

The Emergence of Categorical Norms

Supplementary Information

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

Contents

A	The General Setup	5
B	Analysis of Specific State-Signal Structures	7
B.1	Continuous-Uniform State-Signal Structure	7
B.2	Categorical State-Signal Structure	9
B.3	Discrete-Uniform State-Signal Structure	10
B.4	Normal State-Signal Structure	12
B.5	General Result for Any State-Signal Structure	14
C	Evolutionary Dynamics	15
C.1	Details of the Simulation Reported in Fig. 4 of the Manuscript	15
C.1.1	Fig. 4a: A single, representative simulation	15
C.1.2	Fig. 4b: Average frequencies of each strategy	15
C.1.3	Fig. 4c: A single, representative simulation for the model with $n = 10$ discrete possible values of the state, and $p = .8$	16
C.1.4	Fig. 4d: A single, representative simulation for the model with $n = 10$ discrete possible values of the state, and $p = .67$	16
C.1.5	Fig. 4e: A single, representative simulation for the model with normally distributed states	16
C.2	Additional Simulations	17
C.2.1	Arbitrary Assignment of Starting Strategies for Fig. 4a and 4b	17
C.2.2	N Discrete Values	17
C.2.3	Uniform Distribution with an Atom at $\Omega = \omega_l$	17
C.2.4	State-Dependent Payoffs	17
D	Experimental Evidence: Vignettes	17
D.1	Ethics Compliance and Preregistration	17
D.2	Subjects	19
D.3	Sample size determination	19
D.4	Methods	22
D.5	Statistical Analyses	22
D.6	Vignette 1: Government Killing Protesters	23
D.6.1	Methods	23
D.6.2	Results	24
D.7	Vignette 2: Invading Another Country's Territory	24
D.7.1	Methods	24
D.7.2	Results	25

D.8	Vignette 3: Shoplifting	25
D.8.1	Methods	25
D.8.2	Results	26
D.9	Vignette 4: Murder in a Trolley Problem	26
D.9.1	Methods	26
D.9.2	Results	26
D.10	Vignette 5: Volunteering at a Soup Kitchen	27
D.10.1	Methods	27
D.10.2	Results	27
E	Experimental Evidence: Incentivized Economic Games	27
E.1	Ethics Compliance and Preregistration	27
E.2	Experimental design overview	27
E.3	Sample size determination	28
E.4	Subjects	28
E.5	Procedure	28
E.6	Results	30

List of Figures

1	The Coordination Game - The payoffs are that of Player 1.	5
2	Average Frequency of Strategies, 500 simulations, $p < 1/2$	16
3	Arbitrary Assignment of Starting Strategies for Fig. 4a and 4b	18
4	Average Frequency of Strategies, 100 Discrete Values	19
5	Uniform Distribution with an Atom at $\Omega = \omega_l$	20
6	State-Dependent Payoffs	21

		PLAYER #2	
		U	D
PLAYER #1	U	a	b
	D	c	d
		$a > c, d > b$	

Figure 1: The Coordination Game - The payoffs are that of Player 1.

A The General Setup

For convenience, we repeat the exposition of the model presented in the manuscript. We begin with a standard coordination game, which is the simplest game in which the optimal thing to do depends on what others are doing. In the coordination game, there are two players, each of whom chooses between one of two actions, which in this supplement, we label U and D (in the manuscript, these are labeled S and N , respectively). Each player gets payoff a if she chooses U when the other also chooses U , $b < a$ if she chooses U when the other does not choose U , d if she chooses D when the other chooses D , and $c < d$ if she chooses D when the other does not choose D (the payoffs of Player 1 are given in Fig. 1). The parameter $p = (d - b)/(a - c + d - b)$, which is often referred to as the *risk dominance* (of U), will prove useful. Its interpretation is: if either player expects the other to play U with probability greater than p , then she prefers to play U herself.

We add to this standard coordination game the ability to condition play in the game on a signal that does not directly affect payoffs. Each player observes (a payoff irrelevant) signal that is correlated with the true state of the world before acting in the coordination game. The coordination game together with this information structure is an example of a Bayesian Game.

Before presenting the Bayesian games that we analyze, we provide a general definition of a Bayesian Game and a Bayesian Nash equilibrium, the standard solution concept for such games (see Osborne and Rubinstein (1994) for further details):

Definition 1. A *Bayesian Game* is a tuple $G = \langle N, (A_i)_{i \in N}, \Omega, (T_i)_{i \in N}, (\tau_i)_{i \in N}, (\rho_i)_{i \in N}, (u_i)_{i \in N} \rangle$ where

- N denotes the set of players,
- A_i denotes the set of actions available to player i ,
- Ω denotes the set of all possible states of the world,

- T_i denotes the set of all possible signals (or types) of player i ,
- $\tau_i : \Omega \rightarrow \Delta(T_i)$ denotes the signal function of Player i ,
- $\rho_i \in \Delta(\Omega)$ denotes the prior belief of player i regarding Ω ,
- $u_i : A \times \Omega \rightarrow \mathbb{R}$ denotes the (a possibly state-dependent) payoff function.

We need the following to provide the definition of a Bayesian Nash equilibrium of a Bayesian game G :

- In a Bayesian game, a (pure) strategy describes a complete contingent plan for every possible signal realizations. We denote a strategy of player i as $\sigma_i : T_i \rightarrow A_i$.
- Let $\mathbb{E}[u_i(a_i, \sigma_{-i}(s_{-i}))|s_i]$ denote the expected payoff of player i when he plays action $a_i \in A_i$ and his signal realization is $s_i \in T_i$, i.e., $\mathbb{E}[u_i(a_i, \sigma_{-i}(s_{-i}))|s_i] = \int_{\omega \in \Omega} Pr(\omega|s_i)u_i(a_i, \sigma_{-i}(s_{-i}), \omega)P(ds_{-i}|\omega)$.

Definition 2. We say that a strategy profile $\sigma^* = (\sigma_1^*, \dots, \sigma_n^*)$ is a Bayesian Nash equilibrium of G if for each $i \in N$, and each $s_i \in T_i$,

$$\mathbb{E}[u_i(\sigma_i^*(s_i), \sigma_{-i}^*(s_{-i}))|s_i] \geq \mathbb{E}[u_i(a_i, \sigma_{-i}^*(s_{-i}))|s_i], \text{ for all } a_i \in A_i.$$

In words, a Bayesian Nash equilibrium is a strategy profile where each player maximizes their expected payoff given their belief and the strategy of the other player. That is, there is no profitable deviation for any signal realization (or type) of any of the players. Put differently, “[In a Bayesian Nash Equilibrium] each player chooses the best action available to him given the signal that he receives and his belief about the state and the other player’s actions that he deduces from this signal.” [1].

We consider several different information structures, to be explained further below. These information structures induce different Bayesian games. We refer to each of these Bayesian games as *state-independent Bayesian coordination games*:

Definition 3. A *state-independent Bayesian coordination game* is a Bayesian game $\Gamma = \langle N, (A_i)_{i \in N}, \Omega, (T_i)_{i \in N}, (\tau_i)_{i \in N}, (\rho_i)_{i \in N}, (u_i)_{i \in N} \rangle$ such that

- There are two players, i.e., $N = \{1, 2\}$.
- There are two actions available to each player, i.e., $A_1 = A_2 = \{U, D\}$.
- The set of states of the world is a subset of the real line, i.e., $\Omega \subset \mathbb{R}$.
- The (state-independent) payoff function $u_i : A_1 \times A_2 \times \Omega \rightarrow \mathbb{R}$ is such that $u_1(U, U, \omega) = u_2(U, U, \omega) = a$ for all $\omega \in \Omega$, $u_1(U, D, \omega) = u_2(D, U, \omega) = b$ for all $\omega \in \Omega$, $u_1(D, U, \omega) = u_2(U, D, \omega) = c$ for all $\omega \in \Omega$, and $u_1(D, D, \omega) = u_2(D, D, \omega) = d$ for all $\omega \in \Omega$.
- The two players share a common prior $\rho \in \Delta(\Omega)$, i.e., $\rho_1 = \rho_2 = \rho$ and this is commonly known.
- The range of possible signals is always equal to the set of states. That is, $T_1 = T_2 = \Omega \subset \mathbb{R}$.

- The players' signals are generated according to the same probability distribution function, i.e., $\tau_1 = \tau_2 = \tau$. That is, $\tau_1(\omega) = \tau_2(\omega) = \tau(\omega) \in \Delta(\Omega)$ for all $\omega \in \Omega$.
- The players' signals are independent conditional on the state, i.e., $S_1|_\omega \perp S_2|_\omega$, where S_i denotes the random variable that represents the signal of Player i , and $S_i|_\omega$ denotes the random variable that represents the signal of Player i conditional on the state ω , for both $i = 1, 2$.

We refer to the unspecified parts of a state-independent Bayesian coordination game, namely $\langle \Omega, \rho, \tau \rangle$, as a *state-signal structure*. In each of the following sections, we analyze a particular state-signal structure.

Before moving forward, we define a *threshold strategy*:

Definition 4. A (pure) strategy $\sigma_i^{\bar{s}} : \Omega \rightarrow \{U, D\}$ is said to be a threshold strategy with a threshold at $\bar{s} \in \Omega$ if and only if

$$\sigma_i^{\bar{s}}(s_i) = U \text{ if and only if } s_i > \bar{s}.$$

We refer to the strategy profile $\sigma^{\bar{s}} = (\sigma_1^{\bar{s}}, \sigma_2^{\bar{s}})$ as the threshold strategy profile at \bar{s} .

The crux of our analysis is when a threshold strategy profile can be supported as a Bayesian Nash equilibrium in a state-independent Bayesian coordination game.

B Analysis of Specific State-Signal Structures

B.1 Continuous-Uniform State-Signal Structure

First, we consider the continuous-uniform state-signal structure $\langle \Omega^{C.U.}, \rho^{C.U.}, \tau^{C.U.} \rangle$: In order to avoid edge cases, as is standard in the global games literature, we assume that the state of the world, $\omega \in \Omega^{C.U.}$, is randomly drawn from the real line $\Omega^{C.U.} = (-\infty, +\infty)$ and that the common (improper [2]) prior, $\rho^{C.U.}$, is the uniform distribution over $\Omega^{C.U.}$. Each player observes a signal correlated with the state of the world, independently distributed uniformly within $\epsilon > 0$ of the true state of the world. That is, $\tau^{C.U.}(\omega)$ is the uniform distribution over $[\omega - \epsilon, \omega + \epsilon]$ for each $\omega \in \Omega^{C.U.}$ for both $i = 1, 2$. We refer to the state-independent Bayesian coordination game with continuous-uniform state-structure as $\Gamma^{C.U.}$.

In the manuscript, we claimed that there is no Bayesian Nash equilibrium where the players condition their action on (their signal of) the true state, regardless of how small the error is in observing the true state provided that the risk dominance $p \neq \frac{1}{2}$, under the continuous-uniform state-signal structure. We formalize this claim in Theorem 1. Formally, we show that there is no equilibrium in which players play U if and only if their signal is above some threshold.

Theorem 1. Let $\epsilon > 0$ and $p \neq \frac{1}{2}$. For any $\bar{s} \in (-\infty, \infty)$, the threshold strategy profile at \bar{s} , $\sigma^{\bar{s}}$, is not a Bayesian Nash equilibrium of $\Gamma^{C.U.}$.

Proof. Suppose $\sigma^{\bar{s}}$ is a Bayesian Nash equilibrium of $\Gamma^{C.U.}$ for some fixed threshold $\bar{s} \in (-\infty, +\infty)$.

We calculate the following conditional posterior probability given one's own signal: What is the likelihood that the other's signal is below (or equal to) some threshold \bar{s} given my own signal? Recall that S_i and S_{-i} denote the random variables that represents the signals of players i and $-i$, respectively.

Observe that the conditional p.d.f of S_{-i} given $S_i = s_i$ can be computed as follows: $f(S_{-i}|S_i = s_i) = \int_{-\infty}^{+\infty} f(\omega|S_i = s_i)f(S_{-i}|\omega)d\omega$. By construction, $f(S_{-i}|\omega)$ is the uniform distribution over $[\omega - \epsilon, \omega + \epsilon]$. It is also straightforward to see that $f(\omega|S_i = s_i)$ is the uniform distribution over $[s_i - \epsilon, s_i + \epsilon]$. Hence, $f(S_{-i}|S_i = s_i) = \int_{s_i - \epsilon}^{s_i + \epsilon} f(\omega|S_i = s_i)f(S_{-i}|\omega)d\omega$. This leads to the following conditional p.d.f.

$$f(S_{-i} = s_{-i}|S_i = s_i) = \begin{cases} 0 & \text{if } s_{-i} \leq s_i - 2\epsilon \\ \frac{s_{-i} - s_i + 2\epsilon}{4\epsilon^2} & \text{if } s_i - 2\epsilon < s_{-i} \leq s_i \\ \frac{s_i + 2\epsilon - s_{-i}}{4\epsilon^2} & \text{if } s_i \leq s_{-i} < s_i + 2\epsilon \\ 0 & \text{if } s_i + 2\epsilon \leq s_{-i} \end{cases} \quad (1)$$

Since $Pr(S_{-i} \leq \bar{s}|S_i = s_i) = \int_{-\infty}^{\bar{s}} f(S_{-i}|S_i = s_i)$, we obtain the following (conditional c.d.f):

$$Pr(S_{-i} \leq \bar{s}|S_i = s_i) = \begin{cases} 0 & \text{if } \bar{s} \leq s_i - 2\epsilon \\ \frac{(\bar{s} - s_i + 2\epsilon)^2}{8\epsilon^2} & \text{if } s_i - 2\epsilon < \bar{s} \leq s_i \\ \frac{1}{2} & \text{if } \bar{s} = s_i \\ 1 - \frac{(s_i + 2\epsilon - \bar{s})^2}{8\epsilon^2} & \text{if } s_i \leq \bar{s} < s_i + 2\epsilon \\ 1 & \text{if } s_i + 2\epsilon \leq \bar{s}. \end{cases} \quad (2)$$

From (1) or (2), we see that the distribution of i 's opponent's signal S_{-i} conditional on $S_i = s_i$ is symmetric around s_i . That is, $Pr(S_{-i} \leq s_i|S_i = s_i) = Pr(S_{-i} \geq s_i|S_i = s_i) = \frac{1}{2}$. This is true, in particular for player i with signal realization $s_i = \bar{s}$, i.e., $Pr(S_{-i} \leq \bar{s}|S_i = \bar{s}) = Pr(S_{-i} \geq \bar{s}|S_i = \bar{s}) = \frac{1}{2}$. This means $Pr(\sigma_{-i}^{\bar{s}}(S_{-i}) = U|S_i = \bar{s}) = Pr(\sigma_{-i}^{\bar{s}}(S_{-i}) = D|S_i = \bar{s}) = \frac{1}{2}$. But since, $p \neq \frac{1}{2}$, the player i with signal realization $s_i = \bar{s}$ has a strict best reply (either U or D) to $\sigma_{-i}^{\bar{s}}$. This contradicts the upper-hemi continuity property of the best reply correspondence. \square

What makes the proof above work is the fact that the conditional distribution of the opponent's signal given one's own signal being symmetric, i.e., $Pr(S_{-i} \leq s_i|S_i = s_i) = Pr(S_{-i} \geq s_i|S_i = s_i) = \frac{1}{2}$. The proof applies to any common prior distribution over the continuum $(-\infty, +\infty)$ with this property. We provide the following immediate corollary by referring to any state-independent Bayesian coordination games where signals are obtained from a (not necessarily uniform) distribution over a continuum with the property that $Pr(S_{-i} \leq s_i|S_i = s_i) = Pr(S_{-i} \geq s_i|S_i = s_i) = \frac{1}{2}$ as $\Gamma^{C.S.}$, where $C.S.$ stands for *Continuous-Symmetric*.

Corollary 1. *Let $p \neq \frac{1}{2}$. For any $\bar{s} \in (-\infty, \infty)$, the threshold strategy profile at \bar{s} , $\sigma^{\bar{s}}$, is not a Bayesian Nash equilibrium of any $\Gamma^{C.S.}$.*

A natural question that arises is whether it would be possible that the two individuals use different thresholds. It is also straightforward to see that as long as the conditional PDF $f(S_{-i}|S_i = s_i)$ is symmetric around s_i . This would not be possible. We note this observation as another corollary:

Let $\sigma^{\bar{s}_1, \bar{s}_2} := (\sigma_1^{\bar{s}_1}, \sigma_2^{\bar{s}_2})$ denote the strategy profile such that player 1 is playing according to the threshold strategy at \bar{s}_1 whereas player 2 is playing according to the threshold strategy at \bar{s}_2 .

Corollary 2. *Suppose $\bar{s}_i \in (-\infty, \infty)$, for both $i \in \{1, 2\}$. Then, $\sigma^{\bar{s}_1, \bar{s}_2}$ is not a Bayesian Nash equilibrium of any $\Gamma^{C.S.}$.*

B.2 Categorical State-Signal Structure

Next, we turn to the categorical state-signal structure, $\langle \Omega^{Cat}, \rho^{Cat}, \tau^{Cat} \rangle$: We assume now that $\Omega^{Cat} = \{0, 1\}$, i.e., the true state of the world is either $\omega = 1$ or $\omega = 0$. The common prior ρ^{Cat} over Ω is such that $\rho^{Cat}(1) = q \in (0, 1)$ and $\rho^{Cat}(0) = 1 - q \in (0, 1)$. That is, the state is $\omega = 1$ with probability q and $\omega = 0$ with probability $1 - q$. The players observe the true state of the world with probability $1 - \epsilon$. The signal function τ^{Cat} is such that $\tau^{Cat}(\omega)$ puts $1 - \epsilon$ probability on the signal $s_i = \omega$ and ϵ probability on the signal $s_i = 1 - \omega$ for each $\omega \in \{0, 1\}$. Once again, neither the state nor the signal influences the payoffs to the players. We refer to the state-independent Bayesian coordination game under the categorical state-signal structure as Γ^{Cat} .

We claimed that, so long as the amount of noise, ϵ , is small, there is a Bayesian Nash equilibrium where players condition their play on their signal. In Theorem 2, we show that this is indeed the case.

We refer to a strategy σ_i^c as a *categorical strategy* whenever $\sigma_i^c(s_i) = U$ if and only if $s_i = 1$. We refer to the strategy profile $\sigma^c = (\sigma_1^c, \sigma_2^c)$ as a *categorical strategy profile*.

Theorem 2. *The categorical strategy profile, $\sigma^c = (\sigma_1^c, \sigma_2^c)$, is a Bayesian Nash equilibrium of Γ^{Cat} if and only if*

$$1 - \frac{\epsilon(1 - \epsilon)}{q(1 - \epsilon) + (1 - q)\epsilon} \geq p \geq \frac{\epsilon(1 - \epsilon)}{(1 - q)(1 - \epsilon) + q\epsilon}.$$

Proof. $\sigma^c = (\sigma_1^c, \sigma_2^c)$ is a Bayesian Nash equilibrium of Γ^{Cat} if and only if (i) for each $i \in \{1, 2\}$, it is optimal for player i with signal realization $s_i = 1$ to play U given that the opponent, $-i$, is playing according to the categorical strategy σ_{-i}^c , and (ii) for each $i \in \{1, 2\}$, it is optimal for player i with signal realization $s_i = 0$ to play D given that the opponent, $-i$, is playing according to the categorical strategy σ_{-i}^c .

Recall that for U to be the optimal action of a player the probability that his opponent is to play U must be at least $p = \frac{d-b}{(a-c)+(d-b)} \in (0, 1)$. Similarly, for D to be the optimal action of a player the probability that his opponent is to play D must be at least $1 - p = \frac{a-c}{(a-c)+(d-b)} \in (0, 1)$. This means that the necessary and sufficient condition for (i) and (ii) are $Pr(S_{-i} = 1|S_i = 1) \geq p$ and $Pr(S_{-i} = 0|S_i = 0) \geq 1 - p$,

respectively. That is, (i) holds if and only if

$$\begin{aligned}
Pr(S_{-i} = 1 | S_i = 1) &= \frac{Pr(S_{-i} = 1, S_i = 1)}{Pr(S_i = 1)} \geq p \\
&= \frac{q(1-\epsilon)^2 + (1-q)\epsilon^2}{q(1-\epsilon) + (1-q)\epsilon} \geq p \\
&= 1 - \frac{\epsilon(1-\epsilon)}{q(1-\epsilon) + (1-q)\epsilon} \geq p.
\end{aligned}$$

On the other hand, (ii) holds if and only if

$$\begin{aligned}
Pr(S_{-i} = 0 | S_i = 0) &= \frac{Pr(S_{-i} = 0, S_i = 0)}{Pr(S_i = 0)} \geq 1 - p \\
&= \frac{(1-q)(1-\epsilon)^2 + q\epsilon^2}{(1-q)(1-\epsilon) + q\epsilon} \geq 1 - p \\
&= 1 - \frac{\epsilon(1-\epsilon)}{(1-q)(1-\epsilon) + q\epsilon} \geq 1 - p.
\end{aligned}$$

Combining together we get the desired necessary and sufficient condition:

$$1 - \frac{\epsilon(1-\epsilon)}{q(1-\epsilon) + (1-q)\epsilon} \geq p \geq \frac{\epsilon(1-\epsilon)}{(1-q)(1-\epsilon) + q\epsilon}.$$

□

Observe that as the amount of noise becomes smaller, i.e., $\epsilon \rightarrow 0$, the necessary and sufficient condition stated in Theorem 2 becomes more and more permissive. That is, as $\epsilon \rightarrow 0$, it becomes easier for the categorical threshold, $\sigma^c = (\sigma_1^c, \sigma_2^c)$, to be a Bayesian Nash equilibrium of Γ^{Cat} , as we claimed in the manuscript.

B.3 Discrete-Uniform State-Signal Structure

In the manuscript, we claimed, “...even though the state space is discrete, once it can take on sufficiently many possible values, the results approximate what happens when the state space is continuous.” We now prove this claim. To avoid the effect of endpoints, we again employ a slightly different setup than that described in the manuscript. Instead of having a finite number of possible states, we allow the state to continue indefinitely from $-\infty$ to ∞ . To capture how ‘categorical’ the state space is, we allow the number of possible states within a fixed measure of distance to vary. The more states are possible, the more ‘continuous’ the state space. As before, signals are noisy, and modeled as independent and uniformly distributed about the true state. Our key result is: for a given amount of noise, as the state space gets cut up into smaller pieces, it becomes harder to sustain equilibria in which players condition their action on their signal.

Let $\mathbb{Z} := \{\dots, -2, -1, 0, 1, 2, \dots\}$ denote the set of all integers, $\mathbb{Q} := \{\frac{m}{n} | m, n \in \mathbb{Z}\}$ denote the set of all rational numbers. Fix a positive rational number $q > 0$, which determines the coarseness of the state

space. The discrete state-signal structure associated with q , $\langle \Omega_q^{D.U.}, \rho_q^{D.U.}, \tau_q^{D.U.} \rangle$, is defined as follows: The state space is $\Omega_q^{D.U.} = \{qz | z \in \mathbb{Z}\}$, i.e., the state of the world is always a multiple of q . The common (improper) prior, $\rho_q^{D.U.}$, is the uniform distribution over $\Omega_q^{D.U.}$. Let $\epsilon > q > 0$ denote the noisiness of the signals: Each player observes a signal independently distributed uniformly on $\Omega_q^{D.U.}$ within ϵ of the true state of the world. It will be convenient to define $k := \lfloor \frac{\epsilon}{q} \rfloor \in \mathbb{N}$. For any $\omega \in \Omega_q^{D.U.}$, we note that there are $2k + 1$ possible signals where k of these signals are higher than ω while the other k are lower than ω . Thus, for any $\omega \in \Omega_q^{D.U.}$, $\tau_q^{D.U.}(\omega)$ is the uniform distribution over $\{\omega - kq, \omega - (k-1)q, \dots, \omega - q, \omega, \omega + q, \dots, \omega + (k-1)q, \omega + kq\}$ for both $i = 1, 2$. We refer to the state-independent Bayesian coordination game for any given $\epsilon > q > 0$ and $q \in \mathbb{Q}$ with this signal structure as $\Gamma_q^{D.U.}$.¹

Theorem 3. *Let $\epsilon > q > 0$, $q \in \mathbb{Q}$, and $k := \lfloor \frac{\epsilon}{q} \rfloor$. For any $\bar{s} \in \Omega_q^{D.U.}$, the threshold strategy profile at \bar{s} , $\sigma^{\bar{s}}$, is a Bayesian Nash equilibrium of $\Gamma_q^{D.U.}$ if and only if $\frac{1}{2} - \frac{1}{2k+1} \leq p \leq \frac{1}{2} + \frac{1}{2k+1}$.*

Proof. Let $\sigma^{\bar{s}}$ be a Bayesian Nash equilibrium of $\Gamma_q^{D.U.}$ for some fixed threshold $\bar{s} \in \Omega_q^{D.U.} = \{zq | z \in \mathbb{Z}\}$. As before, let S_i and S_{-i} denote the random variables that represents the signals of players i and $-i$, respectively. The conditional probability distribution of S_{-i} given $S_i = s_i$ can be obtained as follows: $Pr(S_{-i} = s_{-i} | S_i = s_i) = \sum_{\omega \in \Omega_q^{D.U.}} Pr(\omega | s_i) Pr(S_{-i} = s_{-i} | \omega)$. By construction, $Pr(S_{-i} | \omega)$ is the uniform distribution over $\{\omega - kq, \omega - (k-1)q, \dots, \omega - q, \omega, \omega + q, \dots, \omega + (k-1)q, \omega + kq\}$ and it is easy to see that $Pr(\omega | S_i = s_i)$ is the uniform distribution over $\{s_i - kq, \dots, s_i, \dots, s_i + kq\}$. Therefore, $Pr(S_{-i} = s_{-i} | S_i = s_i) = \sum_{\omega \in \{s_i - kq, \dots, s_i, \dots, s_i + kq\}} Pr(\omega | s_i) Pr(S_{-i} = s_{-i} | \omega)$. Hence, we obtain the following conditional probability mass function:

$$Pr(S_{-i} = s_{-i} | S_i = s_i) = \begin{cases} 0 & \text{if } s_{-i} < s_i - 2kq \\ \frac{(2k+1)q - s_i + s_{-i}}{q(2k+1)^2} & \text{if } s_i - 2kq \leq s_{-i} \leq s_i \\ \frac{(2k+1)q - s_{-i} + s_i}{q(2k+1)^2} & \text{if } s_i \leq s_{-i} \leq s_i + 2kq \\ 0 & \text{if } s_i + 2kq < s_{-i} \end{cases} \quad (3)$$

This leads to the following conditional cumulative distribution function:

$$Pr(S_{-i} \leq \bar{s} | S_i = s_i) = \begin{cases} 0 & \text{if } \bar{s} \leq s_i - 2kq \\ \frac{((2k+1) - \frac{s_i - s_{-i}}{q})(2k+2 - \frac{s_i - s_{-i}}{q})}{2(2k+1)^2} & \text{if } s_i - 2kq < \bar{s} < s_i \\ \frac{1}{2} + \frac{1}{2k+1} & \text{if } \bar{s} = s_i \\ 1 - \frac{((2k+1) - \frac{s_{-i} - s_i}{q})(2k+2 - \frac{s_{-i} - s_i}{q})}{2(2k+1)^2} & \text{if } s_i < \bar{s} < s_i + 2kq \\ 1 & \text{if } s_i + 2kq \leq \bar{s}. \end{cases} \quad (4)$$

We see (3) and (4) that the distribution of the (opponent's) signal S_{-i} conditional on $S_i = s_i$ is sym-

¹The astute reader will observe that increasing ϵ and decreasing q has the same effect. Nevertheless, we introduce both variables with their corresponding distinct interpretations in order to better connect this model to our base models and allow us to talk about increasing coarseness for a given amount of noise.

metric around s_i , as before. That is, $Pr(S_{-i} \leq s_i | S_i = s_i) = Pr(S_{-i} \geq s_i | S_i = s_i) = \frac{1}{2}$. This is true for player i with signal realization $s_i = \bar{s}$ as well, i.e., $Pr(S_{-i} \leq \bar{s} | S_i = \bar{s}) = \frac{1}{2} + \frac{1}{2k+1}$. Then, $Pr(\sigma_{-i}^{\bar{s}}(S_{-i}) = U | S_i = \bar{s} + q) = \frac{1}{2} + \frac{1}{2k+1}$. Similarly, $Pr(\sigma_{-i}^{\bar{s}}(S_{-i}) = D | S_i = \bar{s}) = \frac{1}{2} + \frac{1}{2k+1}$. This means we must have $\frac{1}{2} + \frac{1}{2k+1} \geq p$ as well as $\frac{1}{2} + \frac{1}{2k+1} \geq 1 - p$. Hence, we get $\frac{1}{2} - \frac{1}{2k+1} \leq p \leq \frac{1}{2} + \frac{1}{2k+1}$. \square

We now consider what happens when the state space becomes finer. First, notice that, even though this state-signal structure $\langle \Omega_q^{D.U.}, \rho_q^{D.U.}, \tau_q^{D.U.} \rangle$ is always discrete, it converges to the continuous-uniform state-signal structure $\langle \Omega^{C.U.}, \rho^{C.U.}, \tau^{C.U.} \rangle$ of Section B.1 as q gets smaller, in the following sense: As q approaches 0, (i) the Hausdorff distance between $\Omega_q^{D.U.}$ and $\Omega^{C.U.} = (-\infty, +\infty)$ approaches 0; (ii) the common prior converges in distribution to the uniform distribution on $(-\infty, +\infty)$, i.e., $\rho_q^{D.U.}$ converges in distribution to $\rho^{C.U.}$; (iii) for any given $\epsilon > 0$ and $\omega \in \mathbb{Q}$, the distribution of signals conditional on the state being ω converges in distribution to the uniform distribution over $(-\infty, +\infty)$, i.e., $\tau_q^{D.U.}(\omega)$ converges in distribution to $\tau^{C.U.}(\omega)$.

Second, notice that the conditions for existence of a threshold equilibrium also converges to that of Section B.1 in the following sense: Let $P^{C.U.}$ be the set of $p \in (0, 1)$ for which a threshold equilibrium exists in $\Gamma^{C.U.}$. By Theorem 1, $P^{C.U.} = \{\frac{1}{2}\}$. Similarly, let $P_q^{D.U.}$ be the set of $p \in (0, 1)$ for which a threshold equilibrium exists in $\Gamma_q^{D.U.}$ under the current model, i.e., by Theorem 3, $P_q^{D.U.} := [\frac{1}{2} - \frac{1}{2\lfloor \frac{\epsilon}{q} \rfloor + 1}, \frac{1}{2} + \frac{1}{2\lfloor \frac{\epsilon}{q} \rfloor + 1}]$. Observe that for a fixed $\epsilon > 0$, as q approaches 0, $P_q^{D.U.}$ shrinks and converges to $P^{C.U.}$ with respect to the Hausdorff distance.

One interpretation of the above results is as follows: Our key result about the permissibility of threshold equilibria in the continuous-uniform state signal structure is robust with respect to adding a bit of coarseness to the state-signal structure.

B.4 Normal State-Signal Structure

Next, we consider the normal state-signal structure $\langle \Omega^N, \rho_{V_\Omega}^N, \tau_{V_\epsilon}^N \rangle$ where the state and signals are normally distributed as follows: We assume that the state space is $\Omega^N = (-\infty, +\infty)$, and that the common prior, $\rho_{V_\Omega}^N$, is the normal distribution with mean 0 and variance V_Ω . The signal function $\tau_{V_\epsilon}^N$ is such that the distribution $\tau_{V_\epsilon}^N(\omega)$ is normal around the true state of the world $\omega \in \Omega^N$ with mean 0 and variance V_ϵ . That is, if the true state of the world is $\omega \in \Omega^N$, then $S_i = \omega + \epsilon_i$ where $\epsilon_i \sim N[0, V_\epsilon]$, with ϵ_i independent of ω and ϵ_{-i} , for both $i = 1, 2$. We refer to the state-independent Bayesian coordination game under the normal state-signal structure as $\Gamma_{V_\Omega, V_\epsilon}^N$.

The following result shows that a threshold equilibrium always exists in $\Gamma_{V_\Omega, V_\epsilon}^N$. However, there is only one specific \bar{s} that permits such a threshold equilibrium.

Theorem 4. *The threshold strategy profile $\sigma^{\bar{s}}$ is a Bayesian Nash Equilibrium of $\Gamma_{V_\Omega, V_\epsilon}^N$ if and only if*

$$\bar{s} = \frac{\sqrt{\frac{V_\epsilon^2 + 2V_\Omega V_\epsilon}{V_\Omega + V_\epsilon}} \Phi^{-1}(1-p)}{V_\Omega + V_\epsilon}, \text{ where } \Phi(\cdot) \text{ is the CDF of the standard normal distribution.}$$

Proof. Let $\sigma^{\bar{s}}$ be a Bayesian Nash equilibrium of $\Gamma_{V_\Omega, V_\epsilon}^N$ for some fixed threshold $\bar{s} \in (-\infty, +\infty)$. As before, S_i and S_{-i} denote the random variables that represents the signals of players i and $-i$, respectively.

We first note that the random vector $S = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$ has a bivariate normal distribution with mean matrix $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and covariance matrix $\Sigma = \begin{bmatrix} V_\Omega + V_\epsilon & V_\Omega \\ V_\Omega & V_\Omega + V_\epsilon \end{bmatrix}$. This is due to the following facts: (i) for both $i = 1, 2$, S_i is normally distributed with mean 0 and variance $V_\Omega + V_\epsilon$ since $S_i = \omega + \epsilon$ and ω and ϵ are independent normally distributed with mean 0 and variance V_Ω and V_ϵ , respectively; (ii) $Cov(S_1, S_2) = E((S_1 - E(S_1))(S_2 - E(S_2))) = E(S_1 S_2) = E((\omega + \epsilon_1)(\omega + \epsilon_2)) = E(\omega^2) + E(\omega \epsilon_2) + E(\epsilon_1 \omega) + E(\epsilon_1 \epsilon_2) = V_\Omega$ since ω , ϵ_1 , and ϵ_2 are all mutually independent and $E(\omega) = E(\epsilon_1) = E(\epsilon_2) = 0$ implies $E(\omega \epsilon_2) = E(\epsilon_1 \omega) = E(\epsilon_1 \epsilon_2) = 0$ and $E(\omega^2) = V_\Omega$; (iii) for any $\alpha, \beta \in \mathbb{R}$, $\alpha S_1 + \beta S_2$ is normally distributed even though S_1 and S_2 are not independent since $\alpha S_1 + \beta S_2 = \alpha(\omega + \epsilon_1) + \beta(\omega + \epsilon_2) = (\alpha + \beta)\omega + \alpha\epsilon_1 + \beta\epsilon_2$ and ω , ϵ_1 , and ϵ_2 are independent normally distributed random variables; (iv) the random vector $S = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$ has a bivariate normal distribution due to the fact that whenever any linear combination of two (not necessarily independent) normally distributed random variables is normally distributed, the associated random vector has a bivariate normal distribution (see [3] for further details).

Next, we note that the conditional random variable $S_{-i}|S_i$ is normally distributed with mean $\frac{V_\Omega}{V_\Omega + V_\epsilon} S_i$ and variance $V_\Omega + V_\epsilon - \frac{V_\Omega}{V_\Omega + V_\epsilon} V_\Omega = \frac{V_\epsilon^2 + 2V_\Omega V_\epsilon}{V_\Omega + V_\epsilon}$. This follows from the fact that if $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ has a bivariate normal distribution with mean matrix $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, and covariance matrix $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, then, $X_1|X_2$ is distributed normally with mean $\mu_1 + \frac{\Sigma_{12}}{\Sigma_{22}}(X_2 - \mu_2)$ and variance $\Sigma_{11} - \frac{\Sigma_{12}^2}{\Sigma_{22}}$ (see [3] for further details).

As before, the necessary and sufficient condition for σ^s to be a Bayesian Nash equilibrium of $\Gamma_{V_\Omega, V_\epsilon}^{normal}$ is twofolds: (i) for each $s > \bar{s}$, $Pr(S_{-i} > \bar{s} | S_i = s) \geq p$ for both $i = 1, 2$, and (ii) for each $s \leq \bar{s}$, $Pr(S_{-i} \leq \bar{s} | S_i = s) \geq 1 - p$ for both $i = 1, 2$. Condition (i) makes it incentive compatible to play U when one's own signal is over the threshold whereas condition (ii) makes it incentive compatible to play D when one's own signal is below the threshold.

Recall that if $X \sim N[\mu, \sigma^2]$, then $Z = \frac{X - \mu}{\sigma} \sim N[0, 1]$. Thus, condition (i) implies that for each $s > \bar{s}$, it must be that $Pr\left(Z > \frac{\bar{s} - \frac{V_\Omega}{V_\Omega + V_\epsilon} s}{\sqrt{\frac{V_\epsilon^2 + 2V_\Omega V_\epsilon}{V_\Omega + V_\epsilon}}}\right) \geq p$ where $Z \sim N[0, 1]$. On the other hand, condition (ii) implies that, for each $s \leq \bar{s}$, we must have $Pr\left(Z \leq \frac{\bar{s} - \frac{V_\Omega}{V_\Omega + V_\epsilon} s}{\sqrt{\frac{V_\epsilon^2 + 2V_\Omega V_\epsilon}{V_\Omega + V_\epsilon}}}\right) \geq 1 - p$ where $Z \sim N[0, 1]$. Observe that for a fixed \bar{s} , $Pr\left(Z > \frac{\bar{s} - \frac{V_\Omega}{V_\Omega + V_\epsilon} s}{\sqrt{\frac{V_\epsilon^2 + 2V_\Omega V_\epsilon}{V_\Omega + V_\epsilon}}}\right)$ is increasing in s whereas $Pr\left(Z \leq \frac{\bar{s} - \frac{V_\Omega}{V_\Omega + V_\epsilon} s}{\sqrt{\frac{V_\epsilon^2 + 2V_\Omega V_\epsilon}{V_\Omega + V_\epsilon}}}\right)$ is decreasing in s . Therefore, condition (i) and (ii) are satisfied if and only if both conditions hold when $s = \bar{s}$. This is true if and only if $Pr\left(Z \leq \frac{\frac{V_\epsilon}{V_\Omega + V_\epsilon}}{\sqrt{\frac{V_\epsilon^2 + 2V_\Omega V_\epsilon}{V_\Omega + V_\epsilon}}} \bar{s}\right) = 1 - p$ where $Z \sim N[0, 1]$. That is, $\Phi\left(\frac{\frac{V_\epsilon}{V_\Omega + V_\epsilon}}{\sqrt{\frac{V_\epsilon^2 + 2V_\Omega V_\epsilon}{V_\Omega + V_\epsilon}}} \bar{s}\right) = 1 - p$ where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Since every cumulative distribution function is invertible, taking Φ^{-1} of both sides and solving for the threshold \bar{s} gives us the desired conclusion, i.e., we must have $\bar{s} = \frac{\sqrt{\frac{V_\epsilon^2 + 2V_\Omega V_\epsilon}{V_\Omega + V_\epsilon}} \Phi^{-1}(1 - p)}{\frac{V_\epsilon}{V_\Omega + V_\epsilon}}$. \square

B.5 General Result for Any State-Signal Structure

Next, we provide a general result for any state-signal structure that satisfies a simple monotonicity condition—ensuring that states and signals are well-ordered. Our result provides a necessary and sufficient condition for the existence of a threshold equilibrium as long as the state signal structure $\langle \Omega, \rho, \tau \rangle$ satisfies the monotone likelihood ratio property (MLRP). MLRP guarantees that as the signal of a player increases, his posterior distribution over the states shifts weight to the right, which in turn shifts his posterior distribution over the other player's signals to the right. We note that MLRP is satisfied for all the distributions we have investigated so far. This condition allows us to focus on the signals closes to the signals where the temptation to deviate will be highest.

Our result here generalizes the result from previous models: for a threshold equilibrium to exist, the best response needs to change when the player goes from one side of the threshold to the other. This requires that the posterior belief that the other player's signal is above the threshold must go from below p to above p , as one goes from getting a signal below compared to above the threshold. This could happen if the belief varies continuously and traverses p exactly at the threshold (as we have seen with the normal state-signal structure, or in the continuous-uniform state-signal structure when $p = \frac{1}{2}$), or if beliefs discontinuously jump, and p falls anywhere within the jump (as in our categorical and discrete-uniform state-signal structures).

Definition 5. We say that the signal structure satisfies the monotone likelihood ratio property (MLRP) if for any signals s, s' and states ω, ω' , whenever $s' > s$ and $\omega' > \omega$, $\frac{f(\omega'|S_i=s')}{f(\omega'|S_i=s)} \geq \frac{f(\omega|S_i=s')}{f(\omega|S_i=s)}$ where f is the conditional PDF of the state given the signal.

Theorem 5. Let Γ be a state-independent Bayesian coordination game with the state-signal structure $\langle \Omega, \rho, \tau \rangle$ satisfying MLRP. Then, for any $\bar{s} \in \Omega$, a threshold strategy profile $\sigma^{\bar{s}}$ is a Bayesian Nash equilibrium of Γ if and only if, for both $i \in \{1, 2\}$, we have

$$Pr(S_{-i} > \bar{s} | S_i = \bar{s}) \leq p \leq \inf\{Pr(S_{-i} > \bar{s} | S_i = s) : s \in \Omega \text{ with } s > \bar{s}\}.$$

Proof. By definition of Bayesian Nash equilibrium and threshold strategy, $\sigma^{\bar{s}}$ is a Bayesian Nash equilibrium of Γ if and only if (i) for both $i \in \{1, 2\}$, $Pr(S_{-i} > \bar{s} | S_i = s) \geq p$ for each $s \in \Omega$ with $s > \bar{s}$, and (ii) for both $i \in \{1, 2\}$, $Pr(S_{-i} \leq \bar{s} | S_i = s) \geq 1 - p$ for each $s \in \Omega$ with $s \leq \bar{s}$, i.e., $Pr(S_{-i} > \bar{s} | S_i = s) \leq p$ for each $s \in \Omega$ with $s \leq \bar{s}$.

(\Rightarrow) Observe that (ii) immediately implies the left-hand side inequality, when $s = \bar{s}$. Next, we show that the right-hand side inequality holds as well.

Let $p^* = \inf\{Pr(S_{-i} > \bar{s} | S_i = s \in \Omega) : s > \bar{s}\}$. Suppose, for contradiction, $p > p^*$. Let $\epsilon = \frac{p-p^*}{2} > 0$. By definition of infimum, there exists $\tilde{s} \in \Omega$ with $\tilde{s} > \bar{s}$ such that $Pr(S_{-i} > \bar{s} | S_i = \tilde{s}) < p^* + \epsilon$. But, since $p^* + \epsilon = \frac{p+p^*}{2} < p$, this implies $Pr(S_{-i} > \bar{s} | S_i = \tilde{s}) < p$, a contradiction to (i) as $\tilde{s} \in \Omega$ with $\tilde{s} > \bar{s}$.

Thus, $Pr(S_{-i} > \bar{s} | S_i = s) \leq p \leq \inf\{Pr(S_{-i} > \bar{s} | S_i = s) : s \in \Omega \text{ with } s > \bar{s}\}$ for both $i \in \{1, 2\}$.

(\Leftarrow) Suppose $Pr(S_{-i} > \bar{s} | S_i = s) \leq p \leq \inf\{Pr(S_{-i} > \bar{s} | S_i = s) : s \in \Omega \text{ with } s > \bar{s}\}$ for both $i \in \{1, 2\}$. We will show that (i) and (ii) holds.

MLRP implies that $Pr(S_{-i} > \bar{s} | S_i = s)$ is (weakly) increasing in s for all $s \in \Omega$. This follows from applying MLRP twice: when player i gets a higher signal, he puts higher weight on higher states, and in each of those higher states he puts higher weight on the other player getting higher signals. Because $Pr(S_{-i} > \bar{s} | S_i = s)$ is (weakly) increasing in s , for any $s' < \bar{s}$, $Pr(S_{-i} > \bar{s} | S_i = s') \leq Pr(S_{-i} > \bar{s} | S_i = \bar{s})$. Since $Pr(S_{-i} > \bar{s} | S_i = \bar{s}) \leq p$, this implies $Pr(S_{-i} > \bar{s} | S_i = s) \leq p$ for all $s \in \Omega$ with $s \leq \bar{s}$, i.e., (ii) holds.

To see that (i) also holds, observe that as $Pr(S_{-i} > \bar{s} | S_i = s)$ is (weakly) increasing in s for all $s \in \Omega$, we have for any $s' > \bar{s}$, $Pr(S_{-i} > \bar{s} | S_i = s') \geq Pr(S_{-i} > \bar{s} | S_i = \bar{s})$. But since, $p \leq \inf\{Pr(S_{-i} > \bar{s} | S_i = s) : s \in \Omega \text{ with } s > \bar{s}\}$, we must have $Pr(S_{-i} > \bar{s} | S_i = s) \geq p$ for each $s \in \Omega$ with $s > \bar{s}$. \square

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 U , U if and only if $s_i > 0/7$, U if and only if $s_i > 1/7$, U if and only if $s_i > 2/7$, ..., U if and only if $s_i > 7/7$, always D .

Each simulation proceeded as follows. First, we assigned all the players to the play the strategy U 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.

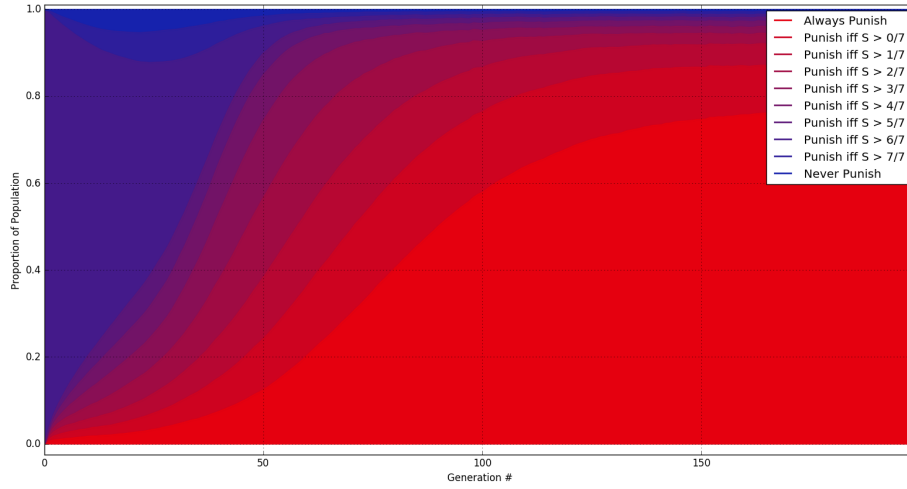


Figure 2: Average Frequency of Strategies, 500 simulations, $p < 1/2$

We also run the same simulations, but start by assigning all players to play the strategy U 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. 2.

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

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

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

The simulations are identical to those described in Section C.1.1 except that the state 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. 3a and 3b present a single run with $p > 1/2$ and $p < 1/2$, respectively. Fig. 3c and 3d present the average frequency of the strategies in each period for $p > 1/2$ and $p < 1/2$, respectively.

C.2.2 N Discrete Values

In Fig. 4 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 $\{1, 2, \dots, 100\}$. We start the entire population at U if and only if $N \geq 10$. The threshold strategy profile is no longer expected to be stable. Indeed, it is not.

C.2.3 Uniform Distribution with an Atom at $\Omega = \omega_l$

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

C.2.4 State-Dependent Payoffs

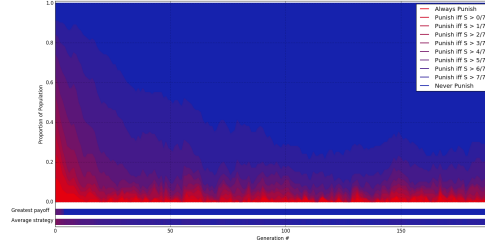
In Fig. 6, we present the same simulations as those in Fig. 4a-b, but we now let payoffs be $a = 4(2\omega + 1/2)$, $b = 2(2\omega + 1/2)$, $c = 0$, $d = 5$. The strategy space includes the following 20 strategies: Sanction if and only if $S_i > \bar{s}$, with $\bar{s} \in \{0, 0.06, \dots, 0.94, 1\}$. We start the population at U if and only if $S_i > 0.12$. We run the simulations for 190 generations. We run the same simulations for $p = 1/3$.

D Experimental Evidence: Vignettes

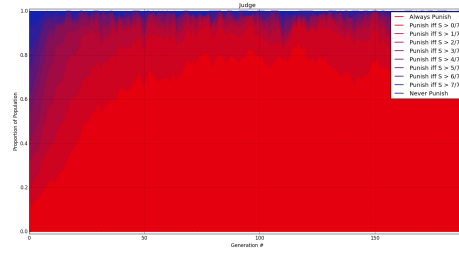
D.1 Ethics Compliance and Preregistration

This research complies with relevant ethical regulations and was approved by the MIT University Institutional Review Board. We obtained informed consent from all participants. Participants were paid the going wage on the online platform that we used to recruit (Amazon Mechanical Turk). This study was preregistered through AsPredicted.com [I will include the link as a final step when we submit and “release” it to be viewable].²

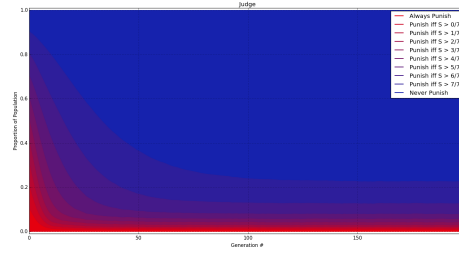
²This study is a near replication of a pilot study with 300 subjects that yielded similar results. Based on the pilot study we omitted one vignette (reducing from six to five), decided generally not to ask questions related to prediction 2 in the control conditions (where they were not so relevant), and made some changes to the control condition for the vignette in which a country invades another country’s territory. Stimuli and data from our pilot study are available upon request.



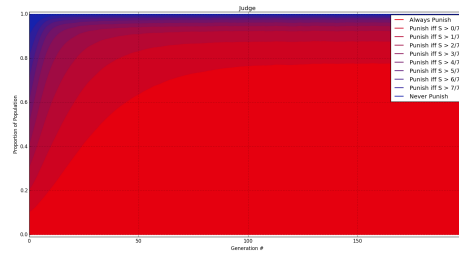
(a) Frequency of Strategies, Random Starting Point, Single Run, $p > 1/2$



(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 3: Arbitrary Assignment of Starting Strategies for Fig. 4a and 4b

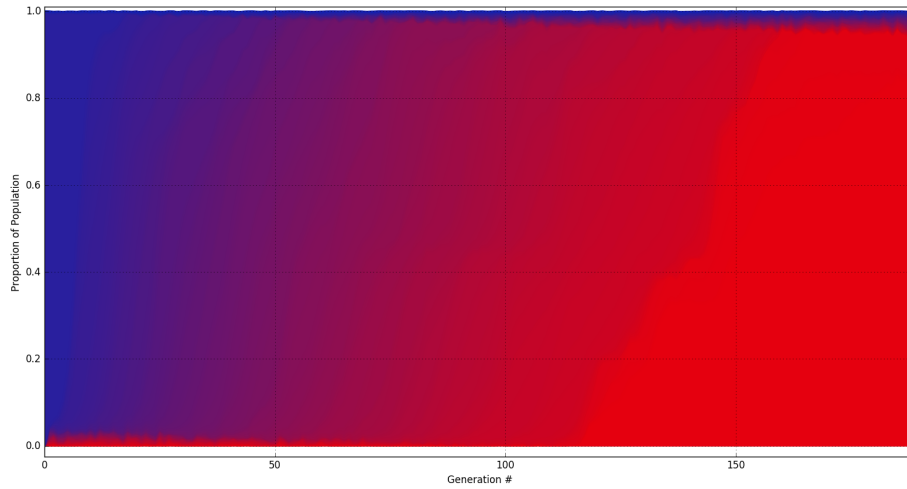


Figure 4: Average Frequency of Strategies, 100 Discrete Values

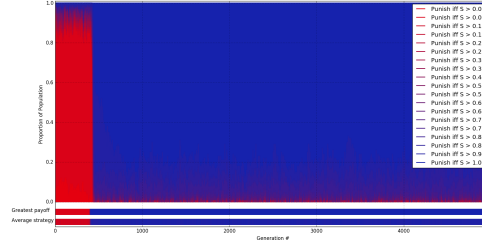
D.2 Subjects

We recruited all subjects for all studies online through Amazon Mechanical Turk (AMT). AMT is an online marketplace where employers pay users for completing short tasks (generally about 10 minutes), usually referred to as Human Intelligence Tasks (HITs)—for a relatively small payment (generally less than a \$1). Workers who have been recruited on AMT receive a baseline payment and can also be paid a bonus depending on their performance and/or choices in the task. This setup lends itself well to incentivized experiments, such as our Study 2: the baseline payment acts as the ‘show up’ fee and the bonus payment is based on the workers’ choices, if applicable.

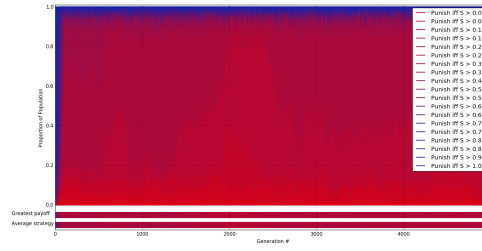
Subjects were 501 respondents on Amazon Mechanical Turk (mean age = 41.0, SD = 12.5, range from 18 to 74). Participation was limited to subjects from the United States and Canada. 45.0% were male, 54.0% female, 1% other/prefer not to say. 74.1% were White/Caucasian, 9.8% African American, 5.0% Hispanic, 6.6% Asian American, 1% marked “other,” and 0.4% marked “prefer not to say.” 90.0% had graduated at least from high school, and 28.3% had graduated at least from college.

D.3 Sample size determination

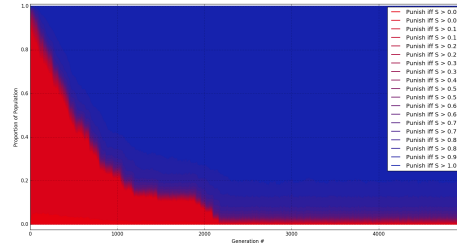
As described above, this study was based on a pilot study that was very similar except for a few changes, giving us confidence that we were requesting ample subjects. As preregistered, we requested 500 subjects from Amazon Mechanical Turk. We set the sample size for each study at the beginning by requesting a certain round number of subjects via Turkprime.com (a website that facilitates running studies on Amazon Mechanical Turk), and then let the study run to completion without ever altering the sample size. Studies typically returned a few more subjects than we requested, likely because some people completed the survey (hosted by Qualtrics) without entering their completion code into Amazon Mechanical Turk and being



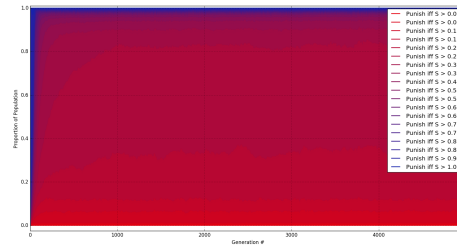
(a) Frequency of Strategies, Starting at Sanction if and only if $S > 0.22$, Single Run, $p > 1/2$



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

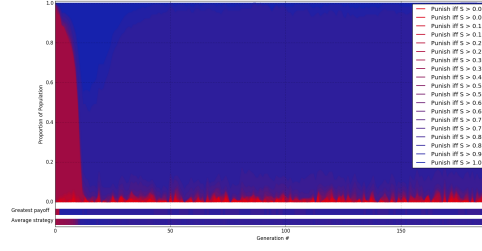


(c) Average Frequency of Strategies, 500 simulations, Starting at Sanction if and only if $S > 0.22$, $p > 1/2$

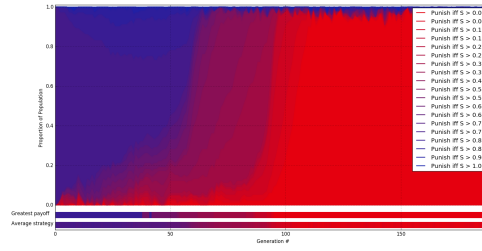


(d) Average Frequency of Strategies, 500 simulation, Start-
ing at Sanction if and only if $S > 0.78$, $p < 1/2$

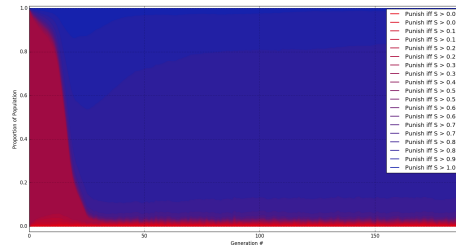
Figure 5: Uniform Distribution with an Atom at $\Omega = \omega_l$



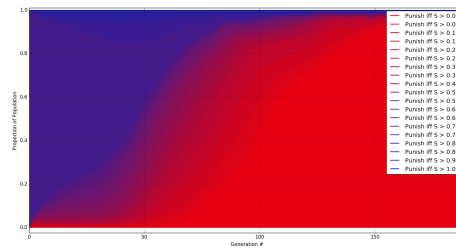
(a) Frequency of Strategies, Starting at Sanction if and only if $S > 0.22$, Single Run, $p > 1/2$



(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 Play U 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 6: State-Dependent Payoffs

counted by that system.

D.4 Methods

This study showed subjects a series of six vignettes about hypothetical transgressions to test two key predictions: (1) whether the psychology of punishment is relatively insensitive to continuous violations compared to categorical violations, and (2) whether there are incentives to treat violations as categorical rather than continuous. In more detail:

Prediction 1: Moral judgments and punishment intuitions will be more sensitive to categorical than continuous distinctions with the same magnitude of impact.

Prediction 2: There are costs to “miscoordination”—to treating norms as continuous, given that others treat them as categorical.

Sub-prediction 1: Deterrence is categorical (subjects perceive the likelihood of a repeat transgression to depend on whether a transgression is punished, but are relatively insensitive to how large that transgression was).

Sub-prediction 2: Judgments of and support for the punisher are categorical. (Subjects view the punisher of a transgression more favorably, and are more likely to support their decision, if they punish a transgression relative to not punishing, but these judgments are relatively insensitive to the magnitude of the transgression. Subjects expect relevant others to behave and feel likewise.)

To test these predictions, subjects saw five vignettes featuring different transgressions. We presented these vignettes in random order. All vignettes had eleven conditions, one control condition in which no transgression (or in one case, good deed) occurred, and ten levels of “impact.” After each vignette, we asked a number of corresponding questions to test our predictions. (The full text of these questions, including specific scale endpoints, can be found in our online-only materials). Based on what made sense for the scenario, the first three vignettes were used to test both predictions, and the last two were used only to test the first prediction. Subjects run in the control conditions of vignettes 1-3 did not see questions intended to test prediction 2, because the contrast with the control condition was not our focus for this second prediction.

After reading each vignette and answering its key questions, subjects completed a few understanding check questions, before moving on to the next vignette. At the end of the study, subjects answered a few demographic questions.

D.5 Statistical Analyses

We used regressions for all analyses. As preregistered, we used one-tailed tests for cases in which we predicted a significant result in a particular direction, and committed in our preregistration not to interpret results in the opposite direction as meaningful. Analyses were between-subjects unless specified as within-subjects.

For all questions testing prediction 1, we wished to obtain two estimates: (1) of the difference in the outcome between the control group and the first level of impact, and (2) of the average difference in the outcome between each level of impact and the one below it, for levels 2-10. We obtained these via a linear regression of the outcome of interest on a ‘dummy’ indicating whether the participant was in the control or any treatment of level, and a variable indicating the impact level (1-10).

We predicted that (1) will be significant, with higher means/more “yes” answers with the lowest impact transgression/good deed, compared to the control condition where there is no transgression/good deed. This reflects finding even the lowest impact transgression immoral and worthy of punishment, and the lowest impact good deed altruistic and worthy of reward.

We also predicted that (2) would not be significant, or would be very small relative to (1). In other words, given a transgression/good deed, the impact of that transgression or good deed would not matter much for perceptions of moral wrongness, altruism, etc., compared to the impact of having a transgression in the first place. We also expected (and preregistered the prediction) that this slope would be quite close to zero, and that the pattern of means would appear random rather than generally increasing or decreasing.

We felt that any specific ratio of the effect of a transgression versus impact would be arbitrary, so we committed (including in our preregistration) to analyzing both (1) and (2) and reporting the results, and showing the graph of the pattern of means (plus error bars). Our readers can determine whether they are convinced that we have found strong evidence for our hypothesis.

For questions testing prediction 2, we asked subjects questions about their expectation for deterrence and support for potential punishers when these potential punishers did or did not punish a transgression (asked within subjects). We calculated the difference between the answers to these two questions, and regressed this difference on the level of impact of the transgression.

For ease of understanding, we will present further details of the methods and results vignette by vignette.

D.6 Vignette 1: Government Killing Protesters

D.6.1 Methods

For subjects who received the transgression treatment, the vignette read:

You will now read a vignette about a foreign government.

Imagine that there are large-scale protests in a foreign country (let’s call it Country X) in response to perceived election fraud. The crowds of protestors are large and chaotic, but nonviolent. In response to the protests, state police arrest and execute [one to ten] unarmed, peaceful protest leader[s]. Incontrovertible evidence of this poisoning is leaked and becomes international news.

Country X is a member of an economic and military alliance with a number of other nations, none of whom has executed a peaceful protestor in recent history. These other member nations are considering whether to sanction the government responsible.

We varied the levels of impact by varying the number of protesters killed by the hypothetical government from 1-10.

For subjects in the control group, the vignette read:

You will now read a vignette about a foreign government.

Imagine that there are large-scale protests in a foreign country (let's call it Country X) in response to perceived election fraud. The crowds of protestors are large and chaotic, but nonviolent. There are some rumors that in response to the protests, state police arrested and executed one or several unarmed, peaceful protest leaders, but these rumors turn out to be unsubstantiated.

Country X is a member of an economic and military alliance with a number of other nations, none of whom has executed a peaceful protestor in recent history. These other member nations are considering whether to sanction the government responsible."

D.6.2 Results

Results for Prediction 1 are summarized in SI Table 1.1, and for prediction 2 are summarized in Table 1.2. CX RTP stands for "Country X's Response to the Protests." Questions here and for the other vignettes may be abbreviated in the tables, but the full question text can be found in our online materials. See Tbl ?? and ??.

D.7 Vignette 2: Invading Another Country's Territory

D.7.1 Methods

For subjects who received the transgression treatment, the vignette read:

You will now read a vignette about two countries.

Country A and Country B both include territory distributed around the world. This territory varies quite a bit in size, natural resources, and population. Country B borders an island that belongs to Country A. This island is called Greener Island. Only about [a thousand people to ten thousand people] live on Greener Island, who are all citizens of Country A. Greener Island also contains no natural resources and has no strategic defense value. Country B decides that it wants to expand its territory to include Greener Island, and attacks Greener Island with the intent of taking it over. Country A has to decide whether to defend Greener Island or let Country B have it. Country A is essentially assured victory, but at considerable expense—the defense will be expensive and involve significant loss of life.

We varied the levels of impact by varying the number of people living on the invaded island, from 1000-10,000, in increments of 1000. The invaded island belongs to one country but is attacked by another.

For subjects in the control group, the vignette read:

You will now read a vignette about two countries.

Country A and Country B both include territory distributed around the world. This territory varies quite a bit in size, natural resources, and population. Some of their citizens also live outside of their territories. About a thousand citizens of Country A live on an island called Greener Island that actually belongs to Country B. Greener Island contains no natural resources and has no strategic defense value. Country B has no citizens living on the island, and has historically ignored it. Country B borders this island, and decides that it wants to have more presence there. Country B lands members of its military on Greener Island. Country A has to decide whether to vie for Greener Island and try to take it over by sending its own military, or let Country B land its military and continue to own it. Country A is essentially assured victory if it vies for Greener Island, but at considerable expense—the effort will be expensive and involve significant loss of life.

D.7.2 Results

Results for Prediction 1 are summarized in SI Tbl ??, and for prediction 2 are summarized in SI Tbl ??.

D.8 Vignette 3: Shoplifting

D.8.1 Methods

For subjects who received the transgression treatment, the vignette read:

You will now read a vignette about a college student named Jenny.

A college student named Jenny enters a local family-owned drugstore on her own and, after looking around carefully, quickly slips an item from the beauty aisle into her purse. The item is worth [\$2 to \$20]. She does not really need this item, but gets a thrill from the idea of taking it without paying. She tries to leave the store without paying for the item. However, she hadn't realized that the bar codes on all products trigger an alarm if someone tries to exit the door without purchasing them. The alarm is triggered, and Jenny is approached by store employees.

We varied the levels of impact by varying the value of the item stolen, from \$2 to \$20, in increments of \$2.

For subjects in the control group, the vignette read:

You will now read a vignette about a college student named Jenny.

A college student named Jenny enters a local family-owned drugstore on her own and, after looking around carefully, notices an item worth \$2. She does not really need this item, but gets a thrill from the idea of taking it without paying. She considers slipping the item into her purse, but thinks better of it and decides not to take it. Jenny leaves the store without taking anything.

D.8.2 Results

Results for Prediction 1 are summarized in SI Tbl ??, and for prediction 2 are summarized in SI Tbl ??.

D.9 Vignette 4: Murder in a Trolley Problem

D.9.1 Methods

For subjects who received the transgression treatment, the vignette read:

You will now read a "trolley problem" vignette.

Let us introduce a contrived scenario that philosophers call a "thought experiment" that helps us gauge people's moral intuitions. First, imagine that a trolley is headed toward a repairman who is stuck on a track. If the trolley continues on its current course, it will hit the repairman and kill him. A control worker nearby sees this. He also sees that there is a fork coming up, and that the trolley can be diverted to a second track at the fork to save the repairman. But the control worker's brand new car is parked on the second track. The control worker also knows that the repairman stuck on the first track is a forty-five-year-old who has unknowingly swallowed a pill that will painlessly kill him in [one to ten] year[s]. The control worker decides not to flip the switch that would divert the train onto the second track. As a result, the repairman is hit and killed, and the control worker's car is unscathed.

We varied the levels of impact by varying the number of years the person stuck on the tracks has left to live, from 1-10. The transgressor kills this person to save his car.

For subjects in the control group, the vignette read:

You will now read a "trolley problem" vignette.

Let us introduce a contrived scenario that philosophers call a "thought experiment" that helps us gauge people's moral intuitions. First, imagine that a trolley is headed toward a repairman who is stuck on a track. If the trolley continues on its current course, it will hit the repairman and kill him. A control worker nearby sees this. He also sees that there is a fork coming up, and that the trolley can be diverted to a second track at the fork to save the repairman. But the control worker's brand new car is parked on the second track. The control worker also knows that the repairman stuck on the first track is a forty-five-year-old who has unknowingly swallowed a pill that will painlessly kill him in one year. The control worker decides to flip the switch that diverts the train onto the second track. As a result, the control worker's car is hit and destroyed, and the repair man is unscathed.

D.9.2 Results

Results for Prediction 1 are summarized in SI Tbl ??.

D.10 Vignette 5: Volunteering at a Soup Kitchen

D.10.1 Methods

For subjects who received the transgression treatment, the vignette read:

You will now read a vignette about a man named Dan.

On the way into his local supermarket, Dan sees a booth representing a local soup kitchen. It's early November at the time, and the representatives at the booth are asking people to donate an hour a week helping to serve the homeless at the kitchen during the holiday season. Dan decides to volunteer. He signs up, and happily serves at the kitchen for [an hour to ten hours] a week through the end of December.

We varied the levels of impact by varying the number of hours volunteered per week, from 1 to 10.

For subjects in the control group, the vignette read:

You will now read a vignette about a man named Dan.

On the way into his local supermarket, Dan sees a booth representing a local soup kitchen. It's early November at the time, and the representatives at the booth are asking people to donate an hour a week helping to serve the homeless at the kitchen during the holiday season. Dan decides that he doesn't have time to volunteer.

D.10.2 Results

Results for Prediction 1 are summarized in SI Tbl ??.

E Experimental Evidence: Incentivized Economic Games

E.1 Ethics Compliance and Preregistration

This research complies with relevant ethical regulations and was approved by the MIT University Institutional Review Board. We obtained informed consent from all participants. Participants were paid the going wage on the online platform that we used to recruit (Amazon Mechanical Turk). This study was preregistered through AsPredicted.com [will fill in as last step] and is a replication of a pilot study that yielded similar results.

E.2 Experimental design overview

This study investigated whether subjects were less sensitive to impact when punishing others for unfair choices than when making decisions about those choices that impacted their own payouts.

E.3 Sample size determination

This preregistered study replicated a pilot study, giving us confidence that we were requesting ample subjects. As preregistered, we requested 600 subjects from Amazon Mechanical Turk.

E.4 Subjects

Subjects were 578 respondents on Amazon Mechanical Turk (mean age = 34.23, SD = 10.22, range from 18 to 67). Participation was limited to subjects from the United States and Canada. 44.3% were male, 54.5% female. 76.6% were White/Caucasian, 7.6% African American, 4.0% Hispanic, 5.5% Asian American, 0.5% marked “other”, and 0.7% marked “prefer not to say”. 88.6% had graduated at least from high school, and 31.1% had graduated at least from college.

E.5 Procedure

This study was a 3x2 design. Subjects read about a “Roller” who made a decision that was unfair with higher impact, unfair with lower impact, or fair. (The impact in the fair choice condition was equivalent to that in the lower impact unfair condition.) Subjects then had the opportunity to either pay to punish the Roller, or to pay to avoid being influenced by the Roller’s decision.

Subjects in the punishment condition read the following description of the Roller’s decision (the Roller is referred to as Player 1):

In a moment, you will participate in an interaction with another MTurk worker, whom we will call Player 1.

This worker had a different interaction with a different MTurk worker, whom we will call Player 2.

In this other interaction, both Player 1 and Player 2 started with a 50-cent bonus. Player 1 then chose which of two virtual dice to roll. These rolls are just like rolling a real die—there is an equal probability that each of the six sides will land “face up.”

The die have some sides that are red, and some sides that are black. If the die lands on a black side, Player 1 loses 25 cents from their 50-cent bonus. If the die lands on a red side, Player 2 loses their entire 50-cent bonus [OR 25 cents of their 50-cent bonus].

Player 1 chose between two die. Die A had three red sides and three black sides, and die B had five red sides and one black side.

Player 1 chose to roll die B, which had five red sides and one black side. [OR Player 1 chose to roll die A, which had three red sides and three black sides.]

Subjects in the avoidance condition read the following description of the Roller’s decision (the Roller is referred to as Player 1, the subject themselves is now in the position of Player 2):

You will now participate in an interaction with another Mturk worker, whom we will call Player 1.

In this other interaction, both you and Player 1 start with a 50-cent bonus. Player 1 then chooses which of two virtual dice to roll. These die rolls are just like rolling a real die—there is an equal probability that each of the six sides will land “face up.”

The die have some sides that are red, and some sides that are black. If the die lands on a red side, you lose 25 cents from your 50-cent bonus. If the die lands on a black side, Player 1 loses their entire 50-cent bonus [OR 25 cents of their 50-cent bonus]. Player 1 chooses between two die. Die A has three red sides and three black sides, and die B has five red sides and one black side.

You are paired with a Player 1 who chose die B, which has five red sides and one black side. [OR you are paired with a player 1 chose to roll die A, which had three red sides and three black sides.]

We varied fairness and impact as follows:

Unfair, higher impact condition: Roller chooses the die with five red sides (five sides favoring them), their partner stands to lose their entire 50 cent bonus if the die lands on red.

Unfair, lower impact condition: Roller chooses the die with five red sides (five sides favoring them), their partner stands to lose 25 cents of their 50-cent bonus if the die lands on red.

Fair condition (also lower impact): Roller chooses the die with three red sides (three sides favoring them), their partner stands to lose 25 cents of their 50-cent bonus if the die lands on red.

Subjects were then given an opportunity to either pay to punish (or reward) the Roller, or pay to avoid being impacted by the decision the Roller made (using a “willingness to pay” measure).

The punishment condition instructions were as follows:

You are also going to play a game with Player 1. (You will not interact with Player 2.) In this second game, you start with a 30-cent bonus. You can then choose to pay to reduce Player 1’s bonus, pay to increase Player 1’s bonus, or neither.

If you choose to pay to reduce Player 1’s bonus, for every 1 cent you pay, Player 1’s bonus is reduced by 3 cents. You can pay up to 15 cents to subtract up to 45 cents from Player 1’s bonus. If you choose to pay to increase Player 1’s bonus, for every 1 cent you pay, Player 1’s bonus is increased by 3 cents. You can pay up to 15 cents to add up to 45 cents to Player 1’s bonus. This interaction is one-way. Player 1 does not make any decisions that affect the bonus you receive.

As an example, if you pay 15 cents to reduce Player 1’s bonus, Player 1’s bonus will be reduced by 45 cents—which is the maximum amount that you can reduce Player 1’s bonus—and you will receive a bonus of 15 cents instead of 30 cents.

If you pay 15 cents to increase Player 1's bonus, Player 1's bonus will be increased by 45 cents—which is the maximum amount that you can increase Player 1's bonus—and you will receive a bonus of 15 cents instead of 30 cents.

If you give up 0 cents, Player 1's bonus will not be reduced or increased at all. So, you always lose money by reducing or increasing Player 1's bonus, and Player 1 does not make any decisions that impact your bonus.

How many cents (if any) would you like to pay to reduce Player 1's bonus?

Remember, you have received a 30-cent bonus, and for every 1 cent that you pay, Player 1's bonus is reduced by 3 cents. Player 1 does not make any decisions that influence your bonus.

How many cents (if any) would you like to pay to increase Player 1's bonus?

Remember, you have received a 30-cent bonus, and for every 1 cent that you pay, Player 1's bonus is increased by 3 cents. Player 1 does not make any decisions that influence your bonus.

The avoidance condition instructions were as follows:

You have the option to pay to exit this game. We are giving you another bonus of 30 cents. Below, please indicate how much of this bonus you would be willing to pay to exit the game, from 0 cents to 30 cents. We will randomly generate a number between 0 and 30. If your number is equal to or larger than this number, you will exit the game and keep your 50-cent bonus. You will also pay the number we generate. Otherwise, Player 1's die roll will determine whether you lose 25 cents from your 50-cent bonus.

For example, if you indicate that you are willing to pay 15 cents and we randomly generate the number 14 you will pay 14 cents from your 30-cent bonus to exit the game and keep your 50-cent bonus. If we generate 15 you will pay 15 cents to exit the game and keep your 50-cent bonus. If we generate 16 you will not exit the game. You will keep your 30-cent bonus, and Player 1's die roll will determine whether you lose 25 cents from your 50-cent bonus. So indicate the maximum amount that you would be willing to pay to exit the game. If you are willing to pay nothing you will definitely not exit the game, and if you are willing to pay all 30 cents you will definitely exit the game.

How much would you be willing to pay from your 30 cent bonus to exit the game? Please enter a number between 0 and 30.

Subjects also answered understanding check questions throughout the survey, and some standard demographic questions at the end.

E.6 Results

We tested two key predictions. Our primary prediction was that the interaction between impact and type of dependent variable would be positive, such that the effect of impact on willingness to pay to avoid the Roller

was greater than the effect of impact on punishment. We tested this prediction with a linear regression of the standardized DVs (WTP to avoid / punishment) on a ‘dummy’ indicating the type of DV, the level of impact, and an interaction between the type of DV and level of impact; this analysis was restricted to participants in the unfair treatment. We found that the interaction between the type of DV and the level of impact was indeed significant: $B = 0.35$, $p = .025$.

Our second prediction was that fairness would significantly impact both willingness to pay and punishment. We tested this prediction with a linear regress of the DVs on a ‘dummy’ indicating whether the unfair die was chosen, the type of DV, and an interaction between these two variables. We restricted this analysis to the high impact treatment. We then evaluated the coefficient on the unfair dummy at both types of DVs. This second prediction was born out for both willingness to pay ($B = 0.30$, $p = .009$) and punishment ($B = 0.26$, $p = .002$). Critically, this shows that punishment can be sensitive to something, just not impact.

Additionally, we would predict that the difference between the effect of impact on punishment versus willingness to pay would be greater than the difference between the effect of fairness on punishment versus willingness to pay. However, we were underpowered to detect this three-way interaction, so we predicted only that it would either go in the predicted direction or be indistinguishable from zero. We tested this triple interaction with a linear regression of the DVs on the type of DV, the impact, and an interaction of the two, as well whether the unfair die was chosen, and an interaction with the type of DV. We then calculated the triple interaction with Stata’s *lincom* command. For this additional prediction, we found that the triple interaction was indistinguishable from zero: $B = 0.07$, $p = .825$.

References

- [1] M. J. Osborne, A. Rubinstein, *A course in game theory* (MIT press, 1994).
- [2] W. Johnson, A. Branscum, T. E. Hanson, R. Christensen, *Bayesian ideas and data analysis: an introduction for scientists and statisticians* (CRC Press, 2010).
- [3] M. L. Eaton, *Multivariate statistics: a vector space approach* (Wiley New York, 1983).

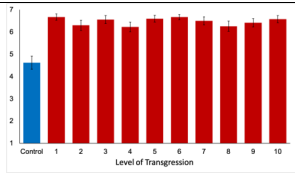
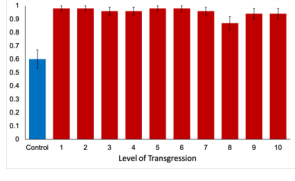
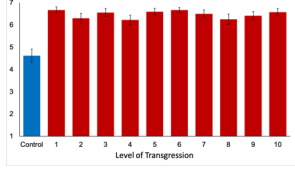
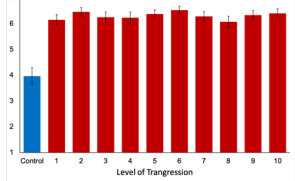
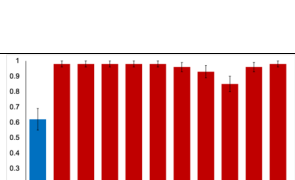
	Graph	First level of transgression vs. Control	Levels 2-10 of transgression	Difference
To what extent was CXRTP a human rights violation?		$B = 1.30$ $p < .001$	$B = -0.00$ $p = .867$	$B = 1.30$ $p < .001$
If you had to choose, would you say that... [repeat] (yes or no)		$B = -1.41$ $p < .001$	$B = -0.01$ $p = .096$	$B = -1.43$ $p < .001$
How morally wrong was CXRTP?		$B = 1.38$ $p < .001$	$B = -0.00$ $p = .833$	$B = 1.38$ $p < .001$
Do you think the other member nations in the trade alliance with Country X should publicly denounce CXRTP?		$B = 1.48$ $p < .001$	$B = 0.01$ $p = .812$	$B = 1.48$ $p < .001$
Would you publicly categorize CXRTP as a human rights violation? (yes or no)		$B = 1.40$ $p < .001$	$B = -0.01$ $p = .149$	$B = 1.42$ $p < .001$

Table 1: Vignette 1 (protestors) - Prediction 1

	Graph	Contrast between punishing versus not	Interaction with level of transgression																																	
If the member nations do nothing to sanction Country X (publicly denounce CX RTP), how likely is Country X's government to kill civilian protesters in the future?	<table><caption>Data for Graph 1: Likelihood of killing protesters</caption><thead><tr><th>Level of Transgression</th><th>Not Punishing (Red)</th><th>Punishing (Blue)</th></tr></thead><tbody><tr><td>1</td><td>6.0</td><td>4.0</td></tr><tr><td>2</td><td>6.0</td><td>4.0</td></tr><tr><td>3</td><td>6.5</td><td>4.5</td></tr><tr><td>4</td><td>6.0</td><td>4.0</td></tr><tr><td>5</td><td>6.5</td><td>4.5</td></tr><tr><td>6</td><td>6.0</td><td>4.0</td></tr><tr><td>7</td><td>6.5</td><td>4.5</td></tr><tr><td>8</td><td>6.0</td><td>4.0</td></tr><tr><td>9</td><td>6.5</td><td>4.5</td></tr><tr><td>10</td><td>6.5</td><td>4.5</td></tr></tbody></table>	Level of Transgression	Not Punishing (Red)	Punishing (Blue)	1	6.0	4.0	2	6.0	4.0	3	6.5	4.5	4	6.0	4.0	5	6.5	4.5	6	6.0	4.0	7	6.5	4.5	8	6.0	4.0	9	6.5	4.5	10	6.5	4.5	$B = -1.92$ $p < .001$	$B = -0.00$ $p = .919$
Level of Transgression	Not Punishing (Red)	Punishing (Blue)																																		
1	6.0	4.0																																		
2	6.0	4.0																																		
3	6.5	4.5																																		
4	6.0	4.0																																		
5	6.5	4.5																																		
6	6.0	4.0																																		
7	6.5	4.5																																		
8	6.0	4.0																																		
9	6.5	4.5																																		
10	6.5	4.5																																		
If the member nations do nothing to sanction Country X (publicly denounce CX RTP), how would you rate their response? (As part of a human rights organization.)	<table><caption>Data for Graph 2: Response rating</caption><thead><tr><th>Level of Transgression</th><th>Not Punishing (Red)</th><th>Punishing (Blue)</th></tr></thead><tbody><tr><td>1</td><td>2.0</td><td>5.5</td></tr><tr><td>2</td><td>2.0</td><td>5.5</td></tr><tr><td>3</td><td>2.0</td><td>5.5</td></tr><tr><td>4</td><td>2.0</td><td>5.0</td></tr><tr><td>5</td><td>2.0</td><td>5.5</td></tr><tr><td>6</td><td>2.0</td><td>5.0</td></tr><tr><td>7</td><td>2.0</td><td>5.0</td></tr><tr><td>8</td><td>2.0</td><td>5.0</td></tr><tr><td>9</td><td>2.0</td><td>5.0</td></tr><tr><td>10</td><td>2.0</td><td>5.5</td></tr></tbody></table>	Level of Transgression	Not Punishing (Red)	Punishing (Blue)	1	2.0	5.5	2	2.0	5.5	3	2.0	5.5	4	2.0	5.0	5	2.0	5.5	6	2.0	5.0	7	2.0	5.0	8	2.0	5.0	9	2.0	5.0	10	2.0	5.5	$B = 3.68$ $p < .001$	$B = -0.02$ $p = .567$
Level of Transgression	Not Punishing (Red)	Punishing (Blue)																																		
1	2.0	5.5																																		
2	2.0	5.5																																		
3	2.0	5.5																																		
4	2.0	5.0																																		
5	2.0	5.5																																		
6	2.0	5.0																																		
7	2.0	5.0																																		
8	2.0	5.0																																		
9	2.0	5.0																																		
10	2.0	5.5																																		
Imagine that another nation in the trade alliance with Country X wants to do nothing in response to Country X's reaction to the protests (publicly denounce Country X). Would you [as an alliance member] support this decision?	<table><caption>Data for Graph 3: Support for decision</caption><thead><tr><th>Level of Transgression</th><th>Not Punishing (Red)</th><th>Punishing (Blue)</th></tr></thead><tbody><tr><td>1</td><td>2.5</td><td>6.0</td></tr><tr><td>2</td><td>2.0</td><td>6.0</td></tr><tr><td>3</td><td>2.5</td><td>6.0</td></tr><tr><td>4</td><td>2.0</td><td>6.0</td></tr><tr><td>5</td><td>2.5</td><td>6.0</td></tr><tr><td>6</td><td>2.5</td><td>5.5</td></tr><tr><td>7</td><td>2.5</td><td>5.5</td></tr><tr><td>8</td><td>2.5</td><td>5.5</td></tr><tr><td>9</td><td>2.5</td><td>5.5</td></tr><tr><td>10</td><td>2.0</td><td>5.5</td></tr></tbody></table>	Level of Transgression	Not Punishing (Red)	Punishing (Blue)	1	2.5	6.0	2	2.0	6.0	3	2.5	6.0	4	2.0	6.0	5	2.5	6.0	6	2.5	5.5	7	2.5	5.5	8	2.5	5.5	9	2.5	5.5	10	2.0	5.5	$B = 3.83$ $p < .001$	$B = -0.03$ $p = .413$
Level of Transgression	Not Punishing (Red)	Punishing (Blue)																																		
1	2.5	6.0																																		
2	2.0	6.0																																		
3	2.5	6.0																																		
4	2.0	6.0																																		
5	2.5	6.0																																		
6	2.5	5.5																																		
7	2.5	5.5																																		
8	2.5	5.5																																		
9	2.5	5.5																																		
10	2.0	5.5																																		

Table 2: Vignette 1 (protestors) - Prediction 2

	Graph	First level of transgression vs. Control	Levels 2-10 of transgression	Difference
How wrong was it for Country B to attack Greener Island?		$B = 1.42$ $p < .001$	$B = -0.00$ $p = .945$	$B = 1.43$ $p < .001$
If you had to choose, would you say that... [repeat] (yes or no)		$B = 1.28$ $p < .001$	$B = 0.02, p = .309$	$B = 1.27$ $p < .001$
Do you think that Country A should defend Greener Island?		$B = 1.06$ $p < .001$	$B = 0.04$ $p = .007$	$B = 1.03$ $p < .001$

Table 3: Vignette 2 (territorial incursion) - Prediction 1

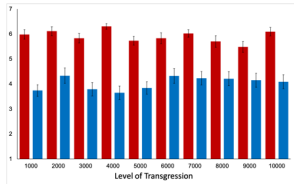
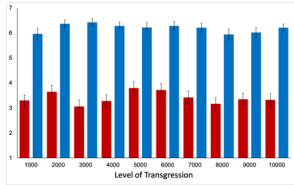
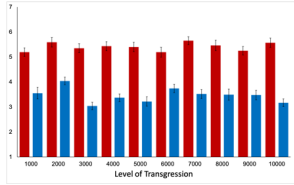
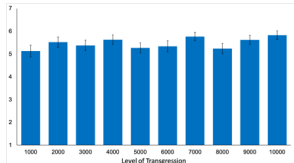
	Graph	Contrast between punishing versus not	Interaction with level of transgression
If Country A does not defend (defends) Greener Island, how likely is Country B to attack Blueiland [a more valuable territory] in the future?		$B = 2.24$ $p < .001$	$B = -0.07$ $p = .071$
If Country A does not defend (defends) Greener Island, do you think Country B would expect Country A to defend Blueiland, should Country B attack it in the future?		$B = -2.84$ $p < .001$	$B = 0.01$ $p = .750$
If Country A does not defend (defends) Greener Island, how likely is another country (besides Country B) to attack one of Country A's resource-rich territories in the future?		$B = 1.74$ $p < .001$	$B = 0.02$ $p = .281$
			Ten levels of transgression
Imagine that you are the head of a country allied with Country A. If Country A decides to defend Greener Island, would you lend military or financial support to the defense?			$B = 0.03$ $p = .088$

Table 4: Vignette 2 (territorial incursion) - Prediction 2

	Graph	First level of transgression vs. Control	Levels 2-10 of transgression	Difference
How morally wrong was Jenny's decision?		$B = 2.28$ $p < .001$	$B = 0.03$ $p = .024$	$B = 2.26$ $p < .001$
If you had to choose, would you say that would you say that... [repeat] (yes or no)		$B = 3.00$ $p < .001$	$B = -0.01$ $p = .356$	$B = 3.02$ $p < .001$
How many hours of community service would be a fair sentence for Jenny? (0-200 hours)		$B = 1.01$ $p < .001$	$B = 0.00$ $p = .912$	$B = 1.02$ $p < .001$

Table 5: Vignette 3 (shoplifting) - Prediction 1

	Graph	Contrast between punishing versus not	Interaction with level of transgression																																																															
If store employees decide not to turn Jenny into the police (to turn Jenny into the police), how likely is she to shoplift again in the future?	<table><caption>Data for Graph 1: Predicted Level of Transgression</caption><thead><tr><th>Level of Transgression</th><th>Punishing (Red)</th><th>Not Punishing (Blue)</th></tr></thead><tbody><tr><td>1</td><td>5.5</td><td>3.0</td></tr><tr><td>2</td><td>5.5</td><td>3.0</td></tr><tr><td>3</td><td>5.5</td><td>3.0</td></tr><tr><td>4</td><td>5.5</td><td>3.0</td></tr><tr><td>5</td><td>5.5</td><td>3.0</td></tr><tr><td>6</td><td>5.5</td><td>3.0</td></tr><tr><td>7</td><td>5.5</td><td>3.0</td></tr><tr><td>8</td><td>5.5</td><td>3.0</td></tr><tr><td>9</td><td>5.5</td><td>3.0</td></tr><tr><td>10</td><td>5.5</td><td>3.0</td></tr><tr><td>11</td><td>5.5</td><td>3.0</td></tr><tr><td>12</td><td>5.5</td><td>3.0</td></tr><tr><td>13</td><td>5.5</td><td>3.0</td></tr><tr><td>14</td><td>5.5</td><td>3.0</td></tr><tr><td>15</td><td>5.5</td><td>3.0</td></tr><tr><td>16</td><td>5.5</td><td>3.0</td></tr><tr><td>17</td><td>5.5</td><td>3.0</td></tr><tr><td>18</td><td>5.5</td><td>3.0</td></tr><tr><td>19</td><td>5.5</td><td>3.0</td></tr><tr><td>20</td><td>5.5</td><td>3.0</td></tr></tbody></table>	Level of Transgression	Punishing (Red)	Not Punishing (Blue)	1	5.5	3.0	2	5.5	3.0	3	5.5	3.0	4	5.5	3.0	5	5.5	3.0	6	5.5	3.0	7	5.5	3.0	8	5.5	3.0	9	5.5	3.0	10	5.5	3.0	11	5.5	3.0	12	5.5	3.0	13	5.5	3.0	14	5.5	3.0	15	5.5	3.0	16	5.5	3.0	17	5.5	3.0	18	5.5	3.0	19	5.5	3.0	20	5.5	3.0	$B = 2.56$ $p < .001$	$B = 0.00$ $p = .847$
Level of Transgression	Punishing (Red)	Not Punishing (Blue)																																																																
1	5.5	3.0																																																																
2	5.5	3.0																																																																
3	5.5	3.0																																																																
4	5.5	3.0																																																																
5	5.5	3.0																																																																
6	5.5	3.0																																																																
7	5.5	3.0																																																																
8	5.5	3.0																																																																
9	5.5	3.0																																																																
10	5.5	3.0																																																																
11	5.5	3.0																																																																
12	5.5	3.0																																																																
13	5.5	3.0																																																																
14	5.5	3.0																																																																
15	5.5	3.0																																																																
16	5.5	3.0																																																																
17	5.5	3.0																																																																
18	5.5	3.0																																																																
19	5.5	3.0																																																																
20	5.5	3.0																																																																
[Repeat]...would the drugstore's owner likely be happy with this response?	<table><caption>Data for Graph 2: Predicted Level of Transgression</caption><thead><tr><th>Level of Transgression</th><th>Punishing (Red)</th><th>Not Punishing (Blue)</th></tr></thead><tbody><tr><td>1</td><td>2.5</td><td>5.5</td></tr><tr><td>2</td><td>2.5</td><td>5.5</td></tr><tr><td>3</td><td>2.5</td><td>5.5</td></tr><tr><td>4</td><td>2.5</td><td>5.5</td></tr><tr><td>5</td><td>2.5</td><td>5.5</td></tr><tr><td>6</td><td>2.5</td><td>5.5</td></tr><tr><td>7</td><td>2.5</td><td>5.5</td></tr><tr><td>8</td><td>2.5</td><td>5.5</td></tr><tr><td>9</td><td>2.5</td><td>5.5</td></tr><tr><td>10</td><td>2.5</td><td>5.5</td></tr></tbody></table>	Level of Transgression	Punishing (Red)	Not Punishing (Blue)	1	2.5	5.5	2	2.5	5.5	3	2.5	5.5	4	2.5	5.5	5	2.5	5.5	6	2.5	5.5	7	2.5	5.5	8	2.5	5.5	9	2.5	5.5	10	2.5	5.5	$B = -3.07$ $p < .001$	$B = -0.04$ $p = .192$																														
Level of Transgression	Punishing (Red)	Not Punishing (Blue)																																																																
1	2.5	5.5																																																																
2	2.5	5.5																																																																
3	2.5	5.5																																																																
4	2.5	5.5																																																																
5	2.5	5.5																																																																
6	2.5	5.5																																																																
7	2.5	5.5																																																																
8	2.5	5.5																																																																
9	2.5	5.5																																																																
10	2.5	5.5																																																																
[Repeat]...would the drugstore's owner likely criticize this response?	<table><caption>Data for Graph 3: Predicted Level of Transgression</caption><thead><tr><th>Level of Transgression</th><th>Punishing (Red)</th><th>Not Punishing (Blue)</th></tr></thead><tbody><tr><td>1</td><td>5.5</td><td>1.5</td></tr><tr><td>2</td><td>5.5</td><td>1.5</td></tr><tr><td>3</td><td>5.5</td><td>1.5</td></tr><tr><td>4</td><td>5.5</td><td>1.5</td></tr><tr><td>5</td><td>5.5</td><td>1.5</td></tr><tr><td>6</td><td>5.5</td><td>1.5</td></tr><tr><td>7</td><td>5.5</td><td>1.5</td></tr><tr><td>8</td><td>5.5</td><td>1.5</td></tr><tr><td>9</td><td>5.5</td><td>1.5</td></tr><tr><td>10</td><td>5.5</td><td>1.5</td></tr></tbody></table>	Level of Transgression	Punishing (Red)	Not Punishing (Blue)	1	5.5	1.5	2	5.5	1.5	3	5.5	1.5	4	5.5	1.5	5	5.5	1.5	6	5.5	1.5	7	5.5	1.5	8	5.5	1.5	9	5.5	1.5	10	5.5	1.5	$B = 3.28$ $p < .001$	$B = -0.02$ $p = .659$																														
Level of Transgression	Punishing (Red)	Not Punishing (Blue)																																																																
1	5.5	1.5																																																																
2	5.5	1.5																																																																
3	5.5	1.5																																																																
4	5.5	1.5																																																																
5	5.5	1.5																																																																
6	5.5	1.5																																																																
7	5.5	1.5																																																																
8	5.5	1.5																																																																
9	5.5	1.5																																																																
10	5.5	1.5																																																																
[Repeat]...would the drugstore's owner likely praise this response?	<table><caption>Data for Graph 4: Predicted Level of Transgression</caption><thead><tr><th>Level of Transgression</th><th>Punishing (Red)</th><th>Not Punishing (Blue)</th></tr></thead><tbody><tr><td>1</td><td>2.0</td><td>6.0</td></tr><tr><td>2</td><td>2.5</td><td>6.0</td></tr><tr><td>3</td><td>2.5</td><td>6.0</td></tr><tr><td>4</td><td>2.5</td><td>6.0</td></tr><tr><td>5</td><td>2.5</td><td>6.0</td></tr><tr><td>6</td><td>2.5</td><td>6.0</td></tr><tr><td>7</td><td>2.5</td><td>6.0</td></tr><tr><td>8</td><td>3.0</td><td>6.0</td></tr><tr><td>9</td><td>2.5</td><td>6.0</td></tr><tr><td>10</td><td>2.5</td><td>6.0</td></tr></tbody></table>	Level of Transgression	Punishing (Red)	Not Punishing (Blue)	1	2.0	6.0	2	2.5	6.0	3	2.5	6.0	4	2.5	6.0	5	2.5	6.0	6	2.5	6.0	7	2.5	6.0	8	3.0	6.0	9	2.5	6.0	10	2.5	6.0	$B = -3.54$ $p < .001$	$B = 0.04$ $p = .273$																														
Level of Transgression	Punishing (Red)	Not Punishing (Blue)																																																																
1	2.0	6.0																																																																
2	2.5	6.0																																																																
3	2.5	6.0																																																																
4	2.5	6.0																																																																
5	2.5	6.0																																																																
6	2.5	6.0																																																																
7	2.5	6.0																																																																
8	3.0	6.0																																																																
9	2.5	6.0																																																																
10	2.5	6.0																																																																
[Repeat]...would the drugstore's owner likely fire employees for this response?	<table><caption>Data for Graph 5: Predicted Level of Transgression</caption><thead><tr><th>Level of Transgression</th><th>Punishing (Red)</th><th>Not Punishing (Blue)</th></tr></thead><tbody><tr><td>1</td><td>2.5</td><td>1.5</td></tr><tr><td>2</td><td>3.0</td><td>1.5</td></tr><tr><td>3</td><td>2.5</td><td>1.5</td></tr><tr><td>4</td><td>3.0</td><td>1.5</td></tr><tr><td>5</td><td>2.5</td><td>1.5</td></tr><tr><td>6</td><td>3.0</td><td>1.5</td></tr><tr><td>7</td><td>2.5</td><td>1.5</td></tr><tr><td>8</td><td>3.0</td><td>1.5</td></tr><tr><td>9</td><td>3.0</td><td>1.5</td></tr><tr><td>10</td><td>2.5</td><td>1.5</td></tr></tbody></table>	Level of Transgression	Punishing (Red)	Not Punishing (Blue)	1	2.5	1.5	2	3.0	1.5	3	2.5	1.5	4	3.0	1.5	5	2.5	1.5	6	3.0	1.5	7	2.5	1.5	8	3.0	1.5	9	3.0	1.5	10	2.5	1.5	$B = 1.93$ $p < .001$	$B = 0.01$ $p = .820$																														
Level of Transgression	Punishing (Red)	Not Punishing (Blue)																																																																
1	2.5	1.5																																																																
2	3.0	1.5																																																																
3	2.5	1.5																																																																
4	3.0	1.5																																																																
5	2.5	1.5																																																																
6	3.0	1.5																																																																
7	2.5	1.5																																																																
8	3.0	1.5																																																																
9	3.0	1.5																																																																
10	2.5	1.5																																																																

Table 6: Vignette 3 (shoplifting) - Prediction 2

	Graph	First level of transgression vs. Control	Levels 2-10 of transgression	Difference
How morally wrong was the control worker's decision?		$B = 2.03$ $p < .001$	$B = -0.01$ $p = .652$	$B = 2.04$ $p < .001$
If you had to choose, would you say that would you say that... [repeat] (yes or no)		$B = -2.05$ $p < .001$	$B = 0.00$ $p = .713$	$B = -2.06$ $p < .001$
What is a fair amount for the control worker to be required to pay the repairman's family in restitution (in dollars, from \$0 to \$1,000,000)?		$B = 1.08$ $p < .001$	$B = 0.01$ $p = .628$	$B = 1.07$ $p < .001$
What is a fair prison sentence for the control worker (in years, from 0 to 100)?		$B = 0.67$ $p < .001$	$B = 0.01$ $p = .484$	$B = 0.67$ $p < .001$

Table 7: Vignette 4 (trolley) - Prediction 1

	Graph	First level of transgression vs. Control	Levels 2-10 of transgression	Difference
How altruistic is Dan?		$B = 1.88$ $p < .001$	$B = 0.03$ $p = .046$	$B = 1.83$ $p < .001$
If you had to choose, would you say that Dan is altruistic?		$B = -2.39$ $p < .001$	$B = 0.00$ $p = .975$	$B = -2.40$ $p < .001$
Imagine that another person, Jennifer, approaches the booth after Dan has left. Jennifer learns about the same request to volunteer, and decides that she doesn't have time. How much more does Dan cares about the welfare of the homeless, compared to Jennifer?		$B = 1.53$ $p < .001$	$B = 0.03$ $p = .053$	$B = 1.50$ $p < .001$
If you could buy Dan a reward as a thank you for his service, like a gift certificate to a restaurant of his choosing, how much would you want to give him (in dollars)? (0-100)		$B = 1.36$ $p < .001$	$B = 0.03$ $p = .059$	$B = 1.32$ $p < .001$

Table 8: Vignette 5 (soup kitchen) - Prediction 1