Online Appendix

for

The Evoluationary Stability of Moral Foundations

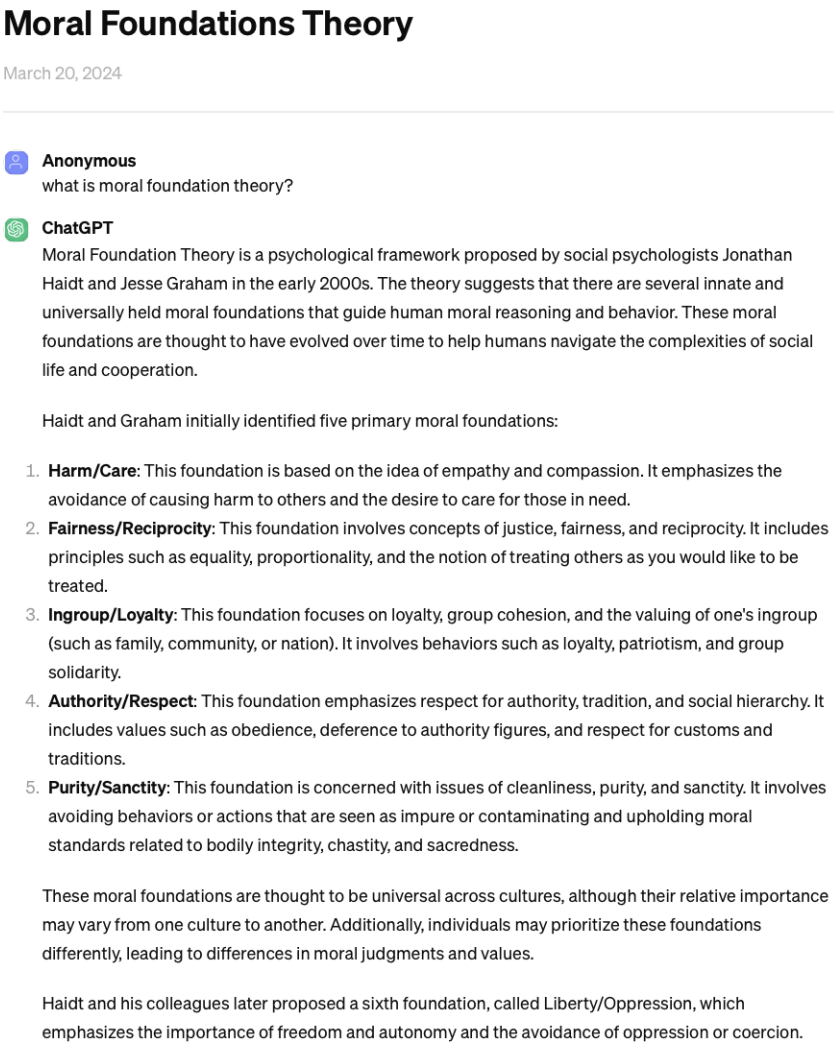
by

Michelle Avataneo

Thomas Norman Nicola Persico

ONLINE APPENDIX I:
REFERENCE MATERIALS FROM THE INTRODUCTION

FIGURE 1
Screenshot from ChatGPT in answer to the query "what is moral foundations theory?"



Moral Foundations Theory

March 20, 2024

Anonymous
what is moral foundation theory?

ChatGPT
Moral Foundation Theory is a psychological framework proposed by social psychologists Jonathan Haidt and Jesse Graham in the early 2000s. The theory suggests that there are several innate and universally held moral foundations that guide human moral reasoning and behavior. These moral foundations are thought to have evolved over time to help humans navigate the complexities of social life and cooperation.

Haidt and Graham initially identified five primary moral foundations:

1. **Harm/Care**: This foundation is based on the idea of empathy and compassion. It emphasizes the avoidance of causing harm to others and the desire to care for those in need.
2. **Fairness/Reciprocity**: This foundation involves concepts of justice, fairness, and reciprocity. It includes principles such as equality, proportionality, and the notion of treating others as you would like to be treated.
3. **Ingroup/Loyalty**: This foundation focuses on loyalty, group cohesion, and the valuing of one's ingroup (such as family, community, or nation). It involves behaviors such as loyalty, patriotism, and group solidarity.
4. **Authority/Respect**: This foundation emphasizes respect for authority, tradition, and social hierarchy. It includes values such as obedience, deference to authority figures, and respect for customs and traditions.
5. **Purity/Sanctity**: This foundation is concerned with issues of cleanliness, purity, and sanctity. It involves avoiding behaviors or actions that are seen as impure or contaminating and upholding moral standards related to bodily integrity, chastity, and sacredness.

These moral foundations are thought to be universal across cultures, although their relative importance may vary from one culture to another. Additionally, individuals may prioritize these foundations differently, leading to differences in moral judgments and values.

Haidt and his colleagues later proposed a sixth foundation, called Liberty/Oppression, which emphasizes the importance of freedom and autonomy and the avoidance of oppression or coercion.

Note: The query is available at the stable link
https://chat.openai.com/share/4f6b83ed-9aa9-4980-a823-16d39a7f638f.

FIGURE 2
Moral Foundations Questionnaire

Moral Foundations Questionnaire

Part 1. When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking? Please rate each statement using this scale:

[0] = not at all relevant (This consideration has nothing to do with my judgments of right and wrong)
[1] = not very relevant
[2] = slightly relevant
[3] = somewhat relevant
[4] = very relevant
[5] = extremely relevant (This is one of the most important factors when I judge right and wrong)

_____Whether or not someone suffered emotionally

_____Whether or not some people were treated differently than others

_____Whether or not someone's action showed love for his or her country

_____Whether or not someone showed a lack of respect for authority

_____Whether or not someone violated standards of purity and decency

_____Whether or not someone was good at math

_____Whether or not someone cared for someone weak or vulnerable

_____Whether or not someone acted unfairly

_____Whether or not someone did something to betray his or her group

_____Whether or not someone conformed to the traditions of society

_____Whether or not someone did something disgusting

_____Whether or not someone was cruel

_____Whether or not someone was denied his or her rights

_____Whether or not someone showed a lack of loyalty

_____Whether or not an action caused chaos or disorder

_____Whether or not someone acted in a way that God would approve of

[continues on the next page]

Part 2. Please read the following sentences and indicate your agreement or disagreement:

[0]	[1]	[2]	[3]	[4]	[5]
Strongly disagree	Moderately disagree	Slightly disagree	Slightly agree	Moderately agree	Strongly agree

_____Compassion for those who are suffering is the most crucial virtue.

_____When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.

_____I am proud of my country's history.

_____Respect for authority is something all children need to learn.

_____People should not do things that are disgusting, even if no one is harmed.

_____It is better to do good than to do bad.

_____One of the worst things a person could do is hurt a defenseless animal.

_____Justice is the most important requirement for a society.

_____People should be loyal to their family members, even when they have done something wrong.

_____Men and women each have different roles to play in society.

_____I would call some acts wrong on the grounds that they are unnatural.

_____It can never be right to kill a human being.

_____ I think it's morally wrong that rich children inherit a lot of money while poor children inherit nothing.

_____ It is more important to be a team player than to express oneself.

_____ If I were a soldier and disagreed with my commanding officer's orders, I would obey anyway because that is my duty.

_____ Chastity is an important and valuable virtue.

The Moral Foundations Questionnaire (full version, July 2008) by Jesse Graham, Jonathan Haidt, and Brian Nosek.
For more information about Moral Foundations Theory and scoring this form, see: www.MoralFoundations.org

[end of questionnaire]

ONLINE APPENDIX II:
CONSISTENCY WITH PRIOR OPERATIONALIZATIONS OF THE MORAL FOUNDATIONS

Each operationalization of the moral foundations we use in this paper has been previously analyzed in either the economics and/or psychology literature. This section explores the connection between our operationalization of the moral foundations and the existing literature.

In what follows we refer to "moral proclivity matrices" as defined in Definition B1.

Care Care is usually understood as selfless service on behalf of others. A preference for serving others for their own benefit is altruism. In economics, we usually say an agent is altruistic if her preferences reflect some concern for other players' welfare. For example, in Bester and Güth (1998), agent i is said to hold altruistic preferences towards agent j if her preference function is

$$P_i = \alpha \Pi_i + (1-\alpha)\Pi_j \,,$$

where Π_i is the fitness of agent i and Π_j is the fitness of agent j. Assuming there are only two possible pure actions (*cooperate* and *defect*), we get that the altruistic preference matrix is:

$$P = \alpha \begin{array}{|c|c|}\hline \text{a} & \text{b} \\ \hline \text{c} & \text{d} \\ \hline \end{array} + (1-\alpha) \begin{array}{|c|c|}\hline \text{a} & \text{c} \\ \hline \text{b} & \text{d} \\ \hline \end{array} \,,$$

which is strategically equivalent to

$$P = \begin{array}{|c|c|}\hline \text{a} & \text{b} \\ \hline \text{c} & \text{d} \\ \hline \end{array} + \frac{1-\alpha}{\alpha} \begin{array}{|c|c|}\hline \text{a} & \text{c} \\ \hline \text{b} & \text{d} \\ \hline \end{array} \,.$$

Hence, a person has altruistic preferences if $P = \Pi + M$, where

$$M = \frac{1-\alpha}{\alpha} \begin{array}{|c|c|}\hline \text{a} & \text{c} \\ \hline \text{b} & \text{d} \\ \hline \end{array} \,.$$

Next, we show that this M is a special case of our *Care* moral principle. Indeed:

1. If $a > b$ and $c > d$, the M moral proclivity matrix above is strategically equivalent to $\begin{array}{|c|c|}\hline \frac{1-\alpha}{\alpha}\text{(a-b)} & \frac{1-\alpha}{\alpha}\text{(c-d)} \\ \hline 0 & 0 \\ \hline \end{array}$, which has the form $\begin{array}{|c|c|}\hline \text{x} & \text{y} \\ \hline 0 & 0 \\ \hline \end{array}$ for suitable values of x and y. That moral proclivity matrix is the one generated by *Care* when $a > b$ and $c > d$.

2. If $a > b$ and $d > c$, the moral proclivity matrix M defined above is strategically equivalent to $\begin{array}{|c|c|}\hline \frac{1-\alpha}{\alpha}\text{(a-b)} & 0 \\ \hline 0 & \frac{1-\alpha}{\alpha}\text{(d-c)} \\ \hline \end{array}$, which is $\begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array}$ for a particular value of x and y. That moral proclivity matrix is the one generated by *Care* when $a > b$ and $d > c$.

3. If $b > a$ and $c > d$, the M moral proclivity matrix above is strategically equivalent to $\begin{array}{|c|c|}\hline 0 & \frac{1-\alpha}{\alpha}\text{(c-d)} \\ \hline \frac{1-\alpha}{\alpha}\text{(b-a)} & 0 \\ \hline \end{array}$, which is $\begin{array}{|c|c|}\hline 0 & y \\ \hline x & 0 \\ \hline \end{array}$ for a particular value of x and y. That moral proclivity matrix is the one generated by *Care* when $b > a$ and $c > d$.

4. If $b > a$ and $d > c$, the M moral proclivity matrix above is strategically equivalent to $\begin{array}{|c|c|}\hline 0 & 0 \\ \hline \frac{1-\alpha}{\alpha}\text{(b-a)} & \frac{1-\alpha}{\alpha}\text{(d-c)} \\ \hline \end{array}$, which is $\begin{array}{|c|c|}\hline 0 & 0 \\ \hline x & y \\ \hline \end{array}$ for a particular value of x and y. That moral proclivity matrix is the one generated by *Care* when $b > a$ and $d > c$.

Thus, when there are only two pure actions $\{cooperate, defect\}$ our operationalization of *Care* is a generalization of the altruism operationalization in Bester and Güth (1998).

Fairness There are several definitions of *Fairness*, one of them being a preference for equality. To our knowledge, in the economics literature, this preference was first operationalized by Fehr and Schmidt (1999), who say that agent i has "inequality aversion" when playing against agent j if

$$P_i = \Pi_i - \alpha_i max\{\Pi_j - \Pi_i, 0\} - \zeta_i max\{\Pi_i - \Pi_j, 0\}.$$

If $\alpha_i = \zeta_i$, the above preference can be rewritten as

$$P_i = \Pi_i - \alpha_i |\Pi_i - \Pi_j|.$$

Assuming there are only two possible pure actions (*cooperate* and *defect*), we get that the inequality averse preference matrix is:

$$P_i = \begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline \end{array} - \alpha_i \begin{array}{|c|c|}\hline |a-a| & |b-c| \\ \hline |c-b| & |d-d| \\ \hline \end{array},$$

which is strategically equivalent to

$$P_i = \begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline \end{array} + \begin{array}{|c|c|}\hline \alpha_i|c-b| & 0 \\ \hline 0 & \alpha_i|b-c| \\ \hline \end{array}$$

Hence, a person has inequality aversion preferences if $P = \Pi + M$, where

$$M = \begin{array}{|c|c|}\hline \alpha_i|c-b| & 0 \\ \hline 0 & \alpha_i|b-c| \\ \hline \end{array}.$$

Note that this M is a special case of $\begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array}$, which is exactly the moral proclivity matrix generated by our *Fairness* moral principle.

Authority Arguably, a natural definition of Authority, and the one Haidt (2012) favors, is respect for the social hierarchy. However, our model is not rich enough to capture a social hierarchy. Instead, the operationalization we use captures only the power aspect of a hierarchical relationship: it captures the drive humans have to be better than others in relative terms. This drive has been discussed for many years in philosophy, psychology, economics and finance. It is colloquially refered to as "the Veblen effect" or the "Staying ahead of the Joneses effect." The same operationalization of the preference for doing better than others appears in Messick and Sentis (1985) and in Ordóñez, Connolly, and Coughlan (2000). It is the following:

$$P_i = f_i(\Pi_i) + g_i(\Pi_i - \Pi_j)$$

If we take $f_i(x) = x$ and $g_i(x) = \alpha_i x$, the above preference can be rewritten as

$$P_i = \Pi_i + \alpha_i(\Pi_i - \Pi_j)$$

Assuming there are only two possible pure actions (*cooperate* and *defect*), we get that the preference above can be written as:

$$P_i = \begin{array}{|c|c|}\hline \text{a} & \text{b} \\ \hline \text{c} & \text{d} \\ \hline \end{array} + \alpha_i \begin{array}{|c|c|}\hline 0 & \text{b-c} \\ \hline \text{c-b} & 0 \\ \hline \end{array}$$

Hence, a person has a preference for doing better than others if $P = \Pi + M$, where

$$M = \begin{array}{|c|c|}\hline 0 & \alpha_i(b-c) \\ \hline \alpha_i(c-b) & 0 \\ \hline \end{array}.$$

Note that M is a special case of our *Authority* moral principle. Indeed:

1. If $b > c$, the moral proclivity matrix M defined above is strategically equivalent to $\begin{array}{|c|c|}\hline \alpha_i\text{(b-c)} & \alpha_i\text{(b-c)} \\ \hline 0 & 0 \\ \hline \end{array}$, which is $\begin{array}{|c|c|}\hline \text{x} & \text{y} \\ \hline 0 & 0 \\ \hline \end{array}$ for a particular value of x and y. That moral proclivity matrix is the one generated by *Authority* when $b > c$.

2. If $c > b$, the moral proclivity matrix M defined above is strategically equivalent to $\begin{array}{|c|c|}\hline 0 & 0 \\ \hline \alpha_i\text{(c-b)} & \alpha_i\text{(c-b)} \\ \hline \end{array}$, which is $\begin{array}{|c|c|}\hline 0 & 0 \\ \hline \text{x} & \text{y} \\ \hline \end{array}$ for a particular value of x and y. That moral proclivity matrix is the one generated by *Authority* when $c > b$.

Loyalty A natural definition of loyalty is behaving kindly towards those who one considers part of an "in-group", and spitefully towards those who one considers part of an "out-group". Our theory does not feature a notion of in- and out-groups, but if we interpret the out-group as "someone who does not choose the Kantian action," then our definition of *Loyalty* is close to a preference for reciprocity as defined by Segal and Sobel (2007). They extend preferences over outcomes to include preference over strategies. In their formulation, an agent has reciprocity preferences if her preferences are of the form:

$$P_i = \Pi_i + a_i^j(\sigma_i, \sigma_j)\Pi_j$$

where $a_i^j(\sigma_i, \sigma_j)$ is the weight agent i places on the fitness of agent j. The weight can be positive or negative and it depends on the strategies. In particular, we could have $a_i^j(\sigma_i, \sigma_j) > 0$ if σ_j is the Kantian action and $a_i^j(\sigma_i, \sigma_j) < 0$ if σ_j is not the Kantian action, in which case Segal and Sobel (2007)'s reciprocity preferences would be the same as the preferences generated by our *Loyalty* moral principle. Assuming there are only two possible pure actions (*cooperate* and *defect*), and $a_i^j(\sigma_i, \sigma_j) = \rho > 0$ when σ_j is the Kantian action, and $a_i^j(\sigma_i, \sigma_j) = -\delta < 0$ when σ_j is not the Kantian action, we get that the reciprocity preference matrix is:

$$P_i = \begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline \end{array} + \begin{array}{|c|c|}\hline \rho\,a & -\delta\,c \\ \hline \rho\,b & -\,\delta\,d \\ \hline \end{array}$$

Hence, an agent has preferences for reciprocity if her preference is $P = \Pi + M$, where

$$M = \begin{array}{|c|c|}\hline \rho\,a & -\delta\,c \\ \hline \rho\,b & -\,\delta\,d \\ \hline \end{array}.$$

Note that this M above is a special case of our *Loyalty* moral principle, because:

1. If $a > b$ and $c > d$, the moral proclivity matrix M defined above is strategically equivalent to $\begin{array}{|c|c|}\hline \rho(a\text{-}b) & 0 \\ \hline 0 & \delta(c\text{-}d) \\ \hline \end{array}$, which is $\begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array}$ for a particular value of x and y. This moral proclivity matrix is the one generated by *Loyalty* when $a > b$ and $c > d$.

2. If $a > b$ and $d > c$, the moral proclivity matrix M defined above is strategically equivalent to $\begin{array}{|c|c|}\hline \rho(a\text{-}b) & \delta(d\text{-}c) \\ \hline 0 & 0 \\ \hline \end{array}$, which is $\begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array}$ for a particular value of x and y. That moral proclivity matrix is the one generated by *Loyalty* when $a > b$ and $d > c$.

3. If $b > a$ and $c > d$, the moral proclivity matrix M defined above is strategically equivalent to $\begin{array}{|c|c|}\hline 0 & 0 \\ \hline \rho(b\text{-}a) & \delta(c\text{-}d) \\ \hline \end{array}$, which is $\begin{array}{|c|c|}\hline 0 & 0 \\ \hline x & y \\ \hline \end{array}$ for a particular value of x and y. That moral proclivity matrix is the one generated by *Loyalty* when $b > a$ and $c > d$.

4. If $b > a$ and $d > c$, the moral proclivity matrix M defined above is strategically equivalent to $\begin{array}{|c|c|}\hline 0 & \delta(\text{c-d}) \\ \hline \rho(\text{b-a}) & 0 \\ \hline \end{array}$, which is $\begin{array}{|c|c|}\hline 0 & \text{y} \\ \hline \text{x} & 0 \\ \hline \end{array}$ for a particular value of x and y. That moral proclivity matrix is the one generated by *Loyalty* when $b > a$ and $d > c$.

Although, not exactly the same, the same flavor of preferences can also be found in Charness and Rabin (2002). In that paper, if the opponent misbehaves, the agent is spiteful, and if the opponent behaves well, the agent is altruistic. However, the opponent is deemed to have behaved well or badly based on outcomes – not strategies.

Sanctity The term *Sanctity* is commonly used in reference to organized religion. However, in the ethnographic literature, the term is also used in reference to notions of purity of behavior such as not defiling the environment, or not eating certain foods, or performing sacrifices. The ethnographic literature is split over whether the actions that are deemed pure (either religiously or culturally) are in fact pro-social or, rather, they emerge arbitrarily out of cultural accident. In our operationalization of *Sanctity*, we adopt the view that *Sanctity* is pro-social.

Our operationalization of *Sanctity* is the same as Alger and Weibull (2013)'s operationalization of "homo moralis". According to Alger and Weibull (2013), an agent is "homo moralis" if her preference is:

$$P_i = \Pi_i(\sigma_i, \sigma_j) + \Theta\Pi_i(\sigma_i, \sigma_i)$$

where σ_i is the action chosen by agent i, and σ_j is the action chosen by agent j.

Assuming there are only two possible pure actions (*cooperate* and *defect*), we get that the "homo moralis" preference matrix is:

$$P_i = \begin{array}{|c|c|}\hline \text{a} & \text{b} \\ \hline \text{c} & \text{d} \\ \hline \end{array} + \Theta \begin{array}{|c|c|}\hline \text{a} & \text{a} \\ \hline \text{d} & \text{d} \\ \hline \end{array}$$

Hence, an agent is homo moralis if her preference is $P = \Pi + M$, where

$$M = \Theta \begin{array}{|c|c|}\hline \text{a} & \text{a} \\ \hline \text{d} & \text{d} \\ \hline \end{array}$$

Note, that when $a > d$, the above moral proclivity matrix M is a special case of $\begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array}$, which is exactly the moral proclivity matrix generated by our *Sanctity* moral principle.

ONLINE APPENDIX III:
PROOFS FOR SECTION IV.A

Exactly as in Dekel, Ely, and Yilankaya (2007), suppose a player observes her opponent's preferences with probability p and, with complementary probability $1-p$, she observes an uninformative signal $\emptyset$. A strategy for an agent with preference P is a rule $\beta_P : supp(\mu) \cup \emptyset \to \Delta$ which specifies a mixed action conditional on what the agent observes. It is assumed that each pair of matched agents plays a Bayes-Nash equilibrium of the game induced by their preferences.

Online Appendix III.A: Scenario A

Each player observes the opponent's preference matrix with independent probability p sufficiently large. A player who fails to observe the opponent's preference matrix best-responds to an expectation based on the population distribution μ. Players do not know whether their own preference matrix is observed by their opponent. These assumptions correspond to the assumptions made in Section 5 of Dekel, Ely, and Yilankaya (2007). The definition of best response, Bayes-Nash equilibrium, and average payoffs extend straightforwardly to this incomplete information setting. However, the definition of stability must be updated to account for an uninformed agent's inability to detect whether she is matched with a mutant. This inability can destabilize those stable preferences that rely on the players' indifference between actions. Substantively, this means that fewer preferences are stable.

A Bayes-Nash equilibrium in this scenario must satisfy:

- If P observes her opponent's preference, she must choose an optimal action conditional on that observation:

$$\beta_P(P') \in \arg\max_{\sigma\in\Delta} \; p\mathcal{P}(\sigma, \beta_{P'}(P)) + (1-p)\mathcal{P}(\sigma, \beta_{P'}(\emptyset))$$

  for each $P' \in supp(\mu)$, and

- If P does not observe her opponent's type, she must best respond to the expectation, that is

$$\beta_P(\emptyset) \in \arg\max_{\sigma\in\Delta} \sum_{P'\in supp(\mu)} [p\mathcal{P}(\sigma, \beta_{P'}(P)) + (1-p)\mathcal{P}(\sigma, \beta_{P'}(\emptyset))] \cdot \mu(P')$$

Note that, unlike the case with perfect observability, the equilibrium depends on μ. Given a population distribution μ and an equilibrium profile $\beta \in B(\mu)$, where $B(\mu)$ is the set of all Bayes-Nash equilibrium profiles when the population distribution is μ, the average fitness of a player with preference $P \in supp(\mu)$ is now:

$$\bar{\Pi}_P(\mu \mid b) = \sum_{P' \in \text{supp}(\mu)} [p^2 \mathscr{T}(\beta_P(P'), \beta_{P'}(P)) + p(1-p)\mathscr{T}(\beta_P(P'), \beta_{P'}(\emptyset)) + (1-p)p\mathscr{T}(\beta_P(\emptyset), \beta_{P'}(P)) + (1-p)^2\mathscr{T}(\beta_P(\emptyset), \beta_{P'}(\emptyset))] \cdot \mu(P').$$

For any potentially stable configuration (μ, β), we say that, given any $\tilde{\mu} \in N_\varepsilon(\mu, \tilde{P})$, an equilibrium strategy profile $\tilde{\beta}(\tilde{\mu}) \in B(\tilde{\mu})$ is *focal to* $\beta(\mu)$ if $\tilde{\beta}_P(P') = \beta_P(P')$ *and* $\tilde{\beta}_P(\emptyset) = \beta_P(\emptyset)$ for all $P, P' \in \text{supp}(\mu)$. The second requirement is new. Due to this additional requirement, now a focal equilibrium might not exist, in which case we say the configuration (μ, β) is not stable. Hence, now a configuration (μ, β) is said to be stable if:

1. it is balanced, and
2. given any ϵ sufficiently small, all $\tilde{P}$, and all $\tilde{\mu} \in N_\epsilon(\mu, \tilde{P})$,
   (a) the set of focal equilibria is nonempty, and
   (b) for all $P \in supp(\mu)$ and all focal equilibria $\tilde{\beta}$, $\bar{\Pi}_P(\tilde{\mu}|\tilde{\beta}) \geq \bar{\Pi}_{\tilde{P}}(\tilde{\mu}|\tilde{\beta})$.

In Scenario A, almost all our results remain the same. Only Theorems 4 and 5 need to be modified. This is formalized in the following proposition.

PROPOSITION C1. When $p < 1$ in Scenario A, the following is true.

a) Theorem 1 continues to hold for p sufficiently close to 1.

b) For any Π, there exists a $\overline{p}$ such that, for $p \in (\overline{p}, 1)$, Theorems 2 and 3 continue to hold.

c) Theorem 4 is amended as follows. For any Π, there is a $\overline{p}$ such that, for $p \in (\overline{p}, 1)$, any moral principle in the $H5$ generates a stable preference, if one exists. In this sense, any single moral principle in the $H5$ represents a minimal moral code.

d) Theorem 5 holds vacuously because, for p large but less than 1, the Prisoner's Dilemma and Hawk-Dove games lack stable preferences, so by Definition 12 no m

is compatible with moral overdrive in these game types; and in the rest of the game types, no strict social improvement is available.

The intuition for part c) above is that introducing incomplete information specifically as done in Scenario A causes the Prisoner's Dilemma and Hawk-Dove games to no longer have a stable preference. Meanwhile, the set of generic stable preferences in all other game types is unchanged. Hence, it is easier for a single element of the $H5$ to be stable in all games in which a stable moral principle exists.

To prove Proposition C1, we must first identify all the preferences P that are stable for a given Π in Scenario A. We do this in the next few pages and then summarize our findings in Table 1 below.

When p is large but smaller than 1 in Scenario A, the set of stable preferences for Π is a subset of the preferences that are stable under perfect observability ($p = 1$). To see this, observe first that, P not being stable means that for all configurations $(\mu, \beta) \in \Theta(P) \times B(\mu)$, where $\Theta(P)$ is the set of all probability distributions with finite support which have P in their support, and $B(\mu)$ is the set of all Bayes Nash profiles for distribution μ, $\bar{\Pi}_P(\tilde{\mu}|\tilde{\beta}) < \bar{\Pi}_{P'}(\tilde{\mu}|\tilde{\beta})$ for some mutant preference $P' \in \mathbb{M}$ and some focal equilibrium $\tilde{\beta}$. From continuity of the real numbers, it follows that if $\bar{\Pi}_P(\tilde{\mu}|\tilde{\beta}) < \bar{\Pi}_{P'}(\tilde{\mu}|\tilde{\beta})$ for $p = 1$, there exists a $\rho \in (0, 1)$ such that for all $p \in (\rho, 1]$, $\bar{\Pi}_P(\tilde{\mu}|\tilde{\beta}) < \bar{\Pi}_{P'}(\tilde{\mu}|\tilde{\beta})$. This implies that for any Π, there exists a $\rho(\Pi)$ such that for any $p \in (\rho(\Pi), 1)$ the set of evolutionarily stable preferences is a subset of Π's evolutionarily stable preferences when $p = 1$. Hence, when p is large, to check whether a preference is stable or not, we can restrict attention to the preference equivalence classes that are stable under perfect observability.

In what follows we characterize, for every Π, the set of evolutionarily stable preferences for any $p \in (\bar{p}(\Pi), 1)$ where $\bar{p}(\Pi) = max\{\rho(\Pi), \frac{2d-b-c}{a+d-b-c}\}$.

Case 1: Πs with $a \geq b, c, d$

From Dekel, Ely, and Yilankaya (2007), we know that when $p = 1$, any preferences P which is strategically equivalent to an element of the set $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0, 1], \mathcal{BA}_1, \theta_0\}$ is stable. For $p \in (\rho(\Pi), 1)$ we know that anything that is not stable under perfect observability is not stable, so we must only check stability of preferences which are strategically equivalent to one element of $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0, 1], \mathcal{BA}_1, \theta_0\}$:

- Any preference which has as a weakly dominant strategy playing *cooperate*, i.e. anything strategically equivalent to an element of $\{\mathcal{AA}, \mathcal{AB}_0, \mathcal{BA}_1$ and $\theta_0\}$ is stable. For any such preferences, playing *cooperate* regardless of who the opponent is or what

is observed is a best response. Take any configuration (μ, β) where $supp(\mu)$ is any subset of $\{\mathcal{AA}, \mathcal{AB}_0, \mathcal{BA}_1$ and $\theta_0\}$, in which all incumbents play *cooperate* against themselves regardless of what is observed, i.e. $\beta_P(P'') = \beta_P(\emptyset) =$ *cooperate*, for all $P, P'' \in supp(\mu)$. Let P' be any mutant. Since it continues being a best response for incumbents to play *cooperate* against each other when they observe each other, and to play *cooperate* when they don't observe anything, focal equilibria exist. Take any $P \in \{\mathcal{AA}, \mathcal{AB}_0, \mathcal{BA}_1$ and $\theta_0\}$. In any focal equilibrium, if $\beta_{P'}(P) \neq$ *cooperate* for any $P \in \{\mathcal{AA}, \mathcal{AB}_0, \mathcal{BA}_1$ and $\theta_0\}$ or $\beta_{P'}(\emptyset) \neq$ *cooperate*, $a > p^2\mathcal{T}(\beta_{P'}(P), A) + p(1-p)\mathcal{T}(\beta_{P'}(P), A) + (1-p)p\mathcal{T}(\beta_{P'}(\emptyset), A) + (1-p)^*\mathcal{T}(\beta_{P'}(\emptyset), A)$, so there exists an $\bar{\epsilon} \in (0, 1)$ such that for all $\epsilon \in (0, \bar{\epsilon})$, all incumbents get a strictly larger average fitness. If $\beta_{P'}(P) =$ *cooperate* and $\beta_{P'}(\emptyset) =$ *cooperate*, incumbents and P' get the same average fitness of a.

- Any preference P which is strategically equivalent to $\mathcal{AB}_\alpha$ with $\alpha \in (0, 1)$ is stable. Take a monomorphic population of P which plays *cooperate* regardless of what is observed. Take any mutant P'. As long as the proportion in which the mutant enters the population is small enough, $\epsilon \leq 1 - \alpha$, it continues being a best response for P to play *cooperate* when it observes it is matched against another P and when it does not observe anything, so focal equilibria exist. Moreover, in all focal equilibrium against any mutant, the average fitness of P is larger because, as with preferences which are strategically equivalent to an element of $\{\mathcal{AA}, \mathcal{AB}_0, \mathcal{BA}_1$ and $\theta_0\}$, P gets the largest possible fitness, a, when matched against its own type.

- Preferences P which are strategically equivalent to $\mathcal{AB}_1$ are not stable for $\frac{2d-b-c}{a+d-b-c} < p < 1$. Take any configuration (μ, β) which has some preference P strategically equivalent to $\mathcal{AB}_1$ in the support.

  (a) Suppose everybody in the support plays *cooperate* regardless of what is observed, that is $\beta_{P'}(P'') = \beta_{P'}(\emptyset) = A$ for all $P', P'' \in supp(\mu)$. Such Bayes Nash equilibrium profile does not have a focal equilibrium profile against any mutant preference which plays *defect*. This is because whenever there is any probability larger than 0 that an opponent might play *defect*, playing *cooperate* stops being a best response for P. There is a probability $\epsilon > 0$ that the opponent might play *defect*, hence in any post entry equilibrium profile $\tilde{\beta}$, $\tilde{\beta}_P(\emptyset) =$ *defect*.

  (b) Hence, any Bayes Nash equilibrium profile which has a focal equilibrium must be such that $\beta_P(\emptyset) =$ *defect*. But $\beta_P(\emptyset) =$ *defect* also means $\beta_P(P) =$ *defect*,

as there is a probability larger than 0 that an agent with preference P is matched against another agent with preference P who doesn't observe anything.

(c) But such Bayes Nash equilibrium profiles are not stable for $p > \frac{2d-b-c}{a+d-b-c}$. A mutant who i) plays against any incumbent other than P, whenever he observes who he is matched against, the same as P plays against them, ii) who plays *defect* when he does not observe anything or when he observes he is matched against P, and iii) who plays *cooperate* when he observes he is matched against his own type would get strictly larger average payoffs than P, for any $\epsilon > 0$. This is because P and the mutant get the same payoff when matched against anyone except a mutant. Against a mutant, another mutant gets expected fitness $p^2a + p(1-p)(b+c) + (1-p)^2d$ and P gets expected fitness d, the former is larger than the latter if

* $p > \frac{2d-b-c}{a+d-b-c}$ when $a+d-b-c > 0$, and
* $p < \frac{b+c-2d}{b+c-a-d}$ when $a+d-b-c < 0$

But note that when $a+d-b-c < 0$, $\frac{b+c-2d}{b+c-a-d} > 1$. So, when $a+d-b-c > 0$, $\mathcal{AB}_1$ is not evolutionarily stable if $p > \frac{2d-b-c}{a+d-b-c}$ and when $a+d-b-c < 0$, $\mathcal{AB}_1$ is not evolutionarily stable for any $p > 0$.

Hence, for any Π with $a \geq b, c, d$, the set of stable preferences when $p \in (\bar{p}(\Pi), 1)$ is $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0,1), \mathcal{BA}_1, \theta_0\}$.

Case 2: Πs with $b \geq a \geq c$ and $2a \geq b + c$

From Dekel, Ely, and Yilankaya (2007), we know that when $p = 1$, any preferences P which is strategically equivalent to an element of the set $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0, \frac{a-d}{b-d}]\}$. Any preference that is not stable for $p = 1$, is also not stable for $p \in (\rho(\Pi), 1)$. Hence, we only need to check stability for preferences which are strategically equivalent to an element of $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0, \frac{a-d}{b-d}]\}$.

- Any preference P which is strategically equivalent to $\mathcal{AA}$ is stable. Take a monomorphic population of $\mathcal{AA}$ which plays *cooperate* regardless of what is observed. Take any mutant P'. Since, playing A no matter what continues being a best response, even if a mutant which plays *defect* enters, focal equilibria exist. Moreover, in all such focal equilibria, the incumbent gets a larger average fitness because $p\mathcal{T}(\beta_{P'}(P), A) + (1-p)\mathcal{T}(\beta_{P'}(\emptyset), A) < a$ unless P' plays *cooperate* when she observes she is matched against $\mathcal{AA}$ and when she does not observe anything. Hence,

there exists a $\bar{\epsilon} \in (0,1)$, such that for all $\epsilon \in (0, \bar{\epsilon})$, $\mathcal{AA}$ gets a larger average fitness.

- Any preference P which strategically equivalent to $\mathcal{AB}_\alpha$ with $\alpha \in [0, \frac{a-d}{b-d})$ is stable. Take a monomorphic population of P which plays *cooperate* regardless of what is observed. As long as the proportion in which the mutant enters the population is small enough, $\epsilon \leq 1-\alpha$, it continues being a best response for P to play *cooperate* against other P and when nothing is observed, so focal equilibria exist. Moreover, in all focal equilibria against any mutant P', the average fitness of P is larger. This is because the largest possible fitness payoff is b, and the only equilibrium in which P' could get b with some probability is the focal equilibrium in which P plays *defect* when she observes she is matched against P' and P' plays *cooperate* with probability α, regardless of what she observes. In that equilibrium, when matched against P the mutant gets $p(\alpha b + (1-\alpha)d) + (1-p)(\alpha a + (1-\alpha)c)$, which is strictly less than a for any $p \in [0,1]$ and any $\alpha \in [0, \frac{a-d}{b-d})$.

Hence, for any Πs with $b \geq a \geq c$ and $2a \geq b+c$, the set of stable preferences when $p \in (\bar{p}(\Pi), 1)$ is $\{\mathcal{AA}, \mathcal{AB}_\alpha \text{ with } \alpha \in [0, \frac{a-d}{b-d})\}$.

Case 3: Πs with $c \geq a$ and $2a \geq b+c$

By the exact same reasoning as for game types Π with $a \geq b, c, d$, for any Π with $c \geq a$ and $2a \geq b+c$, when $p \in (\bar{p}(\Pi), 1)$, preferences P which are strategically equivalent to $\mathcal{AB}_1$ are not stable. However, for this type of games Π, contrary to game types with $a \geq b, c, d$, the only preferences which are stable are strategically equivalent to $\mathcal{AB}_1$. Hence, for this type of games, no stable preferences exist.

Case 4: all other Πs

Any other Π is either non-generic, or it does not have stable preferences when $p = 1$.

The above findings are summarized in Table 1 below.

TABLE 1
EVOLUTIONARILY STABLE PREFERENCES IN SCENARIO A

Π	Evolutionarily stable preferences
Case 1: $a \geq b, c, d$	$\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0,1), \mathcal{BA}_1, \theta_0\}$
Case 2: $b \geq a \geq c$ and $2a \geq b + c$	$\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0, \frac{a-d}{b-d})\}$
Case 3: $c \geq a$ and $2a \geq b + c$	None exists
Case 4: Any other generic Π	None exists

Having characterized all the stable preferences for generic Π's in Scenario A, we now turn to the proof of Proposition C1.

Proof of Proposition C1

Proof. **Part a)** For the proof of Theorem 1 with $p = 1$, which can be seen in Appendix II, only games Π with $a \geq b, c, d$ and only preferences P which are strategically equivalent to $\mathcal{AA}$ or $\mathcal{AB}_\alpha$ with $\alpha \in [\frac{d-b}{d-b+a-c}], 1)$ are used. As can be seen in Table 1, for any Π with $a \geq b, c, d$ (Case 1), there exists a $\bar{p} < 1$ such that these preferences continue being stable. Hence, there exists a $\bar{p}(\Pi)$ such that the proof of Theorem 1 for $p = 1$ continues to go through for all $p \in (\bar{p}(\Pi), 1)$.

Part b) For any Π and any $p \in (\bar{p}(\Pi), 1)$, Theorem 2 continues to hold because:

- As can be seen in Table 1, for any Π in Case 1, the set of stable preferences for $p \in (\bar{p}(\Pi), 1)$ is $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0,1), \mathcal{BA}_1, \theta_0\}$, and as can be seen in Stability Tables 1 and 2 the only preferences in that set that cannot be spanned for Πs with $a \geq b, c, d$ by any moral principle in the $H5$ are those which are strategically equivalent to $\{\mathcal{BA}_1, \theta_0\}$. However $\{\mathcal{BA}_1, \theta_0\}$ are non-generic in the space of stable preferences: the lowest-dimensional vector space containing $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0,1), \mathcal{BA}_1, \theta_0\}$ has dimension 4, as one needs 4 parameters to describe some preferences, but the lowest-dimensional vector space containing $\{\mathcal{BA}_1, \theta_0\}$ had dimension 3, as one needs at most 3 parameters to describe all preferences in that set.

- As can be seen in Table 1, for any Π in Case 2, the set of stable preferences for $p \in (\bar{p}(\Pi), 1)$ is $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0, \frac{a-d}{b-d})\}$. From Stability Table 3, we know that the moral principles in the $H5$ can span all such preferences.

- As can be seen in Table 1, for any other generic Π (Case 3 and Case 4), no stable preference exists.

For any Π and any $p \in (\bar{p}(\Pi), 1)$, Theorem 3 continues to hold. Indeed, for the sufficiency part of the proof when $p \in (\bar{p}(\Pi), 1)$, everything is the same as in the proof of Theorem 3 with $p = 1$, which can be seen in Appendix II, except for the game types Π in Case 3. For those Π, when $p \in (\bar{p}(\Pi), 1)$, no stable preference exists (as can be seen in Table 1). So it is still the case that $\{Care, Loyalty\}$ or $\{Fairness, Sanctity\}$ span all generic evolutionarily stable preferences for all Π. For the necessity part, the proof of Theorem 3 with $p = 1$ only uses fitness games Π in Case 1 and preferences $\mathcal{AA}$ and $\mathcal{AB}_\alpha$ with $\alpha \in (0, 1)$. As can be seen in Table 1, when $p \in (\bar{p}(\Pi), 1)$, $\mathcal{AA}$ and $\mathcal{AB}_\alpha$ with $\alpha \in (0, 1)$ continues being stable for any Π in Case 1. Thus, the proof when $p = 1$ continues to go through when $p \in (\bar{p}(\Pi), 1)$.

Part c) Theorem 4 changes for $p \in (\bar{p}(\Pi), 1)$. As can be seen in Table 1, for the game types Π in Case 3, no stable preference exists for $p \in (\bar{p}(\Pi), 1)$. Hence, the only fitness games that admit stable preferences for $p \in (\bar{p}(\Pi), 1)$ are those in Case 1 and Case 2. As can be seen in Stability Tables 1, 2 and 3, every moral principle in the $H5$ is capable of spanning either $\mathcal{AA}$ or $\mathcal{AB}_\alpha$ for some $\alpha \in (0, 1)$. Thus, any single moral principle in the $H5$ represents a minimal moral code.

Part d) Theorem 5 holds vacuously for $p \in (\bar{p}(\Pi), 1)$ because the only game types that admit a strict social fitness improvement are those in Case 3, which (as seen in Table 1) do not admit evolutionarily stable preferences for $p \in (\bar{p}(\Pi), 1)$. □

Online Appendix III.B: Scenario B

With probability p close to 1, the two players' moral proclivity matrices are common knowledge exactly as described in Section II. With complementary probability $(1 - p)$ neither player observes the other's moral proclivity matrix, and both players best-respond to an expectation based on the population distribution μ. This scenario is *not* studied in Dekel, Ely, and Yilankaya (2007), but it it worth studying here because it suggests that Theorems 4 and 5 are somewhat more robust than suggested by Proposition C1.

A Bayes Nash equilibrium in this environment must satisfy the following equations:

- If P observes her opponent's preference, she knows her opponent also observes her

preference. Hence, she must choose an optimal action conditional on that information:

$$\beta_P(P') \in \arg\max_{\sigma\in\Delta} \mathcal{P}(\sigma, \beta_{P'}(P))$$

for each $P' \in supp(\mu)$.

- If P does not observe her opponent's type, she knows her opponent doesn't observe her type either. So, she must best respond to the expectation, that is

$$\beta_P(\emptyset) \in \arg\max_{\sigma\in\Delta} \sum_{P'\in supp(\mu)} \mathcal{P}(\sigma, \beta_{P'}(\emptyset)) \cdot \mu(P')$$

Given a population distribution μ and an equilibrium profile $\beta \in B(\mu)$, the average fitness of a player with preference $P \in supp(\mu)$ is:

$$\bar{\Pi}_P(\mu \mid b) = \sum_{P'\in \text{supp}(\mu)} \left[p\mathcal{T}(\beta_P(P'), \beta_{P'}(P)) + (1-p)\mathcal{T}(\beta_P(\emptyset), \beta_{P'}(\emptyset))\right] \cdot \mu(P')$$

The definition of stability is the same as in Scenario A.

As in Scenario A, only Theorems 4 and 5 need to be modified, but in a different way.

PROPOSITION C2. When $p < 1$ in Scenario B, the following is true.

a) Theorem 1 continues to hold for p sufficiently close to 1.

b) For any Π, there exists a $\overline{p}$ such that, for $p \in (\overline{p}, 1)$, Theorems 2 and 3 continue to hold.

c) Theorem 4 is amended as follows. For any Π, there is a $\overline{p}$ such that, for $p \in (\overline{p}, 1)$, any moral principle in the $H5$ except *Authority* generates a stable preference, if one exists. In this sense, every single moral principle in theH5 except *Authority* represents a minimal moral code.

d) Theorem 5 continues to hold, but not vacuously because the Prisoner's Dilemma does have stable preferences for p close enough to 1.

The intuition for part c) above is that, in scenario B, some Prisoners' Dilemma games continue having evolutionarily stable preferences, and all moral principles in the $H5$ except

Authority can generate stable moral principles. The intuition for part d) above is that, in scenario B, some Prisoners' Dilemma games continue having evolutionarily stable preferences, hence Theorem 5 holds but, contrary to Proposition C1, not vacuously.

To prove Proposition C2, we must first identify all the preferences P that are stable for a given Π in Scenario B. We do this in the next few pages.

When p is large but smaller than 1 in Scenario B, the set of stable preferences for Π is a subset of the preferences that are stable under perfect observability ($p = 1$). To see this, observe that, for the same reason as in Scenario A, for any Π there exists a $\rho(\Pi)$ such that for any $p \in (\rho(\Pi), 1)$ the set of evolutionarily stable preferences is a subset of Π's evolutionarily stable preferences when $p = 1$. Hence, when p is large, to check whether a preference is stable or not, we can restrict attention to the preference equivalence classes that are stable under perfect observability.

In what follows we characterize, for every Π, the set of evolutionarily stable preferences for any $p \in (\bar{p}(\Pi), 1)$ where $\bar{p}(\Pi) = max\{\rho(\Pi), \frac{2d-b-c}{a+d-b-c}\}$.

Case 1.1: Πs with $a \geq b, c, d$ and $d > b$

From Dekel, Ely, and Yilankaya (2007), we know that when $p = 1$, any preferences P which is strategically equivalent to an element of the set $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0, 1], \mathcal{BA}_1, \theta_0\}$ is stable. For $p \in (\rho(\Pi), 1)$ we know that anything that is not stable under perfect observability is not stable, so we only need to check the stability of preferences which are strategically equivalent to one element of $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0, 1], \mathcal{BA}_1, \theta_0\}$:

- Any preference which has as a weakly dominant strategy playing *cooperate*, i.e. anything strategically equivalent to an element of $\{\mathcal{AA}, \mathcal{AB}_0, \mathcal{BA}_1$ and $\theta_0\}$ is stable. For any such preferences, playing *cooperate* regardless of who the opponent is or what is observed is a best response. Take any configuration (μ, β) where $supp(\mu)$ is any subset of $\{\mathcal{AA}, \mathcal{AB}_0, \mathcal{BA}_1$ and $\theta_0\}$, in which all incumbents play *cooperate* against themselves regardless of what is observed, i.e. $\beta_P(P'') = \beta_P(\emptyset) =$ *cooperate*, for all $P, P'' \in supp(\mu)$. Let P' be any mutant. Since it continues being a best response for incumbents to play *cooperate* against each other when they observe each other, and to play *cooperate* when they don't observe anything, focal equilibria exist. Take any $P \in \{\mathcal{AA}, \mathcal{AB}_0, \mathcal{BA}_1$ and $\theta_0\}$. In any focal equilibrium, if $\beta_{P'}(P) \neq$ *cooperate* for any $P \in \{\mathcal{AA}, \mathcal{AB}_0, \mathcal{BA}_1$ and $\theta_0\}$or $\beta_{P'}(\emptyset) \neq$ *cooperate*, $a > \sum_{P \in supp(\mu)}((p\mathcal{T}(\beta_{P'}(P), cooperate) + (1-p)\mathcal{T}(\beta_{P'}(\emptyset), cooperate))\mu(P)$, so there exists an $\bar{\epsilon} \in (0, 1)$ such that for all $\epsilon \in (0, \bar{\epsilon})$, all incumbents get a strictly larger

average fitness. If $\beta_{P'}(P) =$ *cooperate* and $\beta_{P'}(\emptyset) =$ *cooperate*, incumbents and P' get the same average fitness of a.

- Any preference P which is strategically equivalent to $\mathcal{AB}_\alpha$ with $\alpha \in (0,1)$ is stable. Take a monomorphic population of P which plays *cooperate* regardless of what is observed. Take any mutant P'. As long as the proportion in which the mutant enters the population is small enough, $\epsilon \leq 1-\alpha$, it continues being a best response for P to play *cooperate* when it observes it is matched against another P and when it does not observe anything, so focal equilibria exist. Moreover, in all focal equilibria against any mutant, the average fitness of P is larger because P gets the largest possible fitness, a, when matched against its own type.

- Preferences P which are strategically equivalent to $\mathcal{AB}_1$ are also stable. Take a monomorphic population of P which plays *cooperate* when it observes it is matched against another P and which plays *defect* when nothing is observed. Take any mutant P'. Since, when two P types are matched, there is 0 probability that one of them would play something other than *cooperate*, when both Ps observe each other, playing *cooperate* when observing another P continues being a best response. Playing *defect* when not observing anything also continues being a best response, regardless of what P' plays, as it is a weakly dominant strategy. Furthermore any focal equilibrium gives P a weakly larger fitness:

  Average fitness P: $(1-\epsilon)(pa+(1-p)d) + \epsilon(p\mathscr{T}(\beta_P(P'),\beta_{P'}(P)) + (1-p)\mathscr{T}(\textit{defect},\beta_{P'}(\emptyset)))$

  Average fitness P': $(1-\epsilon)(p\mathscr{T}(\beta_{P'}(P),\beta_P(P')) + (1-p)\mathscr{T}(\beta_{P'}(\emptyset),\textit{defect})) + \epsilon(p\mathscr{T}(\beta_{P'}(P'),\beta_{P'}(P')) + (1-p)\mathscr{T}(\beta_{P'}(\emptyset),\beta_{P'}(\emptyset)))$

  $a \geq \mathscr{T}(\beta_{P'}(P),\beta_P(P'))$ and $d \geq \mathscr{T}(\beta_{P'}(\emptyset),\textit{defect}))$. Hence, unless P' plays *cooperate* when he observes he is matched against P and plays *defect* when he doesn't observe anything, he would get strictly less fitness than P. If he does behave as stated, he gets the same fitness as P. Thus, there exists a $\bar{\epsilon} \in (0,1)$, such that for all $\epsilon \in (\bar{\epsilon},1)$, P gets a larger fitness.

This means that, for any Π with $a \geq b,c,d$ and $d > b$, the set of stable preferences when $p \in (\bar{p}(\Pi),1)$ is $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0,1], \mathcal{BA}_1, \theta_0\}$.

Case 1.2: Πs with $a \geq b,c,d$ and $b > d$

From Dekel, Ely, and Yilankaya (2007), we know that when $p = 1$, any preference P which is strategically equivalent to an element of the set $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0,1], \mathcal{BA}_1, \theta_0\}$ is stable. The analysis is exactly the same as in Case 1.1, except for $\mathcal{AB}_1$. There does not exist a preference P which is strategically equivalent to $\mathcal{AB}_1$ which is stable.

- Take any preference distribution μ which has P in its support. Any bayes nash equilibrium profile in which P plays *cooperate* against itself and when it doesn't observe anything does not have a focal equilibrium, for the same reasons as in Scenario A. Any bayes nash equilibrium profile which has a focal equilibrium must have P playing *defect* when it doesn't observe anything. However, such equilibrium profiles are not stable. Take any configuration (μ, β) with $P \in supp(\mu)$. Take mutant P' which, when preferences are observed, plays against any $Q \in supp(\mu) \backslash P$ whatever P plays against Q when preferences are observed, and which plays *cooperate* when nothing is observed:

  Average fitness P: $\sum_{Q \in supp(\mu)}(1-\epsilon)\mu(Q)(p\mathcal{T}(\beta_P(Q), \beta_Q(P)) + (1-p)\mathcal{T}(\textit{defect}, \beta_Q(\emptyset)) + \epsilon(p\mathcal{T}(\beta_P(P'), \beta_{P'}(P)) + (1-p)c)$

  Average fitness P': $\sum_{Q \in supp(\mu)}(1-\epsilon)\mu(Q)(p\mathcal{T}(\beta_{P'}(Q), \beta_Q(P')) + (1-p)\mathcal{T}(\textit{cooperate}, \beta_Q(\emptyset)) + \epsilon a$

  Note that $\mathcal{T}(\beta_P(Q), \beta_Q(P)) = \mathcal{T}(\beta_{P'}(Q), \beta_Q(P'))$, that since $a \geq c$ and $b \geq d$, $\mathcal{T}(\textit{cooperate}, \beta_Q(\emptyset)) \geq \mathcal{T}(\textit{defect}, \beta_Q(\emptyset))$, and that since a is the largest possible fitness, $a > p\mathcal{T}(\beta_P(P'), A) + (1-p)c)$ for any $p < 1$. So, for any $p < 1$ and $\epsilon > 0$ it is the case that the average fitness of the mutant is strictly larger.

Hence, for any Π with $a \geq b, c, d$, the set of stable preferences when $p \in (\bar{p}(\Pi), 1)$ is $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0,1), \mathcal{BA}_1, \theta_0\}$.

Case 2: Πs with $b \geq a \geq c$ and $2a \geq b + c$

From Dekel, Ely, and Yilankaya (2007), we know that when $p = 1$, any preference P which is strategically equivalent to an element of the set $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0, \frac{a-d}{b-d}]\}$ is stable. Any preference that is not stable for $p = 1$, is also not stable for $p \in (\rho(\Pi), 1)$. Hence, we only need to check stability for preferences which are strategically equivalent to an element of $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0, \frac{a-d}{b-d}]\}$.

- Any preference P which is strategically equivalent to $\mathcal{AA}$ is stable. Take a monomorphic population of $\mathcal{AA}$ which plays *cooperate* regardless of what is observed. Take

any mutant P'. Since, playing *cooperate* no matter what continues being a best response, even if a mutant which plays *defect* enters, focal equilibria exist. Moreover, in all such focal equilibria, the incumbent gets a larger average fitness because $p\mathcal{T}(\beta_{P'}(P), \textit{cooperate}) + (1-p)\mathcal{T}(\beta_{P'}(\emptyset), \textit{cooperate}) < a$ unless P' plays *cooperate* when she observes she is matched against $\mathcal{AA}$ and when she does not observe anything. Hence, there exists a $\bar{\epsilon} \in (0,1)$, such that for all $\epsilon \in (0, \bar{\epsilon})$, $\mathcal{AA}$ gets a larger average fitness.

- Any preference P which is strategically equivalent to $\mathcal{AB}_\alpha$ with $\alpha \in [0, \frac{a-d}{b-d})$ is stable. Take a monomorphic population of P which plays *cooperate* regardless of what is observed. As long as the proportion in which the mutant enters the population is small enough, $\epsilon \leq 1-\alpha$, it continues being a best response for P to play *cooperate* against other P and when nothing is observed, so focal equilibria exist. Moreover, in all focal equilibria against any mutant P', the average fitness of P is larger. This is because the largest possible fitness payoff is b, and the only equilibrium in which P' could get b with some probability is the focal equilibrium in which P plays *defect* when she observes she is matched against P' and P' plays *cooperate* with probability α, regardless of what she observes. In that equilibrium, when matched against P the mutant gets $p(\alpha b + (1-\alpha)d) + (1-p)(\alpha a + (1-\alpha)c)$, which is strictly less than a for any $p \in [0,1]$ and any $\alpha \in [0, \frac{a-d}{b-d})$.

Hence, for any Πs with $b \geq a \geq c$ and $2a \geq b+c$, the set of stable preferences when $p \in (\bar{p}(\Pi), 1)$ is $\{\mathcal{AA}, \mathcal{AB}_\alpha \text{ with } \alpha \in [0, \frac{a-d}{b-d}]\}$.

Case 3.1: Πs with $c \geq a \geq d \geq b$ and $2a \geq b+c$

From Dekel, Ely, and Yilankaya (2007), we know that when $p = 1$, any preference P which is strategically equivalent to $\mathcal{AB}_1$ is evolutionarily stable. By the exact same reasoning as for game types Π in Case 1.1, for any Π with $c \geq a$ and $2a \geq b+c$, when $p \in (\bar{p}(\Pi), 1)$, preferences P which are strategically equivalent to $\mathcal{AB}_1$ are stable.

Case 3.2: Πs with $c \geq a \geq b \geq d$ and $2a \geq b+c$

From Dekel, Ely, and Yilankaya (2007), we know that when $p = 1$, any preference P which is strategically equivalent to $\mathcal{AB}_1$ is evolutionarily stable. However, contrary to Case 3.1, these preferences are not stable. Any stable population only has in its support preferences that are strategically equivalent to $\mathcal{AB}_1$. Hence, to show the absence of stable preferences, it is sufficient to show that a monomorphic population of P which is strategically equivalent

to $\mathcal{AB}_1$ is not stable. For the same reasons as in Case 1.2, the only bayes nash equilibria that have a focal equilibrium must have $\beta_P(\emptyset) = \textit{defect}$.

- Take any mutant P' which always plays *cooperate*, regardless of what it observes. In the focal equilibrium in which P and P' play A when they observe they are matched against each other:

  Average fitness P: $(1-\epsilon)(p\mathcal{T}(\beta_P(P), \beta_P(P)) + (1-p)d) + \epsilon(pa + (1-p)c)$

  Average fitness P': $(1-\epsilon)(pa + (1-p)b) + \epsilon a$

  It is the case that $p\mathcal{T}(\beta_P(P), \beta_P(P)) + (1-p)d < pa + (1-p)b$ because $\mathcal{T}(\beta_P(P), \beta_P(P)) \leq a$ and $d < b$. Hence, for all sufficiently small ϵ, P' gets a strictly larger average fitness. This means $P = \mathcal{AB}_1$ is not stable.

Case 4: all other Πs

All other games are either non-generic, or don't have stable preferences when $p = 1$.

The above findings are summarized in Table 2 below.

TABLE 2
EVOLUTIONARILY STABLE PREFERENCES IN SCENARIO B

Π	Evolutionarily stable preferences
Case 1.1: $a \geq b, c, d$ and $d > b$	$\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0,1], \mathcal{BA}_1, \theta_0\}$
Case 1.2: $a \geq b, c, d$ and $b > d$	$\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0,1), \mathcal{BA}_1, \theta_0\}$
Case 2: $b \geq a \geq c$ and $2a \geq b + c$	$\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0, \frac{a-d}{b-d}]\}$
Case 3.1: $c \geq a \geq d \geq b$ and $2a \geq b + c$	$\mathcal{AB}_1$
Case 3.2: $c \geq a \geq b \geq d$ and $2a \geq b + c$	None exists
Case 4: Any other generic Π	None exists

Having characterized all the stable preferences for generic Π's in Scenario B, we now turn to the proof of Proposition C2.

Proof of Proposition C2

Proof. **Part a)** For the proof of Theorem 1 with $p = 1$, which can be seen in Appendix II, only games Π with $a \geq b, c, d$ and only preferences P which are strategically equivalent to

$\mathcal{AA}$ or $\mathcal{AB}_\alpha$ with $\alpha \in [\frac{d-b}{d-b+a-c}], 1)$ are used. As can be seen in Table 2, these preferences continue being stable for games Π with $a \geq b, c, d$, for any $p \in (\bar{p}(\Pi), 1)$ (Case 1.1 and Case 1.2). Hence, there exists a $\bar{p} < 1$ such that the proof for $p = 1$ continues to go through for all $p \in (\bar{p}, 1)$.

Part b) For any Π and any $p \in (\bar{p}(\Pi), 1)$, Theorem 2 continues to hold because:

- As can be seen in Table 2, for any game types Π in Case 1.1 and Case 1.2, the set of stable preferences for $p \in (\bar{p}(\Pi), 1)$ is either $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0, 1), \mathcal{BA}_1, \theta_0\}$ or $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0, 1], \mathcal{BA}_1, \theta_0\}$. As can be seen in Stability Tables 1 and 2, the only preferences that cannot be spanned by any moral principle in the $H5$ are those which are strategically equivalent to either $\{\mathcal{BA}_1, \theta_0\}$ or $\{\mathcal{BA}_1, \theta_0, \mathcal{AB}_1\}$, but these preferences are non-generic in the space of stable preferences.

- As can be seen in Table 2, for any game types Π in Case 2, the set of stable preferences for $p \in (\bar{p}(\Pi), 1)$ is $\{\mathcal{AA}, \mathcal{AB}_\alpha$ with $\alpha \in [0, \frac{a-d}{b-d})\}$. As can be seen in Stability Table 3, the moral principles in the $H5$ can span all such preferences.

- As can be seen in Table 2, for any game type Π in Case 3.1, the set of stable preferences for $p \in (\bar{p}(\Pi), 1)$ is $\mathcal{AB}_1$. As can be seen in Stability Table 4, the moral principles in the $H5$ can span such preference.

- As can be seen in Table 2, for any other generic Π, i.e. Case 3.2 and 4, when $p \in (\bar{p}(\Pi), 1)$ no stable preference exists.

For any Π and any $p \in (\bar{p}(\Pi), 1)$, Theorem 3 continues to hold. Indeed, for the sufficiency part, when $p \in (\bar{p}(\Pi), 1)$, everything is the same as with $p = 1$, except for the game types in Case 3.2 . For those Π, for $p \in (\bar{p}(\Pi), 1)$ no stable preference exists, so it is still the case that $\{Care, Loyalty\}$ or $\{Fairness, Sanctity\}$ span all generic evolutionarily stable preferences for all Π. For the necessity part, the proof of Theorem 3 with $p = 1$ only uses games represented in Cases 1.1 and 1.2 and preferences $\mathcal{AA}$ and $\mathcal{AB}_\alpha$ with $\alpha \in (0, 1)$. For those games, those preferences continue being stable for $p \in (\bar{p}(\Pi), 1)$, so the proof is exactly the same.

Part c) Theorem 4 changes. For the game types Π in Case 3.2, for $p \in (\bar{p}(\Pi), 1)$ no stable preference exists. Hence, the only fitness games Π that admit stable preferences for $p \in (\bar{p}(\Pi), 1)$ are Case 1.1, Case 1.2, Case 2 and Case 3.1. As can be seen in Stability tables 1, 2, 3 and 4, every moral principle in the $H5$ except *Authority* is capable of spanning either

$\mathcal{AA}$ or $\mathcal{AB}_\alpha$ for some $\alpha \in (0,1)$ for Πs in Case 1.1, Case 1.2 and Case 2, and $\mathcal{AB}_1$ for Πs in Case 3.1.

Part d) Theorem 5 continues to hold. The only games Π which admit social-fitness improving are the ones in Case 3.1. However, as can be seen in Stability Table 4, the only stable moral proclivity matrices require exactly $x = c - a$. Since, $x = c - a$ implies $kx \neq c - a$ for all $k \neq 1$, no moral principle is compatible with moral overdrive. □

ONLINE APPENDIX IV:
PROOFS FOR SECTION IV.B

This appendix proves Theorem 6. We will need the following lemmas.

LEMMA D1. If $((1,0,...,0),(1,0,...,0))$ is an efficient symmetric mixed strategy profile, then $2\pi_{1,1} \geq \pi_{j,1} + \pi_{1,j}$ for all j.

Proof. Suppose, by contradiction, that $((1,0,...,0),(1,0,...,0))$ is an efficient symmetric mixed strategy profile and $2\pi_{1,1} < \pi_{j,1} + \pi_{1,j}$ for some j. The efficient symmetric mixed strategy p solves

$$\max_{p} \; p^2\pi_{1,1} + (1-p)^2\pi_{j,j} + p(1-p)(\pi_{1,j} + \pi_{j,1}).$$

This expression is concave in p because $2\pi_{1,1} - 2\pi_{j,j} - 2(\pi_{1,j} + \pi_{j,1}) < 0$ (this inequality holds because by assumption $\pi_{1,1} + \pi_{1,1} < \pi_{i,j} + \pi_{j,1}$ and $\pi_{1,1} > \pi_{j,j}$). Concavity in p implies that the optimal p equates the first derivative to 0, that is, the optimal p solves

$$2p\pi_{1,1} - 2\pi_{j,j} + 2p\pi_{j,j} + (\pi_{1,j} + \pi_{j,1}) - 2p(\pi_{1,j} + \pi_{j,1}) = 0 \; .$$

Manipulating the equation, we get that the optimal p is

$$p = \frac{\pi_{i,j} + \pi_{j,j} - 2\pi_{j,j}}{2(\pi_{1,j} + \pi_{j,j} - \pi_{1,1} - \pi_{j,j})}.$$

Note that because of our assumption ($\pi_{1,1} + \pi_{1,1} < \pi_{1,j} + \pi_{j,1}$) and because $\pi_{1,1} > \pi_{j,j}$, both the numerator and the denominator of p are positive. Moreover, also because of our assumption, the denominator is larger than the numerator; hence the optimal p is strictly smaller than 1. But this contradicts the fact that $((1,0,...,0),(1,0,...,0))$ is an efficient symmetric mixed strategy profile. □

LEMMA D2. Suppose Π is well-ordered with respect to conflict and has $q = 0$. Any preference such that $p_{1,k} > p_{j,k}$ for all $j \in \{2,...,n\}$ and $k \in \{1,...,n\}$ is evolutionarily stable for Π.

Proof. Since (i) Π is well ordered with respect to conflict and (ii) $q = 0$, Lemma D1 implies that $\pi_{1,1} \geq \pi_{j,1}$ and $\pi_{1,1} \geq \pi_{1,j}$ for all j. Thus, any preference such that $p_{1,k} > p_{j,k}$ for all $j \in \{2,...,n\}$ and $k \in \{1,...,n\}$ is evolutionarily stable because agents with such preference

have as a dominant action playing the first row and, since $\pi_{1,1} \geq \pi_{j,1}$, that guarantees the agent gets a larger average fitness than any opponent. □

Proof of Theorem 6 part 1 Given an ordered set $S = \{s_1, s_2, \ldots s_n\}$, denote by $s[1], s[2], \ldots s[n]$ the smallest, second smallest, ... , largest element of S. Pick a generic $n \times n$ fitness matrix $\Pi' \in \mathbb{D}_n$ such that $\pi'_{1,1} \geq \pi'_{i,j}$ for all $i, j \in \{1, ..., n\}$, and define the following column vectors $\mathbf{h}, \mathbf{h}'$ of real functions.

1. $\mathbf{h}'$ is a vector of strictly increasing functions whose k-th element $h'_k(\cdot)$ is a function that is very large when evaluated at the largest element of the k-th row of Π' (that is, $h'_k(\pi'_{k,\cdot}[n])$ is very large). This vector defines a specific CARE moral principle.

2. $\mathbf{h}$ is a vector whose first element is a strictly increasing function and whose k-th (for $k > 1$) element $h_k(\cdot)$ is a strictly decreasing function that is very large when evaluated at the smallest element of the kth row of Π' (that is, $h_k(\pi'_{k,\cdot}[1])$ is very large). This vector defines a specific LOYALTY moral principle.

We now proceed to show that $\text{CARE}(\mathbf{h}'; \Pi')$ and $\text{LOYALTY}(\mathbf{h}; \Pi')$ are evolutionarily stable for Π' but are not pure-strategy equivalent.

- Evolutionary stability: Since $\pi'_{1,1} \geq \pi'_{i,j}$ for all $i, j \in \{1, ..., n\}$, any preference matrix P for which playing the first row is a best response to the opponent playing the first column is evolutionarily stable (this is because such preferences guarantee that an agent will play the most efficient action against agents with the same preference). This best-response feature is shared by the two preference matrices $\Pi' + \text{CARE}(\mathbf{h}'; \Pi')$ and $\Pi' + \text{LOYALTY}(\mathbf{h}; \Pi')$; indeed, since h'_1 and h_1 are both increasing, we have $\pi'_{1,1} + h_1(\pi'_{1,1}) \geq \pi'_{j,1} + h_1(\pi'_{1,j})$ and $\pi'_{1,1} + h'_1(\pi'_{1,1}) \geq \pi'_{j,1} + h'_1(\pi'_{1,j})$ for all $j \in \{1, ..., n\}$. (Note that the inequalities in the previous sentence correctly feature $\pi'_{1,j}$, not $\pi'_{j,1}$, in the argument of $h_1(\cdot)$: the first element refers to the row player, the second element refers to the column player.)

- Not pure-strategy equivalent: Fix any pure action $k > 1$ for the column player. The best response of a row player with preferences $\Pi' + \text{CARE}(\mathbf{h}'; \Pi')$ is to play the action that maximizes the column player's fitness. This is different, generically, than the action that minimizes the column player's fitness, which is the row player's best response under $\Pi' + \text{LOYALTY}(\mathbf{h}; \Pi')$.

□

Proof of Theorem 6 part 2 Let Π be any generic matrix that is well ordered with respect to conflict. In what follows we show that for any generic preference P that is evolutionarily stable for Π, a pure-strategy equivalent matrix P' exists that is evolutionarily stable and a vector of functions $\mathbf{h}$ such that $P' = \Pi + \textsc{Loyalty}(\mathbf{h}; \Pi)$ or $P' = \Pi + \textsc{Care}(\mathbf{h}; \Pi)$.

Case 1: $\bar{q} > 0$ and $((1, 0, ..., 0), (1, 0, ..., 0))$ is a an efficient symmetric mixed strategy profile

Suppose that Π is well-ordered fitness matrix with $\bar{q} > 0$, and that P is a generic evolutionarily stable matrix for that Π. We will construct a matrix M that, when added to Π, generates a matrix which is strategically equivalent to P. Proving strategic equivalence is sufficient to prove stability and pure-strategy equivalence. If $\bar{q} \in \{1, ..., n-1\}$, then M is generated by LOYALTY; if, instead, $\bar{q} = n$, then M is generated by CARE.

Take any column k of the evolutionarily stable matrix P,

$$P_{\cdot,k} = \begin{bmatrix} p_{1,k} \\ \vdots \\ p_{n,k} \end{bmatrix}.$$

Let r_j denote the row corresponding to the jth smallest element of column vector $P_{\cdot,k}$, so $p_{r_j,k} = p_{\cdot,k}[j]$. Suppose the row with the largest element of $\Pi_{\cdot,k}$ is the row with the ℓth largest element of $P_{\cdot,k}$, so $\pi_{r_\ell,k} = \pi_{\cdot,k}[n]$.

Construct $M_{\cdot,k}$ as follows:

- $M_{r_\ell,k} = A$
- $M_{r_{\ell+t},k} = A + \pi_{\cdot,k}[n] - \pi_{r_{\ell+t},k} + \frac{p_{\cdot,k}[\ell+t] - p_{\cdot,k}[\ell]}{N}$.
- $M_{r_{\ell-t},k} = A + \pi_{\cdot,k}[n] - \pi_{r_{\ell-t},k} - \frac{p_{\cdot,k}[\ell] - p_{\cdot,k}[\ell-t]}{N}$.

for all $t \in \{1, ..., \ell - 1\}$, where A and N are sufficiently large positive numbers.

$M_{\cdot,k}$ satisfies the following properties:

1. $M_{j,k} \geq 0$ for all $j \in \{1, ..., n\}$.
2. $$\begin{aligned} M_{r_j,k} - M_{r_i,k} &= A + \pi_{\cdot,k}[n] - \pi_{r_j,k} + \frac{p_{\cdot,k}[j] - p_{\cdot,k}[\ell]}{N} \\ &\quad - \left(A + \pi_{\cdot,k}[n] - \pi_{r_i,k} + \frac{p_{\cdot,k}[i] - p_{\cdot,k}[\ell]}{N} \right) \\ &= \pi_{r_i,k} - \pi_{r_j,k} + \frac{p_{\cdot,k}[j] - p_{\cdot,k}[i]}{N} \quad . \end{aligned}$$

Thus, for a sufficiently large and positive N, $\pi_{r_i,k} > \pi_{r_j,k}$ implies $M_{r_j,k} > M_{r_i,k}$.

If we let $h_k(.)$ be such that $h_k(\pi_{k,j}) = M_{j,k}$ for all $k \in \{1, ..., n\}$, then $h_k(\pi_{k,r_j}) > h_k(\pi_{k,r_i})$. Therefore,

(a) $k > \bar{q}$: since Π is a matrix that is well ordered with respect to conflict, it is the case that $\pi_{k,r_i} > \pi_{k,r_j}$. Thus $h_k(\cdot)$ is a decreasing function.

(b) $k \leq \bar{q}$: since Π is a matrix that is well ordered with respect to conflict, it is the case that $\pi_{k,r_i} \leq \pi_{k,r_j}$. Thus $h_k(\cdot)$ is an increasing function.

Hence, the constructed matrix M can be the output of the operator in Definition 15, i.e $M = \text{LOYALTY}(\mathbf{h}; \Pi)$ for $\bar{q} \in \{1, ..., n-1\}$ and $M = \text{CARE}(\mathbf{h}; \Pi)$ for $\bar{q} = n$, where $h_k(.)$ is such that $h_k(\pi_{k,j}) = M_{j,k}$ for all $k \in \{1, ..., n\}$.

Next, we prove that $\Pi + M$ is strategically equivalent to P. The elements of column $\Pi_{.,k} + M_{.,k}$ are

- $\pi_{r_\ell,k} + M_{r_\ell,k} = A + \pi_{.,k}[n]$
- $\pi_{r_{\ell+t},k} + M_{r_{\ell+t},k} = A + \pi_{.,k}[n] + \frac{p_{.,k}[\ell+t] - p_{.,k}[\ell]}{N}$
- $\pi_{r_{\ell-t},k} + M_{r_{\ell-t},k} = A + \pi_{.,k}[n] - \frac{p_{.,k}[\ell] - p_{.,k}[\ell-t]}{N}$

for all $t \in \{1, ..., n-\ell\}$. Let P' be the matrix obtained by adding $\frac{p_{.,k}[\ell]}{N} - A - \pi_{.,k}[n]$ to each column of $\Pi + M$. Such matrix is strategically equivalent to $\Pi + M$, and the elements of its k-th column are:

- $P'_{r_\ell,k} = \pi_{r_\ell,k} + M_{r_\ell,k} + \frac{p_{.,k}[\ell]}{N} - A - \pi_{.,k}[n] = \frac{p_{.,k}[\ell]}{N} = \frac{p_{r_\ell,k}}{N}$
- $P'_{r_{\ell+t},k} = \pi_{r_{\ell+t},k} + M_{r_{\ell+t},k} + \frac{p_{.,k}[\ell]}{N} - A - \pi_{.,k}[n] = \frac{p_{.,k}[\ell+t]}{N} = \frac{p_{r_{\ell+t},k}}{N}$ for all $t \in \{1, ..., n-\ell\}$
- $P'_{r_{\ell-t},k} = \pi_{r_{\ell-t},k} + M_{r_{\ell-t},k} + \frac{p_{.,k}[\ell]}{N} - A - \pi_{.,k}[n] = \frac{p_{.,k}[\ell-t]}{N} = \frac{p_{r_{\ell-t},k}}{N}$ for all $t \in \{1, ..., \ell-1\}$

This means that $p'_{i,k} = p_{i,k}/N$ for all i and k, thus P' is strategically equivalent to P, which implies $\Pi + M$ is strategically equivalent to P.

Case 2: $\bar{q} = 0$ and $((1, 0, ..., 0), (1, 0, ..., 0))$ is an efficient symmetric mixed strategy profile

Take any generic Π which is well-ordered with $\bar{q} = 0$, and for which $((1, 0, ..., 0), (1, 0, ..., 0))$ is an efficient symmetric mixed strategy profile. We will construct

a matrix M' such that $\Pi + M'$ is pure-strategy equivalent to P, evolutionarily stable for Π, and such that M' can be the output of the operator $\text{LOYALTY}(\cdot; \Pi)$.

Before we start the construction, we observe that any generic stable preference P must be such that

1. $p_{1,1} \geq p_{j,1}$ for all j, and
2. $p_{j,1} \neq p_{i,1}$ for all $j \neq i$.

The first property follows from the fact that a stable preference must have the first row as a best response to the first column. The second property says that matrices for which $p_{j,1} = p_{i,1}$ for some $j \neq i$ are non-generic in the space of stable preferences. This is because this space has lower dimension than the space that is evolutionarily stable for Π (see Lemma D2).

We are now ready to construct matrix M' such that $\Pi + M'$ is pure strategy equivalent to P and evolutionarily stable for Π, where P is any preference with $p_{j,1} \neq p_{i,1}$ for all $j \neq i$ and $p_{1,1} \geq p_{j,1}$ for all j.

- For columns $k > 1$, let $M'_{j,k} = g_k(\pi_{k,j}) = h_k(\pi_{k,j})$ for all $j \in \{1, ..., n\}$, where $h_k(\cdot)$ is the vector constructed in the above case.
- For column $k = 1$, let $g_1(\cdot)$ be some strictly increasing function where $g_1(\pi_{1,1})$ is finite but very large.

This M' has the following properties:

1. $g_k(\cdot)$ is a decreasing function for $k > 1$ and $g_1(\cdot)$ is a increasing function, thus M' can be the output of operator $\text{LOYALTY}(\mathbf{g}; \Pi)$ from Definition 15.
2. Take any opponent who plays some mixture that places probability 0 on the first column. Whatever best response $\Pi + M'$ induces, it is exactly the same as the one $\Pi + M$ induces (M being the matrix in the previous Case 1) because, for columns $2, ..., n$, the matrices M and M' are constructed in the same way. So, the best response of $\Pi + M'$ to a strictly mixed strategy that places probability 0 on the first column is exactly the same as that of P.
3. Take an opponent who plays any mixture which places a probability strictly larger than 0 on the first column. We can pick $g_1(\pi_{11})$ large enough that the best response

of a player with preferences $\Pi + M'$ will be to play the first row with probability one. Since $\pi_{1,1} \geq \pi_{1,j}$ for all j, this ensures that the expected fitness the opponent gets when matched with the agent is less than $\pi_{1,1}$, the fitness the agent gets when matched against itself. Thus, there exists some $\epsilon > 0$ such that the agent gets a larger average fitness than the opponent. So, the best response of $\Pi + M'$ to a strictly mixed strategy that places probability strictly larger than 0 on the first column is playing the first row with probability 1, which is not necessarily the same best response as that of P. However, that best response guarantees the agent gets a larger average fitness than that of any opponent.

The three items together imply that, for a generic stable preference P, there exists an M' that is generated by LOYALTY such that the best response of $\Pi + M'$ to any pure strategy is exactly the same as that of P, and $\Pi + M'$ is stable for Π.

Case 3: $((1, 0, ..., 0), (1, 0, ..., 0))$ **is NOT an efficient symmetric mixed strategy profile**

We will show that any generic Π which is well-ordered with respect to conflict and for which $((1, 0, ..., 0)(1, 0, ..., 0))$ is not an efficient symmetric mixed strategy profile, does not have evolutionarily stable preferences. Stability requires the incumbent getting an average equilibrium fitness greater than any mutant opponent in every focal Nash equilibrium. Among the mutant opponents is included, in particular, one whose preferences makes that mutant indifferent between every outcome. When playing against this particular mutant, an incumbent may indeed have preferences P that are stable for Π but, we will show, only for non-generic Π's.

For a generic Π, the efficient symmetric mixed strategy profile is unique. Denote this profile by $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)$. Regardless of what the incumbent plays, it is a best response for our hand-picked mutant to play $\boldsymbol{\alpha}$ – therefore, by our definition of stability, the incumbent must get a larger average fitness than the mutant who plays $\boldsymbol{\alpha}$. Moreover, if P is stable, the incumbent must play $\boldsymbol{\alpha}$ against incumbents who also play $\boldsymbol{\alpha}$ (this is because From Dekel, Ely, and Yilankaya (2007) we know that the symmetric efficient payoff is a Nash equilibrium of the game with evolutionarily stable preference); this implies that, given the incumbent's preferences P, $\boldsymbol{\alpha}$ is a best response to our mutant's $\boldsymbol{\alpha}$. If this is the case, then our player with preferences P must be indifferent between playing all the actions in $\boldsymbol{\alpha}$ for which $\alpha_i > 0$. This indifference implies the existence of multiple Nash equilibria in which the incumbent places any positive probability on the actions i for which $\alpha_i > 0$ and the hand-picked mutant plays $\boldsymbol{\alpha}$. Next, we show that, for generic Π, in at least one of these equilibria the mutant

attains a larger fitness than incumbents receive at the efficient symmetric mixed strategy profile; this will contradict the assumption that P is stable.

Let $\boldsymbol{q} = (q_1, ..., q_n)$ represent a generic (i.e., not necessarily extremal) strategy for the incumbent. The mutant attains a larger fitness than incumbents receive at the efficient symmetric mixed strategy profile if

$$\sum_{i=1}^{n} q_i \sum_{j=1}^{n} \alpha_j \pi_{j,i} > \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{n} \alpha_j \pi_{j,i}. \tag{5}$$

The left-hand side is the hand-picked mutant's expected fitness when matched with an incumbent that plays $\boldsymbol{q}$; the right-hand side is the expected fitness incumbents receive at the efficient symmetric mixed strategy profile. If (5) holds for some $\boldsymbol{q}$ that is a best response to $\boldsymbol{\alpha}$, P cannot be stable.

Inequality (5) holds if we pick the $\boldsymbol{q}$ that places probability one on the action i for which $\sum_{j=1}^{n} \alpha_j \pi_{j,i}$ is the largest among the non-zero entries of vector $\boldsymbol{\alpha}$. Such a $\boldsymbol{q}$ happens to be part of the best-response correspondence to $\boldsymbol{\alpha}$. In this case, stability fails. The only way this failure does not happen is if $\sum_{j=1}^{n} \alpha_j \pi_{j,i}$ is the same for all i's that are non-zero entries of vector $\boldsymbol{\alpha}$ (in which case the two sides of (5) are equal). In matrix notation, the condition required for stability is that

$$\underline{\boldsymbol{\alpha}}^T \underline{\Pi} = c \times \mathbf{1}^T \tag{6}$$

where the column vector $\underline{\boldsymbol{\alpha}}$ contains all the strictly positive elements of $\boldsymbol{\alpha}$ and the matrix $\underline{\Pi}$ is the square matrix obtained by removing the rows and columns corresponding to the zero elements of $\boldsymbol{\alpha}$, c is a real number, and $\mathbf{1}$ is an appropriately sized column vector of 1's.

The efficient symmetric strategy profile is the solution to:

$$\max_{x_1,...,x_n} \sum_{i=1}^{n} x_j \left(\sum_{i=1}^{n} x_i \pi_{j,i} \right) \text{ subject to } x_1, ..., x_n \in [0, 1] \text{ and } x_1 + ... + x_n = 1.$$

If $x_i \neq \{0, 1\}$ is part of a solution to this problem, x_i satisfies the first-order conditions

$$\sum_{j=1}^{n} x_j \pi_{j,i} + \sum_{j=1}^{n} x_j \pi_{i,j} = \lambda.$$

By assumption, then, all i's that are non-zero entries of vector $\boldsymbol{\alpha}$ satisfy these conditions. Therefore, this condition implies:

$$(\underline{\boldsymbol{\alpha}}^T \underline{\Pi})^T + \underline{\Pi}\,\underline{\boldsymbol{\alpha}} = \lambda \times \mathbf{1}$$

or equivalently, using (6),

$$\underline{\Pi}\,\underline{\boldsymbol{\alpha}} = (\lambda - c) \times \mathbf{1} \tag{7}$$

For a generic Π, there is no $\boldsymbol{\alpha}$ that solves (6) and (7) simultaneously. Therefore, for a generic Π there exists no stable P.

Proof of Theorem 6 part 4 The only generic well-ordered fitness matrices Π in which a strict social-fitness improvement is possible (refer to Definition 13) are those in which $((1,0,...,0),(1,0,...,0))$ is an efficient symmetric action profile and $\pi_{1,1} < \pi_{\ell,1}$ for at least one $\ell \in \{2,...,n\}$. This is because:

1. generic well-ordered fitness matrices for which $((1,0,...,0),(1,0,...,0))$ is not an efficient symmetric action profile don't have evolutionarily stable preferences (see part 2 case 3 above), and

2. if $((1,0,...,0),(1,0,...,0))$ is an efficient symmetric mixed action and $\pi_{1,1} \geq \pi_{\ell,1}$ for all $\ell \in \{2,...,n\}$, Π is already social-fitness maximizing for itself, so no improvement is possible by definition.

Take an ℓ such that $\pi_{1,1} < \pi_{\ell,1}$. We now show that the P's that are stable for Π are not compatible with moral overdrive.

We begin by showing that any stable P must be such that for some i' such that $\pi_{1,i'} \leq \pi_{1,1}$ and $\pi_{\ell,i'} \leq \pi_{1,1}$ (we know such i' exists because $i' = \ell$ satisfies the inequalities), it is the case that $p_{1,1} = p_{i',1}$ and $p_{i',\ell} > p_{1,\ell}$. To see this, take a mutant who is indifferent between any outcome and who plays column ℓ with probability $\epsilon \in (0,1)$ and column 1 with probability $(1-\epsilon)$ against an incumbent with preference P, and plays action 1 against other mutants. If the incumbent with preference P had as a best response to play row 1 against such mutant, the mutant would get a fitness of $(1-\epsilon)\pi_{1,1} + \epsilon\pi_{\ell,1} > \pi_{1,1}$, which would imply P is not stable. Thus, for P to be stable, it must be that row 1 is not a best response to this mutant's behavior and, moreover, that the best response to the mutant's behavior makes the mutant worse off than the incumbent fitness-wise. That is, we need some row $i' > 1$ such that

(i) $p_{1,1}(1-\epsilon) + p_{1,\ell}\epsilon < p_{i',1}(1-\epsilon) + p_{i',\ell}\epsilon$ for all ϵ (row 1 not a best response), and

(ii) $\pi_{1,i'}(1-\epsilon) + \pi_{\ell,i'}\epsilon \leq \pi_{1,1}$ for all ϵ (the best response i' makes the mutant worse off fitness-wise)

Because (i) must hold for all ϵ, i' must be such that

(i') $p_{1,1} = p_{i',1}$ and $p_{1,\ell} < p_{i',\ell}$ (note that, because P is stable, it couldn't be $p_{1,1} < p_{i',1}$).

Now we conclude by showing that any P that is stable, and therefore satisfies $p_{1,1} = p_{i',1}$ for some i', is not compatible with moral overdrive. By (i'), any moral principle M (from the $H5$ or not) that is stable satisfies $\pi_{1,1} + m_{1,1} = \pi_{i',1} + m_{i',1}$ or, which is the same, $m_{1,1} - m_{i',1} = \pi_{i',1} - \pi_{1,1}$. This equality fails if the left-hand side is multiplied by any $k > 1$.

In sum, we have shown that, if Π is a generic well-ordered fitness matrix in which a strict social-fitness improvement is possible , and $\Pi + M$ is evolutionarily stable, then $\Pi + kM$ cannot be evolutionarily stable for $k > 1$. □

ONLINE APPENDIX V:
STABILITY TABLES

In what follows we refer to "moral proclivity matrices" as defined in Definition B1 in Appendix II.

The stability tables in this section show:

1. for each fitness game, all its stable preferences;
2. for each stable preference in that game, all the moral proclivity matrices that can span it;
3. for each stable moral proclivity matrix in that game, all the *H5*'s moral principles that can produce it.

Before presenting the tables, we explain how they were constructed.

First Column

The first column is taken directly form Dekel, Ely, and Yilankaya (2007)'s Proposition 4. That proposition characterizes, for every 2×2 fitness game, all the evolutionarily stable preferences.

Second Column

The second column is constructed by checking, for each stable preference P, which moral proclivity matrices $M(x,y)$ are such that $\Pi + M(x,y)$ is strategically equivalent to P. There are 4 possible moral proclivity matrices:

$$M_1 = \begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array}, \; M_2 = \begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array}, \; M_3 = \begin{array}{|c|c|}\hline 0 & 0 \\ \hline x & y \\ \hline \end{array}, \; M_4 = \begin{array}{|c|c|}\hline y & y \\ \hline x & 0 \\ \hline \end{array}.$$

For each moral proclivity matrix, we back out which values of x and y make $\Pi + M$ strategically equivalent to P. The moral proclivity matrix M and the values of x and y that achieve strategic equivalence appear in the table. If no value of x or y make $\Pi + M$ strategically equivalent to P, M does not appear in the second column of the stability table of Π.

For example, for Π with $a > c$ and $d > b$, we know from Dekel, Ely, and Yilankaya (2007) that any preference strategically equivalent to $\mathcal{AA} = \begin{array}{|c|c|}\hline 1 & 1 \\ \hline 0 & 0 \\ \hline \end{array}$ is evolutionarily stable. Hence, in the stability table representing those Πs, which is Stability Table 1, the row corresponding to $\mathcal{AA}$ only contains M_1 and M_4 because:

- For $x \geq 0$ and $y > d - b$, $\Pi + M_1 = \begin{array}{|c|c|}\hline a+x & b+y \\ \hline c & d \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AA}$.

- For $0 \leq x < a - c$ and $y > d - b$, $\Pi + M_4 = \begin{array}{|c|c|}\hline a & b+y \\ \hline c+x & d \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AA}$.
- There are no values of $y \geq 0$ such that, for $\Pi + M_2 = \begin{array}{|c|c|}\hline a+x & b \\ \hline c & d+y \\ \hline \end{array}$ and $\Pi + M_3 = \begin{array}{|c|c|}\hline a & b \\ \hline c+x & d+y \\ \hline \end{array}$, $b > d + y$. Hence, there are no values of $x, y \geq 0$ for which $\Pi + M_2$ or $\Pi + M_3$ are strategically equivalent to $\mathcal{AA}$.

Third Column

Every sub-column of the third column is constructed by checking, for each stable moral proclivity matrix, which of the *H5*'s moral principles generate it. To figure which of the $H5$'s moral principles generate the stable moral proclivity matrices, the complete parameter ordering of a, b, c, d is needed. Each sub-column of column 3 represents one such parameter ordering.

For example, in the third sub-column of column 3 in Stability Table 1, which corresponds to parameter ordering $a > d > c > b$, we can find the *H5*'s moral principles which generate the stable moral proclivity matrices M_1 and M_4 for such fitness games. *Sanctity* and *Loyalty* can generate M_1, so they appear in the row corresponding to M_1. For Πs with $a > d > c > b$, no *H5* moral principle can generate moral proclivity matrix M_4, so the word "None" appears in the row corresponding to M_4.

We are now ready to present the tables.[1]

[1] From Dekel, Ely, and Yilankaya (2007) we know only the games presented in the following tables have a stable preference.

TABLE 1
COORDINATION GAME: $a > c, d > b$

Stable preferences	Stable moral proclivities	$H5$s that generate stable moral proclivities *The x and y satisfy the inequalities in the previous column*		
		a>c>d>b	a>d>b>c	a>d>c>b
$\mathcal{AA}$ 1 1 0 0	x y 0 0 x≥0 y>d-b	$Sanctity$(x,y), $Care$(x,y)	$Sanctity$(x,y), $Authority$(x,y), $Loyalty$(x,y)	$Sanctity$(x,y), $Loyalty$(x,y)
	0 y x 0 x<a-c y>d-b	None	None	None
$\mathcal{AB}_\alpha$ with $\alpha \in (0,1)$ $1-\alpha$ 0 0 α	x 0 0 y x≥max{0, $\frac{1-\alpha}{\alpha}$(d-b)-(a-c)} y=$\frac{1-\alpha}{\alpha}$(a-c+x)-(d-b)	$Fairness$(x,y), $Loyalty$(x,y)	$Fairness$(x,y), $Care$(x,y)	$Fairness$(x,y), $Care$(x,y)
	Stable only for $\alpha \in (0, \frac{d-b}{d-b+a-c}]$ x y 0 0 x≤$\frac{1-\alpha}{\alpha}$(d-b)-(a-c) y=(d-b)-$\frac{\alpha}{1-\alpha}$(a-c+x)	$Sanctity$(x,y), $Care$(x,y)	$Sanctity$(x,y), $Authority$(x,y), $Loyalty$(x,y)	$Sanctity$(x,y), $Loyalty$(x,y)
	Stable only for $\alpha \in [\frac{d-b}{d-b+a-c}, 1)$ 0 0 x y x≤(a-c)-$\frac{1-\alpha}{\alpha}$(d-b) y=$\frac{\alpha}{1-\alpha}$(a-c-x)-(d-b)	$Authority$(x,y)	None	$Authority$(x,y)
	0 y x 0 max{0,a-c-$\frac{1-\alpha}{\alpha}$(d-b)}≤x<a-c y=(d-b)-$\frac{\alpha}{1-\alpha}$(a-c+x)	None	None	None
$\mathcal{AB}_1$ 0 0 0 1	0 y x 0 x=a-c y<d-b	None	None	None
	0 0 x y x=a-c y≥0	$Authority$(x,y)	None	$Authority$(x,y)
$\mathcal{AB}_0$ 1 0 0 0	x y 0 0 x≥0 y=d-b	$Sanctity$(x,y), $Care$(x,y)	$Sanctity$(x,y), $Authority$(x,y), $Loyalty$(x,y)	$Sanctity$(x,y), $Loyalty$(x,y)
	0 y x 0 x<a-c y=d-b	None	None	None
$\mathcal{BA}_1$ 0 1 0 0	0 y x 0 x=a-c y>d-b	None	None	None
θ_0 0 0 0 0	0 y x 0 x=a-c y=d-b	None	None	None

TABLE 2

GAMES IN WHICH $(cooperate, cooperate)$ IS THE SOCIAL-FITNESS MAXIMIZING STRATEGY PROFILE, *cooperate* IS A DOMINANT STRATEGY AND A IS THE HIGHEST POSSIBLE PAYOFF: $a > c, b > d, a > b$

Stable preferences	Stable moral proclivities	$H5$s that generate stable moral proclivities *The x and y satisfy the inequalities in the previous column*		
		a>c>b>d	a>b>d>c	a>b>c>d
$\mathcal{AA}$ 1 1 0 0	x y 0 0 x≥0 y≥0	$Sanctity$(x,y), $Care$(x,y)	$Sanctity$(x,y), $Loyalty$(x,y), $Authority$(x.y)	$Sanctity$(x,y), $Care$(x,y), $Authority$(x.y)
	x 0 0 y x≥0 y<b-d	$Fairness$(x.y), $Loyalty$(x,y)	$Fairness$(x.y), $Care$(x,y)	$Fairness$(x.y), $Loyalty$(x,y)
	0 0 x y x<a-c y<b-d	$Authority$(x.y)	None	None
	0 y x 0 x<a-c y≥0	None	None	None
$\mathcal{AB}_\alpha$ with $\alpha \in (0,1)$ $1-\alpha$ 0 0 α	0 0 x y x<a-c y=$\frac{\alpha}{1-\alpha}$(a-c-x)+(b-d)	$Authority$(x,y)	None	None
	x 0 0 y x≥0 y=$\frac{\alpha}{1-\alpha}$(a-c+x)+(b-d)	$Fairness$(x.y), $Loyalty$(x,y)	$Fairness$(x.y), $Care$(x,y)	$Fairness$(x.y), $Loyalty$(x,y)
$\mathcal{AB}_1$ 0 0 0 1	0 0 x y x=a-c y>b-d	$Authority$(x,y)	None	None
$\mathcal{AB}_0$ 1 0 0 0	0 0 x y x<a-c y=b-d	$Authority$(x,y)	None	None
	x 0 0 y x≥0 y=b-d	$Fairness$(x.y), $Loyalty$(x,y)	$Fairness$(x.y), $Care$(x,y)	$Fairness$(x.y), $Loyalty$(x,y)
$\mathcal{BA}_1$ 0 1 0 0	0 0 x y x=a-c y<b-d	$Authority$(x,y)	None	None
	0 y x 0 x=a-c y≥0	None	None	None
θ_0 0 0 0 0	0 0 x y x=a-c y=b-d	$Authority$(x,y)	None	None

TABLE 3

GAMES IN WHICH $(cooperate, cooperate)$ IS THE SOCIAL-FITNESS MAXIMIZING STRATEGY PROFILE AND *cooperate* IS A DOMINANT STRATEGY BUT A IS NOT THE HIGHEST POSSIBLE PAYOFF: $a > c, b > d, b > a$

Stable preferences	Stable moral proclivities	$H5$s that generate stable moral proclivities *The x and y satisfy the inequalities in the previous column*	
		b>a>c>d	b>a>d>c
$\mathcal{AA}$ 1 | 1 0 | 0	x | y 0 | 0 x≥0 y≥0	$Sanctity$(x,y), $Authority$(x.y)	$Sanctity$(x,y), $Authority$(x.y)
	x | 0 0 | y x≥0 y<b-d	$Fairness$(x.y),	$Fairness$(x.y),
	0 | 0 x | y x<a-c y<b-d	$Loyalty$(x,y)	$Care$(x,y)
	0 | y x | 0 x<a-c y≥0	$Care$(x.y)	$Loyalty$(x.y)
$\mathcal{AB}_\alpha$ with $\alpha \in [0, \frac{a-d}{b-d})$ $1-\alpha$ | 0 0 | α	0 | 0 x | y x<a-c y=$\frac{\alpha}{1-\alpha}$(a-c-x)+(b-d)	$Loyalty$(x,y)	$Care$(x,y)
	x | 0 0 | y x≥0 y=$\frac{\alpha}{1-\alpha}$(a-c+x)+(b-d)	$Fairness$(x.y)	$Fairness$(x.y)

TABLE 4

PRISONERS' DILEMMA GAMES IN WHICH (A, A) IS THE SOCIAL-FITNESS MAXIMIZING STRATEGY PROFILE: $c > a > d > b$

Stable preferences	Stable moral proclivities	$H5$s that generate stable moral proclivities *The x and y satisfy the inequalities in the previous column* c>a>d>b
$\mathcal{AB}_1$ 0 0 0 1	x y 0 0 x=c-a y<d-b	$Sanctity$(x,y), $Care$(x,y)
	x 0 0 y x=c-a y≥0	$Fairness$(x,y), $Loyalty$(x,y)

TABLE 5
HAWK-DOVE GAMES IN WHICH (A, A) IS THE SOCIAL-FITNESS MAXIMIZING STRATEGY PROFILE: $c > a > b > d$

Stable preferences	Stable moral proclivities	$H5$s that generate stable moral proclivities *The x and y satisfy the inequalities in the previous column*
		c>a>b>d
$\mathcal{AB}_1$ 0 0 0 1	x 0 0 y x=c-a y>b-d	$Fairness$(x,y), $Loyalty$(x,y)

TABLE 6
PURE-COMMON VALUE HAWK-DOVE GAMES: $c = b > a > d$

Stable preferences	Stable moral proclivities	$H5$s that generate stable moral proclivities *The x and y satisfy the inequalities in the previous column*
		c=b>a>d
$\mathcal{AB}_\alpha$ α 0 0 1-α with $\alpha=\frac{b-d}{c-a+b-d}$	x 0 0 y x=2(c-a) y=2(b-d)	None

ONLINE APPENDIX VI:
VARIABLES FOR FIGURE 4

- **Post-Materialism Index**
  - **Data Source**: World Values Survey waves 5 (2005–2009), 6 (2009–2014), and 7 (2017–2022). Dataset freely available here.
  - **Individual Level**: The variable $Y002$, ranging from 1 (lowest Post-Materialism) to 3 (highest Post-Materialism).
  - **Country Level**: Using $S018$ from the WVS (which normalizes the sample to 1000 individuals per country-wave), we construct a country-wave average of $Y002$. Denoting $y^{\text{PM}}_{i,j,w}$ as the Post-Materialism value for individual i in country j during wave w, weighted by Weight_i, the country-wave mean is:

$$\overline{Y}^{\text{PM}}_{j,w} = \frac{\sum_{i \in N_{j,w}} \left(y^{\text{PM}}_{i,j,w} \cdot \text{Weight}_i \right)}{1000}.$$

    These means are then averaged across waves. For instance:

$$\overline{Y}^{\text{PM}}_{\text{US}} = \frac{\overline{Y}^{\text{PM}}_{\text{US, Wave 5}} + \overline{Y}^{\text{PM}}_{\text{US, Wave 6}} + \overline{Y}^{\text{PM}}_{\text{US, Wave 7}}}{3}.$$

- **Care Index**
  - **Data Source**: World Values Survey waves 5 (2005–2009) and 6 (2010–2014).
  - **Individual Level**: The variable $A193$, ranging from 1 (lowest Care) to 6 (highest Care).
  - **Country Level**: Similarly, employing $S018$ to normalize the sample, the country-wave average of $A193$ is:

$$\overline{Y}^{\text{Care}}_{j,w} = \frac{\sum_{i \in N_{j,w}} \left(y^{\text{Care}}_{i,j,w} \cdot \text{Weight}_i \right)}{1000},$$

    which is then averaged across waves. For example:

$$\overline{Y}^{\text{Care}}_{\text{US}} = \frac{\overline{Y}^{\text{Care}}_{\text{US, Wave 5}} + \overline{Y}^{\text{Care}}_{\text{US, Wave 6}}}{2}.$$

- **Market orientation Index**
    - **Data source**: World Economic Forum: Global Competitiveness Index 4.0 (2007–2017). Dataset freely available here.
    - **Construction of the country-level market orientation index**:
        - For each country, take the value of the following four "pillars" – 6th (goods market efficiency), 7th (labor market efficiency), 8th (financial market development), and 10th (market size) – and calculate the 2007-2017 average score $\overline{X}_{j,p}$ for each pillar p.
        - For each country j, compute the country-level market orientation index $\overline{X}_j$ as the arithmetic average of the four pillars:

$$\overline{X}_j = \frac{\overline{X}_{j,6th} + \overline{X}_{j,7th} + \overline{X}_{j,8th} + \overline{X}_{j,10th}}{4}$$