} C_i(w) - C_i(w_L) & \text{if } C_i(w) > C_i(w_L), \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Inequality Finally, we compute the (change in) inequality due to someone leaving the line, as measured by the difference in finish times between the first and last agents. Let $C_{\text{first}}(w)$ and $C_{\text{last}}(w)$ be the task completion costs for the first and last agents in a world w . Then the (change in) inequality (from the world where everyone stays in line, w_L) is given by:

$$\Delta I(w) = \frac{C_{\text{last}}(w) - C_{\text{first}}(w)}{C_{\text{last}}(w_L)} - \frac{C_{\text{last}}(w_L) - C_{\text{first}}(w_L)}{C_{\text{last}}(w_L)} \quad (6)$$

Experiment

Environment Design We directly test the proposed mechanism by creating stimuli designed to elicit graded responses that our theory predicts with quantitative precision. In particular, when an agent gets out of line, their action varies on how universalizable it is. Sometimes, if everyone were to get out of line and go straight for the water, things would be perfectly fine (or even better than if everyone stood in line). Other times, chaos would ensue and everyone would get slowed down. Still other times, some agents are slowed down while others are able to get their water faster.

In designing our environments, we varied several key dimensions: the number of access points (if any) into an area



Figure 2: Example environments demonstrating how key features were varied – including the number of access points to a space with water, the space available inside the area with water, the number of water sources, and the distribution of water sources across the map. Red arrows indicate the path that the agents in line take.

with water sources, the space available in the access points (only one agent can pass through, or multiple agents can pass through), the number of water sources, and the arrangement of water sources throughout the map. By manipulating these features, we can more easily disambiguate and test whether moral judgments about an agent leaving the line are attributable to universalizability or outcome-based metrics.

For example, in the environments shown in Fig 2a and 2c, both maps lack access points that bottleneck paths towards the water sources, but they vary in the arrangement of water sources; one has a singular water source with a large surface area while the other has multiple water sources distributed across the map which are accessible to only one agent at a time. In both of these cases, the universalizability of line-leaving should be high, since everyone could leave the line without much issue.

As another set of examples, we predicted low universalizability for the environments shown in Fig 2b and Fig 2e. In Fig 2b, this is because there is a single water source accessible to one agent a time, despite the lack of other obstacles. In Fig 2e, this is instead due to the limited number of access points into the space with water, and the limited space available once inside the area, despite the existence of multiple water locations accessible to multiple agents at a time.

Experimental Procedure Participants watched clips of the game being played (in 39 different maps total) and made judgments. One group of participants was asked to make moral judgments of the actions ($n=49$), and another was asked to make universalizability judgments ($n=50$).

Participants were familiarized to the game environment through a series of warm-up tasks that involved watching short videos of the game and answering comprehension questions about the dynamics of the game and the task they were assigned to do. Participants then viewed all 39 videos in

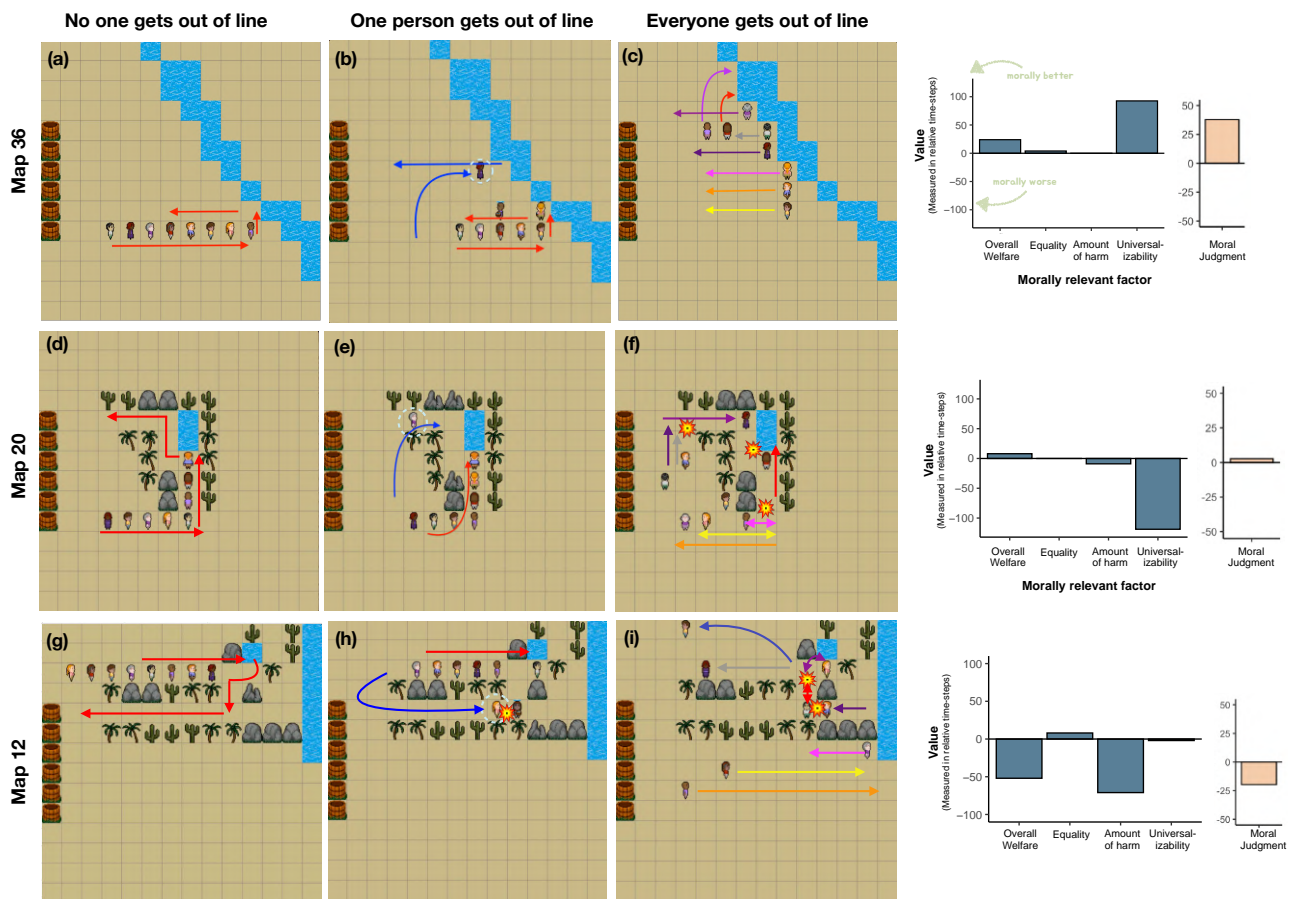


Figure 3: Each row represents different environments (Maps 36, 20, and 12 respectively). The first column depicts how the environment looks when everyone follows the line (red arrow), the second column depicts what happens when one agent leaves the line (blue arrows), and the third column shows model predictions for universalization (each agent's path depicted by a different colored arrow; red explosions indicate collision points). Subjects were only shown the cases where one agent gets out of line; subjects made moral judgments of the line-leaver. Graphs to the right of the maps depict moral judgment (peach bar) and three outcome-based measures and universalizability (blue bars). Outcome-based measures represent the difference in how well things go when one person leaves the line compared to if everyone were to stay in line. These include overall welfare (aggregate welfare ΔW , as given in Eq 3), equality (the negation of inequality ΔI , as given in Eq 6), and amount of harm (cardinal harm CH , Eq 5). Universalizability (aggregate welfare version ($Univ_{\Delta W}$), Eq 2) compares how well things would go if everyone felt free to head directly towards a water source, compared to what would happen if everyone stayed in line. Note that all measures are displayed in raw form and not normalized to C_{last} . The maps vary in universalizability and how good or bad their outcomes are. In Map 1 (a-c), getting out of line is highly universalizable (as shown in (c)) and produces no collisions between agents; when one person gets out of line, outcome metrics improve. Moral judgment of this agent is therefore highly positive. In Map 2 (d-f), many collisions occur when getting out of line is universalized (f), thereby slowing everyone down. When just one person gets out of line, some outcome metrics improve, however. Thus moral judgment of the one line-leaver is neutral. In Map 3 (g-i), some collisions occur when getting out of line is universalized, but there is also space to resolve these collisions (i). However, the one person who gets out of line actually ends up slowing many other players down (h), so outcome based measures decrease dramatically and moral permissibility reflects this fact.

	Full Model	No Univ ΔW	No ΔW	No OH	No CH	No ΔI
(Intercept)	4.33 (7.55)	-4.72 (8.43)	25.85*** (3.75)	9.34 (8.99)	16.88*** (4.11)	-1.09 (7.59)
Universalizability	4.12*** (1.10)		4.80*** (1.21)	5.22*** (1.28)	4.30*** (1.14)	3.59** (1.13)
Aggregate Welfare	41.99** (13.21)	51.74** (15.26)		27.36 (15.34)	17.50*** (4.28)	42.74** (14.00)
Ordinal Harm	-4.22*** (1.04)	-5.19*** (1.18)	-3.32** (1.12)		-3.46** (1.00)	-4.02*** (1.10)
Cardinal Harm	32.68 (16.75)	38.04 (19.65)	-17.90** (5.88)	6.96 (18.72)		30.54 (17.72)
Inequality	67.48* (29.60)	43.81 (34.07)	69.81* (33.32)	57.32 (35.61)	64.24* (30.75)	
AIC	314.28	326.21	322.70	328.13	316.54	317.99
BIC	325.93	336.19	332.68	338.11	326.52	327.97
Log Likelihood	-150.14	-157.10	-155.35	-158.06	-152.27	-152.99
Pearson Correlation (r)	0.891	0.839	0.854	0.830	0.877	0.872
Num. obs.	39	39	39	39	39	39

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Table 1: Six linear regressions models that demonstrate the role of universalizability in participants’ moral judgments. Coefficient values are followed by standard errors in parentheses. The full model (first column), which includes predictors for universalizability, aggregate welfare, ordinal harm, cardinal harm, and inequality fits the data best. Universalizability is significant in all models. The “aggregate welfare” version of universalization is used here; see Table 2 for models containing other versions of universalization.

randomized order. Participants in the Moral Judgment condition were told: “Think about what this character is doing and make a judgment.” They were shown a picture of the agent that got out of line and were asked to make a judgment on a scale from -50 (completely morally unacceptable) to 50 (completely morally acceptable). Participants in the Universalization condition saw the same videos but were posed the following question: “What if everyone felt at liberty to leave the line and try to get water faster for themselves? How much better or worse off would everyone be compared to if they all stayed in line instead? Slide the bar to indicate your answer.” Participants answered on a scale from -50 (much worse off), 0 (the same), 50 (much better off). Participants were then asked a series of demographic questions and given an opportunity to report if there was something confusing or unclear about the study.

Participants Participants were recruited through Prolific. 17 participants were excluded for failing control questions (10 in the moral judgment group and 7 in the universalizability group). Participants were paid a minimum of \$15 per hour. This study was approved by the IRB of the Massachusetts Institute of Technology. Demographic characteristics of the group of participants asked to make moral judgments is as follows. Gender: 49% female, 46% male, 5% non-binary. Mean age: 34.1 years. Race and ethnicity: 69% White, 10% Black/African/Carribbean, 7% Asian, 7% multiple/mixed, 5% prefer not to say. All participants considered

English to be a primary language and all but one identified as growing up in the United States. Demographic characteristics of the group of participants asked to make universalizability judgments is as follows: Gender: 63% female, 35% male, 2% non-binary. Mean age: 37.7 years. Race and ethnicity: 79% White, 5% Black/African/Carribbean, 2% Asian, 12% multiple/mixed, 5% prefer not to say, 2% other. All participants considered English to be a primary language and all but one identified as growing up in the United States.

Results

For each of the 39 maps, we calculated values for the four outcome-based measures (given in Equations (3)-(6)), which describe how well things go for the agents in the game when one person cuts, compared to what would have happened had everyone stayed in line. We also calculated values for multiple versions of universalization, which represent how things would go in the hypothetical world where everyone felt at liberty to get out of line and go straight for the water (as given by the model predictions), compared with what would have happened had everyone stayed in line. (See Figure ??.)

We then compared our universalization model predictions (aggregate welfare version) to participants’ judgments of the universalizability of the action and find a strong correlation, indicating that our objective model is broadly capturing people’s subjective assessment of universalizability (see Fig 4b.)

Next, we evaluate participant moral judgment data against the possibility that some participants are “rule absolutists” and judge all instances of getting out line to be impermissi-

	Full Model	Model 2	Model 3	Model 4
Univ. Criterion M :	ΔW	ΔI	CH	OH
(Intercept)	4.33 (7.55)	-0.47 (8.59)	5.66 (7.86)	-0.44 (8.65)
Univ: Aggregate Welfare	4.12*** (1.10)			
Aggregate Welfare	41.99** (13.21)	52.10** (14.87)	49.98*** (13.21)	61.37*** (16.07)
Ordinal Harm	-4.22*** (1.04)	-4.04** (1.34)	-4.47*** (1.04)	-5.59*** (1.18)
Cardinal Harm	32.68 (16.75)	39.15* (19.16)	41.60* (17.04)	49.43* (20.46)
Inequality	67.48* (29.60)	52.44 (33.58)	65.92* (30.14)	41.30 (33.33)
Univ: Inequality		-0.10 (0.06)		
Univ: Cardinal Harm			-5.12** (1.45)	
Univ: Ordinal Harm				-1.22 (0.75)
AIC	314.28	325.00	315.77	325.26
BIC	325.93	336.65	327.41	336.90
Log Likelihood	-150.14	-155.50	-150.88	-155.63
Pearson Correlation (r)	0.891	0.853	0.886	0.852
Num. obs.	39	39	39	39

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Table 2: Three models that compare different versions of the universalizability metric. The “Full Model” is identical to that reported in Table 1, where we use the “aggregate welfare” version of universalization $Univ_{\Delta W}$. Models 2-4 modify the universalization criterion M , replacing the aggregate welfare version of universalization with three other possibilities: the inequality version $Univ_{\Delta I}$, cardinal harm version $Univ_{CH}$, and ordinal harm version $Univ_{OH}$. The model that uses the aggregate welfare version of universalization, $Univ_{\Delta W}$, captures the data best on AIC and BIC.

ble. To test this hypothesis we used a thresholding approach, where we treated any moral judgment that was negative (< 0) as a judgment of “impermissible” and any positive judgment (≥ 0) as permissible. Calculating the proportion of cases judged permissible by each participant (Fig 4a), we find that *none* of our participants is a rule absolutist: all participants judge that it is *sometimes* permissible to get out of line.

To evaluate the role of universalization and outcome-based metrics, we analyzed the data using a general linear model, predicting mean participant moral judgment for each of the 39 cases. The full model included universalization (aggregate welfare version; Equation 2), aggregate welfare (Equation 3), ordinal harm (Equation 4), cardinal harm (Equation 5), and inequality (Equation 6). Ordinal harm and cardinal harm were re-scaled to a range of -50 to 0. We compared the full model to lesioned versions of the model, each which removed one predictor. The full model best captures the data (*i.e.*, has the lowest AIC and BIC); all predictors are significant in the full model (see Table 1 for details). This suggests that universalization is a critical component of moral judgments, alongside better studied outcome-based mechanisms of moral

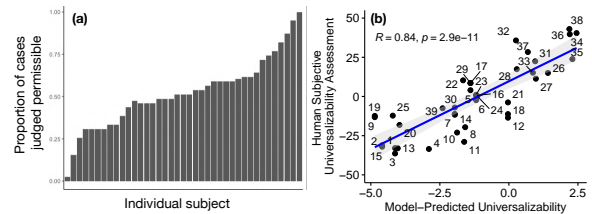


Figure 4: (a) Proportion of cases judged permissible by each individual participant. No participant treats all cases as impermissible, indicating that all participants judge that the “no cutting” rule has at least some exceptions or fine-grained application conditions. (b) Model-predicted universalizability of an instance of cutting strongly correlates with participants’ subjective assessment of universalizability ($R = 0.84$). Points are labeled with a number that corresponds to the mean permissibility judgment of each case (1 = least permissible; 39 = most permissible).

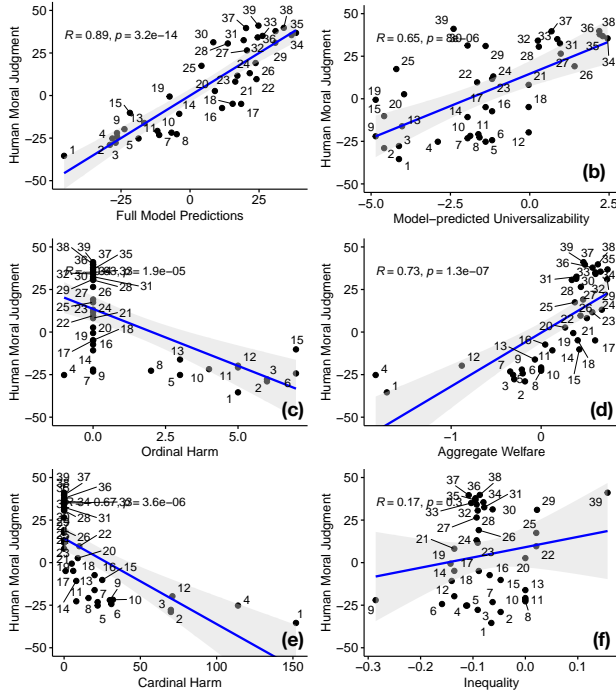


Figure 5: Correlations between participant moral judgment and (a) the full regression model, (b) model-predicted universalization (aggregate welfare version) and (c)-(f) the outcome-based measures. Points are labeled with a number that corresponds to the mean permissibility judgment of each case (1 = least permissible; 39 = most permissible).

judgment (such as welfare and harm).

As mentioned previously (see Equation 2), using the logic of universalization requires establishing which universalization metric (M) to apply to the hypothetical world where everyone feels at liberty to do some action (here, get out of line and go directly towards the water source). In the above models, we calculated aggregate welfare (as given in Equation 3) in the universalized world as the measure of universalization. In Table 2, we present models that use other universalization metrics instead, including inequality (Equation 6), cardinal harm (Equation 5) and ordinal harm (Equation 4). Each metric was calculated based on the outcome in the simulated universalized world. The inequality and cardinal harm versions of universalization are significant predictors, though the ordinal harm version of universalization is not.

Individual Differences Analysis Previous work has shown that not all people use universalization, but that it is used by a substantial minority of the population (Levine et al., 2020). Following up on this finding, we analyzed the judgments of each individual participant, treating each participant’s answer to each scenario as the data to be predicted. We used the model that best fit the aggregate data (including universalization (aggregate welfare version), aggregate welfare, cardinal harm, ordinal harm, and inequality as predictors). We found that 64% of our participants use universalization in their moral judgments (*i.e.*, universalization was a significant

predictor in the linear model), while 51% use aggregate welfare, 59% use ordinal harm, 31% use cardinal harm, and 26% use inequality.

Discussion & Conclusion

In this paper we asked how participants know when the rule “no cutting in line” can be broken. We modeled the universalizability of a particular instance of potential rule-breaking using a multi-agent planner, which allowed us to determine the hypothetical results if all agents in the scene got out of line and went straight for the water. The central finding is that participants’ judgments of the permissibility of getting out of line is best explained by a model that includes the action’s universalizability as well as two objective outcome-based measures (ordinal harm and aggregate welfare). This shows that participants’ judgments are consistent with using a dynamic and context-sensitive simulation of agents acting in a complex world in order to make universalization assessments, and consequently a moral judgment.

A series of questions remain about universalization and its use in moral cognition. First, participants’ subjective universalization judgments (*i.e.*, their answer to the question “What would happen if everyone felt at liberty to get out of line...?”) were strongly but not perfectly correlated with our objective model-based universalization predictions. Is this a “failing” of the model or of people’s judgments? It is possible that people’s subjective assessments are distorted by a variety of factors. For instance, participants watch a video where one person cuts the line and they may notice that it goes well or poorly and then anchor on that fact to make a universalization judgment. (Indeed, there is evidence that this is the case because subjective universalization judgments vary when the map is held constant but the result of one person cutting changes.) On the other hand, our model fails to account for factors that people likely take into account, such as the ability for each agent to predict the paths of the others and smoothly navigate around them. Future work should aim to differentiate between these possibilities.

Our regression models showed that universalization is used alongside outcome-based measures. How are these mechanisms of moral cognition integrated? Are there instances where one of these mechanisms is used to the exclusion of others? Are these processes integrated in a way that is somehow resource-rational? Relatedly, our model allows for the possibility that people may judge the universalizability of an actions according to different criteria, but our experiments thus far have not distinguished them in detail. Do people in fact use some universalization criteria over others, as philosophers have suggested (Forschler, 2017)? For example, can we find cases where an action is universalizable if we only consider aggregate welfare, but non-universalizable because it would consistently harm specific people if universalized?

Lastly, what accounts for the stable finding (*c.f.* Levine et al. (2020)) that a substantial minority (but not all) of participants use universalization? Is the tendency to use universal-

ization a stable trait of an individual? If so, are there demographic features of individuals that predict its use? Or do all (or most) people use universalization probabilistically? Future work should explore individual and cross-cultural differences in the use of this moral mechanism.

Acknowledgements

TZX was supported by the Open Philanthropy AI Fellowship. SL was supported by a grant from the the Templeton World Charity Foundation. The authors would like to thank Sally Zhao for her tireless assistance in data collection for this project, Jean-Francois Bonnefon for his color palette suggestions, and Gottlob Frege for his continual inspiration.

References

- Aeronautiques, C., Howe, A., Knoblock, C., McDermott, I. D., Ram, A., Veloso, M., ... others (1998). PDDL—the Planning Domain Definition Language. *Technical Report, Tech. Rep.*
- Anderson, E. (2001). Unstrapping the Straitjacket of ‘Preference’: A comment on Amartya Sen’s contributions to philosophy and economics. *Economics & Philosophy*, 17(1), 21–38.
- Bentham, J. (1789). *An introduction to the principles of morals and legislation*. T. Payne and Son.
- Braithwaite, R. B. (1969). *Theory of games as a tool for the moral philosopher*. CUP Archive.
- 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.
- Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological science*, 17(12), 1082–1089.
- Forschler, S. (2017). Universal practice and universal applicability tests in moral philosophy. *Philosophical Studies*, 174(12), 3041–3058.
- Gauthier, D. (1987). *Morals by agreement*. Clarendon Press.
- Gert, B. (1998). *Morality: Its nature and justification*. Oxford University Press on Demand.
- Greene, J. D. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.
- Haidt, J. (2003). The moral emotions. *Handbook of affective sciences*, 852–870.
- Hare, R. M. (1981). *Moral thinking: Its levels, method, and point*. Oxford: Clarendon Press; New York: Oxford University Press.
- Harsanyi, J. C. (1977). Morality and the theory of rational behavior. *Social research*, 623–656.
- Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2), 100–107.
- Kant, I. (1785). *Groundwork for the metaphysics of morals*.
- Kohlberg, L. (1969). *Stage and sequence: The cognitive-developmental approach to socialization*. Rand McNally.
- Kwon, J., Tenenbaum, J., & Levine, S. (2022). Flexibility in moral cognition: When is it okay to break the rules? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).
- Levine, S., Chater, N., Tenenbaum, J., & Cushman, F. (2023). Resource-rational contractualism: A triple theory of moral cognition.
- 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.
- Mikhail, J. (2011). *Elements of moral cognition: Rawls’ linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press.
- Mill, J. (1859). *On Liberty*. J. W. Parker and Son.
- Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, 100(3), 530–542.
- Parfit, D. (2011). *On what matters: Volume one* (Vol. 1). Oxford University Press.
- Roemer, J. E. (2015). Kantian optimization: A microfoundation for cooperation. *Journal of Public Economics*, 127, 45–57.
- Scanlon, T. (1998). *What we owe to each other*. Harvard University Press.
- Stich, S. (2018). The quest for the boundaries of morality. In *The routledge handbook of moral epistemology* (pp. 15–37). Routledge.
- Zhi-Xuan, T. (2022). *PDDLjl: An extensible interpreter and compiler interface for fast and flexible AI planning*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Appendix: Materials & Methods

1 Instructions and Exclusion Criteria

1.1 Moral Judgment Group

Participants in the Moral Judgment Group saw the following instructions prior to beginning the experiment.

Instruction Part 1 out of 3

Please read the instructions very carefully.

In this experiment, you will watch videos of a video game. Everyone in the game is trying to collect water from a water source (wells, streams, ponds, etc.), and bring it back to the wooden storage buckets. Each person needs to collect one bucket of water for each "scene" of the video game. The faster a person collects the bucket of water, the more points they receive. In each video, you will focus on one person. When the video is over, you will answer whether you believe that person's actions were morally acceptable or morally unacceptable.

Answer the following comprehension check question and press continue: what is the goal of the people in the game?

Instruction Part 2 out of 3

Please watch the video below to get a sense of what the environment is like, and what actions people can take. In the example video below, there is only 1 active player moving so you can focus on what it looks like when people collect water! But when the actual experiment begins, all the videos you see will have 8 people moving in them. As you progress through the videos and become familiar with the process, you may choose to increase playback speed in the bottom right corner of the video screen.

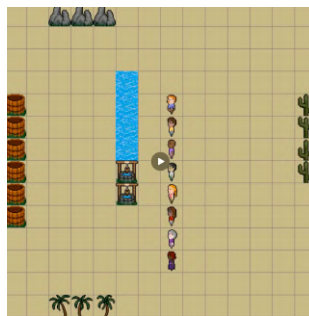


Figure 1: First scene of instruction video.

As you see in the video above, each person can move one square per time step in one of the four cardinal directions (up, down, left, right). They cannot move through squares with objects in them (like rocks, plants, wells, and water). They can collect water if they are in a square that is adjacent to a square with water, and it takes three time steps to collect the water. Something that you'll notice when you see videos with more people in them, is that people can only move to squares that are not occupied by other people.

Answer the comprehension check question and press continue: how many squares can players move in each time step?

Instruction Part 3 out of 3

Preprint. Under review.

In each video, all 8 people will start out waiting in a line in front of one of the water sources. People can leave the line and head straight for the water source without waiting or they can stay waiting in line until it's their turn to collect water. The videos you see will all involve one person leaving the line and heading straight for the water source. Your job is to make a moral judgment about what that character is doing. Feel free to watch the video many times, but please watch it at least once completely and think carefully before you make your judgments.

Answer the following comprehension question: Who will you be making a moral judgment about?

After the experiment is over, you will have an opportunity to give us general feedback and let us know if anything was confusing or unclear.

1.2 Universalization Group

Participants in the Universalization Group saw the following instructions prior to beginning the experiment.

Instruction Part 1 out of 3

Please read the instructions very carefully.

In this experiment, you will watch videos of a video game. Everyone in the game is trying to collect water from a water source (wells, streams, ponds, etc.), and bring it back to the wooden storage buckets. Each person needs to collect one bucket of water for each "scene" of the video game. The faster a person collects the bucket of water, the more points they receive.

Answer the following comprehension check question and press continue: what is the goal of the people in the game?

Instruction Part 2 out of 3

Please watch the video below to get a sense of what the environment is like, and what actions people can take. In the example video below, there is only 1 active player moving so you can focus on what it looks like when people collect water! But when the actual experiment begins, all the videos you see will have 8 people moving in them. As you progress through the videos and become familiar with the process, you may choose to increase playback speed in the bottom right corner of the video screen.

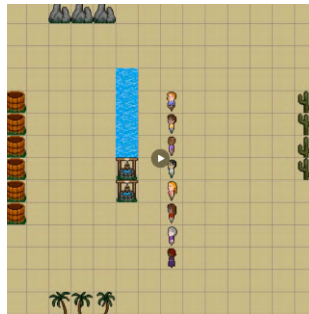


Figure 2: First scene of instruction video.

As you see in the video above, each person can move one square per time step in one of the four cardinal directions (up, down, left, right). They cannot move through squares with objects in them (like rocks, plants, wells, and water). They can collect water if they are in a square that is adjacent to a square with water, and it takes three time steps to collect the water. Something that you'll notice when you see videos with more people in them, is that people can only move to squares that are not occupied by other people.

Answer the comprehension check question and press continue: how many squares can players move in each time step?

Instruction Part 3 out of 3

In each video, all 8 people will start out waiting in a line in front of one of the water sources. People can leave the line and head straight for the water source without waiting, or they can stay waiting in line until it's their turn to collect water. The videos you see will all involve one person leaving the line and heading straight for the water source. Your job is to think about what would happen if EVERYONE felt at liberty to leave the line and try to get water faster for themselves, instead of everyone staying in line. Feel free to watch the video many times, but please watch it at least once completely and think carefully before you make your judgments.

Answer the following comprehension check question. The video shows only one person leaving the line. Your job is to think about what would happen if everyone felt at liberty to do — instead of —? (Fill in the blank)

After the experiment is over, you will have an opportunity to give us general feedback and let us know if anything was confusing or unclear.

2 Main Study

After reading the instructions and answering the attention checks described above, participants proceeded to the main study.

2.1 Moral Judgment Group: Procedure

Each participant saw all 39 maps presented in a random order and answered the following question about each map (see Fig 3).

“Think about what this character is doing and make a judgment.” Participants answered on a scale ranging from completely morally unacceptable (-50) to completely morally acceptable (50)”

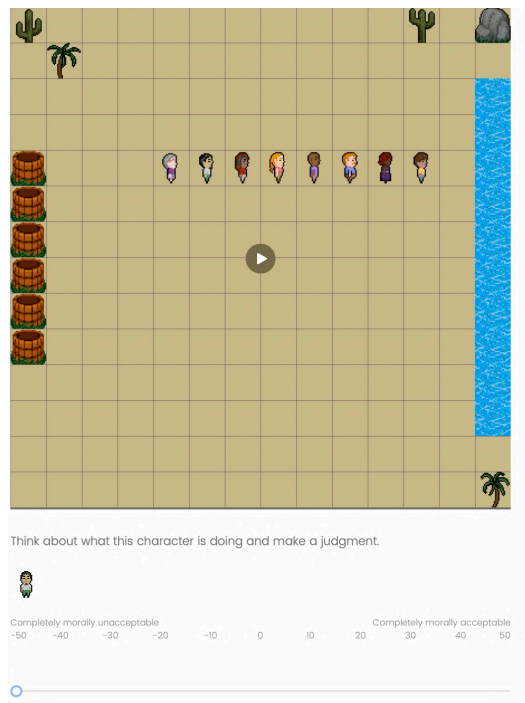


Figure 3: Screen shot of the start of a trial for participants in the Moral Judgment Group.

2.2 Universalization Group: Procedure

Each participant saw all 39 maps presented in a random order and answered the following question about each map (see Fig 4).

“What if everyone felt at liberty to leave the line and try to get water faster for themselves? How much better or worse off would everyone be compared to if they all stayed in line instead?” Participants answered on a scale from much worse off (-50), to the same (0), to much better off (50).

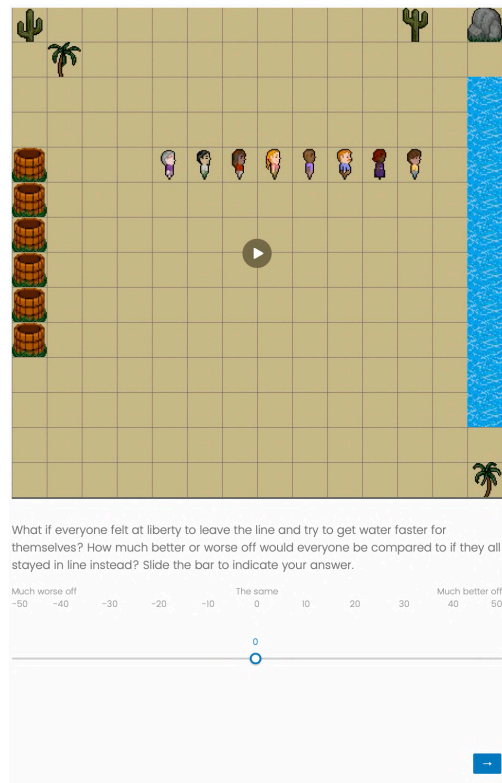








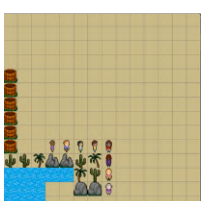
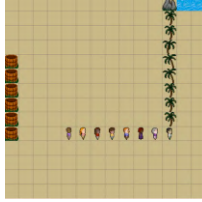



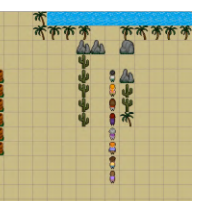









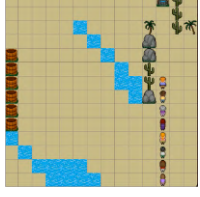


Figure 4: Screen shot of the start of a trial for participants in the Universalization Group.

3 Stimuli

This section contains screen shots of the starting scenes of all maps. Map names correspond to names used in Fig 3-5. Note that some maps have identical starting states (e.g., Maps 5, 17, and 29) but are treated as different contexts because the action unfolds differently in each environment.

Starting Scene	Map Name	Starting Scene	Map Name	Starting Scene	Map Name
	1		6		11
	2		7		12
	3		8		13
	4		9		14
	5		10		15

Starting Scene	Map Name	Starting Scene	Map Name	Starting Scene	Map Name
	16		21		26
	17		22		27
	18		23		28
	19		24		29
	20		25		30

Starting Scene	Map Name	Starting Scene	Map Name
	31		36
	32		37
	33		38
	34		39
	35		