The emergence of categorical norms

Erez Yoeli[1†], N. Aygun Dalkiran[2†], Bethany Burum[3†], Martin A. Nowak[4], and Moshe Hoffman[5*]

[1]Sloan School of Management, Massachusetts Institute of Technology, E94-1502C, 245 First St., Cambridge, MA 02139

[2]Department of Economics, Bilkent University, Ankara, Turkey 06800

[3]Department of Psychology, Harvard University, Cambridge, MA 02138

[4]Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, Cambridge, MA 02138

[5]Max Planck Institute for Evolutionary Biology, August-Thienemann-Straße 2 24306 Plön

[†,‡]These authors contributed equally to this work [*]hoffman.moshe@gmail.com

Abstract

Why are norms unduly sensitive to categorical distinctions compared to continuous variation? For instance, the norm against the use of chemical weapons considers the type of weapon used, not how much suffering was caused; human rights are conditioned on membership in the species *homo spaiens* not on an animal's degree of sentience; norms promoting philanthropy emphasize donating, and are relatively insensitive to the impact a donation will have. Here we present a game theoretic model, which explains why it is easier for norms to be conditioned on categorical distinctions rather than on continuous variation. We explore the robustness of our results and present evolutionary simulations. Then, in a series of experiments, we demonstrate that participants' moral intuitions and willingness to sanction norm violations are influenced by categorical distinctions rather than by continuous variation. We show that the reliance on categorical distinctions weakens when norm enforcement plays less of a role. We discuss various applications, including territoriality, human rights, inefficient altruism, institutionalized racism, and collusion.

Many norms and corresponding moral intuitions depend on categorical distinctions, even though one would expect us to attend moreso to the underlying continuous variation. For instance, norms against the usage of chemical and biological weapons depend on the category of weapon used, even though these norms are presumably motivated by the desire to reduce wanton death or misery. Human rights are granted based on species membership when one might think they would depend on sentience or ability to feel pain. Racist norms, like the 'one-drop rule', often depend on the existence of a known black ancestor, and not the shade of one's skin, or the share of black ancestors. Prohibitions against stealing and murder primarily consider whether another's property was taken or the victim perished, and are relatively insensitive to the value of

the item, neediness of the perpetrator, life-years lost, *etc.* Prescriptions to "do a good turn daily" consider whether a dead was done, not how much good it led to.

Standard models of norms and of our corresponding moral intuitions [1–4], do not explain this seemingly undue dependence on categorical distinctions. Such models typically share the characteristic that arbitrary norms can be sustained, provided norm violators are sanctioned and sanctions are themselves incentivized [2, 5, 6]. Why, of all the arbitrary norms that can be sustained in equilibrium, would we so frequently find ourselves at categorical ones? This is even more puzzling once we incorporate mechanisms that select for efficient norms, such as cultural group selection [5] or deliberative agents who rationally select among norms [7]. Categorical norms are, necessarily, less efficient than norms that condition on continuous variables, which would permit, e.g., the use of chemical weapons only when they are more effective or humane than conventional ones, as Franklin Roosevelt's military advisors argued was the case in Iwo Jima [8], or theft in the case of a particularly needy thief, as prominent economists have advocated [9].

We propose the answer lies with the need for players to coordinate with respect to what counts as a transgression, which is an implicit feature of standard models [1–4]. This makes it difficult to sustain 'threshold norms' in which a transgression is counted only when the continuous variable of interest exceeds a certain threshold (e.g., only once the number of civilian casualties exceeds 100,000, or only if a thief did not sufficiently need the item he or she stole) because when one believes the threshold has just barely been crossed, one cannot be sufficiently sure others believe the same. This leads threshold norms to unravel, and for norms that depend on categorical variables, which do not suffer the same fate, to arise in their place.

We will substantiate this proposition first with the help of a simple game theory model and then with experimental evidence. The game theory model is adapted from the literature on Global Games [10, 11]. To gain a deeper sense of what our results depend on and how robust they are, we will consider various extensions of our model, and present a general theorem that characterizes exact conditions on the distribution of signals and payoffs under which threshold norms cannot be sustained. We will also present evolutionary game theory simulations to demonstrate the instability of threshold norms, and the unraveling alluded to above. Then, we will present evidence for our explanation from a series of experiments. In these, we verify that norms and moral intuitions are (much) more sensitive to categorical distinctions than continuous ones, and find that categorical distinctions are relatively more important when it comes to deterring future transgressions or avoiding the opprobrium of others enforcing the norm. We also find that people are able to rely on continuous variation when norm enforcement plays less of a role. Finally, we will discuss additional applications and relate our results to the existing literature.

Theory

We now present a simple Bayesian Game [3] to formalize the above intuition (Fig 1), which we call 'coordination with private signals'. The game begins when a state of the world is chosen from a set of possible states, according to some 'prior' probability. Players do not observe the state directly. Instead, they each receive an independent signal, chosen from a set of possible signals, according to a distribution that de-

pends on the state. Together, the states, signals, and their respective distributions are known as a state-signal structure.

After observing their signal, players choose their action in a standard coordination game, which is meant to characterize the need for players to coordinate on what they view as a transgression. In the coordination game, each of two players chooses between two actions, which we label S and N. Each player gets payoff a if she chooses S when the other also chooses S, $b < a$ if she chooses S when the other does not choose S, d if she chooses N when the other chooses N, and $c < d$ if she chooses N when the other does not choose N (Fig. 1a). The parameter $p = (d - b)/(a - c + d - b)$ will prove useful. Its interpretation is: if either player expects the other to play S with probability greater than p, then she prefers to play S herself. We readily admit that the coordination game is an oversimplification of the real-world norms we are modeling. We use it to capture an essential feature of sanctioning in norm enforcement models: it is worth sanctioning (playing S) if and only if others expect you to. This is the case, for instance, if you are liable to be sanctioned or exploited if others believe you witnessed a transgression that you did not sanction. This is a common feature of norm enforcement (and of repeated games, more generally, that are used to model the source of cooperative behavior) [1, 3, 4].

We assume payoffs are independent of states and signals. This assumption, is motivated by the fact that 'third-parties' pay costs to sanction others for norm violations, even though these costs are unaltered by the violation [1–5, 12–14].

A strategy in the game specifies which action players will choose in this game for each signal they might receive. We solve such models using the standard solution concept for such settings: Bayesian Nash equilibria (BNE) [3]. BNE is an extension of Nash equilibria (NE) to settings with private information. A strategy profile (specification of each players' strategy) is said to be a NE if each player i's strategy maximizes her payoffs, taking as given that others, $-i$ will play according to their specified strategy. When there is private information, a strategy profile is said to be a BNE if each player's strategy maximizes her expected payoffs, given the signal she received and expects others to have received, applying Bayes' Rule as necessary.

The first state-signal structure we consider in detail models the possibility of receiving continuous signals (Fig. 1.b). For this, we assume the state of the world is randomly drawn from $[0, 1]$ according to the uniform distribution, and each player's signal is independently distributed uniformly within $\epsilon > 0$ of the state. The state of the world might represent the proportion of civilians harmed by a despot and players' signals might represent each country's best assessment of this proportion.

For the continuous case, we consider strategy profiles in which players play according to one strategy for all signals at or below some $\bar{s}$ and another for all signals above it. These are intended to correspond to, e.g., sanctioning only if one's assessment of the harm caused by the despot exceeds a certain threshold. We find that such 'threshold' strategies cannot be a BNE, except in the non-generic case where $p = 0.5$. To see this, note that such a threshold strategy profile requires each player be indifferent when she observes $\bar{s}$, but, at $\bar{s}$ each player believes the other player received a signal below or above $\bar{s}$ with equal probability, and so she expects the other to play S with probability exactly equal to .5, leaving them indifferent between playing

S and N if and only if $p = 0.5$ (fig. **??**a). In the S.I., we prove this result, in fact, holds for any continuous distribution with symmetric posteriors.

The next state-signal structure we consider models the possibility of receiving categorical signals (Fig. 1.c). Now, we assume the state is $\omega = 1$ with probability q and 0 with probability $1 - q$. The players receive the correct signal, $S = \omega$, with probability $1 - \epsilon$, and the incorrect signal, $S = 1 - \omega$, with probability ϵ. The state of the world might represent whether the despot used chemical weapons on his citizens; again, players' signals might represent each observer nation's own intelligence assessment.

For this case, it can be a BNE for players to play S if and only if they receive signal $S = 1$ provided the signals are sufficiently accurate, as determined here by ϵ (Fig. **??**b). When a player receives a signal $S = 1$, she is better off choosing S so long as she believes the other's likelihood of also receiving the same signal is at least $1 - p$, and, when a player receives a signal $S = 0$, she is better off choosing N so long as she believes the others' likelihood of receiving the same signal is at least p. Both of these conditions will hold for sufficiently small ϵ.

We next consider to what extent our results generalize from these simplest cases.

We first consider what happens if we allow the state to take on large but finitely many (n) values instead of a continuum. We find that the larger n gets, the closer p needs to be to 50% in order to allow for a 'threshold equilibrium' (Fig. 2c). Thus, as n increases, the range of coordination games over which such equilibria exist approaches measure 0 as in our continuous case. That is, even though the state space is discrete, once it can take on sufficiently many possible values, the results approximate what happens when the state space is continuous (see S.I. Sec. B2).

The intuition is as follows: for an equilibrium to exist where players sanction if they get a signal above a certain value, then whenever player i gets a signal close to this value, but above it, i must believe that $-i$ has gotten a signal above this threshold with probability greater than or equal to p. For any n and ϵ, we can use Bayes' Rule and the uniform distribution to calculate players' posterior beliefs that $-i$ has also gotten a signal on the same side of the threshold when she herself obtains signals just above the threshold and just below it. For instance, these are $\frac{1}{3}$ and $\frac{2}{3}$ when $n = 10$ and $\epsilon = .3$, respectively. As n increases, the gap between these beliefs shrinks (Fig. 2d). Note that in the S.I., we prove a comparable result for signals distributed uniformly over a 'coarse' subset of the real line, varying the degree of coarseness.

Next, instead of assuming the state is uniformly distributed, we assume that the state is normally distributed about some mean μ, and signals are normally distributed about the state (S.I. Sec. B2). Now, a player no longer expects the other player's signal to be equally likely to be above as below her own signal. Instead, the likelihood that the other player's signal is above one's own monotonically decreases with one's signal. This probability approaches 1 for signals that are arbitrary large, and 0 for signals that are arbitrarily small (hitting .5 exactly at the mean μ). Consequently, for every value of p , there will be a single threshold value, $\bar{s}$, that is sustainable in equilibrium (Fig. SI3). Thus, a threshold equilibrium will be possible when the state is normally distributed. Crucially, though, the location of the threshold cannot be freely chosen and instead is pinned down by the variances of the state and signal distributions, and by p (see Thm. 4 in S.I. Sec. **??**).

Under what conditions does a state-signal structure permit a threshold equilibrium? To answer this more general question, we provide the following theorem. We continue to restrict our attention to cases where there are two players, states are ordered, higher signals are indicative of higher states, and payoffs do not directly depend on the state or signals (see S.I. Sec. B5 for details). We find that a threshold strategy is an equilibrium if and only if p falls between a player's posterior belief that the other's signal is at or below the threshold when her own signal is at the threshold, and the player's posterior belief that the other's signal is above the threshold when her own signal approaches the threshold from above. Formally, a threshold profile $\sigma^{\bar{s}}$ is a Bayesian Nash equilibrium if and only if

$$Pr(S_{-i} > \bar{s}|S_i = \bar{s}) \leq p \leq lim_{s \to \bar{s}^+} Pr(S_{-i} > \bar{s}|S_i = s). \quad (1)$$

Here S_i represents the signal of player i, S_{-i} represents the signal of the other player, $-i$. Notice: the more 'coarse' the distribution of signals are the larger the gap between these two inequalities, allowing for a wider the range of p values where this particular threshold can be used in equilibrium. It is in this sense that our main result generalizes: coarser signal distributions enable specific thresholds to be equilibria for a wider range of coordination games.

In reality, strategy profiles may not be at equilibrium, and players may not be behaving optimally. What would happen if a strategy profile starts off at a threshold and then adjusts according to an evolutionary process? We ran computer simulations that explore this possibility. We employ the Wright-Fischer evolutionary dynamic [15, 16], in which players' strategies update each period according to the payoffs they earned the previous period, with more successful strategies growing in frequency. A small fraction of 'mutants' are randomly assigned different thresholds each period (see S.I. Sec. C for details).

In simulations that employ our continuous setup, we find that the average threshold in the population steadily moves up if $p < .5$ (down if $p > .5$), until eventually everyone plays S for all signals (everyone plays N for all signals). The intuition is as follows. When players in a population are playing according to a given threshold strategy, for a small range of signals near the threshold, the posterior probability that the other player's signal is on the same side of the proposed threshold fall short of p. Mutants with slightly lower threshold strategies then obtain higher payoffs, and their strategy grows in frequency, causing the population's average threshold to slowly migrate down until everyone always plays S. When the posteriors at the proposed threshold exceed p, the logic is analogous, except that the threshold unravels in the opposite direction. Fig 3a presents the case of $p > .5$. A single representative run of our simulations is presented on the left; the average outcomes from many such runs is presented on the right. These results are expected since no threshold strategy is a Nash equilibrium, and Nash is generally a necessary, but not sufficient condition to for stability [17, 18].

Figs. 3b and 3c present simulations for the variation of the model with $N = 10$ discrete levels. In Fig. 3b, p is outside the range for which a threshold is a BNE, and the threshold unravels, much as it did in the continuous case. However, in Fig. **??**c, p is inside the range for which a threshold is a BNE. In this case, the threshold stabilizes. This is perhaps not surprising, since this threshold is a strict Nash equilibrium, and strict Nash are typically evolutionarily stable [19]. Finally, Fig. 3d presents simulations for the variation of

the model where the state is normally distributed. No threshold strategy stabilizes, even when the proposed threshold is a BNE, albeit not a strict Nash allowing for evolutionary instability.

Experimental Evidence

We next provide experimental evidence for our explanation for categorical norms using two paradigms: a series of five naturalistic vignettes, and two incentivized economic games.

In our vignette studies, we first verified that participants treat norms as categorical when judging how "good" or "bad" an action is, and when proscribing punishments or rewards. We chose a range of scenarios closely related to those that are often sanctioned or rewarded via norms. For each, we showed participants one of eleven versions, contrasting no transgression (or good deed) with a transgression (or good deed) that had one of ten levels of impact: a government killing no political protesters versus 1-10, a country adding military to its own territory versus attacking another country's territory that had a population of 1,000-10,000 (in increments of 1,000), a person thinking about shoplifting versus actually shoplifting an item worth \$2-\$20 (in increments of \$2), a person sacrificing his new car to save a man's life versus sacrificing the man's life to save his new car, costing the man his 1-10 more years to live, and a person opting out of volunteering at a soup kitchen versus volunteering for 1-10 hours a week. Across all five vignettes, we found that participants' moral judgments and punishment (or reward) proscriptions increased sharply from no transgression to a small transgression (or good deed), then stayed relatively flat as the magnitude of the transgression increased multiple-fold (see Fig. **??**a).

Second, we used the protest, territorial attack, and shoplifting vignettes to demonstrate that participants expect categorical distinctions to be most important for deterring future transgressions. Participants believed that sanctions were important for deterring governments from killing more protesters, military defense for deterring other nations from attacking more of one's territories (including more valuable territories), and legal consequences for deterring shoplifters from repeat offending. But they believed punishing smaller transgressions was just as important for deterrence as punishing transgressions with ten times the impact (see Fig. **??**b).

Third, in these same three vignettes we showed that participants expected avoiding higher-order punishment (being punished for failing to punish norm violations–a common feature of norm enforcement [5]) or gaining higher-order rewards to depend on punishing a transgression, but not on the transgressions' magnitude. For example, participants expected third party nations to be more supported by their allies and rated more highly by a human rights organization if they sanctioned a government that executed at least one protester than if they failed to sanction, but to gain no extra praise for sanctioning—or censure for not sanctioning—if the government killed ten times the people. (See Fig. **??**c.)

Our incentivized economic games corroborated that people treat norms as categorical, using a very different paradigm, with money on the line. Participants were told of another person (the Roller) who had made a fair or unfair decision that had a relatively large or small impact. Specifically, this Roller chose to roll either a die with three sides favoring himself and three favoring his partner (fair), or a die with five sides

favoring himself and only one favoring his partner (unfair), the Recipient. The Recipient stood to lose either half (lower impact) or all (higher impact) of their bonus if the roll favored the Roller. (All manipulations were between participants.) Our participants were third party observers who could pay money to take money from (or give money to) the Roller. Participants paid to punish the Roller more (and therefore reward less) when the Roller had chosen unfairly, but their punishments were unaffected by the level of impact that this decision would have (see Fig. 6a).

We also used this paradigm to show that participants are capable of taking into account continuous variation when coordination plays less of a role. We expect punishment decisions to typically be shaped in contexts involving coordinated norm enforcement, while personal avoidance requires little coordination. Instead of paying to punish, a second group of participants was randomly assigned to be the Roller's Recipient in a second round of the same "game", in which the Roller would pick a die and they stood to lose money depending on the roll. However, these participants could pay to opt out of this interaction with the Roller. Now that participants were making incentivized choices about how much they were willing to pay to avoid being at the receiving end of a transgression, they were sensitive to impact–participants were willing to pay more to avoid an unfair Roller, but critically were also willing to pay more to opt out when the die roll would have had a larger impact on them (see Fig. 6b).

For additional evidence that participants are capable of taking continuous variation into account when norms and coordination play less of a role, we reference the results from an earlier paper of ours [20] which focuses on another domain: charitable giving. In these studies, which involve both vignettes and incentivized economic games, participants were, again, insensitive to continuous variation in the domain of charitable giving, but quite responsive to analogous continuous variation in other settings, like investment decisions, which are less influenced by coordination and norms. Moreover, other participants' willingness to reward potential donors was again attuned to whether donors gave, but not to how much they gave or how impactful their gifts were.

Discussion

We now discuss some possible applications.

We begin with the puzzle that we motivated the paper with: Why is there a norm against chemical weapons [8, 21], instead of a norm that merely sanctions needless death or suffering? To the extent that international sanctions require coordination—say, because sanctions are only effective if others are also sanctioning, or because failing to sanction when expected is penalized—then it will be comparatively hard to condition sanctioning on, say, having killed more than a certain number of civilians. In contrast, a norm against chemical weapons is easier to sustain, provided each countries' assessment of whether or not a chemical weapon was used is sufficiently unambiguous and false positives are unlikely.

Why do we apply human rights on the basis of membership in the species Homo Sapiens? Why not, instead, assign rights on the basis of, say, sentience or ability to feel pain, which might, for example, lead a chimpanzee to have more rights than a comatose human? The philosopher Peter Singer refers to this

as a bias, which he terms speciesism [22], and equates to other biases like racism. However, if human rights are enforced by coordinated sanctions, say by the coordinated activity of revolutionaries, or of foreign governments, it makes sense that human rights depend on the categorical variable of species membership and not the continuous variable of sentience or ability to feel pain; that makes the norm sustainable.

Likewise, one might wonder why rights are viewed as absolute, immune to trade-offs and off limits to cost-benefit calculations. Again, this makes sense if we understand that violation of rights, say by using torture, is a categorical transgression, and hence is a sustainable norm. For contrast, consider a more utilitarian norm which, say, allows for torture when it prevents more suffering than is inflicted. Such a norm is not sustainable because it depends on a continuous variable: the amount of suffering inflicted by the torture and the amount of suffering prevented by the discovery of the hypothetical ticking bomb.

Why do incursions on a territory of little strategic or economic significance (e.g., the Falkland Islands, or the disputed islands in the South China Sea) invite full-scale retaliation and risk war? Why do nations not retaliate only when the incursion is sufficiently meaningful? If retaliation is modeled as punishment in a repeated game, then punishment is incentivized by the need to deter similar transgressions, which involves coordination, namely on what counts as similar and what type of transgression will incur retaliation. Under these assumptions, it is not possible for retaliation to be triggered by the significance of the incursion–a continuous variable–but only by the presence of an incursion–a categorical variable.

Another application, which we alluded to in our discussion of experimental evidence, is ineffective altruism. People donate a meaningful amount of time and money to charity [23], but are less than careful about ensuring these charities are effective: most donors report that they do not even check a charity's effectiveness before donating [24] and highly ineffective charities persist [25]. One possible reason for this is that efficacy is a continuous variable, but the act of giving is categorical. This makes it possible to have norms that promote charitable giving while making it hard to sustain norms that promote effective giving [20].

Discrimination is often norm enforced, as it arguably was in the Jim Crow South, when mobs of 'third parties' would e.g., punish white-owned businesses that served blacks via boycotts, demonstrations, and violence. In such cases, discrimination is typically based on categorical distinctions such as sex, caste, or having 'one drop' of African blood.

Collusion, like norms and discrimination, is often enforced via coordinated punishment in repeated interactions [26, 27]. Such collusive arrangements may be easier to sustain if they depend on categorical variables like 'only sell through De Beers' than on continuous variables like quantity, price, or costs, particularly when these are not publicly observed.

Standard models of crime and punishment prescribe that punishment ought to increase proportionately to the gains from crime and inversely to the probability of getting caught [**?**]. However, the gains from crime and the probability of getting caught are both continuous variables that are typically somewhat privately assessed. This may explain why these variables do not play as great a role as one might expect in punitive moral intuitions [28], and perhaps also why these models were not intuitive, and remain controversial.

Next, we provide additional evidence for our model by considering what happens when its assumptions

do not hold–an example of a 'comparative static'. Specifically, our model suggests that the reliance on categorical norms is rooted in two problems: the need to coordinate, and the reliance on private information. When these problems are alleviated, we do not expect as much reliance on categorical distinctions. Consistent with this prediction, when coordination is not so pertinent, we do observe conditioning on continuous variables. Choosing who to date or hire is primarily a personal optimization problem–one wants the best partner or employee, largely independently of what others think. In these contexts, we do in fact discriminate based on age [29], attractiveness [9, 30], skin tone [31], weight [32], and height [33], which are continuous variables. Domestic laws are enforced unilaterally by judges or police officers, without the same need for coordination as in international law. Domestic laws do condition on thresholds in continuous variables, as is the case when abortion is precluded after a certain date, or pollution is forbidden above a certain concentration [34]. Altruism towards kin is driven, ultimately, by shared genes, not by reciprocity or norm enforcement. Therefore, altruism toward kin is also more sensitive to the continuous variables of impact and efficacy [20].

When information is commonly observed or easily conveyed, there is also no problem conditioning on continuous variables, even if there remains an element of coordination. Sharecropping norms [35] could fixate at a particular share–a threshold in a continuous variable–because crops presumably were publicly observable. Tipping norms can fixate at 20% because a record of one's tip can be easily verified or shared, if the need for shaming arises. The railroad cartel of the late 1800s was able to use a threshold price to trigger enforcement of its collusive agreement because prices were publicly posted [36, 37]. One benefit of e.g., mediators [38, 39], public trials, and common weights and measures, may be to enable conditioning on continuous variables.

We think there is one more piece of evidence that provides some support for our model: even minor, seemingly insignificant breaches of a categorical distinction that sustains an equilibrium can cause that equilibrium to fall apart. One illustrative example comes to us from the rapid collapse of the norm against bombing cities from aircraft that held for the first months of World War II. This is generally said to have occurred after the accidental bombing of London by German bombers whose intended target were oil storage installations on August 24, 1940. Churchill retaliated by ordering a bombing raid on military targets inside Berlin. Despite minimal casualties, Hitler responded by rescinding the ban on bombing towns and cities, and ordering wide-scale attacks on London and other British cities [40]. By the 15th and 16th of December, when the British area-bombed Mannheim, the norm against bombing civilian targets had fully collapsed. Such episodes, and the logic we set forth herein, provide support for concerns regarding 'slippery slopes'. They also might explain why it's common for lawyers to purposely select sympathetic test cases, which breach a category barrier and lead a categorical norm to unravel, as in *Frontiero v. Richardson* and *Weinberger v. Wiesenfeld* which challenged gender discrimination laws that, on average, harmed women but which, in these cases, harmed men. Once the court ruled in favor of the men, this set a precedent that applied more broadly, leading the categorical norm to unravel.

We next provide two comments on the interpretation of the model. First, although we have focused on norms, as some of our applications likely make clear, we expect our results to extend to any game in which

coordination plays a role. This will be true for any game with multiple equilibria, such as the Hawk-Dove game and the Repeated Prisoners' Dilemma, which have been used to model animal territoriality [41], our sense of rights [42, 43], and dyadic relations [3, 44]. Thus, for instance, we expect animal territoriality to only condition on categorical variables like 'who arrived first' or 'who built the nest', and not on continuous variation like which animal appears larger or more desperate.

Likewise, we can ask: what kind of events trigger revolutions and protests? Consider the American Revolution. One turning point was the Boston Tea Party, a protest launched not after a gradual tax hike, but after categorically new tax–the Tea Tax–was imposed. Interestingly, this tax was imposed concurrently with other tax reductions which led to an overall lower tax rate [45]. Clearly the actual tax rate (continuous) did not foment the revolution; instead it was the addition of a new tax (categorical). Likewise, the Arab Spring–a series of political revolutions in North Africa–was launched by the public self immolation of the Tunisian street vendor Mohamed Bouazizi, not, directly at least, the increasing poverty, abuse, or corruption in the Tunisian government that led Bouazizi to act.

Second, we emphasize that the difference between the continuous and categorical cases depends on a difference in the information available to agents–what signals they have and can condition on–and not on how this information might subsequently be interpreted or described, and, in particular, categorized after the fact. That is, when we say the signal has to be categorical to support coordination, we do not mean that the signal can still be continuous (the number of people who were killed) so long as it is interpreted in a categorical way (was this number more than 100,000?). In this case, players would still be better off deviating at signals close to 100,000. Perhaps counter-intuitively, the fact that categorical signals contain less information–and, specifically, make it impossible to obtain signals near a threshold–is what makes it possible to condition on them.

We conclude with a discussion of related literatures. The literatures on repeated games [3, 44], norm enforcement [1, 46], and indirect reciprocity [4] motivated many of our applications. To our knowledge, these literatures do not home in on the important role of coordination, and its implications for categorical distinctions. Some modelers have noted that conventions like sharecropping and tipping rules often fall on non-random values, like two-thirds or 20 percent [35, 47]. On first inspection, these examples appear to contradict our model, but, as discussed above, we believe they can be reconciled with our model by noting that such conventions apply to situations where information is either commonly observed or can be easily verified. Indeed, the models used to explain these conventions typically assume that players observe a common signal.

Another closely related literature inspired our setup: "global games" [10, 11]. These models differ from ours in that the state does influence payoffs, and it is never commonly known that players are playing a coordination game. Consequently, the typical result in that literature highlights is the role of idiosyncratic noise in reducing the multiplicity of equilibria, whereas our results highlight the role of idiosyncratic noise in precluding threshold equilibria. Our focus on both the coarseness of signals and on state-independent payoffs is, as far as we know, unique.

A literature in psychology explores the ways that people tend to categorize [48], which provides an

alternative explanation for why people might pay undue attention to categorical distinctions: they are easier to notice, remember, measure, *etc.* Our analysis suggest an additional reason for categorization, which we believe is needed to explain its heightened occurrence in settings involving coordination and private information.

Our work contributes to a growing literature that uses insights from game theory to explain puzzling aspects of our moral intuitions and otherwise puzzling social behaviors [4, 35, 42–44, 47, 49–60], as well as the literature that uses laboratory experiments to provide evidence for such models [12, 20, 42, 55, 61–68].

Acknowledgements

We thank Andrew Ferdowsian for research assistance.

Conflicts of Interest

The authors have no conflicts of interest to disclose.

References

[1] Panchanathan, K. & Boyd, R. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* **432**, 499–502 (2004).

[2] Fudenberg, D. & Maskin, E. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica: Journal of the Econometric Society* 533–554 (1986).

[3] Osborne, M. J. & Rubinstein, A. *A course in game theory* (MIT press, 1994).

[4] Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).

[5] Boyd, R. *A different kind of animal* (Princeton University Press, 2017).

[6] Henrich, J. & Muthukrishna, M. The origins and psychology of human cooperation. *Annual Review of Psychology* **72**, 207–240 (2021).

[7] Pinker, S. *The better angels of our nature: The decline of violence in history and its causes* (Penguin UK, 2011).

[8] Harris, R. & Paxman, J. *A higher form of killing: the secret history of chemical and biological warfare* (Random House Incorporated, 2002).

[9] Becker, G. S. A theory of marriage: Part i. *The Journal of Political Economy* 813–846 (1973).

[10] Carlsson, H. & Van Damme, E. Global games and equilibrium selection. *Econometrica: Journal of the Econometric Society* 989–1018 (1993).

[11] Morris, S. & Shin, H. S. Global games: Theory and applications. *Advances in Economics and Econometrics* 56 (2001).

[12] Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evolution and human behavior* **25**, 63–87 (2004).

[13] Bicchieri, C. *The grammar of society: The nature and dynamics of social norms* (Cambridge University Press, 2005).

[14] Bicchieri, C. *Norms in the wild: How to diagnose, measure, and change social norms* (Oxford University Press, 2016).

[15] Ewens, W. J. *Mathematical population genetics: theoretical introduction*, vol. 1 (Springer, 2004).

[16] Imhof, L. A. & Nowak, M. A. Evolutionary game dynamics in a wright-fisher process. *Journal of mathematical biology* **52**, 667–681 (2006).

[17] Weibull, J. W. *Evolutionary game theory* (MIT press, 1997).

[18] Nowak, M. A. *Evolutionary dynamics: exploring the equations of life* (Harvard university press, 2006).

[19] Hofbauer, J., Sigmund, K. *et al. Evolutionary games and population dynamics* (Cambridge university press, 1998).

[20] Burum, B., Nowak, M. & Hoffman, M. An evolutionary explanation for ineffective altruism. *Nature Human Behavior* (Forthcoming).

[21] The Economist. The history of chemical weapons: The shadow of Ypres. `http://www.economist.com/news/briefing/21584397-how-whole-class-weaponry-came-b` (2013). Online; accessed October 16, 2016.

[22] Singer, P. *The expanding circle* (Clarendon Press Oxford, 1981).

[23] Trust, N. P. Charitable giving statistics. `http://www.nptrust.org/philanthropic-resources/charit` (2014). Online; accessed February 13, 2014.

[24] Baltussen, R. M., Sylla, M., Frick, K. D. & Mariotti, S. P. Cost-effectiveness of trachoma control in seven world regions. *Ophthalmic epidemiology* **12**, 91–101 (2005).

[25] Times, T. B. America's 50 worst charities rake in nearly $1 billion for corporate fundraisers. `http://www.tampabay.com/news/business/americas-50-worst-charities-rake-in-nearl` (2013). Online; accessed February 16, 2014.

[26] Stigler, G. J. A theory of oligopoly. *The Journal of Political Economy* 44–61 (1964).

[27] Carlton, D. W. & Perloff, J. M. *Modern industrial organization* (Pearson Higher Ed, 2015).

[28] Carlsmith, K. M., Darley, J. M. & Robinson, P. H. Why do we punish? deterrence and just deserts as motives for punishment. *Journal of personality and social psychology* **83**, 284 (2002).

[29] Macnicol, J. *Age discrimination: An historical and contemporary analysis* (Cambridge University Press, 2006).

[30] Mobius, M. M. & Rosenblat, T. S. Why beauty matters. *The American Economic Review* **96**, 222–235 (2006).

[31] Goldsmith, A. H., Hamilton, D. & Darity Jr, W. Shades of discrimination: Skin tone and wages. *American Economic Review* **96**, 242–245 (2006).

[32] Baum, C. L. & Ford, W. F. The wage effects of obesity: a longitudinal study. *Health Economics* **13**, 885–899 (2004).

[33] Case, A. & Paxson, C. Stature and status: Height, ability, and labor market outcomes. *Journal of Political Economy* **116** (2008).

[34] Agency, U. E. P. Drinking water requirements for states and public water systems. `https://www.epa.gov/dwreginfo/chemical-contaminant-rules`. Online; accessed September 17, 2017.

[35] Young, H. P. The economics of convention. *Journal of economic perspectives* **10**, 105–122 (1996).

[36] Porter, R. H. A study of cartel stability: the joint executive committee, 1880-1886. *The Bell Journal of Economics* 301–314 (1983).

[37] Grossman, P. Z. Why one cartel fails and another endures: The joint executive committee and the railroad express. *How cartels endure and how they fail: Studies of industrial collusion* 111–129 (2004).

[38] Boyd, R. & Mathew, S. Arbitration supports reciprocity when there are frequent perception errors. *Nature Human Behaviour* **5**, 596–603 (2021).

[39] Singh, M. & Garfield, Z. H. Evidence for third-party mediation but not punishment in mentawai justice. *Nature Human Behaviour* 1–11 (2022).

[40] Read, A. *The devil's disciples: Hitler's inner circle* (WW Norton & Company, 2004).

[41] Smith, J. M. *Evolution and the Theory of Games* (Cambridge university press, 1982).

[42] DeScioli, P. & Wilson, B. J. The territorial foundations of human property. *Evolution and Human Behavior* **32**, 297–304 (2011).

[43] DeScioli, P. & Karpoff, R. People's judgments about classic property law cases. *Human Nature* **26**, 184–209 (2015).

[44] Trivers, R. L. The evolution of reciprocal altruism. *Quarterly review of biology* 35–57 (1971).

[45] Thorndike, J. J. Four things you should know about the boston tea party. http://www.taxhistory.org/thp/readings.nsf/ArtWeb/1BB0C8F894BB490B852577020083A Online; accessed December 28, 2017.

[46] Boyd, R. How humans became outliers in the natural world. http://econ.as.nyu.edu/docs/IO/41614/Outliers.pdf. Online; accessed October 16, 2016.

[47] Young, H. P. The evolution of conventions. *Econometrica: Journal of the Econometric Society* 57–84 (1993).

[48] Harnad, S. Categorical perception. In *Encyclopedia of Cognitive Science*, vol. 67:4 (MacMillan: Nature Publishing Group, 2003).

[49] Frank, R. H. *Passions within reason: the strategic role of the emotions.* (WW Norton & Co, 1988).

[50] Binmore, K. G. The evolution of fairness norms. *Rationality and Society* **10**, 275–301 (1998).

[51] Binmore, K. *Natural justice* (Oxford University Press, 2005).

[52] Axelrod, R. M. *The evolution of cooperation* (Basic books, 2006).

[53] Pinker, S., Nowak, M. A. & Lee, J. J. The logic of indirect speech. *Proceedings of the National Academy of sciences* **105**, 833–838 (2008).

[54] Lee, J. J. & Pinker, S. Rationales for indirect speech: the theory of the strategic speaker. *Psychological review* **117**, 785 (2010).

[55] DeScioli, P., Christner, J. & Kurzban, R. The omission strategy. *Psychological Science* **22**, 442–446 (2011).

[56] DeScioli, P., Gilbert, S. S. & Kurzban, R. Indelible victims and persistent punishers in moral cognition. *Psychological Inquiry* **23**, 143–149 (2012).

[57] DeScioli, P. & Kurzban, R. A solution to the mysteries of morality. *Psychological Bulletin* **139**, 477 (2013).

[58] Chwe, M. S.-Y. *Rational ritual: Culture, coordination, and common knowledge* (Princeton University Press, 2013).

[59] Hoffman, M., Yoeli, E. & Nowak, M. A. Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences* **112**, 1727–1732 (2015).

[60] Hoffman, M., Yoeli, E. & Navarrete, C. D. Game theory and morality. In *The Evolution of Morality*, 289–316 (Springer, 2016).

[61] Barclay, P. Trustworthiness and competitive altruism can also solve the "tragedy of the common". *Evolution and Human Behavior* **25**, 209–220 (2004).

[62] Barclay, P. Reputational benefits for altruistic punishment. *Evolution and Human Behavior* **27**, 325–344 (2006).

[63] Barclay, P. & Raihani, N. Partner choice versus punishment in human prisoner's dilemmas. *Evolution and Human Behavior* **37**, 263–271 (2016).

[64] Jordan, J. J., Hoffman, M., Nowak, M. A. & Rand, D. G. Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences* **113**, 8658–8663 (2016).

[65] Jordan, J. J., Hoffman, M., Bloom, P. & Rand, D. G. Third-party punishment as a costly signal of trustworthiness. *Nature* **530**, 473–476 (2016).

[66] Jordan, J. J., Sommers, R., Bloom, P. & Rand, D. G. Why do we hate hypocrites? evidence for a theory of false signaling. *Psychological science* **28**, 356–368 (2017).

[67] Arnocky, S., Piché, T., Albert, G., Ouellette, D. & Barclay, P. Altruism predicts mating success in humans. *British Journal of Psychology* **108**, 416–435 (2017).

[68] Pleasant, A. & Barclay, P. Why hate the good guy? antisocial punishment of high cooperators is greater when people compete to be chosen. *Psychological science* **29**, 868–876 (2018).

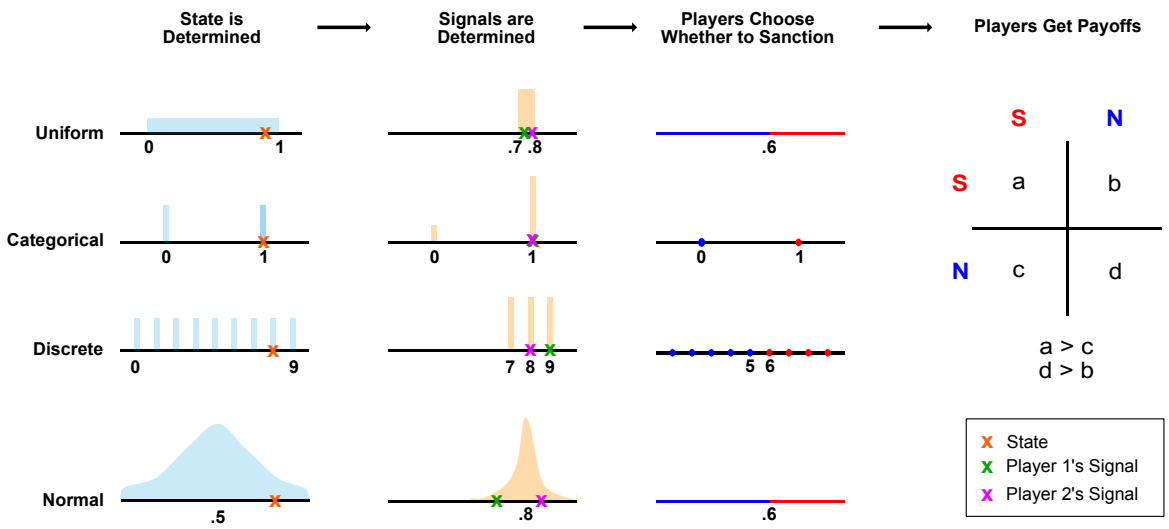


Figure 1: Stylized models of coordination with private signals

We explore a series of two-player coordination games with private signals, each of which proceeds as follows. Column 1: the state is randomly determined from the set of possible states according to the prior distribution shown in light blue. From top to bottom, these distributions are: uniform between 0 and 1; either 0 or 1 with equal probability, uniform on $\{1, 2,...9\}$, normal with mean .5. In each row, the dark orange 'X' marks one, randomly chosen state. Column 2: Players then receive a private signal drawn independently from a distribution that depends on the chosen state, shown in light orange. The signal distributions are: uniform within .5 of the chosen state, equal to the chosen state with probability .75, equally likely to be the chosen state or one value above or below, normally distributed about the chosen state). Green and magenta 'X's each mark one, randomly chosen signal for player 1 and 2, respectively. Column 3: Players then choose between two actions, S (for sanction; red) and N (for not sanction; blue), depending on their signal. Strategies are represented by coloring possible signals according to the action to be played at that signal. We focus on threshold strategies in which both players play S (red) only if their signal is above some value, which in the figure is arbitrarily selected to be .6. Column 4: Players then receive payoffs according to the standard coordination game; each player gets payoff a if she chooses S when the other also chooses S, $b < a$ if she chooses S when the other does not choose S, N if she chooses N when the other chooses N, and $c < d$ if she chooses N when the other does not choose N. The parameter $p = (d-b)/(a-c+d-b)$, which is known as the 'risk dominance' of the coordination game, will prove to be a useful summary statistic of the game's payoffs.

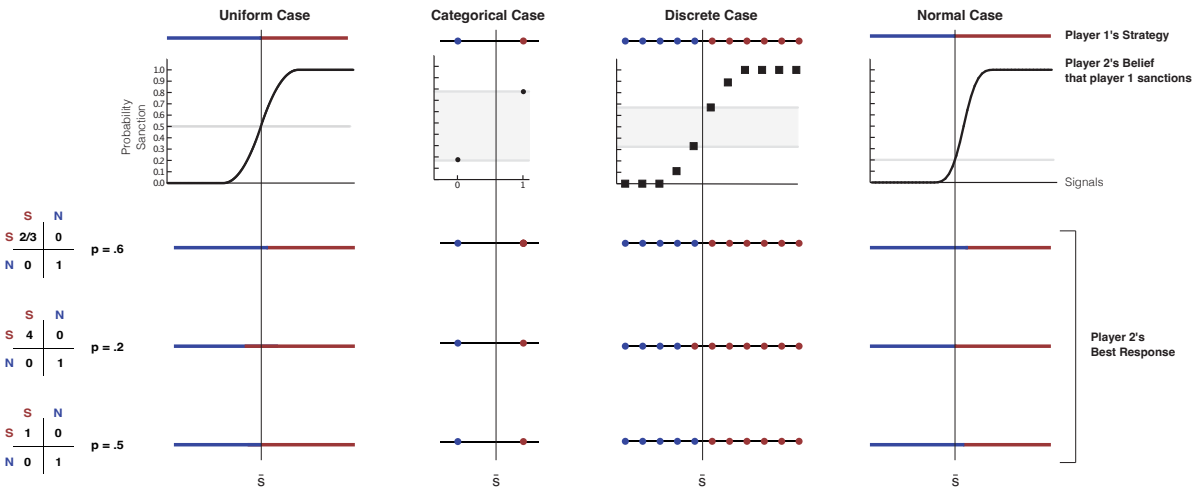


Figure 2: Nash Equilibrium Analyses

Threshold strategies will generally not be Bayesian Nash equilibria in the uniform or normal case but will be in the categorical case. The discrete case falls somewhere in between. To understand each case, first consider a threshold strategy for player 1 (top line). Then, for any given signal, calculate player 2's beliefs that 1 is going to be sanctioning (black curve) if they play according to the specified threshold strategy. Given these beliefs, and given a particular coordination game with risk dominance p, player 2 best responds by selecting a threshold such that her beliefs are below p for all signals below this threshold and above p for all signals above this threshold (illustrated for particular values of p; bottom three lines). A Nash equilibrium exists only when player 2's best response threshold matches player 1's threshold (which can be seen by following the vertical black bar). We mark in grey the values of p for which this will be true. For the uniform case, such a Nash exists only for the non-generic case where $p = .5$. For the categorical case, this will hold for a large range around .5, where the exact bounds depend on ϵ, the error rate. For the discrete case, there will be a smaller range around $p = .5$, where the size of the range depends on the number of possible signal realizations (in the figure, there are 10 possible realizations); the size of this range shrinks as the number of possible signal realizations grows, approaching measure zero in the limit. For the normal case, each threshold strategy will be an equilibrium for exactly one value of p, but that exact value of p will depend on where the threshold is located and does not appear to be evolutionary stable (next figure).

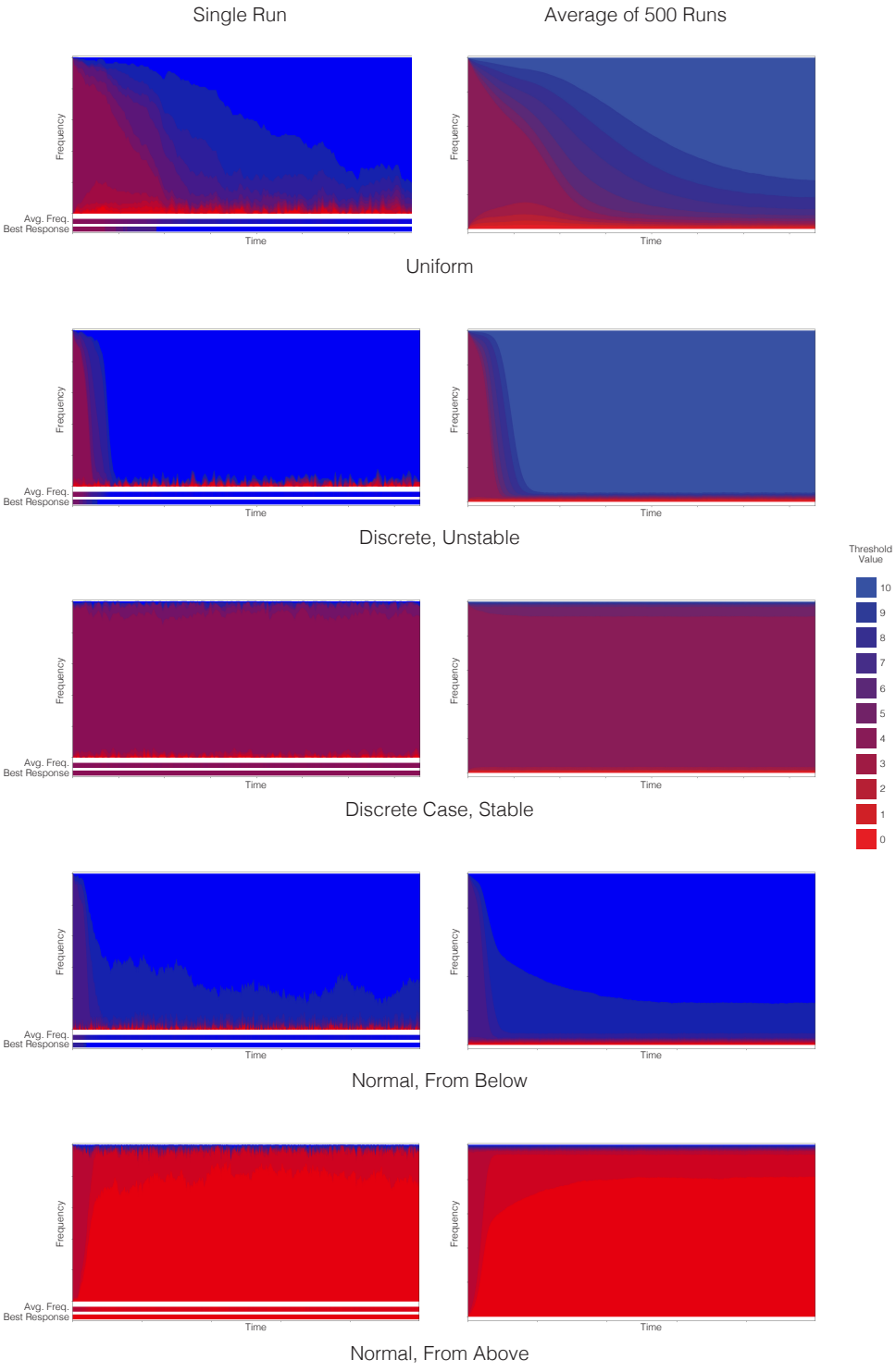


Figure 3: Evolutionary Dynamics of Threshold Strategy Profiles

We ran computer simulations that illustrate the evolutionary dynamics of threshold strategies. Each set of simulations appears on a row. These sets vary in terms of the distribution of states and signals. In each set of simulations, we considered eleven thresholds, whose position always coincided with deciles in the state distribution; a threshold at the 0th decile corresponds to always playing S and the threshold at the 10th decile corresponds to never playing S. We represent these strategies on a spectrum between red (0; always S) and blue (10; never S). At the beginning of each simulation, everyone in the population ($N = 75$) is assigned to play according to the same threshold. We employ a Wright-Fisher dynamic in which strategies update each period according to the payoffs they earned the previous generation, with more successful strategies growing in frequency, and a small fraction of individuals are randomly assigned to 'mutate' to different thresholds. This is allowed to continue for 200 generations (rows 1-3) or 500 generations (rows 4-5). On the left, we show the results of a single, representative run. For each generation, we present the frequencies of each of the ten strategies (top image), the threshold corresponding to the average threshold strategy, given these frequencies (bar immediately below image), and the threshold strategy with the highest payoff, given these frequencies (bottom-most bar). On the right, we show average frequencies of each strategy over 500 simulation runs. Top row ('uniform' case; $p = 2/3$; starting at threshold 3): The average threshold steadily moves up (becomes 'blue-er') leading players to sanction less, until eventually everyone plays never S. This happens because the strategy with the highest payoff always has a higher threshold (i.e. it is blue-er) than the average strategy. This is true anytime $p > .5$. Had we chosen $p < .5$, the population would evolve toward always S. On the right, we see that the single run of our simulations just discussed was indeed representative: in the right image, too, the average threshold in the population steadily moves up until eventually everyone plays never S. Second row (discrete case; $n = 10, p = .8$; starting at threshold 3): For these values of n and p, a threshold strategy is not supported in equilibrium. Once again, the threshold in the population steadily moves up (becomes 'blue-er') until eventually everyone plays 'never S'. Third row (discrete case; $n = 10, p = .67$; starting at threshold 3): Now p is inside the range of values for which a threshold strategy at 3 is supported in equilibrium. When the population starts out with everyone playing according to this equilibrium threshold strategy, it stays there because the threshold strategy with the highest payoff is the one everyone, except the mutants, are already playing. Fourth row (normal case; $p = .556$, variance of errors $= 1/2$ variance of the state; starting at threshold 2): For the parameter values chosen, a single threshold strategy is supported in equilibrium at threshold 3. We start the population with everyone playing below this threshold strategy, at threshold 2. The population does not stabilize: the average threshold in the population moves up (becomes 'blue-er') until eventually everyone plays never S. This is because, at thresholds below the equilibrium, the likelihood that others punish is above p due to reversion towards the mean. Fifth row (normal case; $p = .556$, variance of errors $= 1/2$ variance of the state; starting at threshold 7): This set of simulations is identical, except we now start the population with everyone playing above the equilibrium threshold strategy. Again, the population does not stabilize, though this time the average threshold moves down (becomes 'red-er') until eventually everyone plays always S, since, at thresholds above the equilibrium, reversion to the mean causes the likelihood that others punish to be below p.

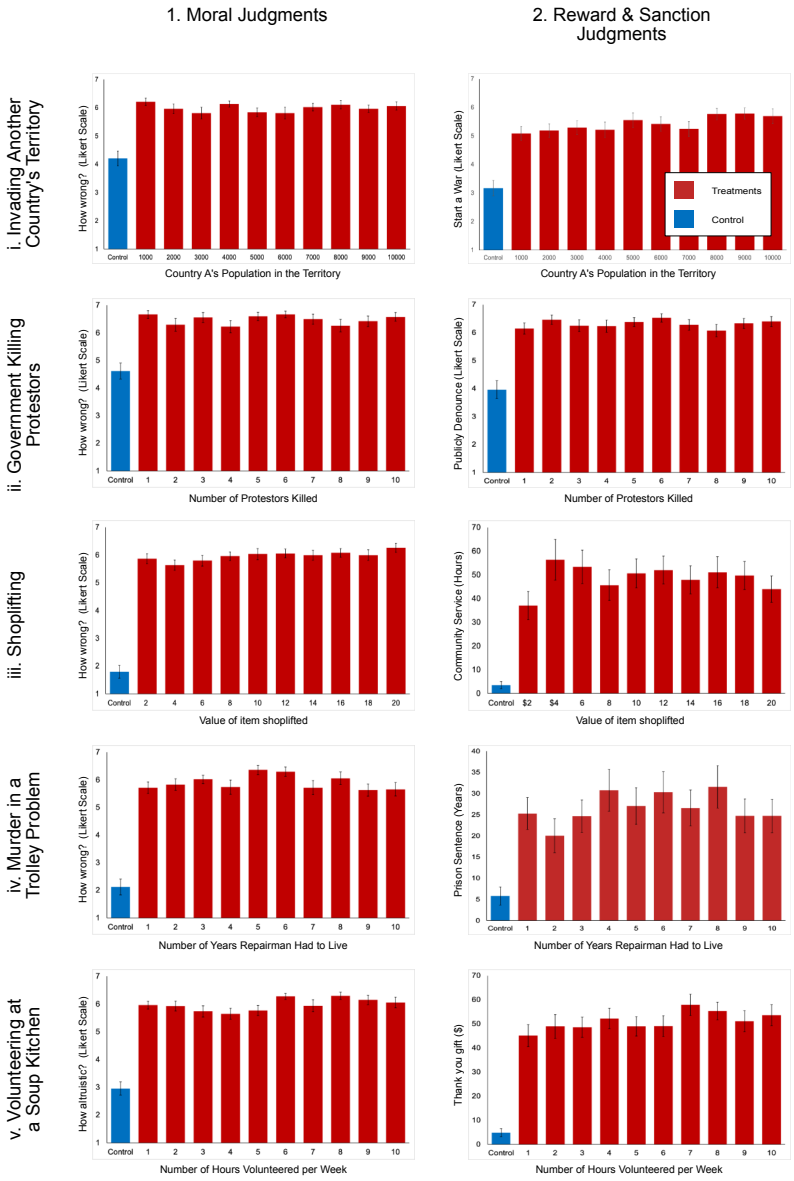


Figure 4: Experimental Evidence - Vignette Studies - Moral Judgements; Reward & Sanction Judgements

Using five different vignettes (top to bottom), we investigated the role of categorical distinctions (blue vs. red bars), compared to continuous variation (comparison across red bars), on participants' moral judgments (column 1) and judgement of appropriate reward and sanctions (column 2) of an actor. In the first four vignettes (rows i-iv), the actor either violated a norm (red bars) or did not (blue bar) and the magnitude of the violation was varied (red bars). In the fifth vignette (row v), the actor either acted pro-socially (red bars) or did not (blue), and the magnitude of the pro-social act was again varied (red bars). Across all vignettes and dependent variables, participants showed a significant effect of the categorical distinction but no effect of the continuous variation (there is a sizeable gap between red and blue bars, but not within the red bars).

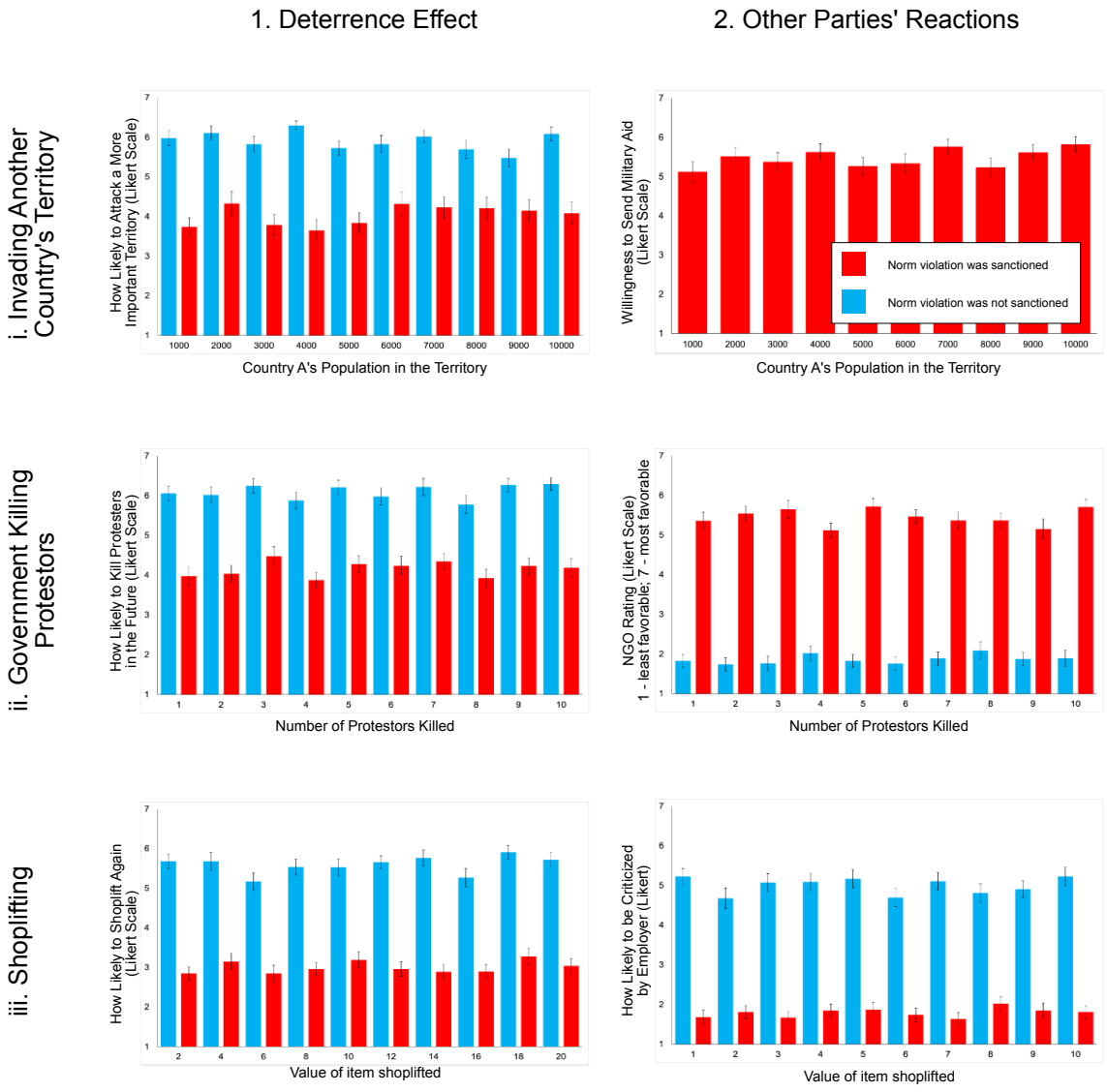
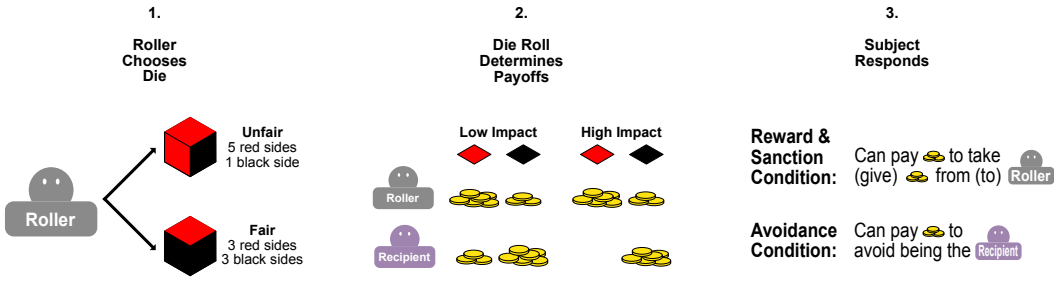
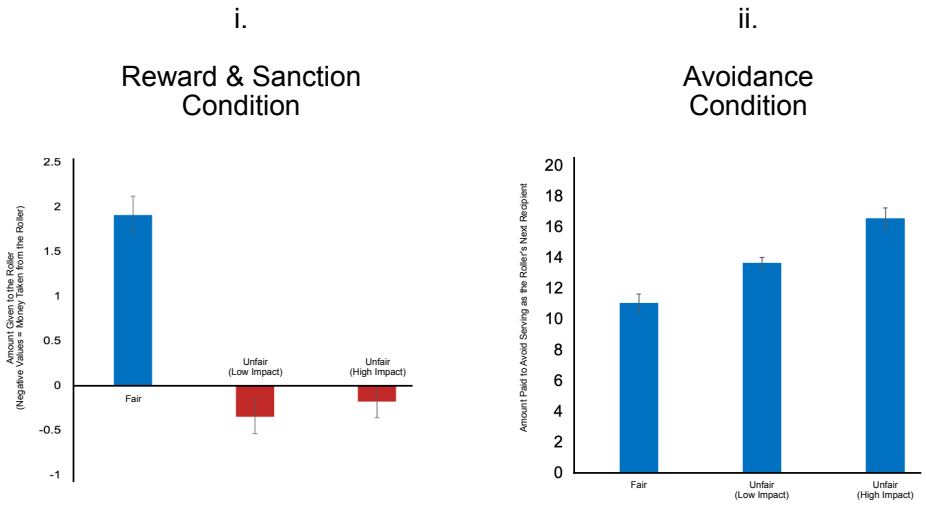


Figure 5: Experimental Evidence - Vignette Studies - Deterrence; Other Parties' Reactions

Using the same vignettes, we also investigated participants' expectations of how subsequent potential transgressors (column 1) or other interested parties (column 2) might respond to transgressions of various magnitudes that were either sanctioned (light red; absent in one plot where the question was not sensible) or not (light blue). Participants expected potential future transgressors and third party on-lookers to take into account whether past transgressions were sanctioned, regardless of the magnitude of past transgressions (there is a sizeable gap between, but not within, red and blue bars).



(a)



(b)

Figure 6: Experimental Evidence - Economic Games Experiments

A: Design. Participants were told of another person (the Roller) who had made a fair or unfair decision (a categorical difference). Specifically, this Roller chose between two die: one with five red sides (unfair), and one had three red sides (fair), where rolling red resulted in payoffs that were favorable to the Roller and unfavorable to a Recipient. We further varied whether rolling red had a relatively low or high impact (a difference along a continuous dimension). Specifically, the Recipient stood to lose either half (low impact) or all (high impact) if the die landed on red. Participants assigned to the 'reward and sanctioning' condition played the role of third party observers who could pay money to take money from (or give money to) the Roller. Participants assigned to the 'avoidance condition' were, instead, assigned to be the Recipient, but they could pay to opt out. They had to make this choice after knowing what die was selected but before knowing the outcome of the die roll. All manipulations were between participants. **B: Results.** In the reward and sanctioning condition, participants paid to take more (give less) to the Roller when the roller had chosen unfairly, but the amount they paid to take (give) was unaffected by the level of impact that the roller's decision had. Whereas, in the avoidance condition, participants were sensitive to both fairness and impact: they were not only willing to pay more to avoid an unfair Roller, but, critically, were also willing to pay more to opt out of the unfair die roll when the die roll would have had a larger impact on them.