Why norms are categorical

Moshe Hoffman[a], **Erez Yoeli**[a,b], **N. Aygun Dalkiran**[c], **and Martin A. Nowak**[a,d]

[a]Program for Evolutionary Dynamics, Department of Mathematics, Harvard University, Cambridge, MA 02138; [b]Applied Cooperation Team, Department of Psychology, Yale University, New Haven, CT 06511; [c]Department of Economics, Bilkent University, Ankara, Turkey 06800; [d]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138

This manuscript was compiled on February 17, 2018

Many social norms are categorical: they are sensitive to category membership instead of underlying continuous variables. For instance, the norm against chemical weapons sanctions sovereigns based on the type of weapon used irrespective of the number of civilians killed. While standard game theory models show that nearly any norm can be sustained in equilibria (1), it is unclear why categorical norms are so prevalent. We explain the prevalence of categorical norms by incorporating an insight from the game theory literature on global games (2): small perceptual errors impede coordination on the basis of continuous variables. Witnesses to a transgression receive a noisy signal of the magnitude (the number of civilians killed) or type of transgression (whether chemical weapons were used), and then choose whether to sanction the transgression. Payoffs to sanctioning are modeled using a coordination game, which captures the fundamental feature of norm enforcement: each witness only wants to act if she expect others will act, too. We show that there is no equilibrium where sanctions are conditioned on the magnitude of the transgression, but there are equilibria where sanctions are conditioned on the type of transgression. We consider various model extensions, prove a general theorem, and investigate evolutionary dynamics. We discuss various additional applications, including human rights, inefficient altruism, institutionalized racism, territorial disputes, revolutions, and collusion.

Social norms | Cooperation | Game theory | Evolutionary dynamics

Many norms depend on categorical variables even though the underlying variable we care about is in fact continuous. For instance, we have norms against the usage of chemical and biological weapons, even though these norms are presumably intended to reduce wanton death or misery. Why do we not simply have a norm against excessive civilian casualties or needlessly painful deaths? Likewise, human rights are applied to all human beings, regardless of their degree of sentience or intelligence. A particularly clever chimp might be smarter than a young child, or feel more pain than an adult in a coma. Why do we not apply rights proportionately to the intelligence of the individual or their ability to feel pain? Less admirable norms also often share this puzzling feature: the Jim Crow South's norm that African Americans should give up their seat to Caucasians did not require that anyone with darker skin tone had to give up their seat to someone with lighter skin tone. Rather, race was defined categorically, based on the infamous 'one drop' rule.

Standard models used to explain the evolution and maintenance of norms have a hard time explaining this puzzling but ubiquitous feature. Such models typically share the characteristic that arbitrary norms can be sustained, provided norm violators are sanctioned and those who do not sanction when expected are themselves sanctioned (1). Why, of all the arbitrary norms that can be sustained in equilibria, would we so frequently find ourselves at categorical norms? Categorical norms are even more puzzling once we incorporate cultural group selection (3) or deliberative agents who select among norms (4) which typically select socially optimal norms. Categorical norms are, necessarily, less efficient than the threshold norms, which would allow one to, say, use 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 (5).

We explain the absence of threshold norms as follows. Consider a norm which dictates that any government that kills more than say 1% of its civilians ought to be sanctioned. Suppose that two countries, say France and the U.S., each assess how many civilians have been killed in Syria, then decide whether they wish to impose sanctions on President Assad. The countries make this decision without first discussing, perhaps because such discussions are hampered by conflicts of interests. Moreover, suppose that neither country wishes to impose sanctions if they are the only one doing so, or if they are the only ones who think that sanctions are warranted, as is often the case (1, 6). For this norm to be upheld, the U.S. needs to be willing to impose sanctions if and only if her best estimate is that more than 1% of Syrians have been killed. However, there will be estimates quite close to 1% at which the U.S. will want to deviate. At such estimates, the U.S. believes France is roughly 50% likely to impose sanctions. If the U.S.'s risk tolerance for sanctioning when France does not sanction is higher than 50%, the U.S. will prefer to deviate and avoid sanctioning even for estimates slightly higher than 1%. If the U.S.'s risk tolerance is lower than 50%, the U.S. will deviate and sanction even for estimates slightly lower than

Significance Statement

Many social norms are surprisingly categorical. For example, we sanction sovereigns based on the category of weapon used, and assign rights based on the category of species membership. Why do we not sanction based on the number of civilians killed or condition rights on sentience, intelligence, or ability to feel pain? We use game theory to show that, if norms are socially enforced–so that sanctions and rewards must be coordinated–then these sanctions and rewards can depend on categorical variables (type of weapon, species), but not continuous ones (number of civilians killed, intelligence). Applications include human rights, inefficient altruism, institutionalized racism, territorial disputes, revolutions, and collusion.

M.H., E.Y., N.A.D., and M.A.N. designed research and performed research. M.H., E.Y., and M.A.N. wrote the paper.

[1]M.H. and E.Y. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: hoffman.moshe@gmail.com and eyoeli@gmail.com

1%. Either way, there is an incentive to deviate, making such a norm unstable.

In contrast, if the norm is to sanction any government that uses chemical weapons, then this can be sustainable. Consider a norm which dictates that any government that uses chemical weapons ought to be sanctioned. Suppose the two countries send surveyors to collect chemical residues and either does or does not detect residue, perhaps with some error rate. And suppose, again, that the U.S. wishes to sanction if and only if she thinks it is sufficiently likely that France will sanction. What happens when the U.S. detects chemical residues? So long as France is abiding by the norm and comes to a similar assessment, and the U.S. is sufficiently risk tolerant, the U.S. will strictly prefer to sanction. And, had the U.S. not discovered any traces of chemical weapons, the U.S. would think it likely France had not discovered any traces either, so strictly prefer to not sanction. Hence, the norm is self-enforcing. The crucial difference between categorical norms and threshold norms is that, for categorical norms, one cannot receive a signal that yields posterior beliefs (that the other will punish) which approach 50%. Whereas for threshold norms, signals close to the threshold yield beliefs that approach 50%.

We first present a stylized model that formalizes the above logic. We then consider various extensions to show that our main result is robust, and to gain a deeper sense of what the result depends on. We also model and simulate the evolutionary dynamics of threshold norms in order to more clearly show that threshold norms are not sustainable, and fully 'unravel'. We present a general theorem which indicates exact conditions on the distribution of signals and payoffs under which threshold norms can be sustained. Finally, we discuss additional applications and relate our results to the existing literature. In so doing, we contribute to a growing literature that uses insights from game theory to explain puzzling aspects of our beliefs and preferences (7–10), which, itself, is part of an important literature that offers functional, evolutionary based accounts to otherwise puzzling social behaviors (6, 11–19).

We begin with a stylized model of a threshold norm (Fig. 1a). First, a transgression occurs. It has a randomly distributed magnitude; for now, we assume the magnitude is distributed uniformly. There are two witnesses to the transgression. Each receives a noisy signal of the magnitude, uniformly distributed about the true magnitude with noise $\epsilon > 0$. Then, each witness decides whether to sanction. The payoffs to sanctioning depend on whether the other witness sanctions. In particular, each witness gets a if she sanctions when the other also sanctions, $c < a$ if she sanctions when the other does not sanction, d if she does sanction when the other is not sanctioning, and $b < d$ if she does not sanction when the other is sanctioning (Fig. 1c). Note that, for now, the payoffs to sanctioning are presumed to not depend on the magnitude of the transgression. The parameter $p = (d-b)/(a-c+d-b)$ will prove useful. Its interpretation is: if one player sanctions with probability greater than p, then the other prefers to sanction.

We solve such models by identifying their Bayesian Nash equilibria (BNE). BNE is a standard extension of Nash's equilibrium to probabilistic settings. Players are assumed to maximize expected payoffs, and form beliefs according to Bayes' Rule.

Suppose that each player sanctions whenever they receive a signal of $\bar{s}$ or higher. Such a norm is only an equilibrium

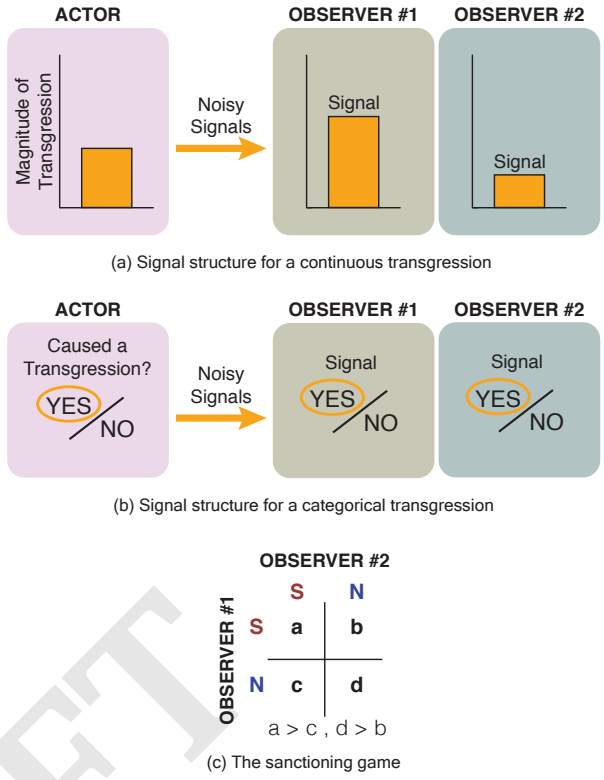


Fig. 1. Stylized models of norms

for the non-generic case where $p = 0.5$. If $p > 0.5$, then any witness who receives a signal above $\bar{s}$ but sufficiently close to $\bar{s}$, is better off 'playing it safe' and deviating to not sanctioning (Fig. 2a). If $p < 0.5$, then anyone who receives a signal below $\bar{s}$ but sufficiently close to $\bar{s}$ benefits from deviating to sanctioning. (See S.I. for all proofs.) For example, suppose that $a = 4, b = 0, c = 2$, and $d = 4$, which implies $p = 0.\bar{6}7$. A player who gets a signal $s_i = \bar{s}$ gets $0.5 \cdot 4 + 0.5 \cdot 0 = 2$ if she sanctions and $0.5 \cdot 2 + 0.5 \cdot 4 = 3$ if she does not sanction, and thus prefers not to sanction. We will soon show that, even though players benefit from deviating only in a small range, this results in a 'slippery slope' and we will never observe a threshold norm.

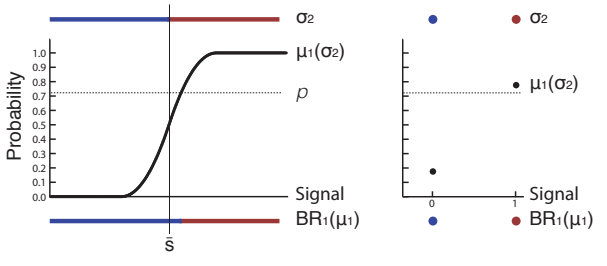


Fig. 2. Nash Equilibrium Analyses of Threshold Norms and Categorical Norms

We model categorical norms similarly (Fig. 1b), except that, now, the transgression is $H = 1$ with probability q and 0 with probability $1-q$, and witnesses' signal is incorrect, $S = 1-H$,

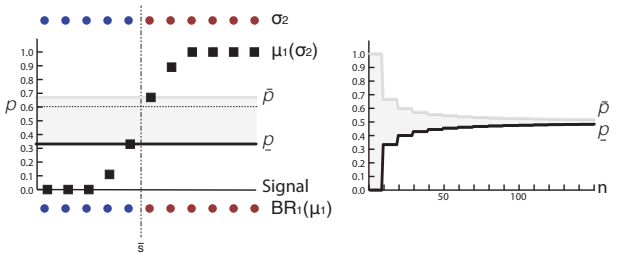
with probability ϵ and correct, $S = H$, with probability $1 - \epsilon$. In contrast to threshold norms, now there is a Bayesian Nash equilibrium where sanctions depend on the signal received, provided the error in the signal is sufficiently small (Fig. 2b). This is because when a player receives a signal $S = 1$, she is better off sanctioning so long as she believes the other was sufficiently likely to also receive a signal of 1. And, when a player receives a signal $S = 0$, she is better off not sanctioning so long as as she believes the other was sufficiently unlikely to receive a signal of 1. For example, for $q = 0.2, \epsilon = 0.05$, if one player gets signal of 1, she believes the other got the same signal with probability 0.79, and will be best off sanctioning so long as $p < 0.79$.

Note that a threshold norm cannot be supported by treating the continuous signal as discrete, I.e. by ignoring the value of the signal and focusing on whether it was above or below the threshold. This is because witnesses still have access to the more-informative continuous signal, and would deviate from the threshold strategy when their signal is close to the threshold.

We next investigate to what extent these results generalize. First, it should be noted that if signals are observed without noise, or if the payoffs to sanctioning depend on the magnitude of the transgression but not on what each expects others to do, then threshold norms can be sustained in equilibrium. For example, suppose that after observing a noisy signal of a transgression, a witness decides whether to partner with the transgressor. If the witness prefers not to match with a transgressor who is more likely to select higher magnitude transgressions, the witness can–and will prefer to–partner with the transgressor if and only if her signal of the transgression was below a certain threshold (Fig. S1). Similarly, threshold norms are possible if witnesses can commit in advance to taking an action. Thus, coordination is key to obtaining our result. We note that many games have a coordination component, such as repeated games that are often used to model norm enforcement (20, 21).

Next, we consider alternative distributions of signals, as well as different payoffs for sanctioning. We generally find that our main result holds. Even when there exist equilibria where players condition their behavior on their signal, the thresholds depend on the particulars of the signal and the witnesses' payoffs. If one wanted another threshold, say one determined by the social cost of the transgression, it would not be an equilibrium. Moreover, any threshold equilibria that do exist are unstable–they unravel under standard learning or evolutionary processes.

First, consider what happens if we replace the continuous variable with a discrete variable that is equally likely to take on one of many (n) values. The larger n gets, the closer p needs to be to 50% in order to allow for an equilibrium that depends on the signal (Fig. 3a). Thus, for larger values of n, the range of coordination games over which such equilibria exist gets smaller and approaches measure 0 (Fig. 3b). That is, even though harm is discrete, once it can take on sufficiently many possible values, the results approximate what happens when harm is continuous. 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 closest to this value, but above it, i must believe that $-i$ has gotten a signal above this threshold with probability $> p$. For example, for $n = 10$ and an error of $\epsilon = 1$, this belief will be $.\bar{6}7$. And, when i get a signal close to the threshold but below, she must believe $-i$ has probability $< p$ of having received a signal above the threshold. For $n = 10, \epsilon = 1$ this belief will be $.\bar{3}3$. And hence such an equilibrium exists for p between $.\bar{3}3$ and $.\bar{6}7$. However, the larger n is, the closer these two beliefs will be, closing the gap of permissible values of p, with both approaching .5. For instance, for $n = 20, \epsilon = 2$, these values are .4 and .6, so only p within $[.4, .6]$ will work.



(a) Nash equilibrium analysis for n=10 discrete signal values

(b) Values of p that can support threshold norms

Fig. 3. Transgression is Uniformly Distributed Over n Discrete Values

Next, instead of assuming that harm is uniformly distributed, we assume that it is normally distributed. Now, a player who gets a signal that is higher than the mean harm level will think it is more likely that the other player got a signal below hers–how much depends on her signal and the variance of h. If the variance of h is relatively large, then the likelihood the other player got a signal below hers will be relatively close to 50%. In this case, there still will be no threshold equilibrium. If the variance of h is relatively small, then, for signals far from the mean of h, the likelihood the other player got a signal below hers can diverge meaningfully from 50%. In such cases, it can be possible for a threshold norm to be an equilibrium. However, there will be only one, highly specific threshold that can be sustained in equilibrium, and that threshold will depend on p and the variance of h, and not on what is socially optimal (Fig. SI4). We will, momentarily, use evolutionary dynamics to show that this threshold is not stable (Fig. 4e).

We now consider what happens if the continuous variable directly influences the payoffs to sanctioning. In particular, we consider the case where the payoffs are as before but we add an additional benefit to sanctioning that is an increasing function of the magnitude of the transgression. Now p depends on the magnitude of the transgression, and if the dependency is sufficiently strong, then, as with normally distributed harm, it is possible to support a single threshold norm in equilibrium. But, once again, the threshold will be determined by the strength of this dependency, and not all socially relevant considerations.

Next, we provide a more general characterization of which thresholds can be equilibria, given any continuously differentiable distributions of signals and harm. We restrict our attention to cases where there are two players and payoffs are independent of harm. We find that a threshold norm can be an equilibrium if and only if p falls between the witness's posterior probability that the others' signal is below the threshold when her own signal approaches the threshold from below, and the witness's posterior probability that the others' signal is above

the threshold when her own signal approaches the threshold from above.

In reality norms may not be at equilibrium. What would happen if a norm starts off at a threshold and then adjusts according to an evolutionary process? We ran computer simulations that model such an evolutionary process (see methods section and S.I. for details). Strategies update each period according to the payoffs they earned the previous period, with more successful strategies growing in frequencies. A small fraction are randomly assigned different thresholds to mimic experimentation. We find that the average threshold in the population steadily moves up if $p < .5$, and down if $p > .5$, until eventually everyone either always sanctions or never sanctions. Fig. 4a illustrates a single representative run of our simulations, for the first model we presented (uniformly distributed harm), with $p > .5$. Fig. 4b illustrates the average outcomes from many such runs. Figs. 4c and 4d present a single representative simulation for the variation of the model with $N = 10$ discrete levels of harm. In Fig. 4c, p is inside the range for which a threshold norm is supported, and thus, the threshold norm is stable. Whereas, in Fig. 4d, p is outside this range, and thus the threshold norm is unstable. Fig. 4e presents a single representative simulation for the variation of the model with normally distributed harm. These simulations consistently confirm our equilibria analysis. This is to be expected since our equilibria analysis is actually based on a solution concept that requires fewer assumptions than Nash: iterative elimination of strictly dominated strategies (see S.I.). And, this, more permissive solution concept has the property that evolutionary processes will always eventually converge to those strategies that satisfy iterative elimination of strictly dominated strategies (in contrast to Nash equilibria, to which evolutionary processes do not always converge).

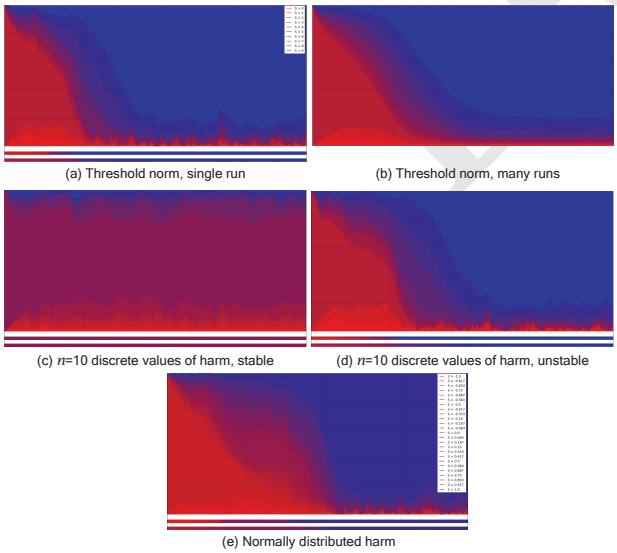


Fig. 4. Evolutionary Dynamics of Norms

Finally we consider a couple of additional model extensions which make the model more realistic, which are presented in more detail in the S.I.. First, we extend our model to take into account the fact that norms are often enforced by large groups of potential sanctioners, not just two, and again obtain comparable results. Second, we recognize that sanctions can themselves vary in magnitude, perhaps with each witness's payoffs depending on the difference between how much she sanctioned and how much others sanctioned. Therefore, we extend the sanctioning decision to permit variation in magnitude of sanctions and show that this, too, does not permit threshold norms, provided players's payoffs do not happen to have a particular, non-generic structure.

We now discuss some applications. 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 intelligence or sentience, 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 why human rights depend on the categorical variable of species membership and not the continuous variable of intelligence or sentience: that is the only way to make the norm sustainable.

Likewise, one might wonder why rights are viewed as absolute, immune to tradeoffs 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 the continuous variable: the amount of suffering inflicted by the torture and the amount of suffering prevented by the discovery of the hypothetical ticking bomb.

We have strong norms against discriminating against well defined categorical groups like women or African Americans. Of course, many people are discriminated against for other reasons, such as being old (23), unattractive (24, 25), overweight (26), or short (27). Why are there not equally strong norms against such discrimination? On the flip side, there also has never been institutionalized discrimination against older, less attractive, heavier, or shorter individuals in the same way there has been against women or African Americans. We would explain this distinction as follows: in their own private decisions, people are able to discriminate on the basis of continuous variables like age, attractiveness, weight, and height, but such discrimination cannot be enforced–or prevented–based on coordinated sactions. In contrast, it is possible to enforce or prevent discrimination against groups that are categorically defined, like women or those with 'one drop' of African blood.

One consequence of this relates to measures of subconscious discrimination, which persists, but is perhaps hidden in the presence of countervailing norms. An example of such a measure is the Implicit Association Task (IAT) (28). Since subconscious discrimination does not require coordination, we expect measures like the IAT to be less influenced by categorical distinctions than measures of overt discrimination, which is intended for others to see and reward, and which we expect to display categorical distinctions.

Another puzzling behavior that may have to do with categorical norms: people donate a lot of time and money to charity (29), 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 (30) and highly ineffective charities persist (31). 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. We expect people to consider effectiveness more with kin than with friends or strangers, since altruism towards kin is sometimes not driven by norm enforcement, reputations, or reciprocity, but rather by kin selection, for which coordination is not relevant.

One more phenomenon that may relate to categorical norms: 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 (32). 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. Our model suggests that protests and revolutions, which must be coordinated, can be triggered by categorical events like an all-new tax category or public immolation, but not by changes in continuous conditions like increased tax or poverty rates.

A final example of a categorical norm: small disputes often lead to escalated conflicts. Examples include the Falkland Islands War, Operation Paul Bunyan, and the constant risk of war over disputed islands in the South China Sea. In each, the contested resource which led to, or risks conflict is of trivial value. Why do countries not ignore tiny infractions, instead of risking wars that are much more costly than the resources under dispute? Intuitively, we think this is because any country that does not defend its borders and its resources will be up for the taking, but why do we not presume countries will defend their valuable resources, while not risking war over insignificant ones? We believe the answer again lies in the fact that norms that disavow stealing are sustainable, but norms that disavow stealing of *valuable* resources are not, because the former norm is categorical while the latter is continuous.

We now use the model to elucidate when our norms or judgments will be more categorical, and when they will be more continuous.

First, we expect norms and judgments to be more continuous when norms are not enforced communally through coordinated sanctions and rewards, but rather agents can act unilaterally. One setting in which we act more unilaterally is when we evaluate partners, as elucidated in our model of partner choice (Fig. 1c). So, individuals can consider height and earnings when dating, and even employ thresholds like, 'I only date others who are taller than I am.' By the same logic, we expect that our judgments will be more continuous when we are judging character and not deciding sanctions. For instance, we expect our avoidance behaviors and fear responses to vary continuously with moral transgressions, even if our anger and outrage depend more on categorical distinctions. Another setting in which we act more unilaterally is in the enforcement of domestic laws. Most modern states grant enforcement authority to agencies, judges, police, *etc.* While there is some check on these agents by the masses (as evidenced by the recent 'black lives matter' protests against police departments that were perceived as too-readily shooting young black constituents), they do, largely, act unilaterally. So, civil penalties can be conditioned on harm, and the Environmental Protection Agency can–and does–sue companies whose emissions result in pollution levels above a threshold (33). As previously discussed, we also expect norms and judgments to be more continuous for kin-based altruism.

Second, we expect norms to be more continuous when parties do not have private information, or, they are able to credibly commit to communicating all private information. France and the U.S. could enforce a threshold norm conditioned on the number of civilian casualties if they could commit to convening at the U.N. and jointly reviewing all estimates of the number of civilian casualties. Similarly, it can be a norm for Harvard's students to wear flip flops only when the temperature on the large digital thermometer overlooking Harvard Square is 95 or greater.

Our model helps clarify the costs and benefits of categorical norms. Since threshold norms more closely track the underlying variable of interest, there is a concrete benefit to having unilateral enforcement and open communication. However, to the extent that communication is hampered by incentives or logistical problems, and to the extent that unilateral enforcement is limited by conflicts of interest, or checks and balances, threshold norms will nevertheless remain untenable. Moreover, even in settings where continuous norms can be sustained, the psychology that has been learned or evolved from settings where they cannot can spillover and influence our moral intuitions.

For instance, when President Obama was considering declaring war against Assad after Assad used chemical weapons, many commentators criticized Obama for enforcing the norm against chemical weapons after Obama had done little to prevent Assad from killing roughly 100,000 civilians. This criticism misses the key logic that we believe motivated Obama: the norm against chemical weapons is sustainable while a norm against wantonly killing civilians is harder to sustain, and while a norm against chemical weapons may be a poor proxy for minimizing civilian casualties, it is still a proxy that does to some extent limit the number of civilian deaths. Hence, there is some benefit to maintaining this norm. Another telling anecdote: in his book 'Nuclear Weapons and Foreign Policy' (34), Henry Kissinger advocated for "limited nuclear war," a position he later retracted out of concerns that such war would spiral out of control (35).

One might proffer that we have categorical norms not because of coordinated sanctions, but because it is easier to notice and/or communicate categorical transgressions, or to encode categorical norms. While categorical distinctions are in fact easier to notice, communicate, and encode, and this no doubt plays some role in promoting categorical norms, it seems unlikely to explain the entire phenomenon. In particular, it cannot explain why categorical norms persist in situations where stakes are high and parties are highly deliberative, as was the case in Iwo Jima or in Syria. Or, in situations in which our emotional response is unresponsive to argument; we challenge the reader to attempt to convince friends that one should grant more rights to especially smart

chimpanzees than to especially incapable people. Moreover, our model uniquely predicts the comparative static that we will emphasize categorical distinctions more when norms are socially enforced.

We will conclude with a discussion of some related literatures. The classical game theory literature on repeated games typically focuses on achievable payoffs, and minimal requirements on the signal structure to achieve cooperation (21), but not on explaining the features of equilibria, and, in particular, the inefficiencies of existing norms. Likewise, the literatures on cooperation (11) and norm enforcement (3) focus on the ability to support cooperation through reciprocity, or a variety of norms through higher order sanctions, respectively. But, as far as we know, they do not attempt to explain constraints on cooperation and the norms that can be enforced, such as the inability to condition cooperation and norms on continuous variables.

In developing our model, we borrowed heavily from the literature on global games (2). In such games, payoffs depend on a continuous variable. If players observe this variable without noise, then, for some range in this variable, there are multiple equilibria. However, if players observe the continuous variable with arbitrarily small amounts of noise, then there is only a single, threshold equilibrium. These results have been used to explain currency attacks and bank runs, and to draw into question multiplicity of equilibria in such contexts. We adjust that framework to focus on state independent payoffs, and the contrast between continuous and categorical variables. Instead of reducing multiplicity of equilibria to a single threshold equilibrium, we ask when will there exist a threshold equilibria, and use this result to explain the prevalence of categorical norms.

Another potentially related literature is in Industrial Organization. Therein, researchers have investigated when collusion can be maintained among oligopolies (36, 37). That literature focuses on the need for observability of defections, or the difficulty of enforcing agreements when there are many players. It does not focus on the variables oligopolies can collude on. We contribute to that literature by arguing that whenever there is noisy private information about critical continuous variables, like each producer's costs or output, it is easier for colluders to sustain norms that are categorical in nature, like 'only sell through De Beers', and harder to sustain norms that depend on continuous variables, like 'make sure prices are marked up at least 20% above costs'.

Our model also has implications for moral realism and group selection. To the extent that there are not alternative explanations for categorical norms, this paper provides further evidence (6–19) that incentives–as characterized by Nash equilibria–act as a constraint on our sense of morality, and that morality cannot be solely explained by logic or group level benefits. This raises into question the general approach of analytic moral philosophy, which attempts to explain our sense of morality without recourse to incentives and Nash, but based on logic alone (4, 38), as well as, group selection models that attempt to explain our morality on the basis of group-level benefits alone, without considering individual incentive compatibility constraints (39, 40). More generally, it highlights the role of incentives in shaping our preferences and ideologies, and the added value of using game theory to explore the constraints that incentives will impose upon our norms and moral intuitions.

ACKNOWLEDGMENTS. We thank Andrew Ferdownsian for research assistance.

1. Fudenberg D, Maskin E (1986) The folk theorem in repeated games with discounting or with incomplete information. *Econometrica: Journal of the Econometric Society* pp. 533–554.
2. Morris S, Shin HS (2001) Global games: Theory and applications. *Advances in Economics and Econometrics* p. 56.
3. Boyd R (year?) How humans became outliers in the natural world (http://econ.as.nyu.edu/docs/IO/41614/Outliers.pdf). Online; accessed October 16, 2016.
4. Pinker S (2011) *The better angels of our nature: The decline of violence in history and its causes.* (Penguin UK).
5. Economist T (year?) The history of chemical weapons: The shadow of ypres (http://www.economist.com/news/briefing/21584397-how-whole-class-weaponry-came-be-seen-indecent-shadow-ypres). Online; accessed October 16, 2016.
6. DeScioli P, Kurzban R (2013) A solution to the mysteries of morality. *Psychological Bulletin* 139(2):477.
7. DeScioli P, Christner J, Kurzban R (2011) The omission strategy. *Psychological Science* 22(4):442–446.
8. DeScioli P, Wilson BJ (2011) The territorial foundations of human property. *Evolution and Human Behavior* 32(5):297–304.
9. Hoffman M, Yoeli E, Nowak MA (2015) Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences* 112(6):1727–1732.
10. Hoffman M, Yoeli E, Navarrete CD (2016) Game theory and morality in *The Evolution of Morality.* (Springer), pp. 289–316.
11. Trivers RL (1971) The evolution of reciprocal altruism. *Quarterly review of biology* pp. 35–57.
12. Frank RH (1988) *Passions within reason: the strategic role of the emotions.* (WW Norton & Co).
13. Axelrod RM (2006) *The evolution of cooperation.* (Basic books).
14. Nowak MA (2006) Five rules for the evolution of cooperation. *science* 314(5805):1560–1563.
15. Pinker S, Nowak MA, Lee JJ (2008) The logic of indirect speech. *Proceedings of the National Academy of sciences* 105(3):833–838.
16. Lee JJ, Pinker S (2010) Rationales for indirect speech: the theory of the strategic speaker. *Psychological review* 117(3):785.
17. DeScioli P, Gilbert SS, Kurzban R (2012) Indelible victims and persistent punishers in moral cognition. *Psychological Inquiry* 23(2):143–149.
18. Chwe MSY (2013) *Rational ritual: Culture, coordination, and common knowledge.* (Princeton University Press).
19. DeScioli P, Karpoff R (2015) People's judgments about classic property law cases. *Human Nature* 26(2):184–209.
20. Panchanathan K, Boyd R (2004) Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* 432(7016):499–502.
21. Osborne MJ, Rubinstein A (1994) *A course in game theory.* (MIT press).
22. Singer P (1981) *The expanding circle.* (Clarendon Press Oxford).
23. Macnicol J (2006) *Age discrimination: An historical and contemporary analysis.* (Cambridge University Press).
24. Becker GS (1973) A theory of marriage: Part i. *The Journal of Political Economy* pp. 813–846.
25. Mobius MM, Rosenblat TS (2006) Why beauty matters. *The American Economic Review* 96(1):222–235.
26. Baum CL, Ford WF (2004) The wage effects of obesity: a longitudinal study. *Health Economics* 13(9):885–899.
27. Case A, Paxson C (2008) Stature and status: Height, ability, and labor market outcomes. *Journal of Political Economy* 116(3).
28. Greenwald AG, McGhee DE, Schwartz JL (1998) Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74(6):1464.
29. Trust NP (2014) Charitable giving statistics (http://www.nptrust.org/philanthropic-resources/charitable-giving-statistics). Online; accessed February 13, 2014.
30. Baltussen RM, Sylla M, Frick KD, Mariotti SP (2005) Cost-effectiveness of trachoma control in seven world regions. *Ophthalmic epidemiology* 12(2):91–101.
31. Times TB (year?) 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-nearly-1-billion-for-corporate/2124083). Online; accessed February 16, 2014.
32. Thorndike JJ (year?) Four things you should know about the boston tea party (http://www.taxhistory.org/thp/readings.nsf/ArtWeb/1BB0C8F894BB490B852577020083A6F6?OpenDocument). Online; accessed December 28, 2017.
33. Agency UEP (year?) Drinking water requirements for states and public water systems (https://www.epa.gov/dwreginfo/chemical-contaminant-rules). Online; accessed September 17, 2017.
34. Kissinger HA (1984) *Nuclear weapons and foreign policy.* (Westview Press, Inc., Boulder, CO).
35. Isaacson W (2005) *Kissinger: A biography.* (Simon and Schuster).
36. Stigler GJ (1964) A theory of oligopoly. *The Journal of Political Economy* pp. 44–61.
37. Carlton DW, Perloff JM (2015) *Modern industrial organization.* (Pearson Higher Ed).
38. Shermer M (2015) *The moral arc: How science and reason lead humanity toward truth, justice, and freedom.* (Macmillan).
39. Fehr E, Fischbacher U, Gächter S (2002) Strong reciprocity, human cooperation, and the enforcement of social norms. *Human nature* 13(1):1–25.
40. Haidt J (2012) *The righteous mind: Why good people are divided by politics and religion.* (Vintage).

Figure Legends.

Fig. 1: Stylized Models of Norms. A: We begin with a stylized model of a threshold norm. First, a transgression occurs. It has a randomly distributed magnitude; for now, we assume the magnitude is distributed uniformly, $H \sim U[h_L, h_H]$. There are two observers to the transgression. Each receives a noisy signal of the magnitude, uniformly distributed about the true magnitude, $S_i \sim U[h-\epsilon, h+\epsilon]$ where $\epsilon > 0$.

B: We model categorical norms similarly, except that, now, the transgression is $H = 1$ with probability q, and 0 with probability $1-q$, and observers' signal is incorrect, $S_i = 1-h$, with probability ϵ and correct, $S_i = h$, with probability $1-\epsilon$.

C: In either case, after the observers receive their signals, each observer decides whether to sanction, P, or not, N. The payoffs to sanctioning depend on whether the other observer sanctions. In particular, each observer gets a if she sanctions when the other also sanctions, $c < a$ if she sanctions when the other does not sanction, d if she does sanctions when the other is not sanctioning, and $b < d$ if she does not sanction when the other is sanctioning. The parameter $p = (d-b)/(a-c+d-b)$ will prove useful. Its interpretation is: if one observer sanctions with probability greater than p, then the other prefers to sanction.

Fig. 2: Nash Equilibrium Analyses of Threshold Norms, Categorical Norms, and Partner Choice. A: We show that a threshold norm cannot be a Bayesian Nash equilibrium when harm is continuous by presuming observers play according to a threshold norm with some threshold $\bar{s}$ and showing a beneficial deviation. (1) Suppose that each observer punishes whenever they receive a signal of $\bar{s}$ or higher. We represent this strategy for observer 2 by a line shaded blue to the left of $\bar{s}$ and red to its right. (2) Below it, we represent observer 1's belief that observer 2 is punishing, for any signal, s_1. (3) Finally, we represent observer 1's best response, which deviates from the proposed threshold norm for signals just above $\bar{s}$. Why? For $p > .5$, when observer 1 receives signals above $\bar{s}$ but sufficiently close to $\bar{s}$, her belief that observer 2 is punishing is below the dotted line representing p. Therefore, she is better off 'playing it safe' and deviating to not punishing.

B: In contrast to threshold norms, now there is a Bayesian Nash equilibrium where punishment depends on the signal received, provided the error in the signal is not too large. (1) Suppose observer 2 punishes if and and only if her signal is $s_2 = 1$. We represent this strategy by a red dot at 1 and a blue dot at 0. (2) Below it, we represent observer 1's belief that observer 2 is punishing, for $s_1 = 0$ and 1, when $q = .2$ and $\epsilon = .05$. (3) Finally, we again represent observer 1's best response. When she receives a signal $s_1 = 1$, she is better off punishing since she believes observer 2 was sufficiently likely (relative to p) to also receive a signal of 1. And, when a observer receives a signal $s_1 = 0$, she is better off not punishing since observer 2 is sufficiently unlikely to receive a signal of 1.

Fig. 3: Transgression is Uniformly Distributed Over n Discrete Values. We next consider cases in between the purely categorical and purely continuous case in Fig. 1 in order to demonstrate that the result does not depend on signals being purely continuous, but rather just having many possible realizations.

(1) Now, there is a range of values of p for which a threshold norm can be supported in equilibrium. The bounds of the range are determined by the observers' beliefs just above and below the proposed threshold. For an equilibrium to exist where observers punish if they get a signal above a certain value, then whenever observer 1 gets a signal closest to this value, but above it, 1 must believe that 2 has gotten a signal above this threshold with probability $> p$, and similarly for observer 2. For example, for $n = 10$ and an error of $\epsilon = .3$, this belief will be $.\bar{6}7$. And, when i get a signal close to the threshold but below, she must believe $-i$ has probability $< p$ of having received a signal above the threshold. For $n = 10, \epsilon = .3$, this belief will be $.\bar{3}3$. And, hence, a threshold norm is an equilibrium for p between $.\bar{3}3$ and $.\bar{6}7$. (2) However, the larger n is, the smaller this range will be. For instance, for $n = 20, \epsilon = 2$, only p within $[.4, .6]$ will work. As n grows, this range will converge to a single point at $p = .5$.

Fig. 4: Evolutionary Dynamics of Threshold Norms. A: We ran computer simulations that model the evolutionary dynamics of threshold norms. We limit the strategy to 'always punish', 'punish iff $S_i > 1/9$', 'punish iff $S_i > 2/9$', ..., 'punish iff $S_i > 8/9$', 'never punish', and represent these strategies on a spectrum between red (always punish) and blue (never punish). At the beginning of the simulation, everyone in the population is assigned to punish if and only if their signal is greater than 1/9. Strategies update each period according to the payoffs they earned the previous period, with more successful strategies growing in frequencies. A small fraction are randomly assigned different thresholds to mimic experimentation. Here, we show the results of a single, representative run of our simulations of the base model (with uniformly distributed harm) and $p > .5$. We present the frequencies of all the strategies, the strategy with the greatest payoff, and the average strategy. The average threshold in the population steadily moves up (becomes 'blue-er') until eventually everyone never punishes. This happens because the strategy with the highest payoff is always 'blue-er' than the average strategy.

B: Average frequencies of each strategy over 500 simulations for the base model, with the same parameter values as in Fig. 5a. We see that the single run of our simulations presented in Fig. 5a was indeed representative: the average threshold in the population always steadily moves up until eventually everyone never punishes.

C: We now show results of a single simulation for the model with $n = 10$ discrete possible values of harm, and $p = .67$, which is inside the range of values for which a threshold norm can be supported. We present the frequencies of all the strategies, the strategy with the greatest payoff, and the average strategy. For these values of n and p, there exists a threshold norm. When the population starts out with everyone playing according to a threshold norm, it stays there (because the strategy with the highest payoff is always the average strategy).

D: We now show results of a single, representative simulation for the model with $n = 10$ discrete possible values of harm, and $p = .8$, which is outside the range of values for which a threshold norm can be supported. We present the frequencies of all the strategies, the strategy with the greatest payoff, and the average strategy. For these values of n and p, there does not exist a continuos norm. Again, we start the population with everyone playing according to a threshold norm. Now, the threshold in the population steadily moves up (becomes 'blue-er') until eventually everyone never punishes.

E: Finally, we show results of a single, representative simulation for the model with normally distributed harm. For the parameter values chosen, there is a single threshold norm at 'punish iff $S_i > -1.83$'. We present the frequencies of all the strategies, the strategy with the greatest payoff, and the average strategy. We start the population with everyone playing according to this threshold norm. As this simulation illustrates, the threshold norm is not stable: the threshold in the population moves up (becomes 'blue-er') until eventually everyone never punishes.

Supplementary Information:
Why norms are categorical

Moshe Hoffman, Erez Yoeli, N. Aygun Dalkiran, and Martin A. Nowak

February 17, 2018

Contents

List of Figures

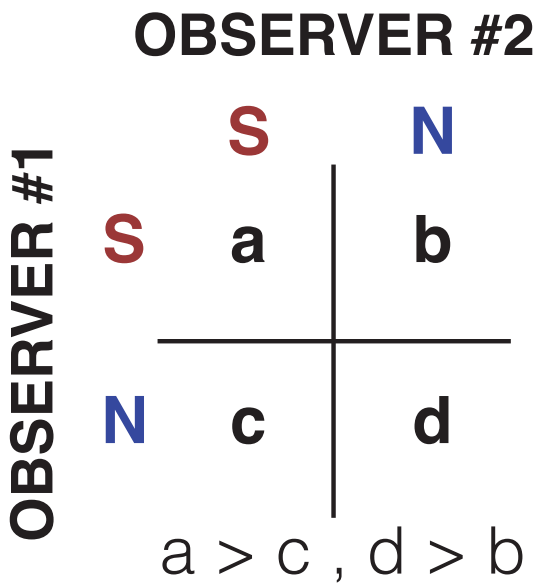


Figure 1: The Sanctioning Game

A Base Model

For convenience, we repeat the exposition of the model presented in the manuscript. We suppose there are two witnesses to a transgression who each simultaneously choose whether to 'sanction', **S**, or 'not sanction', **N**, a norm violation. For simplicity, we assume that sanctioning is optimal if and only if you expect others to sanction. That is, we assume players play the Sanctioning Game (Fig. 1) –a special case of the symmetric Coordination Game, which is the simplest game in which whether a player is best off sanctioning a transgression depends not just on whether she believes a transgression has occurred, but on whether she believes others believe a transgression has occurred. In the Sanctioning Game, a player receives a if both choose **S**, b if she chooses **S** and the other chooses **N**, c if she chooses **N** and the other chooses **S**, and d if both choose **N**, where $a > c$ and $d > b$. The game's pure equilibria are $(\mathbf{S}, \mathbf{S})$ and $(\mathbf{N}, \mathbf{N})$: if both players are playing **S**, neither can benefit from deviating, and similarly for **N**. The parameter $p = (d - b)/(a - c + d - b)$ will prove useful. Its interpretation is: if one player plays **S** with probability greater than p, then the other prefers to play **S**. As is sometimes the convention for symmetric games, we omit the column player's payoffs from the figure.

A.1 No Threshold Norm Can Be Sustained as a Bayesian Nash Equilibrium

We model a continuous transgression as follows: we presume that the harm from the transgression, H, varies continuously according to a random variable with known distribution. For simplicity, we presume this distribution is uniform over some range, $H \sim U[h_l, h_h]$, where h is a realization of H.

We then presume that each witness obtains a signal of the harm, $S_i|H$. For simplicity, we again presume the signal is uniform within a small range, ϵ around the true harm, $S_i \sim U[H-\epsilon, H+\epsilon]$, where s_i is a realization of S_i. We will consider how our results generalize to other distributions for H and $S_i|H$ in Section B.2. Also, for now, we assume the payoffs of sanctioning are independent of H (e.g., the material costs of sanctioning a country does not depend on the magnitude of its human rights violations); we will generalize to the case where the payoffs depend on H in Section B.4.

In the manuscript, we claimed that there is no Bayesian Nash equilibrium where the players condition their sanctioning decision on (their signal of) the magnitude of the harm, regardless of how small the error is in observing the harm level. We formalize this claim in Thm. 1. Technically, we show that there cannot exist a norm in which players sanction whenever their signal of the harm is above some threshold (we call this a threshold norm). As one might notice, the claim is true so long as the threshold is not 'too near' (within 2ϵ) of h_l or h_h. So long as the noise, ϵ, is small, we interpret any threshold within this distance of the endpoints as being observationally equivalent or qualitatively similar to the norm 'always sanction' and 'never sanction'.

We begin by defining a threshold strategy and a threshold norm.

Definition 1. *We say player i is playing a* threshold strategy $\sigma_i^{\bar{s}}$ *with a threshold at $\bar{s} \in [h_l, h_h]$ if and only if $\forall s_i \in [h_l - \epsilon, h_h + \epsilon], \sigma_i^{\bar{s}}(s_i) = \mathbf{S}$ if and only if $s_i > \bar{s}$. We say that players are playing a* threshold norm *at $\bar{s}$ if and only if for every player i, $\sigma_i = \sigma_i^{\bar{s}}$.*

To aid with proving Thm. 1, Lemma 2, and with the generation of some of our figures, we first calculate the following posterior probability: for any signal S_i, what is the likelihood that the other's signal is below some threshold, $S_{-i} < s$? The result of this calculation is presented in Lemma 1. Note that these calculations are not essential for the proofs.

Lemma 1. *For any $s \in [h_l + 2\epsilon, h_h - 2\epsilon]$*

$$Pr\left(S_{-i} < s | S_i\right) = \begin{cases} 0 & \text{for } s < S_i - 2\epsilon \\ \frac{(s-S_i)^2 + 4\epsilon(s-S_i)}{8\epsilon^2} + 1/2 & \text{for } S_i - 2\epsilon < s < S_i \\ \frac{-(s-S_i)^2 + 4\epsilon(s-S_i)}{8\epsilon^2} + 1/2 & \text{for } S_i < s < S_i + 2\epsilon \\ 1 & \text{for } S_i + 2\epsilon < s \end{cases}$$

Proof. First, it will be convenient to define $u_i = S_i - H$ and $z = S_i - s$.

Then, after some simplification, we see that, $Pr\left(S_{-i} < s | S_i\right) = Pr\left(u_1 - u_2 < z\right)$. Since $u_1 \sim -u_1$, $Pr\left(u_1 - u_2 < z\right) = Pr\left(u_1 + u_2 < z\right)$.

Next, we determine the *P.D.F.* of $u_1 + u_2$. As a sum of two uniform distributions, the *P.D.F.* is the convolution of two uniform density functions, where the density function of the uniform distribution is:

$$f(u_i) = \begin{cases} \frac{1}{2\epsilon} & \text{for } u_i \in [-\epsilon, \epsilon] \\ 0 & \text{otherwise} \end{cases}$$

For $-2\epsilon < 4z < 0$:

$$\begin{aligned} f_{u_1+u_2}(z) &= \int_{-\infty}^{\infty} f_u(z-y) f_u(y) dy \\ &= \int_{-\epsilon}^{\epsilon} f_u(z-y) \frac{1}{2\epsilon} dy \\ &= \int_{-\epsilon}^{z+\epsilon} \frac{1}{4\epsilon^2} dy \\ &= \frac{z+2\epsilon}{4\epsilon^2} \end{aligned}$$

Next, we calculate the *C.D.F.*:

$$\begin{aligned} Pr\left(u_1+u_2 < z\right) &= \int_{-\infty}^{z} f_{u_1+u_2}(z) \\ &= \int_{-2\epsilon}^{z} \frac{x+2\epsilon}{4\epsilon^2} dx \\ &= \frac{1/2z^2 + 2\epsilon z}{4\epsilon^2} - \frac{1/2(2\epsilon)^2 - 4\epsilon^2}{4\epsilon^2} \\ &= \frac{z^2+4\epsilon z}{8\epsilon^2} + 1/2 \end{aligned}$$

Thus, $Pr\left(S_{-i} < s | S_i\right) = \frac{(s-S_i)^2+4\epsilon(s-S_i)}{8\epsilon^2} + 1/2$ for $-2\epsilon < z < 0$.
Similarly, for $0 < z < 2\epsilon$ we find that $Pr\left(S_{-i} < s | S_i\right) = \frac{-(s-S_i)^2+4\epsilon(s-S_i)}{8\epsilon^2} + 1/2$. □

Now we are prepared to prove the claim in the manuscript.

Theorem 1. *For any $\bar{s} \in [h_l + 2\epsilon, h_h - 2\epsilon]$, if $\epsilon > 0$ and $p \neq \frac{1}{2}$, the threshold norm at $\bar{s}$ is not a Bayesian Nash Equilibrium.*

Proof. First, suppose that $p > 1/2$. We begin by using the distribution in Lemma 1 to find the signal, $s_i^p = \bar{s} - 2\epsilon + 2\epsilon\sqrt{2p}$ such that $Pr\left(S_{-i} > \bar{s} | s_i^p\right) = p$. Notice s_i^p exists for any $\epsilon > 0$, $p > 0$ and any $\bar{s} \in [h_l + 2\epsilon, h_h - 2\epsilon]$. Also, since $p > 1/2$, $-2\epsilon + 2\epsilon\sqrt{2p} > 0$, so $s_i^p > \bar{s}$. And, the distribution in Lemma 1 is strictly increasing in s_i, so for any $s_i < s_i^p$, $Pr\left(S_{-i} > \bar{s} | s_i\right) < p$.

Now suppose player i receives a signal just above the threshold. In particular, $s_i \in [\bar{s}, s_i^p)$. The strategy pair under consideration prescribes that player i play **S**. But, $Pr\left(S_{-i} > \bar{s} | S_i\right) < p$, so player i benefits from deviating to **N**.

A similar argument holds when $p < 1/2$: player i benefits from deviating for any $s_i \in (\bar{s} - 2\epsilon + 2\epsilon\sqrt{2p}, \bar{s}]$, since in this region, $Pr\left(S_{-i} < \bar{s} | S_i\right) < 1 - p$. □

In the manuscript, we claimed, "... our equilibria analysis is actually based on a solution concept that requires fewer assumptions than Nash: iterated elimination of strictly dominated strategies (see SI). And this weaker solution concept has the property, unlike Nash, that evolutionary processes will eventually converge to those strategies that satisfy iterative elimination of strictly dominated

strategies ().” We now argue that the only strategy profile that survives iterated elimination of dominated strategies is never sanction when $p > 1/2$, and always sanction when $p < 1/2$.

Lemma 2. *If $p > 1/2$, the only threshold norm which survives iterated elimination of strictly dominated strategies is at $h_h - \epsilon$. If $p < 1/2$, it is the threshold norm at $h_l + \epsilon$.*

Proof. Let $p > 1/2$.

The best response to the threshold strategy $\sigma_{-i}^{\bar{s}}$ is $\sigma_i^{\bar{s}+2\epsilon\sqrt{2p}-2\epsilon}$, since at this threshold strategy, i plays **S** if and only if $Pr(S_{-i} > \bar{s}) \geq p$. (To see that the best response to a threshold strategy must be another threshold strategy, first note that probability the other player sanctions is weakly increasing in one's own signal, S_i. Then, for a given signal s, if $\mathbb{E}U(\mathbf{N}) > \mathbb{E}U(\mathbf{S})$, then $Pr(S_{-i} > \bar{s}|S_i = s) < p$, and any best response must involve **N** when $S_i <= s$. A similar argument can be made for when $\mathbb{E}U(\mathbf{N}) < \mathbb{E}U(\mathbf{S})$. These arguments together imply that if **N** is strictly optimal for a given signal, then it is also strictly optimal for any smaller signal.)

This implies that the strategies $\sigma_i^{\bar{s}}|\bar{s} \in [h_l + \epsilon, h_l + \epsilon + (2\epsilon\sqrt{2p} - 2\epsilon))$ are strictly dominated and can be eliminated. By the same logic, we next eliminate strategies $\sigma_i^{\bar{s}}|\bar{s} \in [h_l + \epsilon + (2\epsilon\sqrt{2p} - 2\epsilon), h_l + \epsilon + 2(2\epsilon\sqrt{2p} - 2\epsilon))$. *Etc.* Notice that we are eliminating a range of constant size each time. Hence, we can repeat this until we are left with a subset of $\{\sigma_i^{\bar{s}}|\bar{s} \in [h_h - \epsilon - (2\epsilon\sqrt{2p} - 2\epsilon), h_h - \epsilon]\}$. $\sigma_i^{h_h - \epsilon}$ strictly dominates all remaining strategies: even if player $-i$ is playing according to the threshold strategy at $h_h - \epsilon - (2\epsilon\sqrt{2p} - 2\epsilon)$, which maximizes the likelihood that $-i$ sanctions, then player i prefers not to sanction anywhere in $[h_h - \epsilon - (2\epsilon\sqrt{2p} - 2\epsilon), h_h - \epsilon]$.

When $p < 1/2$, the argument is symmetric. The best response to the threshold strategy $\sigma_{-i}^{\bar{s}}$ is now $\sigma_i^{\bar{s}-(2\epsilon\sqrt{2p}-2\epsilon)}$, since at this threshold strategy, i plays **S** if and only if $Pr(S_{-i} > \bar{s}) \geq p$. This time, it's the strategies $\sigma_i^{\bar{s}}|\bar{s} \in (h_h - \epsilon - (2\epsilon\sqrt{2p} - 2\epsilon), h_h - \epsilon]$ that can first be eliminated. We next eliminate strategies $\sigma_i^{\bar{s}}|\bar{s} \in (h_h - \epsilon - 2(2\epsilon\sqrt{2p} - 2\epsilon), h_h - \epsilon - (2\epsilon\sqrt{2p} - 2\epsilon)]$. *Etc.* We repeat until we are left with a subset of $\{\sigma_i^{\bar{s}}|\bar{s} \in [h_l + \epsilon, h_l + \epsilon + (2\epsilon\sqrt{2p} - 2\epsilon)]\}$. $\sigma_i^{h_l + \epsilon}$ strictly dominates all remaining strategies. □

A.2 A Categorical Norm Can Be Sustained as a Bayesian Nash Equilibrium

We model a categorical transgression as a binary random variable, H, where the transgression occurs ($H = 1$) with probability q. We allow for the possibility that the witnesses observe the transgression with some error, ϵ. That is, $S_i = H$ with probability ϵ and $S_i = 1 - H$ with probability $1 - \epsilon$.

We claimed that, so long as the amount of noise, ϵ, isn't too large, there is a Bayesian Nash equilibrium where players condition their sanctioning on their signal of whether the transgression occurred. In Thm. 2, we show this.

Theorem 2. *When $q \neq 1/2$, $p \neq 1/2$, and $\epsilon < 1/2$, the strategy pair in which player i plays **S** if and only if $S_i = 1$ is a Bayesian Nash equilibrium if and only if:*

$$\epsilon \leq min\left(\frac{1 + p(1-2q) - \sqrt{p^2(2q-1)^2 - 2p + 1}}{2}, \frac{p + 2q(1-p) - \sqrt{4q^2(p-1)^2 - 4q(p-1)^2 + p^2}}{2}\right)$$

Proof. First, we present the eight possible states of the world, and their priors:

1. No transgression, neither detected a transgression, $(1-q)(1-\epsilon)^2$
2. No transgression, only witness 1 detected a transgression, $(1-q)\epsilon(1-\epsilon)$
3. No transgression, only witness 2 detected a transgression, $(1-q)\epsilon(1-\epsilon)$
4. No transgression, both witnesses detected a transgression, $(1-q)\epsilon^2$
5. Transgression, neither detected a transgression, $(1-q)\epsilon^2$
6. Transgression, only witness 1 detected a transgression, $(1-q)\epsilon(1-\epsilon)$
7. Transgression, only witness 2 detected a transgression, $(1-q)\epsilon(1-\epsilon)$
8. Transgression, both witnesses detected a transgression, $(1-q)(1-\epsilon)^2$

We'll refer to these states and their prior probabilities in the proof.

The proof proceeds as follows. We determine the conditions under which the threshold norm is an equilibrium, we require the player's best response is to sanction upon observing harm and to not sanction upon observing no harm. This generates multiple conditions, some of which are not so relevant either because they require that parameters take very precise values ($p = 1/2$ or $q = 1/2$) or that the amount of noise be large ($\epsilon \geq 1/2$), so we rule these conditions out.

We start by checking the two deviations: not sanctioning when observing harm, and sanctioning when observing no harm.

Suppose player i does not detect the transgression, $s_i = 0$. The strategy under consideration prescribes that i play $\mathbf{N}$. She will not benefit from deviating so long as $\frac{\epsilon(1-\epsilon)}{(1-q)(1-\epsilon)^2+q\epsilon^2+\epsilon(1-\epsilon)} \leq p$. The left hand side of this equation is the conditional likelihood $-i$ detected a transgression, given that i did not detect a transgression. The numerator is the sum of the priors on states 3 and 7, and the denominator is the sum of the priors on states 1, 3, 5, and 7. With some simplification, we see that this condition holds whenever one of the following three conditions is true:

1. $\epsilon \leq \frac{1+p(1-2q)-\sqrt{p^2(2q-1)^2-2p+1}}{2}$
2. $\epsilon \geq \frac{1+p(1-2q)+\sqrt{p^2(2q-1)^2-2p+1}}{2}$
3. $\frac{1-2\sqrt{(1-q)q}}{(2q-1)^2} < p < \frac{1+2\sqrt{(1-q)q}}{(2q-1)^2}$

Suppose player i detects the transgression, $s_i = 1$. The strategy under consideration prescribes that i play $\mathbf{S}$. She will not benefit from deviating so long as $\frac{q(1-\epsilon)^2+(1-q)\epsilon^2}{q(1-\epsilon)^2+(1-q)\epsilon^2+\epsilon(1-\epsilon)} \geq p$. The left hand side of this equation is the conditional likelihood $-i$ detected a transgression, given that i detected the transgression. The numerator is the sum of the priors on states 4 and 8. The denominator is the sum of the priors on states 2, 4, 6, and 8. Similar to above, with some simplification, we see that this condition holds whenever one of the following three conditions is true:

4. $\epsilon \leq \frac{p+2q(1-p)-\sqrt{4q^2(p-1)^2-4q(p-1)^2+p^2}}{2}$

5. $\epsilon \geq \frac{p+2q(1-p)+\sqrt{4q^2(p-1)^2-4q(p-1)^2+p^2}}{2}$

6. $\frac{4q(q-1)-2\sqrt{q(1-q)}}{(2q-1)^2} < p < \frac{4q(q-1)+2\sqrt{q(1-q)}}{(2q-1)^2}$

Now, we'll show that some of these conditions are not relevant once we assume $p \neq 1/2$, $q \neq 1/2$, and $\epsilon < 1/2$.

We begin with condition 2. Suppose $\epsilon < 1/2$. Condition 2 therefore implies $1/2 \geq \frac{1+p(1-2q)+\sqrt{p^2(2q-1)^2-2p+1}}{2}$. We simplify and find that it can hold only if $q \geq 1/2$.

Now consider condition 5. Again, suppose $\epsilon < 1/2$. Condition 5 therefore implies $1/2 \geq \frac{p+2q(1-p)+\sqrt{4q^2(p-1)^2-4q(p-1)^2+p^2}}{2}$. This can hold only if $q \leq 1/2$. Thus, when $\epsilon < 1/2$, conditions 2 and 5 can only both hold if $q = 1/2$.

Lastly, we consider when the conditions 3 and 6 hold. We will show the two are mutually exclusive when $p \neq 1/2$.

First, we show that the left hand side of condition 3 is less than or equal to $1/2$ only if $q = 1/2$, where it holds with equality. This implies when $p < 1/2$, equation 3 is violated.

$$\begin{aligned} \frac{1-2\sqrt{(1-q)q}}{(2q-1)^2} &\leq 1/2 \\ \implies (2q-1)^4 &\leq 0 \\ \implies q &= 1/2 \end{aligned}$$

Next, we show the right hand side of equation 6 is greater than or equal to $1/2$ only if $q = 1/2$ where it holds with equality. This implies that when $p > 1/2$, equation 6 is violated.

$$\begin{aligned} 1/2 &\leq \frac{4q(q-1)+2\sqrt{q(1-q)}}{(2q-1)^2} \\ \implies (2q-1)^4 &\leq 0 \\ \implies q &= 1/2 \end{aligned}$$

Then, the left hand side of condition 3 cannot be less than $1/2$ and the right hand side of condition 6 cannot be greater than $1/2$. This implies that both conditions can be satisfied only if $p = 1/2$ (and $q = 1/2$). □

A.3 Partner Choice

Next, we present a simple setting where sanctioning is incentive compatible and does not involve a coordination element, in which case, a threshold norm is possible. In this setting–a stylized 'partner choice' game–there is variation in the types of the perpetrators, where 'good' types have

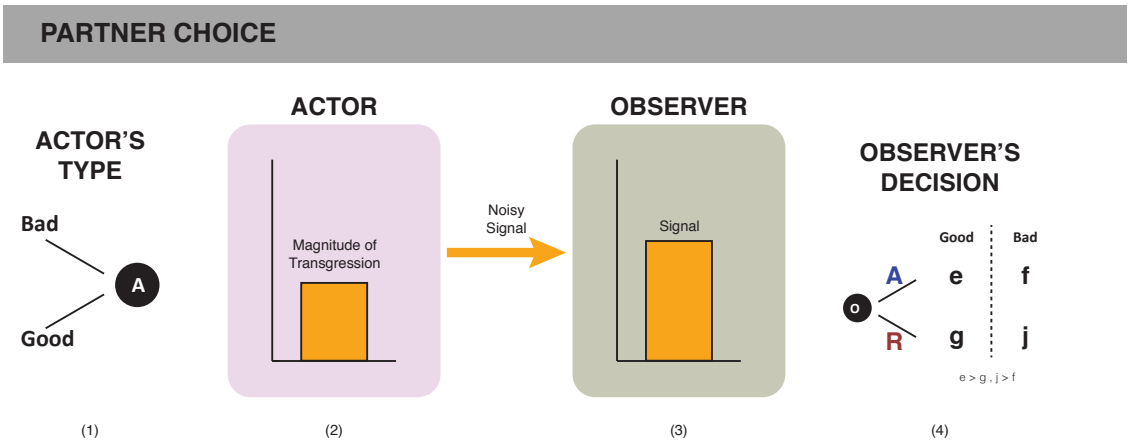


Figure 2: A Partner Choice Game

distributions of transgressions that tend to be less harmful, and these types of perpetrators make more desirable 'partners'. The game proceeds as follows (Fig. 2). (1) First, the type of the perpetrator is determined: with probability q he is good, and with probability $1-q$ he is bad. (2) A transgression occurs, and its distribution is determined by the perpetrator's type. If the perpetrator is good, the harm is distributed according to the probability distribution function $f_g(h)$; if the perpetrator is bad, the harm is distributed $f_b(h)$. Let $F_i(h)$ represent the *C.D.F.* of $f_i(h)$ for $i = \{b, g\}$. We assume these distributions are continuous and have support over $[h_l, h_h]$. (3) A single witness receives a noisy signal of the magnitude, uniformly distributed about the true magnitude, $S \sim U[h-\epsilon, h+\epsilon]$ where $\epsilon > 0$. (4) Then, the witness decides whether to accept, A, or reject, R, the perpetrator as a partner. Her payoffs depend on the perpetrator's type: if the perpetrator is good and she accepts him, she gets e, whereas if she rejects him she gets $g < e$; if the perpetrator is bad and she accepts him, she gets f, whereas if she rejects him she gets $j > f$.

We show that it is a Bayesian Nash equilibrium for the witness to play a threshold strategy, $\sigma^{\bar{s}}$, as defined in definition 1.

Lemma 3. *Let f_g, f_b be continuous with support over $[h_l, h_h]$, and $(f_b(S+\epsilon) - f_b(S-\epsilon))(F_g(S+\epsilon) - F_g(S-\epsilon)) > (F_b(S+\epsilon) - F_b(S-\epsilon))(f_g(S+\epsilon) - f_g(S-\epsilon))$ for all $S \in [h_l, h_h]$.*

Also, let $Pr(B|h_l) \leq \frac{j-g}{j-g+e-f}$ and $Pr(B|h_h) > \frac{j-g}{j-g+e-f}$, where B represents the 'bad' type. Then, there exists a threshold $\bar{s} \in [h_l, h_h]$ such that, in equilibrium, the witness rejects if and only if $S > \bar{s}$. Moreover, this equilibrium is unique.

Proof. Consider the witness's decision problem. As in the original setup (Section A.1), If the perpetrator is bad with probability greater than $\frac{j-g}{j-g+e-f}$, the witness prefers to reject. If the perpetrator is bad with lower probability the witness prefers to accept. (And, if with probability equal to $\frac{j-g}{j-g+e-f}$, the witness is indifferent.)

Given a signal $S \in [h_l + \epsilon, h_h - \epsilon]$, their posterior regarding the type of the perpetrator is

$Pr(B|S) = \frac{F_b(S+\epsilon)-F_b(S-\epsilon)}{F_b(S+\epsilon)-F_b(S-\epsilon)+F_g(S+\epsilon)-F_g(S-\epsilon)}$, by Baye's rule. This will also hold in the 'edge' cases. When the signal is near the lower boundary, $S \in [h_l, h_l+\epsilon]$ $Pr(B|S) = \frac{F_b(S+\epsilon)-F_b(h_l)}{F_b(S+\epsilon)-F_b(h_l)+F_g(S+\epsilon)-F_g(h_l)} = \frac{F_b(S+\epsilon)-F_b(S-\epsilon)}{F_b(S+\epsilon)-F_b(S-\epsilon)+F_g(S+\epsilon)-F_g(S-\epsilon)}$), (both f_g and f_b are zero outside $[h_l, h_h]$). Similarly, in the 'edge' case where the signal is near the upper boundary, $Pr(B|S) = \frac{F_b(S+\epsilon)-F_b(S-\epsilon)}{F_b(S+\epsilon)-F_b(S-\epsilon)+F_g(S+\epsilon)-F_g(S-\epsilon)}$).

We wish to show that $Pr(B|S)$ is increasing in S. We begin by taking the partial with respect to S. When $S \in [h_l, h_h]$, we find:

$$\frac{(f_b(S+\epsilon)-f_b(S-\epsilon))(F(S+\epsilon)-F(S-\epsilon))-(F_b(S+\epsilon)-F_b(S-\epsilon))(f(S+\epsilon)-f(S-\epsilon))}{(F(S+\epsilon)-F(S-\epsilon))^2}$$

This simplifies to:

$$(f_b(S+\epsilon)-f_b(S-\epsilon))(F_g(S+\epsilon)-F_g(S-\epsilon)) > (F_b(S+\epsilon)-F_b(S-\epsilon))(f_g(S+\epsilon)-f_g(S-\epsilon))$$

We simplify and find that this is strictly positive if and only if:

$$(f_b(S+\epsilon)-f_b(S-\epsilon))(F_g(S+\epsilon)-F_g(S-\epsilon)) \geq (F_b(S+\epsilon)-F_b(S-\epsilon)(f_g(S+\epsilon)-f_g(S-\epsilon))$$

Finally, suppose $Pr(B|h_l) \leq \frac{j-g}{j-g+e-f}$ and $Pr(B|h_h) > \frac{j-g}{j-g+e-f}$. Then, given that f_b and f_g are both continuous, and $Pr(B|h_l) < \frac{j-g}{j-g+e-f} < Pr(B|h_h)$, there exists a unique $\bar{s}$ s.t. $Pr(B|\bar{s}) = \frac{j-g}{j-g+e-f}$ by the Intermediate Value Theorem. Then, the witness's optimal strategy is to reject if and only if $S > \bar{s}$. □

In Theorem 3, we show that the condition in Lemma 3 is met–and thus a threshold strategy is optimal–if the distributions f_b and f_g satisfy the monotone likelihood ratio property.

Theorem 3. *Let f_g, f_b be continuous with support on $[h_l, h_h]$ and $\frac{f_g(h)}{f_b(h)}$ is decreasing for $h \in [h_l, h_h]$. Also, let $Pr(B|h_l) \leq \frac{j-g}{j-g+e-f}$ and $Pr(B|h_h) > \frac{j-g}{j-g+e-f}$. Then, there exists $\bar{s} \in [h_l, h_h]$ s.t. the witness rejects if and only if $S > \bar{s}$.*

Proof. First, let $S+\epsilon \geq x$. Then, by the monotone likelihood ratio property *(MLRP)*, $\frac{f_g(S+\epsilon)}{f_b(S+\epsilon)} \leq \frac{f_g(x)}{f_b(x)}$.

We perform the following algebra:

$$\begin{aligned}
\frac{f_g(S+\epsilon)}{f_b(S+\epsilon)} &\leq \frac{f_g(x)}{f_b(x)} \\
f_g(S+\epsilon)f_b(x) &\leq f_g(x)f_b(S+\epsilon) \\
\int_{S-\epsilon}^{S+\epsilon} f_g(S+\epsilon)f_b(x) &\leq \int_{S-\epsilon}^{S+\epsilon} f_g(x)f_b(S+\epsilon) \\
f_g(S+\epsilon)[F_b(S+\epsilon)-F_b(S-\epsilon)] &\geq f_b(S+\epsilon)[F_g(S+\epsilon)-F_g(S-\epsilon)] \qquad (1)
\end{aligned}$$

Next, we let $S-\epsilon \leq x$. Then, by the MLRP, $\frac{f_g(S-\epsilon)}{f_b(S-\epsilon)} \geq \frac{f_g(x)}{f_b(x)}$. We perform a similar algebraic

transformation, and find:

$$f_g(S-\epsilon)[F_b(S+\epsilon)-F_b(S-\epsilon)] \leq f_b(S-\epsilon)[F_g(S+\epsilon)-F_g(S-\epsilon)] \tag{2}$$

Subtracting equation 2 from equation 1 gives the condition in Lemma 3:

$$(f_g(S+\epsilon)-f_g(S-\epsilon))(F_b(S+\epsilon)-F_b(S-\epsilon)) \leq (f_b(S+_\epsilon)-f_b(S-\epsilon))(F_g(S+\epsilon)-F_g(S-\epsilon))$$

Thus, by Lemma 3, there exists a threshold equilibrium. □

B Robustness and Extensions

B.1 N Discrete Values

In the manuscript, we stated, "...consider what happens if we replace the continuous variable with a discrete variable that is equally likely to take on one of many (n) values. We find that the larger n gets the closer p needs to be to 50% in order to allow for an equilibrium that depends on the signal (Fig. 2)." To demonstrate this, we solve the model presented in Section A.2, except:

- Instead of assuming that H is binary, we assume that H is uniformly distributed over the domain $\{1,\ldots,n\}$ (technically, $H \sim F(h) = \frac{\lfloor h \rfloor}{n}$), where $h \in \{1,\ldots,n\}$ is the realization of H.
- We assume players' signals are uniformly distributed over the domain $\left\{H - \lfloor \frac{n}{k} \rfloor, H - \lfloor \frac{n}{k} \rfloor + 1, \ldots, H + \lfloor \frac{n}{k} \rfloor\right\}$ (technically, $F(h) = \frac{\lfloor h \rfloor - \left(H - \lfloor \frac{n}{k} \rfloor\right)+1}{2\lfloor \frac{n}{k} \rfloor + 1}$), where $k \in \{1,\ldots,n\}$ represents the amount of noise.

Theorem 4. *The threshold norm $\sigma^{\bar{s}}$ is a Bayesian Nash equilibrium if and only if $\frac{\lfloor \frac{n}{k} \rfloor + 1}{2\lfloor \frac{n}{k} \rfloor + 1} > p$, for $\bar{s} \in \{1 + 2\lfloor \frac{n}{k} \rfloor, \ldots, n - 2\lfloor \frac{n}{k} \rfloor\}$.*

Proof. For a given h, there are $2\lfloor \frac{n}{k} \rfloor + 1$ possible signals, each of which is chosen with probability $\frac{1}{2\lfloor \frac{n}{k} \rfloor + 1}$. The posterior probability of $-i$'s signal given i's signal is $\#\{x, y \in \{\lfloor \frac{n}{k} \rfloor, \ldots, 2\lfloor \frac{n}{k} \rfloor\} | x + y = S_i + a\}$:

$$Pr\{S_{-i} = S_i + a\} = \begin{cases} 0 & \text{for } a < -2\lfloor \frac{n}{k} \rfloor \\ \frac{2\lfloor \frac{n}{k} \rfloor + 1 + a}{(2\lfloor \frac{n}{k} \rfloor + 1)^2} & \text{for } -2\lfloor \frac{n}{k} \rfloor \leq a \leq 0 \\ \frac{2\lfloor \frac{n}{k} \rfloor + 1 - a}{(2\lfloor \frac{n}{k} \rfloor + 1)^2} & \text{for } 0 < a \leq 2\lfloor \frac{n}{k} \rfloor \\ 0 & \text{for } 2\lfloor \frac{n}{k} \rfloor < a \end{cases}$$

Suppose $p > 1/2$ and consider what happens when a player receives signal just above the threshold, $S_i = \bar{s} + 1$. We use the posterior to calculate the likelihood the other's signal is also above the threshold, and therefore expected to play **S**:

$$Pr\{S_{-i} > \bar{s} | s_i = \bar{s}+1\} = \frac{\sum_{i=1}^{2\lfloor \frac{n}{k} \rfloor + 1} i}{(2\lfloor \frac{n}{k} \rfloor + 1)^2}$$
$$= \frac{\lfloor \frac{n}{k} \rfloor + 1}{2\lfloor \frac{n}{k} \rfloor + 1}$$

A threshold equilibrium exists if and only if $\frac{\lfloor \frac{n}{k} \rfloor + 1}{2\lfloor \frac{n}{k} \rfloor + 1} \geq p$.

Next, let $p < 1/2$. Suppose player i receives a signal at the threshold. We calculate the probability that the opposing player receives a signal above the threshold:

$$Pr\{S_{-i} > S_i | s_i = \bar{s}\} = \frac{\sum_{i=1}^{2\lfloor \frac{n}{k} \rfloor} i}{(2\lfloor \frac{n}{k} \rfloor + 1)^2}$$
$$= \frac{\lfloor \frac{n}{k} \rfloor}{2\lfloor \frac{n}{k} \rfloor + 1}$$

A threshold equilibrium exists if and only if $\frac{\lfloor \frac{n}{k} \rfloor}{2\lfloor \frac{n}{k} \rfloor + 1} \leq p$, since player i would not benefit by deviating to $\mathbf{S}$. □

This argument is illustrated in Fig. 3. First, suppose $n = 10$, $S_i \sim U[h-1, h+1]$, and $\bar{s} = 4$. For this to be a Nash equilibrium, when player i gets a signal of 5, she must believe player $-i$ got a signal of at least 5 with at least probability p, which happens if $p > 1/3$. And, when player i gets a signal of 4, she must believe $-i$ got a signal of 4 or below with probability $1-p$, which happens if $p < 2/3$. Thus, the threshold equilibrium exists for sanctioning games with payoffs such that p is within a moderate range of $\frac{1}{2}$, namely $p \in [\frac{1}{3}, \frac{2}{3}]$.

Now, suppose the number of states is expanded and the amount of noise in the signals remains proportionately the same: $n = 100$, $S_i \sim U[h-10, h+10]$, and $\bar{s} = 49$. Then, the range of values of p around $\frac{1}{2}$ that permit the threshold equilibrium shrinks to $p \in [\frac{10}{21}, \frac{11}{21}]$.

Notice that the range in which the threshold equilibrium exists approaches $p = \{\frac{1}{2}\}$ as $n \to \infty$.

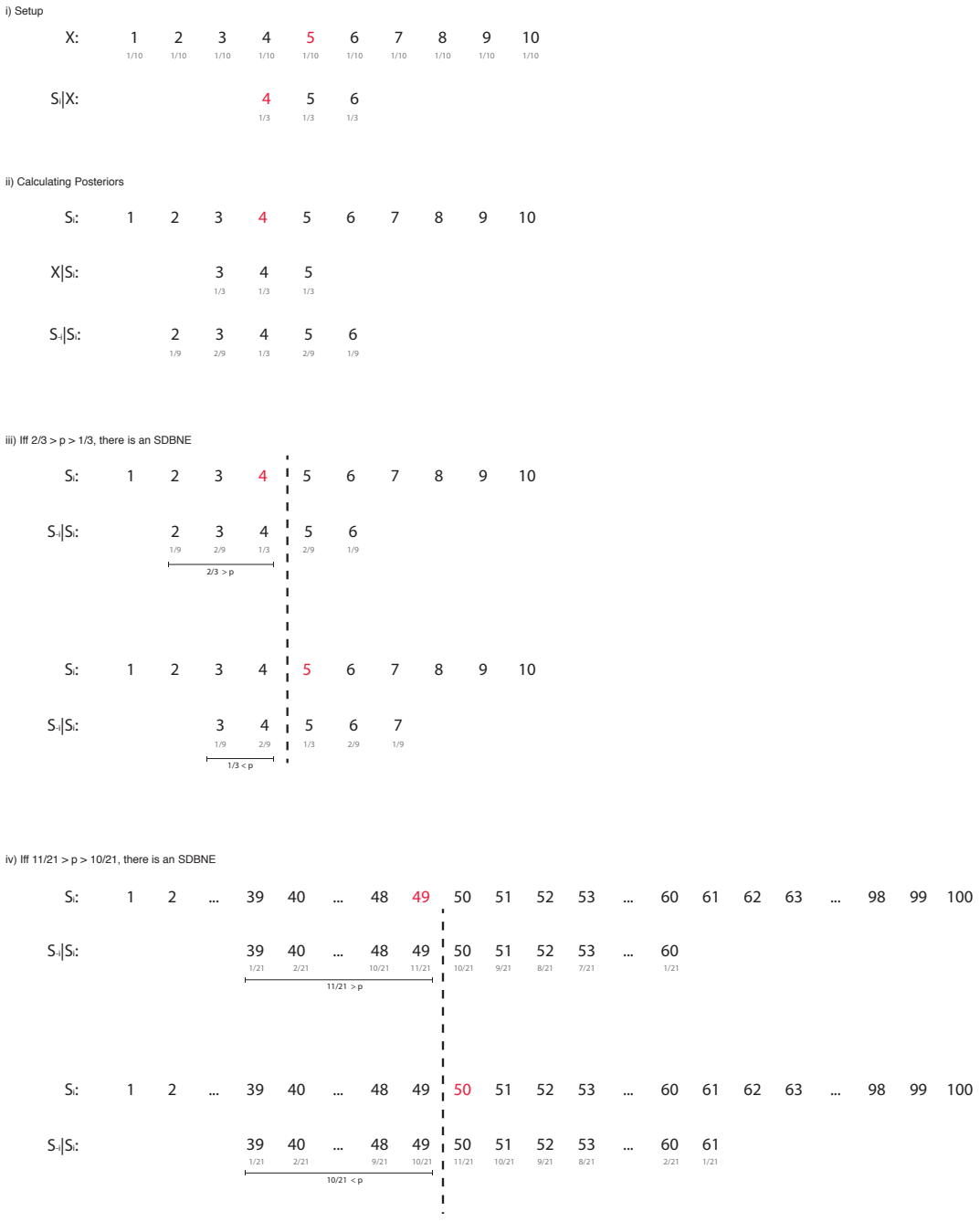


Figure 3: Transgressions With More Than Two Possible Values

B.2 General Result for Any Signal Distributions

The following result holds for any continuously differentiable distribution on the harm, H, and the witnesses' signals, S_i.

Theorem 5. *A threshold norm $\sigma^{\bar{s}}$ is a Bayesian Nash equilibrium if and only if*

$$\lim_{S_i \to \bar{s}^-} Pr\{S_{-i} \leq \bar{s} | S_i\} \leq p \leq \lim_{S_i \to \bar{s}^+} Pr\{S_{-i} \leq \bar{s} | S_i\}$$

Proof. $\sigma^{\bar{s}}$ is a Bayesian Nash equilibrium if and only if for all received signals S_i such that $\sigma(S_i) = \mathbf{S}$, $Pr\{S_{-i} \geq \bar{s} | S_i\} \geq p$, and for all received signals S_i such that $\sigma(S_i) = \mathbf{N}$, $Pr\{S_{-i} \geq \bar{s} | S_i\} \leq p$.

Since for all $\bar{s} \in [h_l + 2\epsilon, h_h - 2\epsilon]$ and all $\langle H, S_1, S_2 \rangle \in \Omega$, $\frac{\partial Pr\{S_{-i} \geq \bar{s} | S_i\}}{\partial S_i} \geq 0$, then for all $S_i \leq \bar{s}$, $Pr\{S_{-i} \leq \bar{s} | S_i\} \geq \lim_{S_i \to \bar{s}^-} Pr\{S_{-i} \leq \bar{s} | S_i\}$ and for all $S_i \geq \bar{s}$, $Pr\{S_{-i} \leq \bar{s} | S_i\} \leq \lim_{S_i \to \bar{s}^+} Pr\{S_{-i} \leq \bar{s} | S_i\}$.

Since a threshold strategy $\sigma(S_i) = \mathbf{S}$ if and only if $S_i > \bar{s}$.

$$\lim_{S_i \to \bar{s}^-} Pr\{S_{-i} \leq \bar{s} | S_i\} \leq p \leq \lim_{S_i \to \bar{s}^+} Pr\{S_{-i} \leq \bar{s} | S_i\}$$

□

B.3 Other Specific Distributions of Harm

Next, we consider some alternative distributions on H. First, let H be distributed as follows. There exists an atom at h_l, where $Pr(h = h_l) = a$. For any other value $h \in (h_l, h_h]$, the remaining $1 - a$ probability is uniformly distributed over the interval $(h_l, h_h]$. As before $s_i \sim U[h - \epsilon, h + \epsilon]$. We show that there is at most one, non-generic threshold equilibrium.

Lemma 4. *For any (p, a, h_l), there only exists at most one threshold norm which is an equilibrium.*

Proof. We begin by solving for the posterior probability that $H = h$ given i's signal: First, for $s_i \in (h_l + \epsilon, h_h - \epsilon)$: $Pr\{H < h | s_i\} = \begin{cases} 0 & h < s_i - \epsilon \\ \frac{h + \epsilon - s_i}{2\epsilon} & h_l < h < s_i + \epsilon \\ 1 & s_i + \epsilon < h \end{cases}$

For $s_i \in [h_l, h_l + \epsilon]$:

$$Pr\{H < h | s_i\} = \begin{cases} 0 & h < h_l \\ \frac{a}{a + \frac{(1-a)(s_i + \epsilon - h_l)}{h_h - h_l}} + \frac{\frac{(1-a)(s_i + \epsilon - h_l)}{h_h - h_l}}{a + \frac{(1-a)(s_i + \epsilon - h_l)}{h_h - h_l}} \cdot \frac{h - h_l}{s_i + \epsilon - h_l} & s_i - \epsilon < h < s_i + \epsilon \\ 1 & s_i + \epsilon < h \end{cases}$$

To ease notation, we let P_l represent the probability $h = h_l$, given $h \in [h_l, h_l + \epsilon]$, $P_l = \frac{a}{a + \frac{(1-a)(s_i + \epsilon - h_l)}{h_h - h_l}}$.

Next, we can solve for the likelihood the other's signal is less than some threshold, given one own's signal, $Pr\{S_{-i} < \bar{s} | S_i = s_i\}$.

First, for $s_i \in (h_l + \epsilon, h_h - \epsilon]$, the probability is the same as in section A.

Next, for $s_i \in [h_l, h_l + \epsilon]$:

We first calculate the PDF: $f(z) = \begin{cases} 0 & z < -2\epsilon \\ \frac{P_l}{2\epsilon} + (1 - P_l)\frac{z+\epsilon-h_l+s_i}{2\epsilon(s_i-h_l+\epsilon)} & -\epsilon + h_l - s_i < z < 0 \end{cases}$

Then the corresponding CDF:

$F(z) = \begin{cases} 0 & z < -2\epsilon \\ \frac{P_l(z-(-\epsilon+h_l-s_i))}{2\epsilon} + (1 - P_l)\frac{\frac{z^2-(h_l-s_i-\epsilon)^2}{2}+(z-h_l+s_i+\epsilon)(\epsilon-h_l+s_i)}{2\epsilon(s_i-h_l+\epsilon)} & -2\epsilon < z < 0 \end{cases}$

In particular, for $z = 0, s_i \in [h_l, h_l + \epsilon]$:

$Pr\{s_{-i} < s_i | s_i\} = \frac{(s_i+\epsilon-h_l)(1+P_l)}{4\epsilon}$

Notice, the derivative of this with respect to s_i is $1 - P_l^2$. Then, this function is monotonically increasing in s_i from $\frac{1+P_l}{4}$ at h_l to $\frac{1+P_l}{2}$ at $h_l + \epsilon$, where there is a discontinuity (for $a > 0$). For $(h_l + \epsilon, h_h - \epsilon]$ the function is exactly equal to half.

Begin by considering threshold norms $\sigma^{\bar{s}} | \bar{s} \in (h_l + \epsilon, h_h)$. Since, in this range, $Pr\{S_{-i} < \bar{s} | S_i = s_i\} = 1/2$, no threshold norm is a Bayesian Nash equilibrium by the same logic we employed in Thm. 1.

Next, we consider threshold norms $\sigma^{\bar{s}} | \bar{s} \in [h_l, h_l + \epsilon)$.

Let $\frac{1+P_l}{4} < p < \frac{1+P_l}{2}$. If $Pr\{S_{-i} < h_l | S_i = h_l\} > p$ and $Pr\{S_{-i} < h_l + \epsilon | S_i = h_l + \epsilon\} < p$, by the Intermediate Value Theorem and Thm. 5, there is exactly one threshold norm that is a Bayesian Nash equilibrium. Otherwise, by Thm. 5, no threshold norm can be a Bayesian Nash equilibrium. □

Corollary 1. *The threshold norm at $\bar{s} = \frac{(1+P_l)(s_i+\epsilon-h_l)}{4\epsilon}$ is the unique Bayesian Nash Equilibrium if and only if $\frac{1+P_l}{4} < p < \frac{1+P_l}{2}$. Otherwise, no threshold norm is a Bayesian Nash equilibrium.*

Proof. This follows directly from Lemma 4 and Theorem 5. The probability of player $-i$ sanctioning given signal $s_i = \bar{s}$ is given by $\frac{(1+P_l)(s_i+\epsilon-h_l)}{4\epsilon}$. Therefore, $\frac{(1+P_l)(s_i+\epsilon-h_l)}{4\epsilon} = p$ is a requirement for a Nash equilibrium. □

Recall, in the manuscript, we claimed, "We could instead replace the uniform distributions with normal distributions and we would find that provided the variance in the noise is not too large, there will not be an equilibria where the threshold depends on the state. The reason for this is that in normal distributions, the probability that the other player has a signal below yours will not be precisely 50%, but will depend on your signal, albeit it will still remain close to 50% if the noise is not too large. Thus, for small noise, there still won't be a threshold equilibria except for a tiny range of coordination games, where that range again will approach measure 0 as the level of noise approaches 0. Even if, however, the noise level is sufficiently large that there is an equilibria where sanctioning occurs above a certain signal, only a very specific threshold will be possible–the threshold at which my posterior that you got a signal above that threshold, given that I got a signal at that threshold, is precisely p. Thus, the threshold will be determined by p and the variance of the two distributions, not what's socially optimal (see figure 3b and SI)." To substantiate these claims, we let H be normally distributed, $H \sim N[0, 1]$, and let the noise also be

normally distributed, $S_i \sim N[h, \sigma^2]$.

Theorem 6. *If harm is distributed $H \sim N[0,1]$ and each player receives a signal $S_i = h + \epsilon_i$, where $\epsilon_i \sim N[0, \sigma^2]$ there exists at most one threshold norm at $\bar{s}$, where $\bar{s}$ is such that $Pr\left(z > \frac{\bar{s} - \frac{S_1}{\sigma^2+1}}{\sqrt{\frac{(\sigma^2+1)^2-1}{\sigma^2+1}}}\right) = p$ where $z \sim N[0,1]$.*

Proof. We begin by calculating the mean and variance of S_2 given S_1. Witnesses' signals are distributed according to the multivariate normal distribution $S = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$. The mean matrix is given by $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. The covariance matrix is given by $\Sigma = \begin{bmatrix} \sigma^2+1 & 1 \\ 1 & \sigma^2+1 \end{bmatrix}$. We note that $\mu_{S_2|S_1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(S_1 - \mu_1)$ which, upon simplifies to $\mu_{S_2|S_1} = \frac{S_1}{1+\sigma^2}$. We similarly calculate $\Sigma_{S_2|S_1} = \frac{(\sigma^2+1)^2-1}{\sigma^2+1}$. Finally, since the conditional distribution of a multivariate normal distribution is also normal [1], $S_2|S_1$ is distributed $N[\frac{S_1}{\sigma^2+1}, \frac{(\sigma^2+1)^2-1}{\sigma^2+1}]$.

By Theorem 5, there exists a threshold equilibrium $\sigma^{\bar{s}}$ if and only if $Pr(S_2 > \bar{s}|S_1 = \bar{s}) = p$. We calculate $Pr(S_2 > \bar{s}|S_1 = \bar{s}) = Pr\left(z > \frac{\bar{s} - \frac{S_1}{\sigma^2+1}}{\sqrt{\frac{(\sigma^2+1)^2-1}{\sigma^2+1}}}\right)$. Thus we conclude there exists a threshold equilibrium $\sigma^{\bar{s}}$ if and only if $Pr\left(z > \frac{\bar{s} - \frac{S_1}{\sigma^2+1}}{\sqrt{\frac{(\sigma^2+1)^2-1}{\sigma^2+1}}}\right) = p$ where $z \sim N[0,1]$. □

The intuition behind this result is represented in Fig. 4. (1) When harm is normally distributed, an observer who gets a signal that is higher than the mean harm level will think it is more likely that the other observer got a signal below hers–exactly how much depends on her signal and the variance of h–and vice versa for signals lower than the mean level (we have drawn the former case). Critically, observer 1's posterior is still not equal to p, so her best response involves deviating from the threshold norm. (2) If the variance of h is relatively large, then the likelihood the other observer got a signal below hers will be relatively close to 50%. In this case, there still will be no threshold equilibrium. If the variance of h is relatively small (shown), then, for signals far from the mean of h, the likelihood the other observer got a signal below hers can diverge meaningfully from 50%. In such cases, it can be possible for a threshold norm to be an equilibrium. However, there will be only one, highly specific threshold, s^*, that can be sustained in equilibrium, and that threshold will depend on p and the variance of h, and not on what is socially optimal. Moreover, this equilibrium is unstable.

B.4 State-Dependent Payoffs

In the manuscript, we wrote, "We now consider what happens if the continuous variable directly influences the payoffs to sanctioning. In particular, we consider the case where the payoffs are as before but we add an additional benefit to sanctioning that is an increasing function of the magnitude of the transgression. Now $a(h)$ and $b(h)$ both depend on the magnitude of the transgression, and if the dependency is sufficiently strong, then, as with normally distributed harm, it is possible to

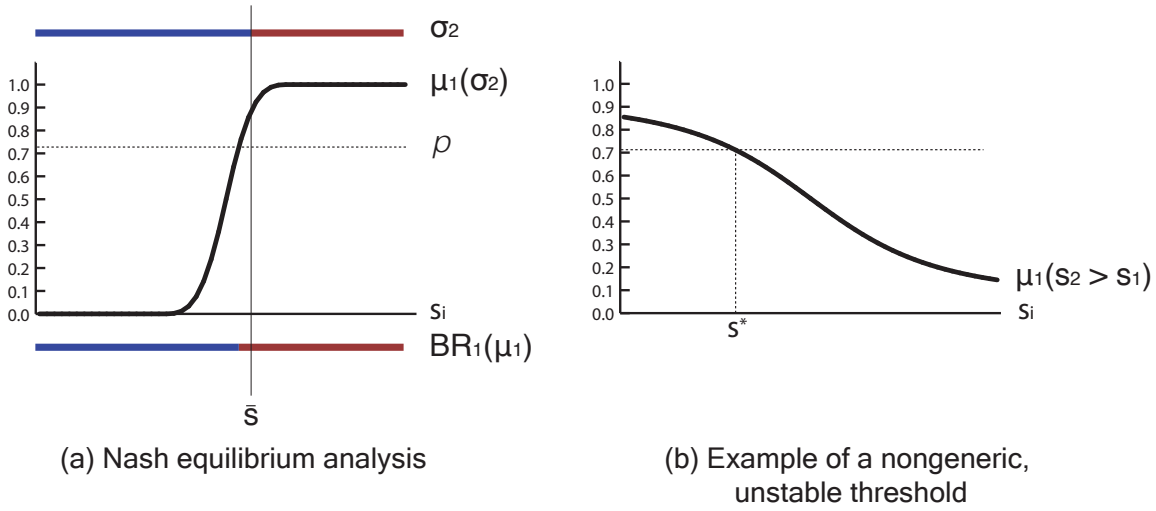


Figure 4: Normally Distributed Transgression.

support a single continuous norm in equilibrium. But, once again, the threshold will be determined by the strength of this dependency, and not all socially relevant considerations (Fig. 4)."

To show this, we solve the model presented in Section A.1, except we allow the cost of sanctioning to depend on the size of the transgression. In particular, we make the payoff when a player sanctions a monotonic and differentiable function of h: $u_1(\mathbf{S},\mathbf{S}) = u_2(\mathbf{S},\mathbf{S}) = a(h)$ and $u_1(\mathbf{S},\mathbf{N}) = u_2(\mathbf{N},\mathbf{S}) = b(h)$. We assume, for all $h \in [0,1]$: $a'(h) > 0$, $b'(h) > 0$, $a(h) > b(h)$, $b > 0$, and $a(h) > c, b(h) < d$.

We will find it useful to define $p(h)$ as the state-dependent analog of the risk-dominance, p:

$$p(h) = \frac{d - b(h)}{d - b(h) + a(h) - c}$$

To prove that there is at most one uniquely defined equilibrium, we will show that $p(h)$ is strictly decreasing. By Theorem 5, a threshold norm is a threshold equilibrium only if the probability of the other sanctioning is equal to $p(h)$. Then, given the probability of the other sanctioning is equal to $1/2$ (for any interior h), there can be at most one threshold equilibrium where $p(h)$ intersects with $1/2$.

We begin by showing that $p(h)$ is decreasing in h.

Lemma 5. $p'(h) < 0$.

Proof. We begin by calculating $p'(h)$:

$$p'(h) = \frac{-b'(h)[d - b(h) + a(h) - c] - (d - b(h))(a'(h) - b'(h))}{(d - b(h) + a(h) - c)^2} \tag{3}$$

The assumptions $d - b(h) > 0$ and $a(h) - c > 0$ imply $d - b(h) + a(h) - c > 0$. We then multiply equation 3 by the denominator, $(d - b(h) + a(h) - c)^2$, a strictly positive number. Therefore the sign does not change. This yields:

$$-b'(h)[d - b(h) + a(h) - c] - (d - b(h))(a'(h) - b'(h))$$

We can rearrange this to:

$$-[a'(h)(d - b(h)) + b'(h)(a(h) - c)]$$

The assumptions $a'(h), b'(h) > 0$ and $d > b(h), a(h) > c$, imply that the overall expression is negative. Therefore $p'(h) < 0$. □

Now we are ready to find conditions under which a threshold equilibrium exists.

Theorem 7. *There exists $\bar{s} \in [h_l + \epsilon, h_h - \epsilon]$ such that $\sigma^{\bar{s}}$ is a threshold equilibrium if and only if $p(h_l + \epsilon) \geq 1/2$ and $p(h_h - \epsilon) \leq 1/2$. Furthermore, such a $\bar{s}$ must be unique.*

Proof. Notice that the posterior $Pr\left(S_{-i} < s | s_i\right)$ is unchanged from Lemma 1, since the change in payoffs doesn't influence the posterior.

By the same logic as in Theorem 5, there exists a threshold equilibrium at $\bar{s}$ if and only if $p(\bar{s}) = 1/2$.

Suppose $p(h_l + \epsilon) \geq 1/2$ and $p(h_h - \epsilon) \leq 1/2$. Then, by the Intermediate Value Theorem and because $p(h)$ is continuous (continuity is preserved by multiplication), there exists a single $\bar{s} \in [h_l + \epsilon, h_h - \epsilon]$ such that $p(\bar{s}) = 1/2$. This $\bar{s}$ is a threshold equilibrium.

Next, suppose $p(h_l + \epsilon) < 1/2$. As $p(h)$ is strictly decreasing, there cannot exist $\bar{s} \in (h_l + \epsilon, h_h - \epsilon]$ such that $p(\bar{s}) = 1/2 > p(h_l + \epsilon)$. Therefore, there exists no threshold equilibrium. Similarly, if $p(h_h - \epsilon) > 1/2$ there cannot exist a threshold equilibrium.

Lastly, we prove uniqueness. Let there exist $\bar{s}, \bar{s}'$ such that both $\sigma^{\bar{s}}, \sigma^{\bar{s}'}$ are threshold equilibria. Without loss of generality, let $\bar{s} < \bar{s}'$. Then $p(\bar{s}) = p(\bar{s}') = 1/2$. But $p'(h) < 0$ implies that $p(\bar{s}) > p(\bar{s}')$, a contradiction. Therefore, there exists at most one threshold equilibrium. □

B.5 More Than Two Witnesses

In the manuscript, we stated, "...we extend our model to take into account the fact that norms are often enforced by large groups of potential sanctioners, and not just two, and again obtain comparable results."

We retain the setup in Section A, with the following modifications:

1. Instead of assuming there are two witnesses, we assume there are n witnesses.

2. Witnesses' payoffs are: $u_i(\mathbf{S}, n_S) = a(n_S)$ and $u_i(\mathbf{N}, n_S) = b(n_S)$, where n_S is the number of witnesses who play $\mathbf{S}$, $a(n_S) : \{0, \ldots, n\} \to \mathbb{R}$, and $b(n_S) : \{0, \ldots, n\} \to \mathbb{R}$. We assume a is weakly increasing in n_S and b is weakly decreasing in n_S. Note that the identity of those who sanction is irrelevant–only the number matters.

First, suppose the harm is continuous and witnesses receive signals just as in Section A.1. We show that there only exists a threshold norm when there is a non-generic relationship between $a(n_S)$ and $b(n_S)$ that occurs with measure zero.

Theorem 8. *If harm is uniformly distributed $H \sim [h_l, h_h]$ and each player receives an independent uniformly distributed signal $S_i \sim [H+\epsilon, H-\epsilon]$, then $\sigma^{\bar{s}}$ is a threshold equilibrium if and only if:*

$$\sum_{n_S=0}^{n-1} a(n_S+1) = \sum_{n_S=0}^{n-1} b(n_S)$$

Proof. As in Theorem 5, a threshold norm is an equilibrium if and only if, at $S_i = \bar{s}$, witnesses are indifferent between sanctioning and not sanctioning.

To calculate witnesses' payoffs at $S_i = \bar{s}$, we must calculate $Pr(n_S = k | S_i = \bar{s})$ for any k. This involves the following steps.

First, we calculate the probability generating function of n_S when $S_i = \bar{s}$. In order to do so, we consider the information of player i after observing signal $S_i = \bar{s}$. This player's posterior distribution of harm is uniform over $[\bar{s}-\epsilon, \bar{s}+\epsilon]$. Let the realization of H be h. Then, the probability an arbitrary player $-i$ sanctions is $Pr(S_{-i} > \bar{s}|h) = \frac{h+\epsilon-\bar{s}}{2\epsilon}$ and the number of (other) players which sanction (n_S) is $n_S \sim Bin(n-1, \frac{h+\epsilon-\bar{s}}{2\epsilon})$.

Second, we can use these distributions–and the distribution of H given S_i to find the distribution of n_S given a signal $\bar{s}$. In particular, n_S is distributed $Bin(n-1, X)$, where X is uniformly distributed over $[0,1]$.

Third, using the above facts, we can now calculate the probability generating function of n_S, which will yield the probability of a given number of players sanctioning.

$$\begin{aligned} E(z^{n_S}) &= E[E(z^{n_S}|U[0,1])] \\ &= \int_0^1 Pr(n_S = k | h = u) f_h(u) du \\ &\dots \\ &= \frac{1}{n}(1 + z + \dots + z^{n-1}) \end{aligned}$$

Finally, using the probability function found above, we can calculate the probability of k players sanctioning. Simply take the coefficient of z^k, which is always $\frac{1}{n}$.

We can now use this result to calculate witnesses' payoffs at $S_i = \bar{s}$. The payoffs from sanctioning is $\sum_{n_S=0}^{n-1} a(n_S+1)$. The payoffs from not sanctioning are $\sum_{n_S=0}^{n-1} b(n_S)$. So, witnesses will only be indifferent at $S_i = \bar{s}$ if:

$$\sum_{n_S=0}^{n-1} a(n_S+1) = \sum_{n_S=0}^{n-1} b(n_S)$$

This occurs only for a non-generic set of payoffs, and therefore this implies a continuous norm

can only occur in rare circumstances. □

Now, suppose harm is discrete and witnesses receive signals just as in Section A.2. Again, there is a Bayesian Nash equilibrium in which players conditionally sanction so long as the amount of noise, ϵ, is not too large.

Theorem 9. *The strategy profile in which player i plays* **S** *if and only if $S_i = 1$ is a Bayesian Nash equilibrium if and only if:*

$$\sum_{k=0}^{n-1} a(k+1)\binom{n-1}{k}\left[\frac{q(1-\epsilon)}{q(1-\epsilon)+(1-q)\epsilon}\epsilon^{n-1-k}(1-\epsilon)^k + (1-\frac{q(1-\epsilon)}{q(1-\epsilon)+(1-q)\epsilon})\epsilon^k(1-\epsilon)^{n-1-k}\right]$$

$$\geq$$

$$\sum_{k=0}^{n-1} b(k)\binom{n-1}{k}\left[\frac{q(1-\epsilon)}{q(1-\epsilon)+(1-q)\epsilon}\epsilon^{n-1-k}(1-\epsilon)^k + (1-\frac{q(1-\epsilon)}{q(1-\epsilon)+(1-q)\epsilon})\epsilon^k(1-\epsilon)^{n-1-k}\right]$$

and

$$\sum_{k=0}^{n-1} b(k)\binom{n-1}{k}\left[\frac{q\epsilon}{q\epsilon+(1-q)(1-\epsilon)}\epsilon^{n-1-k}(1-\epsilon)^k + (1-\frac{(1-q)(1-\epsilon)}{q\epsilon+(1-q)(1-\epsilon)})\epsilon^k(1-\epsilon)^{n-1-k}\right]$$

$$\geq$$

$$\sum_{k=0}^{n-1} a(k+1)\binom{n}{k}\left[\frac{q\epsilon}{q\epsilon+(1-q)(1-\epsilon)}\epsilon^{n-1-k}(1-\epsilon)^k + (1-\frac{(1-q)(1-\epsilon)}{q\epsilon+(1-q)(1-\epsilon)})\epsilon^k(1-\epsilon)^{n-1-k}\right]$$

Proof. First, suppose a player receives $S_i = 1$. Consider the probabilities with which they expect other players to sanction. With probability $\frac{q(1-\epsilon)}{q(1-\epsilon)+(1-q)\epsilon}$, harm was actually done, in which case each other player has independent probability $1-\epsilon$ with which they will sanction. With probability $\frac{(1-q)\epsilon}{q(1-\epsilon)+(1-q)\epsilon}$, harm was not done, in which case each player has independent probability ϵ with which they will sanction. The probability with which k players sanction is $\binom{n-1}{k}[\frac{q(1-\epsilon)}{q(1-\epsilon)+(1-q)\epsilon}\epsilon^{n-1-k}(1-\epsilon)^k + (1-\frac{q(1-\epsilon)}{q(1-\epsilon)+(1-q)\epsilon})\epsilon^k(1-\epsilon)^{n-1-k}]$.

For a player to sanction given $S_i = 1$, the payoff from sanctioning, $\sum_{k=0}^{n-1} a(k+1)\binom{n-1}{k}[\frac{q(1-\epsilon)}{q(1-\epsilon)+(1-q)\epsilon}\epsilon^{n-1-k}(1-\epsilon)^k + (1-\frac{q(1-\epsilon)}{q(1-\epsilon)+(1-q)\epsilon})\epsilon^k(1-\epsilon)^{n-1-k}]$, must be greater than the payoff from not sanctioning, $\sum_{k=0}^{n-1} b(k)\binom{n-1}{k}[\frac{q(1-\epsilon)}{q(1-\epsilon)+(1-q)\epsilon}\epsilon^{n-1-k}(1-\epsilon)^k + (1-\frac{q(1-\epsilon)}{q(1-\epsilon)+(1-q)\epsilon})\epsilon^k(1-\epsilon)^{n-1-k}]$.

Similarly, the probability k players sanction given $S_i = 0$ is $\binom{n-1}{k}[\frac{q\epsilon}{q\epsilon+(1-q)(1-\epsilon)}\epsilon^{n-1-k}(1-\epsilon)^k + (1-\frac{(1-q)(1-\epsilon)}{q\epsilon+(1-q)(1-\epsilon)})\epsilon^k(1-\epsilon)^{n-1-k}]$. The payoff from not sanctioning, $\sum_{k=0}^{n-1} b(k)\binom{n-1}{k}[\frac{q\epsilon}{q\epsilon+(1-q)(1-\epsilon)}\epsilon^{n-1-k}(1-\epsilon)^k + (1-\frac{(1-q)(1-\epsilon)}{q\epsilon+(1-q)(1-\epsilon)})\epsilon^k(1-\epsilon)^{n-1-k}]$, must be greater than the payoff from sanctioning, $\sum_{k=0}^{n-1} a(k+1)\binom{n}{k}[\frac{q\epsilon}{q\epsilon+(1-q)(1-\epsilon)}\epsilon^{n-1-k}(1-\epsilon)^k + (1-\frac{(1-q)(1-\epsilon)}{q\epsilon+(1-q)(1-\epsilon)})\epsilon^k(1-\epsilon)^{n-1-k}]$. □

B.6 Continuous Sanctions

In the manuscript, we stated, "...we recognize that sanctions can itself vary in magnitude, perhaps with each witness's payoffs depending on the difference between how much she sanctioned and how much others sanctioned. Therefore, we extend the sanctioning decision to permit variation in magnitude of sanctioning and show that this, too, does not permit threshold norms, provided the payoffs to differences in sanctioning don't happen to have a particularly non-generic structure."

We retain the setup in Section A, with the following modifications:

1. Instead of assuming the range of players' strategy functions is binary, we assume it is $[0, 1]$. I.e. $\sigma_i : [h_l - \epsilon, h_h + \epsilon] \to [0, 1]$. The strategy functions are also assumed to be weakly increasing.

2. Witnesses' payoffs are decreasing in the difference between their chosen level of sanction. I.e. $u_1(\sigma_1, \sigma_2) = u_2(\sigma_1, \sigma_2)$ is such that $u_i'(|\sigma_i - \sigma_{-i}|) < 0$, where u_i is continuous and differentiable. For example, payoffs might be, $u_i' = \mathbb{E}(\sigma_1(s_1) - \sigma_2(s_2))^2$.

First, suppose harm is continuous and witnesses receive signals just as in section A. We show that threshold strategies cannot be an equilibrium. There can be equilibria in which players sanction more except in a unique, linear, case.

Theorem 10. *Suppose for $i = \{1, 2\}$, $\sigma_i(h_l) = \underline{\sigma}$ and $\sigma_i(h_h) = \overline{\sigma}$, and also σ_i is not constant almost everywhere. Then there exists at most one Bayesian Nash equilibrium, where $\sigma_i(s_i) = (\overline{\sigma} - \underline{\sigma})s_i - \underline{\sigma}$ for $i = \{1, 2\}$.*

Proof. We will begin by showing that in any Bayesian Nash equilibrium, $\sigma_1 = \sigma_2$. That is, the strategy σ_i must be a best response to itself. This is due to the fact that i's best response is an averaging of σ_{-i} weighted appropriately by the posterior. When noise is uniformly distributed, this implies that both σ_is must be linear. Given two points, there exists only one line passing through both, so there exists a single Bayesian Nash equilibrium.

First, we will show that, in any Bayesian Nash equilibrium, $\sigma_1 = \sigma_2$. Consider some arbitrary $s \in [h_l - \epsilon, h_h + \epsilon]$. Imagine player 1 observes this signal. Her payoff at this signal is $\int u_1(\sigma_1(s), \sigma_2(s_2))Pr(s_2|s)ds_2$. We take the derivative and set it equal to zero and solve for the best response: it is the weighted average of player 2's sanctioning decisions, $\sigma_1(s) = \mathbb{E}(\sigma_2(s_2)|s)$ The same is true for player 2. The only way for this to be possible is for the two players to make the same choice at s, I.e. $\sigma_1(s) = \mathbb{E}(\sigma_2(s_2)|s) = \mathbb{E}\left(\mathbb{E}(\sigma_1(s_1)|s)|s\right) = \mathbb{E}(\sigma_1(s_1)|s) = \sigma_2(s)$.

Next we will show that when the noise is uniform, the resulting best response function is always linear. We take the first order condition for player 1, $\int \frac{\partial u_1(\sigma_i(s_1), \sigma_i(s_2))}{\partial \sigma_i(s_1)} Pr(s_2|s_1)ds_2 = 0$. This implies $\int u_1'(\sigma_i(s_1), \sigma_i(s_2))\sigma_i'(s_1)Pr(s_2|s_1)ds_2 = 0$. Because when player one receives signal s_1, her posterior probability of player 2 receiving a signal outside $[s_1 - 2\epsilon, s_1 + 2\epsilon]$ is zero, we can rewrite this integral as $\int_{s_1-2\epsilon}^{s_1+2\epsilon} u_1'(\sigma_i(s_1), \sigma_i(s_2))\sigma_i'(s_1)Pr(s_2|s_1)ds_2 = 0$, which implies the following condition:

$$\int_{s_1-2\epsilon}^{s_1} u_1'(\sigma_i(s_1), \sigma_i(s_2))\sigma_i'(s_1)Pr(s_2|s_1)ds_2 = -\int_{s_1}^{s_1+2\epsilon} u_1'(\sigma_i(s_1), \sigma_i(s_2))\sigma_i'(s_1)Pr(s_2|s_1)ds_2 \quad (4)$$

Since u_i is defined as a function of the distance, or difference, between two values, $u_i(\sigma_i(s_1), \sigma_i(s_1 + x)) = u_i(\sigma_i(s_1), \sigma_i(s_1 - x))$. Similarly, given that the distribution of noise is symmetric, $Pr(s_1 - x|s_1) = Pr(s_1 + x|s_1)$. For condition 4 to hold, it must therefore be that $\sigma_i'(s_1 - x) = \sigma_i'(s_1 + x)$ for all $x < 2\epsilon$, and for all s_1. Then, σ_i' is constant everywhere. I.e. σ_i is linear.

Lastly, if strategies are linear, and $\sigma_i(h_l) = \underline{\sigma}$ and $\sigma_i(h_h) = \overline{\sigma}$, then there exists only a single possible such line. □

C Evolutionary Dynamics

The simulations in this manuscript were performed using DyPy, which is available at `https://github.com/aaandrew152/dynamics_sim`. The code for these simulations and all others in this section is available for download at `https://github.com/aaandrew152/CtsDisc`.

C.1 Details of the Simulation Reported in Fig. 4 of the Manuscript

C.1.1 Fig. 4a: A single, representative simulation

We analyzed a single population of players playing the game described in Section A. The parameters we employed were: $N = 7$, $a = 4$, $b = 2$, $c = 0$, and $d = 4$, so that $p = 2/3$. The strategy space was restricted to the following ten strategies: always sanction, sanction if and only if $s_i > 0/7, s_i > 1/7$, sanction if and only if $s_i > 2/7$, ..., sanction if and only if $s_i > 7/7$, never sanction.

Each simulation proceeded as follows. First, we assigned all the players to the play the strategy sanction if and only if $s_i > 5/7$. In each round:

1. Players receive the expected payoffs from playing against another player randomly selected from the population with signals uniformly selected from $[0, 1]$.

2. Strategies are re-assigned proportionally to payoffs, $\delta_{i,t+1} = \delta_{i,t} \cdot e^{u_{i,t}}$ where $\delta_{i,t}$ is the proportion of the population playing strategy i in round t and $u_{i,t}$ is the expected payoff from strategy i in round t.

3. Players are randomly selected with probability 0.05 to "mutate". That is, if they are selected, they are assigned a strategy randomly selected from the ten strategies.

At the end of each round, we recorded the frequency of and payoffs associated with each strategy. Each simulation lasted for 190 rounds. In Fig. 4a, we present a single simulation such simulation.

C.1.2 Fig. 4b: Average frequencies of each strategy

In Fig. 4b, we ran the simulations described in Section C.1.1 500 times, and presented the average frequency of the strategies in each period.

We also run the same simulations, but start by assigning all players to play the strategy sanction if and only if $s_i > 1/7$, and let payoffs equal $a = 4$, $b = 0$, $c = 2$, and $d = 4$, so that $p = 1/3$. We present the average frequency of the strategies in each period in Fig. 5.

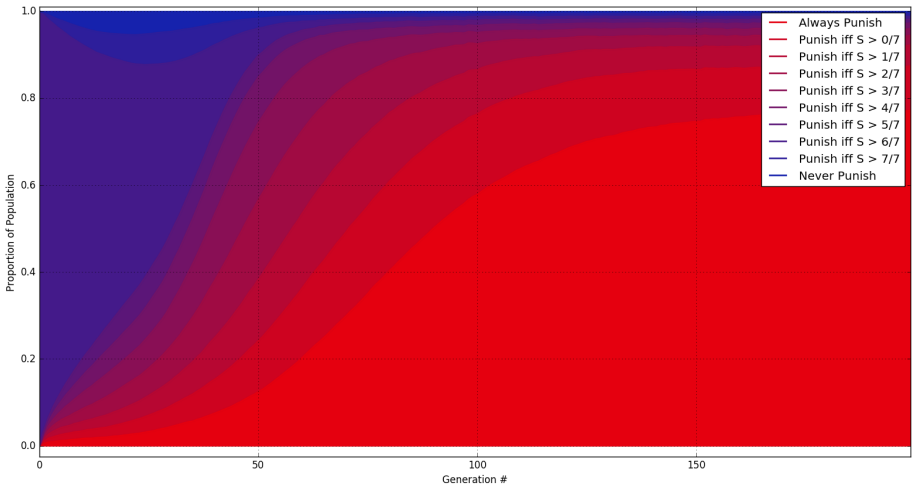


Figure 5: Average Frequency of Strategies, 500 simulations, $p < 1/2$

C.1.3 Fig. 4c: A single, representative simulation for the model with $n = 10$ discrete possible values of harm, and $p = .8$

The simulations are identical to those described in Section C.1.1 except that harm can take 10 possible values, $\{1, 2, \ldots, 10\}$, and the set of possible strategies is {Sanction if and only if $S_i > 0$, ..., Sanction if and only if $S_i > 10$}.

C.1.4 Fig. 4d: A single, representative simulation for the model with $n = 10$ discrete possible values of harm, and $p = .67$

The simulations are identical to those described in Section C.1.3, except that $p = .67$.

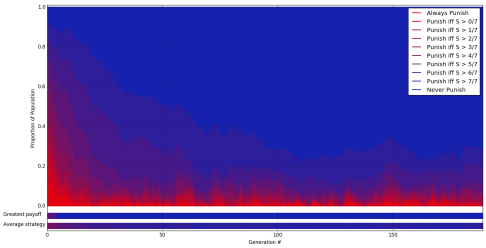
C.1.5 Fig. 4e: A single, representative simulation for the model with normally distributed harm

The simulations are identical to those described in Section C.1.1 except that harm is distributed $H \sim N[0, 1]$.

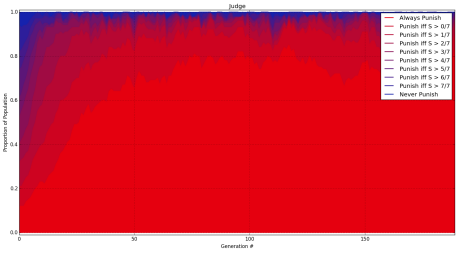
C.2 Additional Simulations

C.2.1 Arbitrary Assignment of Starting Strategies for Fig. 4a and 4b

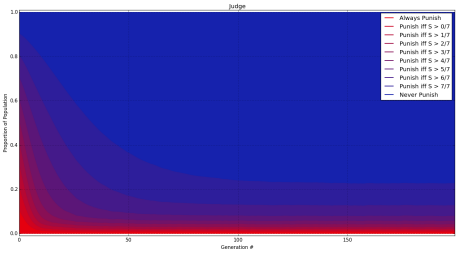
We run the same simulations as in Fig. 4a and 4b of the manuscript, but, instead of starting the entire population off at the same strategy, we start by assigning strategies randomly. Fig. 6a and 6b present a single run with $p > 1/2$ and $p < 1/2$, respectively. Fig. 6c and 6d present the average frequency of the strategies in each period for $p > 1/2$ and $p < 1/2$, respectively.



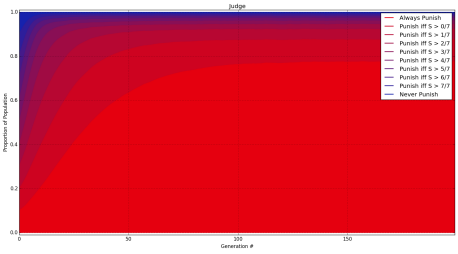
(a) Frequency of Strategies, Random Starting Point, Single Run, $p > 1/2$



(b) Frequency of Strategies, Random Starting Point, Single Run, $p < 1/2$



(c) Average Frequency of Strategies, 500 simulations, Random Starting Point, $p > 1/2$



(d) Average Frequency of Strategies, 500 simulations, Random Starting Point, $p < 1/2$

Figure 6: Arbitrary Assignment of Starting Strategies for Fig. 4a and 4b

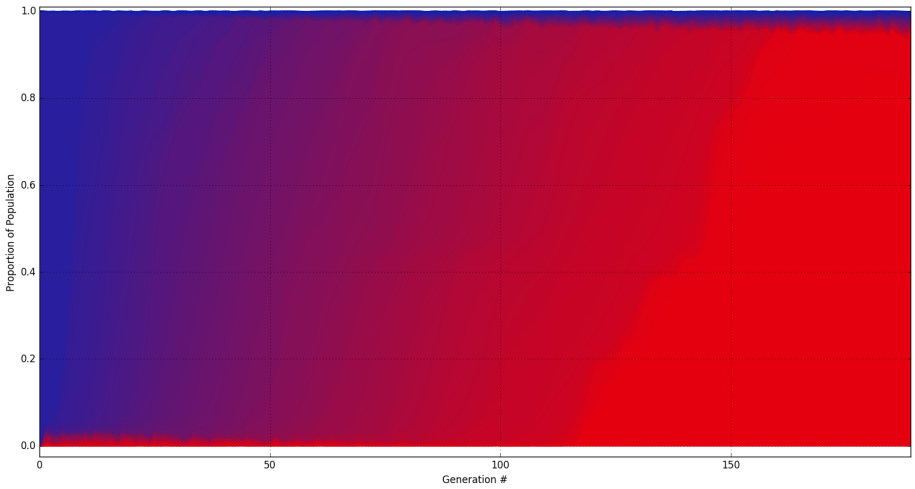


Figure 7: Average Frequency of Strategies, 100 Discrete Values

C.2.2 N Discrete Values

In Fig. 7 we present analogous simulations to those in Fig. 4d of the manuscript, with identical parameters, except that we now let the domain of H be $\{1, 2, \ldots, 100\}$. We start the entire population at sanction if and only if $N >= 10$. The norm is no longer expected to be stable. Indeed, it is not.

C.2.3 Uniform Distribution with an Atom at $H = h_l$

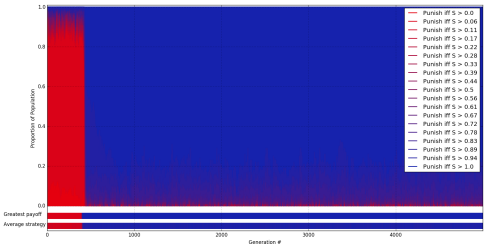
In Fig. 8, we present the same simulations as those in Fig. 4a-b, but we now let harm be distributed $H \sim F(h) = 1/5 + 4/5h$. The strategy space includes the following 20 strategies: Sanction if and only if $S_i > \bar{s}$ with $\bar{s} \in \{0, 0.06, \ldots, 0.94, 1\}$. We start the population at sanction if and only if $S_i > 0.12$. We run the simulations for 5000 generations.

C.2.4 State-Dependent Payoffs

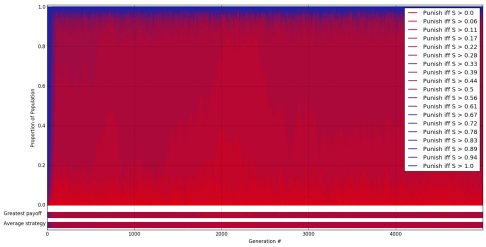
In Fig. 9, we present the same simulations as those in Fig. 4a-b, but we now let payoffs be $a = 4(2h + 1/2), b = 2(2h + 1/2), c = 0, d = 5$. The strategy space includes the following 20 strategies: Sanction if and only if $S_i > \bar{s}$, with $\bar{s} \in \{0, 0.06, \ldots, 0.94, 1\}$. We start the population at sanction if and only if $S_i > 0.12$. We run the simulations for 190 generations. We run the same simulations for $p = 1/3$.

References

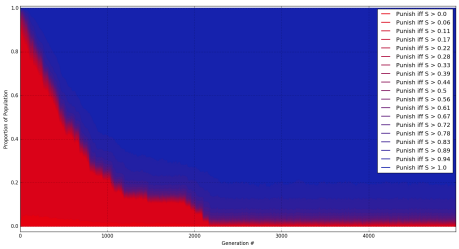
[1] M. L. Eaton, *Multivariate statistics: a vector space approach* (Wiley New York, 1983).



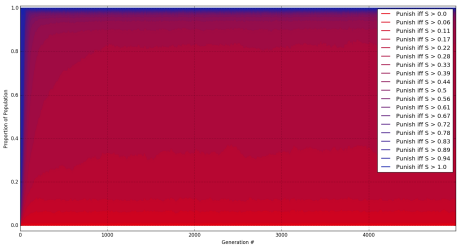
(a) Frequency of Strategies, Starting at Sanction if and only if $S > 0.22$, Single Run, $p > 1/2$



(b) Frequency of Strategies, Starting at Starting at Sanction if and only if $S > 0.78$, Single Run, $p < 1/2$

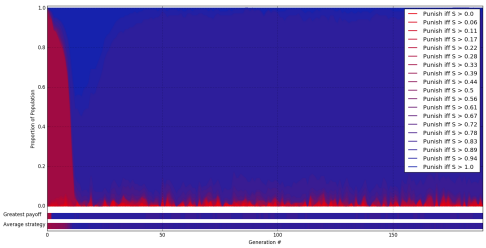


(c) Average Frequency of Strategies, 500 simulations, Starting at Sanction if and only if $S > 0.22$, $p > 1/2$

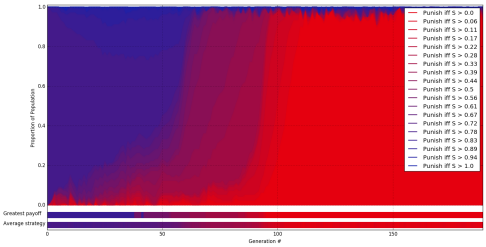


(d) Average Frequency of Strategies, 500 simulation, Starting at Sanction if and only if $S > 0.78$, $p < 1/2$

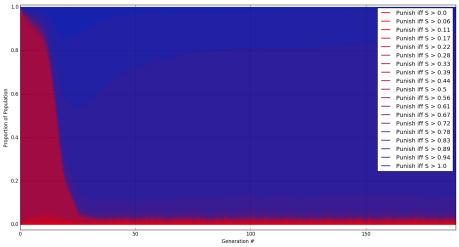
Figure 8: Uniform Distribution with an Atom at $H = h_l$



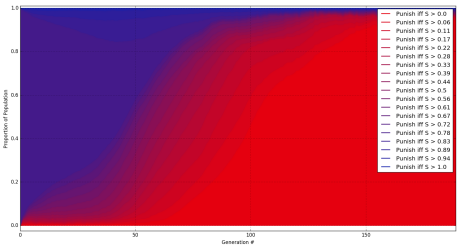
(a) Frequency of Strategies, Starting at Sanction if and only if $S > 0.22$, Single Run, $p > 1/2$



(b) Frequency of Strategies, Starting at Sanction if and only if $S > 0.78$, Single Run, $p < 1/2$



(c) Average Frequency of Strategies, 500 simulations, Starting at Starting at Sanction if and only if $S > 0.22$, $p > 1/2$



(d) Average Frequency of Strategies, 500 simulation, Starting at Starting at Sanction if and only if $S > 0.78$, $p < 1/2$

Figure 9: State-Dependent Payoffs