The Evolution of Moral Foundations

Michelle Avataneo, Thomas Norman
and Nicola Persico*

November 20, 2023

Abstract

Moral Foundations Theory is an influential empirical description of moral perception. According to this theory, individuals make moral judgments based on five distinct "moral foundations:" Care, Fairness, Loyalty, Authority and Sanctity. We provide a theory that explores the claimed evolutionary basis for these moral foundations. The theory conceptualizes these five moral foundations as specific modifications of fitness preferences in a 2×2 game. We find that the five foundations are distinguishable from each other and evolutionarily stable. However, the five foundations are not a minimal set: strict subsets of the five moral foundations suffice to describe all preferences that are evolutionarily stable. Not all foundations that are evolutionarily stable need deliver fitness improvement over the equilibrium in the fitness game: we characterize which do. Finally, we study moral overdrive, i.e., the situation in which the moral component of preferences totally dominates fitness and drives decision making entirely. While every one of the five foundations is compatible with moral overdrive in at least one fitness game, there is no fitness game in which moral overdrive is compatible with fitness improvement.

Journal of Economic Literature Classification: C73; D91.

Key Words: morality; preference evolution; indirect evolution.

*Avataneo: Northwestern University; Norman: Magdalen College, Oxford; Persico: Northwestern University. The paper benefited from discussions with Jeff Ely, Tim Feddersen, Peyton Young.

1 Introduction

There are reasons to believe that human morality has been shaped at least partly by evolutionary pressures. Very young babies show proto-moral behavior, suggesting that certain moral principles are partly innate and, hence, potentially inheritable.[1] In addition, twin studies find that several distinct moral principles are inherited.[2] This paper asks: what kind of moral principles emerge from and survive an evolutionary process?

We focus on five distinct moral principles that have been identified by recent empirical scholarship. According to Moral Foundations Theory (Graham et al., 2013), five distinct "moral foundations" – *Care*, *Fairness*, *Authority*, *Loyalty*, and *Sanctity* – anchor moral judgments and influence behavior. These five principles are said to be selected by evolution,[3] however, no rigorous theoretical justification has been offered for this claim. Empirically, these principles are elicited through surveys,[4] and individuals vary in the emphasis they place on each of these principles.[5] Though some scholarship argues that this variation can be reduced to just two principal components rather than five separate foundations, there is agreement that, empirically, individuals vary in their moral makeup along *at least two* dimensions.[6]

At odds with this multidimensional moral landscape, the seminal game-theoretic contributions on the evolution of morality, (Bester and Guth, 1998; Alger and Weibull, 2013, 2016, 2017; Alger et al., 2020) identifies a single dimension, i.e., a single moral principle that, moreover, all individuals who interact with each other possess in the same degree. Existing theory, therefore, is mono-factor and predicts no individual

[1] For example, babies as young as 34 hours old cry reflexively when exposed to crying sound, suggesting a concern for others; and 5-month-old infants prefer (reach to) an adult Helper of a puppet to a Hinderer, and prefer appropriately antisocial characters (who harm Hinderers) over inappropriately prosocial ones (who help Hinderers), actions which are consistent with moral evaluation. See Hamlin (2013).

[2] Zakharin and Bates (2023) show that monozygotic twins are more "morally similar" than dizygotic twins.

[3] Graham et al. (2013) p. 60 state that the five foundations are "the concerns, perceptions, and emotional reactions that consistently turn up in moral codes around the world, and for which *there are already-existing evolutionary explanations*" (emphasis added).

[4] Various survey instruments are made available at https://moralfoundations.org/questionnaires/. See also Graham et al. (2011).

[5] This variation across people is referred to as "moral pluralism." Intriguingly, the degree to which people emphasize certain principles has been found to correlate with political leanings, with Liberals (in the US sense of the term) most focused on Care and Fairness whilst Conservatives valuing each foundation more evenly (Graham et al., 2009).

[6] Zakharin and Bates (2021) review the "number of dimensions" literature.

variation. In this paper, we build on the empirical groundwork laid by Moral Foundations Theory in order to produce a mathematical theory in which several moral principles can coexist, meaning that different individuals who interact with each other can hold different moral principles, and where *at least two* moral principles are evolutionarily stable. Our exercise, then, provides a theoretical underpinning for the empirically observed multidimensionality of the moral landscape.

The model is as follows. Players repeatedly play a two-by-two symmetric game with randomly drawn opponents. The symmetric game is fully characterized by the row player's fitness matrix, denoted by Π. However, in playing the game, players do not maximize Π but, rather, $\Pi + m(\Pi)$. The matrix-valued function $m(\cdot)$ will be referred to as a "moral principle." Different functions m capture different moral principles; for example, a function m captures the *Fairness* principle if, in the game induced by Π, the matrix $m(\Pi)$ adds rewards (positive utils) on the outcomes in which players have more-similar payoffs. The evolutionary stability concept operates on moral principles, and it is the same as in Dekel et al. (2007): we say that "m *is evolutionarily stable for* Π" if $\Pi + m(\Pi)$ is evolutionarily stable. [Note: evolutionary fitness is based on Π only, not on $\Pi + m(\Pi)$].

If m is evolutionarily stable for Π, a population that maximizes $\Pi + m(\Pi)$ achieves a fitness level that is no lower, and sometimes higher, than a population that maximizes Π. It may be surprising that non-fitness maximizing behavior could be evolutionarily stable; consider, however, that m provides players with potentially pro-social commitment power. The reason that this commitment power is not exploited by mutants endowed with some arbitrary $\tilde{m}$, is that incumbents are able to detect mutants and play differently against them. For commitment power to be evolutionarily stable, then, the players' (including the mutants') moral principles must be observable.[7] We defend the assumption that moral principles are observable at page 25, based on the notion that, in the small communities where morality may have evolved historically, a person's moral principles may have been observable for practical purposes. Empirical evidence supports the idea that moral principles are common knowledge among acquaintances.[8]

[7] Indeed, Dekel et al. (2007), Section 4 show that, if moral principles are not observable, all evolutionarily stable equilibrium behavior reduces to the Nash equilibrium set of Π.

[8] Helzer et al. (2014) asked subjects to rate their own moral character using a multi-trait measure; then, the subjects' friends, family members, and acquaintances rated the subjects on the same traits. The authors find substantial self/other and inter-judge agreement.

In some game forms, however, m may be evolutionarily stable and yet a population that maximizes $\Pi + m(\Pi)$ behaves the same as a population that maximizes Π alone. In other words, m does not affect behavior except in play against mutants. In these games, morality evolves not because it improves the fitness of the outcome, but because it is not a hindrance in the competition against mutants.

We allow evolution to operate on the space of *all* moral principles, i.e., all matrix-valued functions $m(\cdot)$. But this is a very rich space, and not all different functions $m(\cdot)$ give rise to different behavior. A focal (and interpretable) subset of this space is represented by the five foundations from Moral Foundations Theory. Accordingly, we define five different mathematical functions $m(\cdot)$ that operationalize *Care*, *Fairness*, *Authority*, *Loyalty*, and *Sanctity* generating, for each different Π, five potentially different $m(\Pi)$'s – each of which may or may not be stable. We refer to these five principles collectively as the "5m's" (m stands for "moral principle"). We ask:

1. (distinguishability) Are the 5m's distinct from each other, i.e., do different elements m in the 5m's give rise to strategically different matrices $\Pi + m(\Pi)$? In other words, for each pair in the 5m's, is there an opponent that induces different behavior?

2. (spanning) For every fitness matrix Π, is every evolutionarily stable matrix $m(\Pi)$ generated by a moral principle m that is in the 5m's?

3. (minimal spanning) Is the set of the 5m's minimal, or are there strict subsets of the 5m's that generate all evolutionarily stable matrices $m(\Pi)$?

4. (moral code design) Not all evolutionarily stable matrices $m(\Pi)$ lead to fitness improvement in equilibrium play. Suppose a planner sought to design a stable minimal moral code that guaranteed all possible fitness improvements: which subset of the 5m's would the planner need to include in the code?

5. (moral overdrive) Can the matrix $m(\Pi)$ be very "large" relative to Π and still be evolutionarily stable? That is, what moral principles can be "blown out of proportion" and still be evolutionary stable?

The answer to question 1 is yes: for any pair of principles m and m' in the 5m's, there exists a fitness matrix Π such that $\Pi + m(\Pi)$ is *not* strategically equivalent to $\Pi + m'(\Pi)$. Therefore, any two elements of the 5m's are pairwise distinguishable

(Theorem 1). This is non-obvious because it goes against the "counting intuition" that there are many 5m's (to wit, five) compared to the small complexity of behavior that can be encoded in a 2×2 matrix $m(\Pi)$. If any two elements of the 5m's were not distinguishable, our 2×2 setting would be too low-dimensional to allow for five *different* moral principles.

The answer to question 2 is yes as well; this means that the 5m's are, generically, "rich enough" to generate all $m(\Pi)$'s that are evolutionarily stable for any given Π (Theorem 2). But, the 5m's are *not the smallest set* that is rich enough: Theorem 3 identifies two proper subsets of the 5m's each of which is rich enough to generate all $m(\Pi)$'s that are evolutionarily stable for any given Π. Therefore, the answer to question 3 is no: the 5m's are not the minimal set that generates all evolutionarily stable moral principles.

Question 4 is answered in Theorem 4: a designer who seeks to maximize fitness by designing a moral code based on the fewest number of principles in the 5m's will only need to include *Fairness* or *Loyalty* in the moral code.

The answer to question 5 is yes: in some games, moral principles can be "blown out of proportion" and still be evolutionary stable (Lemma 2). However, when this is the case, these moral principles do not improve fitness relative to equilibrium play with Π (Theorem 5).

This paper contributes to the literature on moral psychology by proving theoretical support for three commonly made claims: that people can be moved by several (i.e., more than one) distinct moral principles; that people in the same society vary in the emphasis they place on these principles; and that these principles emerge from an evolutionary process. We provide a rigorous theoretical model in which these three claims hold simultaneously. It's worth noting that our analysis relies on an individual selection mechanism, and not on group selection *à la* Bowles and Gintis (2011): we elaborate on this in Section 4. There, we also discuss the way that our mechanism differs from "internalized norms."

Our paper also contributes to the empirical debate concerning how many distinct moral principles there are (see, e.g., Zakharin and Bates (2021)). The creators of Moral Foundations Theory (henceforth, MFT) hold that there are five foundations, but others hold that, empirically, the majority of cross-person variation in the response to the MF Questionnaire can be summarized by just two high-level factors: a so-called individualizing one (overlapping with Care and Fairness) and a so-called binding

one (overlapping with Loyalty, Authority, and Sanctity). We find that, while all five foundations postulated by MFT withstand the test of an evolutionary process, nevertheless the set of five foundations is not minimal in the sense that all stable behavior in all games can be generated by a proper subset of the five foundations. In fact, we identify two different proper subsets of the five foundations that are sufficient to generate all stable behavior in all games. Intriguingly, each of these two subsets happens to contain one individualizing and one binding foundation, arguably consistent with the two-factor reading of the empirical evidence.

In this paper, we adopt the "indirect evolutionary approach" (Güth, 1995; Güth and Yaari, 1992; Dekel et al., 2007) to preference evolution.[9] This approach assumes that individuals play rationally for given preferences, but that these preferences are subject to evolutionary selection according to the "biological fitness" of the behavior that they induce. The indirect evolutionary approach can generate departures from fitness-maximizing preferences because of their strategic effect on opponents (Heifetz et al., 2007). But this breaks down if preferences are unobservable (Ok and Vega-Redondo, 2001; Ely and Yilankaya, 2001; Güth and Peleg, 2001), at least if the standard assumption of random matching is maintained. By contrast, if assortative matching is assumed under incomplete information, then preferences depart from fitnesses once more, in the direction of "Kantian" preferences: players place some weight on the action that, if played by both players, would result in maximal fitness (Alger and Weibull, 2013). This finding has spawned a literature exploring what moral behavior is stable (Alger and Weibull, 2016, 2017; Alger et al., 2020), but its operation is quite different from the complete-information analysis of the current paper and, as a result, these papers yield a mono-factor theory of morality. Our analysis builds on, but is different to the framework laid out by Dekel et al. (2007) because their analysis focuses on the preference matrix P, which they interpret as a generic "subjective preference." In contrast, our analysis focuses on the matrix $P - \Pi$: we define functions m that produce $P - \Pi$ and are interpretable as moral foundations. Therefore, we are able to ask (and answer) questions 1-5, which are based on m and not on P.

[9] On preference evolution in general, see Robson and Samuelson (2011) and Alger (2023).

2 The Model

Members of a large (but finite) population are repeatedly matched at random to play a two-player two-action normal-form game G, which has action set $\{A, B\}$ with mixed action being denoted by Δ. The 2×2 *fitness matrix*

$$\Pi = \begin{array}{|c|c|} \hline a & b \\ \hline c & d \\ \hline \end{array}$$

measures a player's fitness, or reproductive success, resulting from the players' action profiles. Every agent in the population has the same Π, and we examine all possible configurations of Π. A particular ordering of $\{a, b, c, d\}$ induces a *fitness game* or, with some abuse of notation, a *game*; for example, if $c > a > d > b$, we refer to Π as "a Prisoners' Dilemma game." By convention, and without loss of generality, we assume $a \geq d$. We think of Π as capturing the "material" payoffs earned by the players, separate from their moral principles.

The players do not care directly about fitnesses, but instead have the more familiar (von Neumann–Morgenstern) preferences over G's *outcomes*, which are probability distributions on $\{A, B\} \times \{A, B\}$. The space of all 2×2 matrices $\mathbb{M}$ captures the set of all possible utility functions on $\{A, B\} \times \{A, B\}$. A generic element $P \in \mathbb{M}$ captures a player's preferences. If a player with *preference matrix* P plays the mixed action σ against an opponent playing σ', then she receives expected utility $P(\sigma, \sigma')$. We think of P as capturing the player's overall motivation, which reflects both material payoffs and any moral principles.

The preferences of a player's opponent are randomly drawn from the population's preference distribution, generically denoted by the probability distribution μ on $\mathbb{M}$. We assume that matched players observe one another's preference matrices; this assumption is discussed at page 25.[10]

A *strategy* for a player with preference matrix P is a function $\sigma_P : \mathbb{M} \to \Delta$ that specifies a mixed action conditional on the preference matrix of the matched opponent. The choice of strategies by members of the population μ then defines a complete-information population game $\Gamma(\mu)$, and we assume that the players play a Nash equilibrium of this game, i.e., $b_P(P') \in \arg\max_{\sigma \in \Delta} P(\sigma, b_{P'}(P))$ for each $P, P' \in$

[10] Dekel et al. (2007) also analyze scenarios with incomplete and partial information.

$\mathbb{M}$.[11] Let $B(\mu)$ denote the set of Nash equilibrium strategy profiles of $\Gamma(\mu)$.

Given a population distribution μ and an equilibrium strategy profile $b \in B(\mu)$, the average fitness of a player with preference matrix $P \in \text{supp}(\mu)$ is

$$\bar{\Pi}_P(\mu \mid b) = \sum_{P' \in \text{supp}(\mu)} \Pi(b_P(P'), b_{P'}(P)) \cdot \mu(P').$$

2.1 Moral proclivity matrices and moral principles

Our object of interest is the difference between P and Π, which we will interpret as a "moral proclivity." To this end, we introduce notation for a matrix M which, when added to Π, produces P.

Definition 1 (**moral proclivity matrix**). *A 2×2 matrix with non-negative entries $\{0, x, y\}$ is a moral proclivity matrix if it has at least one zero in each column and, by convention, x in the left-hand column. We denote a moral proclivity matrix by $M(x, y)$.*

The moral proclivity matrix $M(x, y)$ will be added to Π, a player's fitness matrix, to obtain P, the player's preference matrix. The moral proclivity matrix captures the "moral" aspect of a game that make a player behave differently than would be dictated by Π alone. Since Definition 1 implies that x and y are nonnegative, $M(x, y)$ represents rewards, not punishments. Note that there is no loss of generality (and considerable economy of parameters) in requiring that two entries of a moral proclivity matrix equal zero. Indeed, the sole function of the moral proclivity matrix is to change a player's best response and two strategically placed numbers, x and y, suffice to achieve *any* best response.

Moral principles are rules that take as input any fitness matrix Π and output a moral proclivity matrix.

Definition 2 (**moral principle**). *A moral principle $m(x, y; \cdot)$ is a function that maps a fitness matrix Π into moral proclivity matrix $M(x, y)$.*

Figure 1 illustrates the relationship between: the fitness matrix Π; a moral principle $m(x, y; \cdot)$; the moral proclivity matrix $M(x, y)$ generated by the moral principle; and preferences P. The fitness matrix Π induces a symmetric game whose payoffs

[11] Note that this definition implies that if two players have the same P, they play the same action in equilibrium.

drive reproductive success. However, instead of maximizing Π, the agent maximizes preferences P, which result from a combination of fitness and moral proclivities. Moral proclivities, in turn, are generated by moral principles, i.e., general rules that apply to any given Π.

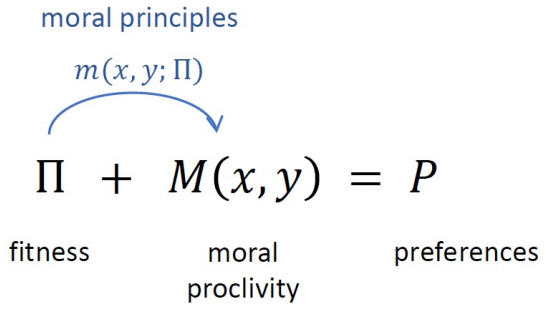

Figure 1: Relationship between Π, $m(x, y; \cdot)$, $M(x, y)$, and P.

We shall be interested in which moral principles m survive the evolutionary process in the game induced by a given Π. For example, we will be interested in whether the moral principle $m =$*Care* survives the evolutionary process when Π induces a Prisoners' Dilemma game.

2.2 The 5m's

Moral principles could be very complicated functions of x, y, and Π. Next, we describe five principles that were empirically identified by Moral Foundations Theory: *Care*, *Fairness*, *Authority*, *Loyalty*, and *Sanctity*. We call these principles the "5m's."

To operationalize the 5m's as functions $m(x, y; \Pi)$, we need to introduce the concept of a Kantian action.

Definition 3 (**Kantian action**). *The Kantian action is the pure-strategy action which, when taken by both players, yields the highest total fitness.*

Since, by assumption, $a \geq d$, the Kantian action is A. Table 1 illustrates how we operationalize the 5m moral principles. Appendix A provides mathematical definitions of the 5m's.

Operationalizing the "5m" moral principles

Moral principle	**Operational rule: boost with x or y the entry which ...**
Care $(x, y; \Pi)$	for each given opponent's action makes the opponent best off
Fairness $(x, y; \Pi)$	for each given opponent's action has the most-equal payoffs
Authority $(x, y; \Pi)$	for each given opponent's action makes me best off relative to the opponent
Loyalty $(x, y; \Pi)$	rewards the opponent who took the Kantian action and punishes the opponent who did not
Sanctity $(x, y; \Pi)$	corresponds to the Kantian action regardless of the opponent's action

Table 1: Operationalizing Moral Foundations Theory. Appendix A provides mathematical definitions of the 5m's.

A few comments on how we operationalize the moral foundations. In Table 1, *Care* is operationalized as a concern for the material well-being (i.e., fitness) of the other player, and *Fairness* is operationalized as a preference for equality of outcomes. We regard these operationalizations as consistent with the literature and relatively uncontroversial.[12] Our definition of *Authority* captures a preference for high relative status.[13] Admittedly, this preference is only part of what goes on in a hierarchical or power relationship; the element of respect for hierarchy, or deference to power, is missing from the definition. However, the model is too stylized to embed a hierarchical relationship so our definition is, of necessity, limited. Turning to *Loyalty*, the most natural definition would be a preference for behaving kindly toward those who one considers part of an "in-group," and spitefully towards those who one considers part

[12]For example, Bester and Guth (1998) and Miettinen et al. (2020) operationalize "altruism" in the same way as we do *Care*. And Fehr and Schmidt (1999) operationalize "inequality aversion" in the same way as we do *Fairness*. Refer to Appendix B for a proof of these statements.

[13]Note that *Authority* is *not* the opposite of *Fairness*: the latter is about the *absolute value* of the difference in the players' fitness, whereas the former is about the sign of the difference. Another way of seeing the difference: the opposite of *Fairness* is a preference for inequality, whereas *Authority* is a preference for *favorable* inequality.

of an "out-group". Lacking a definition of in- and out-groups, we approximate this notion as for behaving kindly toward players who *behave virtuously from the group's perspective* and, conversely, being spiteful toward players who don't.[14] Finally, in the context of Moral Foundations Theory, the term *Sanctity* is used not necessarily in reference to organized religion but, rather, 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 in a given culture are in fact pro-social or, rather, they emerge arbitrarily out of cultural accident. Our operationalization of *Sanctity* adopts the pro-social view.

The set **5m** is comprised of all the moral principles listed in Table 1. Formally,

$$\mathbf{5m} = \{m(x, y; \cdot) \mid m \in \{\textit{Care, Fairness, Loyalty, Authority, Sanctity}\};\, x, y \geq 0\}\,.$$

Appendix A provides mathematical definitions of the 5m's. Appendix B highlights previous papers in which our operationalizations of different moral principles have previously surfaced, suggesting that our operationalizations are not entirely idiosyncratic. The next example illustrates the 5m's in the context of a coordination game.

Example 1 (the 5m's moral principles in a coordination game). *The fitness matrix*

$$\Pi = \begin{array}{|c|c|}\hline 2 & -1 \\ \hline 1 & 0 \\ \hline \end{array}$$

induces the coordination game

$$\begin{array}{c|c|c|} & A & B \\ \hline A & 2,2 & -1,1 \\ \hline B & 1,-1 & 0,0 \\ \hline \end{array}\quad .$$

[14] If we interpret the out-group as "someone who does not choose the Kantian action," then our definition of *Loyalty* may be interpreted as "loyalty to ideals of behavior that our group deems to be fitness-maximizing," where "our group" means "those whose fitness is determined by Π." In this interpretation, the in-group are those whose follow these ideals of behavior, and the out-group are those who follow different ideals (those where B is fitness-maximizing). Our definition of *Loyalty* prescribes that the former should be rewarded, and the latter punished. There is some evidence that young children behave this way: Hamlin (2013) presents experimental evidence that children reach out for puppets that behave pro-socially, and punish puppets that behave anti-socially.

The moral principle Care gives extra utils to the entry which, for each given opponent's action, makes the opponent best off. Since $2 > -1$ *and* $1 > 0$*, applying this moral principle to the coordination game induced by* Π *produces the following moral proclivity matrix:*

$$Care\,(x, y; \Pi) = \begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array}.$$

The moral principle Fairness gives extra utils to the entry which, for each given opponent's action, has the most-equal payoffs. Since, $|2-2| < |1-(-1)|$ *and* $|0-0| < |-1-1|$*, applying this moral principle to the coordination game induced by* Π *produces the following moral proclivity matrix:*

$$Fairness\,(x, y; \Pi) = \begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array}.$$

The moral principle Authority gives extra utils to the entry which, for each given opponent's action, makes the row player best off relative to the opponent. Since, $1 - (-1) > 2 - 2$ *and* $0 - 0 > -1 - 1$*, applying this moral principle to the coordination game induced by* Π *produces the following moral proclivity matrix:*

$$Authority\,(x, y; \Pi) = \begin{array}{|c|c|}\hline 0 & 0 \\ \hline x & y \\ \hline \end{array}.$$

Applying the moral principle Loyalty requires identifying the Kantian action, which is A by assumption. Loyalty requires boosting the utils of the row player who: (a) conditional on the opponent playing A, picks the action that makes the opponent best off; and (b) conditional on the opponent playing B, picks the action that makes the opponent worse off. Since $2 > -1$ *and* $0 < 1$*,*

$$Loyalty\,(x, y; \Pi) = \begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array}.$$

Finally, the moral principle Sanctity requires boosting the utils of the row player

who plays the Kantian action A, so:

$$Sanctity\,(x, y; \Pi) = \begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array}.$$

Note that, for this Π*, not all moral principles generate different moral proclivity matrices: for example, Fairness and Loyalty happen to produce the same moral proclivity matrix. Therefore, for this* Π*, Fairness and Loyalty are not distinguishable. We will return to the issue of distinguishability in Section 3.1.*

2.3 Stability

We adopt the notion of stability introduced by Dekel et al. (2007). Technically, this definition applies jointly to population preference distributions and equilibrium strategy profiles, but the gist of it is that preferences are stable if they survive the entry of players with new preferences. We lay out their definition next.

The first ingredient of a stable configuration (μ, b) is that all incumbents must receive the same fitness, a property referred to as *balance*. Throughout the paper, agents with preferences in $supp(\mu)$ are referred to as an incumbents, and an agents with preferences not in $supp(\mu)$ are referred to as a mutants.

Definition 4 (**balanced configurations**). *A configuration* (μ, b) *is* balanced *if*

$$\bar{\Pi}_P(\mu|b) = \bar{\Pi}_{P'}(\mu|b) \text{ for all } P, P' \in supp(\mu).$$

Intuitively, a configuration (μ, b) is *balanced* if the average fitness of all incumbents is the same. The second ingredient is the stabilizing equilibrium selection for incumbent play. Given an original configuration (μ, b) and a mutant preference $\tilde{P}$, let $N_\varepsilon(\mu, \tilde{P}) = \{\mu' : \mu' = (1 - \varepsilon')\mu + \varepsilon'\tilde{P}, \varepsilon' < \varepsilon\}$ denote the set of all preference distributions resulting from entry by at most ε mutants.

Definition 5 (**focal strategy profiles**). *Given* $\tilde{\mu} \in N_\varepsilon(\mu, \tilde{P})$*, an equilibrium strategy profile* $\tilde{b} \in B(\tilde{\mu})$ *is* focal *if* $\tilde{b}_P(P') = b_P(P')$ *for all* $P, P' \in supp(\mu)$*.*

Intuitively, an equilibrium profile is focal if, when playing against each other, incumbents do not switch Nash equilibria as a function of the prevailing mutant population. Let $B(\tilde{\mu} \mid b)$ be the set of all focal equilibrium profiles relative to b if the population distribution is $\tilde{\mu}$.

Definition 6 (**stable configurations**). *A configuration* (μ, b) *is* stable *if it is balanced and if there exists* $\varepsilon > 0$ *such that, for every* $\tilde{P} \in \mathbb{M}$ *and* $\tilde{\mu} \in N_\varepsilon(\mu, \tilde{P})$,

$$\bar{\Pi}_P(\tilde{\mu} \mid \tilde{b}) \geq \bar{\Pi}_{\tilde{P}}(\tilde{\mu} \mid \tilde{b}) \quad \textit{for all } \tilde{b} \in B(\tilde{\mu} \mid b) \textit{ and } P \in \textit{supp}(\mu).$$

Thus, at a stable configuration, incumbents must do no worse than any mutant in any post-mutation focal equilibrium.[15] Next, we define stable preferences (Definition 7 below, part 1). In addition, since we want to talk about morality, it helps to define stable moral principles and stable moral proclivity matrices (Definition 7, parts 2 and 3).

Definition 7 (**stability**).

1. *A preference* P *is stable for* Π *if* $P \in \textit{supp}(\mu)$ *for some stable configuration* (μ, b).
2. *A moral proclivity matrix* $M(x, y)$ *is stable for* Π *if the preference* $\Pi + M(x, y)$ *is stable for* Π.
3. *A moral principle* $m(x, y; \cdot)$ *is stable for* Π *if the moral proclivity matrix* $m(x, y; \Pi)$ *is stable for* Π.

For any given Π, Dekel et al. (2007) characterize all preferences P that are stable; we list them in appendix C. In this appendix we also report, for each Π, the stable moral proclivity matrices and the moral principles that generate them.[16]

[15] Since the set $B(\tilde{\mu}|b)$ of focal equilibrium profiles (relative to b under $\tilde{\mu}$) is always nonempty for observable preferences, the second part of Dekel et al. (2007) stability Definition 3 does not apply, leading to the complete-information version of stability stated above.

[16] Dekel et al. (2007)'s is a static stability concept is in the spirit of evolutionary stability (Maynard Smith and Price, 1973; Maynard Smith, 1974), but with some modifications to alleviate (though not remove) existence problems in the preference evolution setting. Specifically: given the equivalent play frequently induced by multiple types, "neutral stability" (Maynard Smith, 1982) is much better suited to identifying stable preferences, so that mutants may not have *strictly* larger payoffs than incumbents in the equilibrium that their entry induces; and given the inherently destabilizing force of multiple equilibria for given preferences, it is natural to make a stabilizing equilibrium selection for incumbent play.

3 Results

3.1 Distinguishability of the 5m's

According to Moral Foundations Theory, the 5m's as empirically distinguishable from each other. But are they, within our model? That is, do they generate different behavior? For a given Π, the answer is no: Example 1 shows that several different 5m moral principles generate the same moral proclivity matrix. This makes sense: in any given 2×2 game, we cannot expect five *different* best responses. In what sense, then, can the 5m's be said to generate different behavior? Theorem 1 below provides the answer.

In order to state Theorem 1, we first need to define strategic equivalence. Here is a formal definition.

Definition 8 (**strategic equivalence**). *Two preferences P and P' are said to be strategically equivalent if their best response correspondences are the same.*

Strategic equivalence means that, for every action taken by the opponent, the player's best response(s) under P and P' are the same. We are now ready to state Theorem 1.

Theorem 1 (**pairwise distinguishability of the 5m's moral principles**). *For any pair of the 5m's moral principles, there exists a fitness matrix Π such that the two moral principles span stable preferences that are not strategically equivalent.*

Proof. See the appendix D.2. □

Theorem 1 says that any two 5m principles generate different behavior *in some game* – even though they cannot generate different behavior *in all games*. This is a good sanity check because, if any two elements of the 5m's were not distinguishable, our 2×2 setting would be too low-dimensional to allow for five *different* moral principles.

The empirical content of Theorem 1 is revealed in equilibrium play. When two moral principles m and m' generate preferences that are not strategically equivalent, there exists a mutant (or possibly an incumbent) against which m prescribes a different equilibrium play than m'.

3.2 The 5m's span the set of stable preferences

Let's start by setting expectations: it is not obvious that the 5m's should be able to generate all evolutionarily stable preferences P for all Π's, because the 5m's are just a few of all possible moral principles. Indeed, the next lemma shows that the 5m's are not rich enough to generate all possible preferences P. In order to state the lemma, we must first define the concept of spanning.

Definition 9 (**spanning**). *A moral principle m spans P for a given Π if $\Pi + m(x, y; \Pi)$ is strategically equivalent to P for some $x, y \geq 0$.*

Intuitively, the moral principle m spans P if an agent with preference $\Pi + m$ best-responds exactly like an agent with preference P would against any possible opponent.

Lemma 1 (**preference matrices generated by the 5m's do not span the space of all matrices**). *There exists a pair Π, P such that no moral principle in the 5m's spans P.*

Proof. See appendix D.3. □

This lemma says that, for the purpose of generating preferences P, there *is* loss of generality in restricting attention to the 5m's. Hence, it is not obvious that the 5m's would be able to generate *all* the evolutionarily stable preferences P *for all* Π's. The fact that they do (Theorem 2) is, therefore, not trivial.

Theorem 2 is the main result of this subsection. It shows that generically, i.e., for almost any fitness matrix Π, the 5m's moral principles are sufficient to span almost all of its *stable* preferences. To state Theorem 2, we must first introduce a notion of genericity.[17]

Definition 10 (**non-genericity**). *Given two sets A, B belonging to a vector space V with $A \cap B \neq \emptyset$, A is non-generic in B if it has lower dimension than B. If B is the parent space to which A belongs, A is said to be simply non-generic.*

A property holds non-generically in a set if it holds only on a subset of lower dimension than the whole set.[18] For example, a point is non-generic in the real line, but an

[17] A similar notion of genericity is used by Govindan and Wilson (2012).

[18] Dimension here is the standard notion of dimension in a vector space, namely the cardinality of its basis; or in the case of a subset A of V, the cardinality of the basis of the smallest subspace of V that contains A.

interval $[a, b]$ is not; the real line is non-generic in $\mathbb{R}^2$, but a square $[a, b] \times [a, b]$ is not. In the space $\mathbb{M}$ of 2×2 matrices, meanwhile, whereas the set of all Hawk-Dove games[19] is generic because it has the same dimension (four) as $\mathbb{M}$, the following game has dimension three and is hence non-generic.

Example 2 (**pure-common value Hawk-Dove game is non-generic**). *The fitness matrix*

$$\begin{array}{|c|c|}\hline a & b \\ \hline b & d \\ \hline \end{array},$$

with $b > a \geq d$, is a pure-common value Hawk-Dove game. The set of all such matrices is non-generic in the space $\mathbb{M}$ of 2×2 matrices.

Rather than applying the concept of genericity to set of matrices in $\mathbb{M}$, it will be expositionally convenient to apply this notion to graphs in $\mathbb{M}^2$ that map fitness matrices into preference matrices.[20] We use this graph as a convenient litmus test for non-genericities both in the space of fitness matrices and in the space of preference matrices. This is because a graph is non-generic in $\mathbb{M}^2$ if either its domain or its range are non-generic in $\mathbb{M}$; so a graph "inherits" any non-genericities in its domain and range.[21] To this end, define the following graph:

$$\mathcal{ES} = \{(\Pi, P) \in \mathbb{M}^2 \mid P \text{ is stable for } \Pi\},$$

which is the graph of the set of evolutionarily stable preferences. Let $5\mathcal{M}$ be the graph of the set of preferences spanned by the 5m's, i.e.

$$5\mathcal{M} = \{(\Pi, P) \in \mathbb{M}^2 \mid P = \Pi + m(x, y; \Pi),\ m \in \mathbf{5m},\ x, y \geq 0\}.$$

We are now ready to state this subsection's main result.

Theorem 2 (**the 5m's generically span the stable preferences**). *The graph $\mathcal{ES} \backslash 5\mathcal{M}$ is non-generic in $\mathcal{ES}$.*

Proof. See appendix D.3. □

[19] This is the set of all fitness matrices with $c > a$ and $b > d$.

[20] Note that the dimension of a set in the product space $\mathbb{M}^2$ is simply the sum of the dimension of its projection on each constituent copy of $\mathbb{M}$.

[21] In light of Example 2, therefore, any mapping from the set of pure-common value Hawk-Dove games to any subset of $\mathbb{M}$ is non-generic.

The interpretation of Theorem 2 is that the 5m's are "rich enough" to transform *almost all* fitness matrices Π into *almost all* their associated stable preferences P. Put differently, the 5m's are capable of spanning almost all stable preference matrices for almost all Π's. This result stands in contrast to Lemma 1, which shows that the 5m's are *not* capable of spanning all (including non-stable) preference matrices.[22]

Next, we develop some intuition for why just a few moral principles suffice to span all stable preferences (Theorem 2). Very broadly, the argument goes like this. Preference matrices come in four classes: A is a dominant strategy, B is a dominant strategy, main-diagonal dominant, anti-diagonal dominant. Conveniently, however, just two of these classes suffice to represent all evolutionarily stable preferences: the classes in expression (1). This reduction, which follows from the fitness maximization properties of Dekel et al. (2007)'s notion of stability, considerably reduces the set of moral principles necessary to span all stable preferences – so much so that the 5m suffice to span. This intuitive argument is developed next.

Evolutionary stability in the sense of Definition 7 is tightly connected to the notion of fitness. Articulating this connection requires stating the following definition.

Definition 11 (**fitness-maximizing preferences**)**.** *A preference matrix P is fitness-maximizing for Π if the Nash equilibrium set of P includes the mixed action σ that, when played by both players, maximizes the sum of their fitness payoffs.*

Intuitively, this definition says that a preference matrix is fitness-maximizing if its Nash equilibrium implements the "fittest" action. From Dekel et al. (2007), we know that a preference can only be stable for Π if it is fitness-maximizing for Π in the sense of Definition 11. Moreover, for almost all Π's that have a stable preference, the fitness-maximizing mixed action is in fact the pure strategy A,[23] and every preference matrix whose equilibrium set includes A is strategically equivalent to one of the following two matrices:

[22] The contrast between Lemma 1 (the 5m's can't span) and Theorem 2 (the 5m's span) is substantive – it is not a technical artifact of the fact that Theorem 2 only holds generically, whereas Lemma 1 is stated uniformly. Indeed, the non-spanning property in Lemma 1 happens to have "positive measure." For example, for Πs with $a > d > b > c$, which are generic in the space of 2×2 matrices, no principle from the 5m's can span any preference with $c > a$ and $b > d$, which are also generic in the space of 2×2 matrices.

[23] The only Π that has a stable preference for which the fitness-maximizing mixed action is not the pure strategy A is the common-value Hawk-Dove game of Example 2, which happens to be non-generic.

$$\begin{array}{c|cc} & A & B \\ A & x & y \\ B & 0 & 0 \end{array} \quad \text{or} \quad \begin{array}{c|cc} & A & B \\ A & x & 0 \\ B & 0 & y \end{array} \quad \text{for some } x, y \geq 0. \tag{1}$$

So, for almost all Π's that have a stable preference, every stable preference must look like one of the two matrices in (1). The 5m's are capable of spanning almost all these matrices for any Π. (In fact, even fewer than the 5m's suffice: the intuition for this will be explained in Section 3.3 below). Therefore, the 5m's are capable of spanning almost all stable matrices for almost all Π's.

Remark 1 (**non-genericities**). *What (non-generic) games are not covered by Theorem 2? The proof of Theorem 2 shows that there is only one class of games for which the entire set of stable preferences is not spanned by the 5m's. This is the pure-common value Hawk-Dove game of Example 2 which, of course, is non-generic in* $\mathbb{M}$. *In addition, there are (generic) games* Π *for which some, but not all of the stable preferences are not spanned by the 5m's. The non-spanned preferences all have the form*

$$\begin{array}{|c|c|} \hline e & f \\ \hline e & h \\ \hline \end{array},$$

which is, of course, non-generic in $\mathbb{M}$.[24] *Furthermore, for those* Π*'s, the non-spanned preferences are even non-generic in* $\mathcal{ES}(\Pi)$ *meaning, intuitively, that the quasi-totality of the stable preferences are in fact spanned by the 5m's.*

3.3 Minimal spanning moral principles

Theorem 2 shows, roughly, that the 5m's are sufficient to span the set of almost all stable preferences for almost all fitness matrices. But it does not say that the 5m's constitute a minimal set: a smaller set could, in principle, suffice to span almost stable preferences. This subsection shows that this is indeed the case.

Let $\mathcal{FS}$ be the graph of the set of preferences generated by $\{Fairness, Sanctity\}$, that is:

$$\mathcal{FS} = \{(\Pi, P) \in \mathbb{M}^2 \mid P = m(x, y; \Pi),\ m \in \{Fairness, Sanctity\},\ x, y \geq 0\};$$

[24] Depending on the values of f and h, this preferences specialize to, in the notation of Dekel et al. (2007), $\mathcal{BA}_1$, $\mathcal{AB}_1$, and θ_0.

similarly, let $\mathcal{CL}$ denote that of $\{Care, Loyalty\}$. Finally, let $\mathcal{D}$ denote the graph of the set of preferences generated by some subset D of the 5m's, that is:

$$\mathcal{D} = \{(\Pi, P) \in \mathbb{M}^2 \mid P = m(x, y; \Pi),\ m \in D \subseteq \mathbf{5m},\ x, y \geq 0\}.$$

The next theorem shows that there are two subsets D of the 5m's, each of which is sufficient to span almost all the evolutionary stable preferences for almost all fitness matrices. Moreover, any other subset with the same property must include at least one of these two subsets. In these sense, these two subsets are "minimal."

Theorem 3 (**the two minimal spanning pairs**). *If $\mathcal{D} \in \{\mathcal{FS}, \mathcal{CL}\}$, the graph $\mathcal{ES} \backslash \mathcal{D}$ is non-generic in $\mathcal{ES}$. Moreover, if the graph $\mathcal{ES} \backslash \mathcal{D}$ is non-generic in $\mathcal{ES}$, then $\mathcal{D}$ must contain $\mathcal{FS}$ or $\mathcal{CL}$.*

Proof. See appendix D.4 □

The theorem says that we can rationalize all evolutionarily stable moral behavior by appealing to fewer than five moral principles, to wit: either *Sanctity* and *Fairness*; or *Care* and *Loyalty*. Intriguingly, each of these two subsets happens to contain at least one *individualizing* component (including *Care* and *Fairness*) and one *binding* component (including *Loyalty*, *Authority*, and *Sanctity*), potentially consistent with the strand of the empirical literature that reduces moral pluralism to these two principal components, instead of five foundations.

Next, we present some intuition for why each spanning pair in Theorem 3 spans all stable preferences. The intuition is more transparent for the pair $\{Fairness, Sanctity\}$. For almost all Πs, *Fairness* boosts the main diagonal of Π,[25] just like the right-hand matrix in (1); and *Sanctity* boosts the top row of Π (recall that, by convention, A is the Kantian action) – just like the left-hand matrix in (1). Therefore, intuitively, *Sanctity* and *Fairness* "morph" the matrix Π toward the matrices in (1), which we know are the only stable preferences for all Π. The intuition is the same for the pair $\{Care, Loyalty\}$, the difference being that, depending on the specific Π, *Care* morphs the matrix Π toward either one of the two matrices in (1) – and *Loyalty* morphs Π toward the other matrix.

This intuition suggests that, in order to span all the stable preferences for a given Π, one needs two moral principles: one that generates *unconditional* moral proclivities (as

[25] The exception being the common value Hawk-Dove game of Example 2, which is non-generic.

in the the left-hand matrix in (1)), and another one that generates moral proclivities that are *conditional* on the opponent's action (as in the the right-hand matrix in (1)).

Finally, we note that *Authority* is not featured in Theorem 3. The reason is that, for every Π, *Authority* is only necessary to generate certain non-generic stable preferences, and Theorem 3 does not cover these non-generic preferences. In this sense, *Authority* plays a less-important role in our theory.

3.4 The design of fitness-improving moral codes

Consider a designer who, for any game Π, can pick an initial distribution over the 5m moral principles, i.e., an initial condition μ_0 from which a natural selection process drives the evolution of μ toward its stable rest points. We interpret μ_0 as a set of Π-specific moral values that is instilled early on in a person's life (perhaps through schooling) but, after being set, is subject to evolutionary pressures throughout that person's life.[26]

We know that, starting from any μ_0, the evolutionary process will necessarily achieve a fitness-maximizing point if it converges at all.[27] But suppose this designer wants to maximize fitness during the transition: what should μ_0 be set at? Intuitively, this means that the designer seeks to endow young people with a subset of 5m moral principles – we call this a *moral code* – that will serve them well through life and is evolution-proof, meaning that future experiences will not drive the young people away from the moral code. Furthermore, the designer seeks to minimize the number of distinct moral principles that are necessary to generate the right μ_0's across all games Π. We say that this designer seeks a *minimal* moral code.

What does a minimal moral code look like? For each Π, the moral code must include at least one stable moral principle in the 5m. In the Hawk-Dove game, it turns out, the only stable moral principles are *Fairness* and *Loyalty*, so any moral code must include at least one of these principles. Furthermore, *Fairness* and *Loyalty* happen to be stable in all games. Therefore, there are only two minimal moral codes: one based on *Fairness*, the other based on *Loyalty*. This observation is recorded in the following theorem.

[26] In this section, evolutionary selection is interpreted as an adaptive learning process that takes place throughout each person's life, boosting the moral values that increase fitness and withering those that don't, rather than as a process of genetic transmission across generations.

[27] This follows from Proposition 2 in Dekel et al. (2007).

Theorem 4 (**fitness-improving moral codes**). *There are only two minimal moral codes: one based on Fairness, the other based on Loyalty.*

Proof. See appendix D.5. □

It is worth contrasting Theorem 4 with the results in Alger and Weibull (2013). In their paper, the minimal moral code, i.e., the moral code that is stable in all fitness games, is referred to as *homo moralis*, corresponding, in our terminology, to *Sanctity*.[28] Of note, *Sanctity* is not featured in Theorem 4. This difference is not unexpected, of course, due to the very different settings.

3.5 Intensity of moral principles and moral overdrive

One can ask whether, given a particular game Π, the stable moral proclivity matrices M are "large," "small," or must be "just right." The case of interest here is when stable moral proclivity matrices M can be arbitrarily large. This is the gist of the next definition.

Definition 12 (**moral overdrive**). *Take a moral principle $m(x, y; \cdot)$ that, for some x, y, is evolutionarily stable for Π. If, for all numbers $k > 1$, $k \cdot m(x, y; \cdot)$ is evolutionarily stable for Π, we say that m is compatible with moral overdrive in game Π.*

Intuitively, moral overdrive means that it is evolutionarily stable for the moral component of preferences to swamp fitness as a decision-making criterion, and to totally drive decision making in game Π. Next, we show that every moral principle in the 5m is compatible with moral overdrive in at least one game Π.

Lemma 2. *Every moral principles in the 5m's is compatible with moral overdrive in at least one game Π.*

Proof. See appendix D.6. □

Even though all of the 5m's are compatible with moral overdrive in some game, not all stable moral principles (whether in the 5m or not) are compatible with moral overdrive in all games. In particular, there is an interesting class of games in which no stable moral principle is compatible with moral overdrive. This is the class of games

[28] In Appendix B.5 we trace the formal connection between *Sanctity* and *homo moralis*.

where moral principles actually improve equilibrium fitness. Fitness-improving moral principles are defined next.

Definition 13 (**strictly fitness-improving moral principles**). *A moral principle m is strictly fitness improving for Π if, for some $x, y \geq 0$, $\Pi + m(x, y; \Pi)$ is fitness-maximizing for Π whereas Π is not fitness-maximizing for itself.*

Intuitively, a moral principle is strictly fitness-improving if the symmetric Nash equilibrium among two players who adopt this principle generates more fitness than any symmetric Nash equilibrium among two players who play according to Π. Not many game types Π have strictly fitness-improving moral principles for the simple reason that, often, the best Nash equilibrium of Π is fitness-maximizing and, thus, not susceptible to improvement.[29] Only in the Prisoners' Dilemma and Hawk-Dove games is there a mixed action σ that, when played by both players, delivers a strictly higher fitness than the best symmetric Nash equilibrium.[30] If we restrict attention to these games, it turns out that no stable moral principle (including those not in the 5m's), is compatible with moral overdrive. Hence the next theorem.

Theorem 5 (**moral overdrive is incompatible with fitness improvement**). *There is no game Π in which a strictly fitness-improving moral principle is compatible with moral overdrive.*

Proof. See the appendix D.6. □

In a sense, Theorem 5 could be interpreted as saying that moral overdrive is "not needed" or "unhelpful." However, it's important to keep in mind that, in our theory, evolutionary stability is not predicated on "being needed" or "being helpful" in the sense of *strictly* improving fitness.

[29] Of course, even when the best Nash equilibrium of Π is fitness-maximizing, the presence of moral principles is detectable because it changes equilibrium play against mutants. For example, suppose $\Pi = \begin{array}{|c|c|}\hline 2 & -1 \\ \hline 1 & 0 \\ \hline \end{array}$. Π's fitness-maximizing equilibrium is (A, A). $\Pi + \textit{Sanctity}(x \geq 0, y > 1; \Pi) = \begin{array}{|c|c|}\hline 2+x & -1+y \\ \hline 1 & 0 \\ \hline \end{array}$, which from Appendix C.1 we know is stable for Π, does not induce a coordination game, it induces a game in which A is the dominant action. Hence, $\textit{Sanctity}(x \geq 0, y > 1; \Pi)$ changes the best response function relative to Π, but it does not change the fitness-maximizing equilibrium, which continues being (A, A).

[30] Definition 13 restricts attention to symmetric equilibria: this restriction is standard in the literature and is adopted by Dekel et al. (2007), among others.

4 Interpretation

In this section, we discuss several interpretive points of our theory.

Moral principles can evolve through genetic transmission, individual development, or interpersonal transmission The concept of evolutionary stability laid out in Definition 7 is abstract enough to encompass several possible channels of evolutionary transmission. First, and conceptually simpler, individuals who adopt a moral principle m that is unfit in some game Π will have below-average fitness on average across all the games they play, potentially fewer offspring, and so evolution will select against moral principle m in game Π. This genetic transmission mechanism is one way to account for the inheritability of moral principles documented by Zakharin and Bates (2023). For this mechanism to work, all individuals in society must play the same games with the same frequency.

Also, moral principles could be selected through individual development by reinforcement, from youth through adulthood, of those moral principles which perform well in each game type. This mechanism has a counterpart in theories such as Piaget's theory of moral development. We adopted this interpretation in Section 3.4. Finally, the moral principles of successful (in fitness terms) peers could be adopted through a process of social imitation. The theory we develop does not seek to discriminate between these three transmission mechanisms.

Morality is game-specific Regardless of the specific mechanism through which evolution operates, the model allows a moral principle to be stable in one fitness game type and not in another. For example, *Authority* is stable in some cooperation games but not in the Prisoners' Dilemma game. In this sense, stable moral principles are game-specific. Empirically, this means that evolution has predisposed us to apply different moral principles depending on the environment in which we operate. This seems realistic. With this being said, there are two moral principles, *Fairness* and *Loyalty*, each of which is stable for every Π. So, the accurate statement is that the *set* of moral preferences that are stable varies with Π.

What makes a moral principle evolutionarily stable Moral principles are not stable *because* they improve fitness. Indeed, it is possible for a moral principle to be stable without improving fitness: as mentioned following Definition 13, in many

fitness game types Π, the moral principles that are stable do not improve fitness above the best Nash equilibrium for Π. Hence, moral principles can be evolutionarily stable even if they do not improve fitness.[31] However, stable moral principles can never have lower fitness than the best Nash equilibrium for Π.[32] This observation helps put into perspective the role that morality plays in this paper.

While fitness maximization is to some extent "built into" Definition 7, our definition of stability, fitness maximization is not a sufficient condition for stability. For moral principles to be stable, they must also not be a liability in the competition against mutants.

Moral pluralism Moral pluralism is the notion that different individuals *in the same situation* (meaning, in our theory, the same fitness game) might hold different moral principles. Our results support this view because, in all fitness games, many moral principles in the 5m's are simultaneously evolutionarily stable. However, our theory does not support moral agnosticism: not all moral principles are evolutionarily stable in all fitness games.

Justifying the observability of moral principles An important assumption in our theory is that moral principles are observable to the opponent.[33] How valid is this assumption? If, as is plausible, the evolutionary process took place in small communities over many millennia, it makes sense that individuals in these communities might have a good sense of the moral makeup of the people they were interacting with. Not suprisingly, perhaps, Helzer et al. (2014) provide experimental support for the hypothesis that one's moral makeup is accurately assessed by those that one interacts with: the authors document that experimental subjects assessed their own moral character in a way that aligned with the assessment of the subjects' friends, family members, and acquaintances.[34]

We acknowledge that this justification for observability raises the possibility of

[31] For example, for Π's with $a > c$ and $b > c$, the only Nash equilibrium is fitness maximizing. Nonetheless, *Fairness* is stable for Π and, for some values of x, y, creates other Nash equilibria that are not fitness maximizing.

[32] This follows from Proposition 2 in Dekel et al. (2007).

[33] This is an important difference with Alger and Weibull (2013), who do not make this assumption. Rather, they assume assortative matching along similar "moral types."

[34] Unfortunately, the *Moral Character Questionnaire* used to elicit moral traits in Helzer et al. (2014) is somewhat different from Haidt's *Moral Foundations Questionnaire*. We are unaware of a similar study done using the latter questionnaire.

repeated interaction. Our analysis assumes that players play the one-shot Nash equilibrium, but if players interact repeatedly, more complex dynamic strategies are available. In a repeated interaction setting, our analysis continues to hold verbatim if players who interact repeatedly play (the same) "stage Nash" equilibrium in every game. In ongoing research, Avataneo explores a repeated interaction setting where moral principles evolve among players who play dynamic strategies, including trigger strategies.

Telling moral principles apart based on surveys An interpretation of Theorem 1 is that, in the real world, one could find a setting Π where a survey respondent would be able to articulate the difference between any two given moral principles in the 5m's. Moreover, Theorem 1 also says that the two moral principles may be chosen to be evolutionarily stable given Π. We interpret this property as saying that, in the real world, for any two elements of the 5m's, one could find a setting Π and two survey subjects in the population of evolutionary incumbents, who would answer the question "what do I feel is the right thing do in this case" differently.

Telling moral principles apart based on equilibrium play Although all evolutionarily stable moral principles prescribe the same action when playing against an opponent who holds an evolutionarily stable moral principle in that game,[35] the difference between stable moral principles can be detected in equilibrium play against mutants. This was shown in Section 3.1. In our theory, then, the difference between stable moral principles manifests, in action, when playing against individuals holding "uncommon" or "deviant" moral principles.

Moral dilemmas Moral dilemmas might be defined as situations in which two people holding different moral principles would choose different actions. In our theory, it is possible for two players i and i' holding moral principles m and m' to act differently

[35] This is because, by definition, all stable preferences must i) generate the same average fitness and ii) that average fitness must be larger than that of any mutant who enters the population in a small proportion. For these conditions to be satisfied, all stable preferences must play the fitness-maximizing outcome against each other. If the stable preferences where not all playing the fitness-maximizing action against each other, i) the average fitness wouldn't be the same, violating the same average fitness condition, or ii) there would be a mutant that does play the fitness-maximizing action against itself and plays against the incumbents whatever the incumbents are playing against each other, violating the larger average fitness than any mutant condition. For more detail see Dekel et al. (2007)'s Proposition 2.

in the same game Π, provided that they are playing against a (judiciously chosen) mutant.[36] In this admittedly abstract sense, our theory allows for the existence of moral dilemmas.

Different from internalized norms We define "internalized norms" as incentives that modify a player's fitness-maximizing behavior and exist only in the player's mind, not in the physical world. According to Coleman (1994),[37] not all norms are, or can be, internalized: for a norm to be internalized by agent i, there must be other agent(s) $-i$ who benefit from, and promote the internalization. In this view, norms are internalized by agent i "in the shadow" of physical-world pressure by other players who are affected by player i's behavior. This pressure may come from formal laws, and/or from informal punishment meted out in repeated interaction.

In a sense, moral principles in our theory could be interpreted as internalized norms, in that they modify a player's fitness-maximizing behavior and exist only in the player's mind. However, in our theory, it is not the presence of externalities that causes these psychological incentives to "embed" in the players' minds. In fact, in our theory, moral principles can be evolutionarily stable even if they do not improve the opponents' (or society's) equilibrium fitness. Moreover, in our theory, repeated interaction is not present.

One advantage of our evolution-based theory, over a theory where internalization pressure comes from punishment in repeated interaction, is predictive power. If the internalization pressure is formalized as out-of-equilibrium punishment by the opponent(s) in a repeated game,[38] a potential concern is that too many of player i's non-fitness maximizing strategies can be supported by a suitable choice of punishment by players $-i$. That is, if the set of internalizable norms is taken to be "all the equilibrium strategies that can be sustained in a repeated game setting," then this set is vast because of the multiplicity of self-enforcing strategies in a repeated game. By contrast, in our theory, the mechanism of evolutionary selection limits the set of moral principles that are stable for any given Π. Indeed, many moral principles (whether in the 5m's or not) are *not* stable.[39]

[36] The reason why differences in equilibrium play only manifest themselves in play against mutants is discussed in the previous paragraph; refer also to footnote 35.

[37] Our discussion of internalized norms follows Coleman (1994), page 292 and ff.

[38] As in Voss (2001).

[39] These are the moral principles that only span non-stable preferences in the theory of Dekel et al. (2007).

Different from group selection We use a concept of evolutionary stability that is based on individual selection, and not on group selection *à la* Bowles and Gintis (2011). This is good because group selection is controversial.[40]

Ranking the 5m moral principles according to simplicity Which moral principles among the 5m's are "informationally simpler," in the sense that less information about the fitness matrix Π is required in order to generate the moral proclivity matrix, i.e., to know where to place x and y in M? The simplest principles are *Sanctity*, *Fairness*, and *Authority* because knowledge of only two numbers, (a and d for *Sanctity*, b and c for *Fairness* and *Authority*), is required in order to know where in the moral proclivity matrix to place x and y. *Care* and *Loyalty* require full knowledge of the matrix Π.

Robustness to perturbations of x and y In some fitness games Π, there are moral principles $m(x, y; \cdot)$ that are stable only for very specific values of x, y; in other games, or for other moral principles, stability holds for generic $x, y \geq 0$. This property of robustness to perturbations of x and y is, perhaps unsurprisingly, linked to a similar property: compatibility with moral overdrive. Indeed, it can be shown that any $m(x, y; \cdot) \in \mathbf{5m}$ that is compatible with moral overdrive for game Π, is also stable for game Π for any $x, y \geq 0$.[41]

5 Conclusions

It is commonly argued that human morality has been shaped, at least in part, by evolutionary pressures. This paper asks: what kind of moral principles emerge from and survive an evolutionary process?

We have focused on five distinct moral principles that have been identified by recent empirical scholarship (Moral Foundations Theory, Graham et al. 2009): *Care*, *Fairness*, *Authority*, *Loyalty*, and *Sanctity*. These principles are elicited through surveys, and individuals vary in the emphasis they place on each of these principles. We have provided a theory in which these five moral principles are operationalized as mathematical functions that, for each distinct "fitness" game that agents play, produces a "moral proclivity" which shapes behavior.

[40] See Robson (2017).

[41] See Lemma 4 in the appendix

We have shown that any two moral principles are distinguishable from each other in that they shape behavior differently; and that, together, these five moral principles are sufficient to generate all moral proclivities that can ever be evolutionarily stable. However, we have also shown that these five moral principles are somewhat redundant, in that two proper subsets of them are each rich enough to generate all evolutionarily stable moral proclivities.

We have shown that a designer who seeks to maximize fitness by designing a moral code based on the fewest number of principles will only need to include *Fairness* or *Loyalty* in the moral code. Finally, we have asked whether any of these moral principles can be "blown out of proportion" and still be evolutionary stable: the answer is yes, but, when this is the case, these moral principles do not improve fitness relative to fitness-maximizing equilibrium play.

The evolutionary model analyzed here need not apply only to genetic transmission across generations: it could also capture cultural transmission across generations, or to the adoption within the lifetime of a single individual of values that are rewarded by society. Indeed, the analysis relies on an individual selection mechanism, and not on group selection.

This paper provides theoretical support for three commonly made (but not rigorously justified) claims: that people can be moved by several (i.e., more than one) distinct moral principles; that people in the same society vary in the emphasis they place on these principles; and that these principles emerge from an evolutionary process.

Appendices

A Mathematical definition of the 5m moral principles

This section provides the mathematical definition of each of the *5m*'s moral principles. Let the fitness matrix be denoted by

$$\Pi = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{array}{|c|c|} \hline \text{a} & \text{b} \\ \hline \text{c} & \text{d} \\ \hline \end{array} \end{array},$$

which induces the fitness game

$$G = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{array}{|c|c|} \hline \text{a,a} & \text{b,c} \\ \hline \text{c,b} & \text{d,d} \\ \hline \end{array} \end{array}.$$

Without loss of generality, it is assumed $a \geq d$. Furthermore, in what follows we stipulate $x, y \geq 0$.

A.1 Care

Care is defined as *boost, with x or y, the entry which for a given opponent's action, makes the opponent best off.* Hence, the comparisons that need to be made are $a \lessgtr b$ and $c \lessgtr d$. The formal definition is:

$$Care(x, y; \Pi) = \begin{cases} \begin{array}{|c|c|} \hline \text{x} & \text{y} \\ \hline 0 & 0 \\ \hline \end{array} & \text{if } a > b \text{ and } c > d \\ \\ \begin{array}{|c|c|} \hline \text{x} & 0 \\ \hline 0 & \text{y} \\ \hline \end{array} & \text{if } a > b \text{ and } d > c \\ \\ \begin{array}{|c|c|} \hline 0 & \text{y} \\ \hline \text{x} & 0 \\ \hline \end{array} & \text{if } b > a \text{ and } c > d \\ \\ \begin{array}{|c|c|} \hline 0 & 0 \\ \hline \text{x} & \text{y} \\ \hline \end{array} & \text{if } b > a \text{ and } d > c \end{cases}$$

In the non-generic case of $a = b$, replace x with 0, and in the non-generic case of $c = d$, replace y with 0.

A.2 Fairness

Fairness is defined as *boost, with x or y, the entry which for a given opponent's action has the most equal payoffs*. Hence, the comparisons that need to be made are $|a-a| \gtreqless |c-b|$ and $|b-c| \gtreqless |d-d|$, which reduce to checking if $|c-b|$ is equal or different from 0. The formal definition is:

$$Fairness(x,y;\Pi) = \begin{cases} \begin{array}{|c|c|}\hline \text{x} & 0 \\ \hline 0 & \text{y} \\ \hline \end{array} & \text{if } |c-b| \neq 0 \\ \\ \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 0 & 0 \\ \hline \end{array} & \text{if } |c-b| = 0 \end{cases}$$

A.3 Authority

Authority is defined as *boost, with x or y, the entry which for a given opponent's action makes the agent best off relative to her opponent*. Hence, the comparisons that need to be made are $a-a \gtreqless c-b$ and $b-c \gtreqless d-d$, which reduce to checking $b-c \gtreqless 0$. The formal definition is:

$$Authority(x,y;\Pi) = \begin{cases} \begin{array}{|c|c|}\hline \text{x} & \text{y} \\ \hline 0 & 0 \\ \hline \end{array} & \text{if } b-c > 0 \\ \\ \begin{array}{|c|c|}\hline 0 & 0 \\ \hline \text{x} & \text{y} \\ \hline \end{array} & \text{if } b-c < 0 \\ \\ \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 0 & 0 \\ \hline \end{array} & \text{if } b-c = 0 \end{cases}$$

A.4 Loyalty

Loyalty is defined as *boost, with x or y, the entry which rewards the opponent who takes the Kantian action and punishes the opponent who does not*. Hence, the comparisons that need to be made are $a \gtreqless b$ and $c \gtreqless d$. The formal definition is:

$$Loyalty(x, y; \Pi) = \begin{cases} \begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array} & \text{if } a > b \text{ and } c > d \\ \\ \begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array} & \text{if } a > b \text{ and } d > c \\ \\ \begin{array}{|c|c|}\hline 0 & 0 \\ \hline x & y \\ \hline \end{array} & \text{if } b > a \text{ and } c > d \\ \\ \begin{array}{|c|c|}\hline 0 & y \\ \hline x & 0 \\ \hline \end{array} & \text{if } b > a \text{ and } d > c \end{cases}$$

In the non-generic case of $a = b$, replace x with 0, and in the non-generic case of $c = d$, replace y with 0.

A.5 Sanctity

The sanctity moral principle is defined as *boost, with x or y, the entry which corresponds to the Kantian action regardless of the opponent's action.* Since $a > d$ by assumption, the Kantian action is A. The formal definition is:

$$Sanctity(x, y; \Pi) = \begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array}$$

B Online Appendix: 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.

B.1 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 Guth (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 (A and B), 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 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 *Care* when $a > b$ and $c > d$.

2. **If** $a > b$ **and** $d > c$, the M moral proclivity matrix 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 $\begin{array}{|c|c|}\hline \text{x} & 0 \\ \hline 0 & \text{y} \\ \hline \end{array}$ 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 $\begin{array}{|c|c|}\hline 0 & \text{y} \\ \hline \text{x} & 0 \\ \hline \end{array}$ for a particular value of x and y. That moral proclivity matrix is the one generated by *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 $\begin{array}{|c|c|}\hline 0 & 0 \\ \hline \text{x} & \text{y} \\ \hline \end{array}$ 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 $\{A, B\}$ our operationalization of *Care* is a generalization of the altruism operationalization in Bester and Guth (1998).

B.2 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

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

If $\alpha_i = \beta_i$, 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 (A and B), we get that the inequality averse preference matrix is:

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

which is strategically equivalent to

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

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 this 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.

B.3 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 Verben 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 et al. (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 (A and B), 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 M moral proclivity matrix 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 M moral proclivity matrix 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 $c > b$.

B.4 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 to our *Loyalty* moral principle. Assuming there are only two possible pure actions (A and B), 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 this M above is a special case of our *Loyalty* moral principle, because:

1. **If** $a > b$ **and** $c > d$, the M moral proclivity matrix 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 M moral proclivity matrix 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 M moral proclivity matrix above is strategically equivalent to $\begin{array}{|c|c|}\hline 0 & 0 \\ \hline \rho\text{(b-a)} & \delta\text{(c-d)} \\ \hline \end{array}$, 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 M moral proclivity matrix 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 $\begin{array}{|c|c|}\hline 0 & y \\ \hline x & 0 \\ \hline \end{array}$ 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 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.

B.5 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 (A and B), we get that the "homo moralis" preference matrix is:

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

Hence, an agent is homo moralis if her preference is $P = \Pi + M$, where $M = \Theta\begin{array}{|c|c|}\hline a & a \\ \hline d & d \\ \hline \end{array}$. 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.

C Online Appendix: Stability Tables

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 $5m$'s moral principles that can generate it.

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

First Column

The first column is taken directly form Dekel et al. (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 stability table of Π.

For example, for Π with $a > c$ and $d > b$, we know from Dekel et al. (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 C.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 *5m*'s moral principles generate it. To figure which of the $5m$'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 orderings.

For example, in the third sub-column of column 3 in Stability Table C.1, which corresponds to parameter ordering $a > d > c > b$, we can find the *5m*'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 *5m* 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.

C.1 Coordination game: a>c, d>b

Stable preferences	Stable moral proclivities	5*ms* 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

C.2 Games in which (A, A) is the fitness-maximizing strategy profile, A is a dominant strategy and a is the highest possible payoff: a>c, b>d, a>b

Stable preferences	Stable moral proclivities	5*ms* 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

C.3 Games in which (A, A) is the fitness-maximizing strategy profile and A is a dominant strategy but a is not the highest possible payoff: a>c, b>d,b>a

Stable preferences	Stable moral proclivities	5*m*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)

C.4 Prisoners' Dilemma games in which (A, A) is the fitness-maximizing strategy profile: c>a>d>b

Stable preferences	Stable moral proclivities	5*m*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)

C.5 Hawk-Dove games in which (A, A) is the fitness-maximizing strategy profile: c>a>b>d

Stable preferences	Stable moral proclivities	5*ms* 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)

C.6 Pure-common value Hawke-Dove games: c=b>a>d

Stable preferences	Stable moral proclivities	5*ms* 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 $\alpha=\frac{\text{b-d}}{\text{c-a+b-d}}$	x 0 0 y x=2(c-a) y=2(b-d)	None

C.7 All other games

From Dekel et al. (2007) we know no other game has a stable preference.

D Online Appendix: Proofs

D.1 Proofs for Section 2

Lemma 3 (**moral proclivity matrices span the set of all preferences**). *Fix any pair of* 2×2 *matrices* Π, P. *There exists a matrix* P' *that is strategically equivalent to* P *and a moral proclivity matrix* $M(x,y)$ *such that* $\Pi + M(x,y) = P'$.

Proof We proceed by constructing, for each possible P, a suitable matrix $M(x,y)$.

Case 1: P is strategically equivalent to $\mathcal{AB}_\alpha$ with $\alpha \in [0,1]$

Let $M(x,y) = \begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array}$ with $x = L(1-\alpha) - (a-c)$ and $y = L\alpha - (d-b)$. If $L > \max\{\frac{a-c}{1-\alpha}, \frac{d-b}{\alpha}, 0\}$, x and y are both positive and $\Pi + M(x,y) = \begin{array}{|c|c|}\hline L(1-\alpha)+c & b \\ \hline c & b+L\alpha \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AB}_\alpha$.

Case 2: P is strategically equivalent to $\mathcal{BA}_\alpha$ with $\alpha \in [0,1]$

Let $M(x,y) = \begin{array}{|c|c|}\hline 0 & y \\ \hline x & 0 \\ \hline \end{array}$, with $x = L(1-\alpha) - (c-a)$ and $y = L\alpha - (b-d)$. If $L > \max\{\frac{c-a}{1-\alpha}, \frac{b-d}{\alpha}, 0\}$, both x and y are positive and $\Pi + M(x,y) = \begin{array}{|c|c|}\hline a & d+L\alpha \\ \hline a+L(1-\alpha) & d \\ \hline \end{array}$ is strategically equivalent to $\mathcal{BA}_\alpha$.

Case 3: P is strategically equivalent to $\mathcal{AA}$

Let $M(x,y) = \begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array}$. If $x > |c-a|$ and $y > |d-b|$, both x and y are positive and $\Pi + M(x,y) = \begin{array}{|c|c|}\hline a+x & b+y \\ \hline c & d \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AA}$.

Case 4: P is strategically equivalent to $\mathcal{BB}$

Let $M(x,y) = \begin{array}{|c|c|}\hline 0 & 0 \\ \hline x & y \\ \hline \end{array}$. If $x > |a-c|$ and $y > |b-d|$, both x and y are positive and $\Pi + M(x,y) = \begin{array}{|c|c|}\hline a & b \\ \hline c+x & d+y \\ \hline \end{array}$ is strategically equivalent to $\mathcal{BB}$.

Case 5: P is strategically equivalent to θ_0

1. If $a \geq c$ and $b \geq d$, let $M(x,y) = \begin{array}{|c|c|}\hline 0 & 0 \\ \hline a-c & b-d \\ \hline \end{array}$. Then $\Pi + M(x,y) = \begin{array}{|c|c|}\hline a & b \\ \hline a & b \\ \hline \end{array}$ is strategically equivalent to θ_0.
2. If $a \geq c$ and $d \geq b$, let $M(x,y) = \begin{array}{|c|c|}\hline 0 & d-b \\ \hline a-c & 0 \\ \hline \end{array}$. Then $\Pi + M(x,y) = \begin{array}{|c|c|}\hline a & d \\ \hline a & d \\ \hline \end{array}$ is strategically equivalent to θ_0.

3. If $c \geq a$ and $b \geq d$, let $M(x,y) = \begin{array}{|c|c|}\hline c-a & 0 \\ \hline 0 & b-d \\ \hline \end{array}$. Then $\Pi + M(x,y) = \begin{array}{|c|c|}\hline c & b \\ \hline c & b \\ \hline \end{array}$ is strategically equivalent to θ_0.

4. If $c \geq a$ and $d \geq b$, let $M(x,y) = \begin{array}{|c|c|}\hline c-a & d-b \\ \hline 0 & 0 \\ \hline \end{array}$. Then $\Pi + M(x,y) = \begin{array}{|c|c|}\hline c & d \\ \hline c & d \\ \hline \end{array}$ is strategically equivalent to θ_0.

□

D.2 Proofs for Section 3.1

Proof of Theorem 1 From Stability Table C.1, we know the following:

1. For fitness games with $a > c > d > b$: only *Sanctity* and *Care* can span $\mathcal{AA}$ and only *Fairness*, *Authority* and *Loyalty* can span $\mathcal{AB}_\alpha$ with $\alpha \in [\frac{d-b}{d-b+a-c}, 1)$.

2. For fitness games with $a > d > c > b$: only *Sanctity* and *Loyalty* can span $\mathcal{AA}$ and only *Fairness*, *Authority* and *Care* can span $\mathcal{AB}_\alpha$ with $\alpha \in [\frac{d-b}{d-b+a-c}, 1)$. Furthermore, all foundations except *Authority* can span $\mathcal{AB}_\alpha$ with $\alpha \in [0.\frac{d-b}{d-b+a-c})$, but only *Authority* can span $\mathcal{AB}_1$.

Together, points 1 and 2 above imply every $5m$ moral principle is pairwise distinguishable. □

D.3 Proofs for Section 3.2

Proof of Lemma 1 Let $\Pi = \begin{array}{|c|c|}\hline 2 & 0 \\ \hline -1 & 1 \\ \hline \end{array}$. This is a coordination game with $a > d > b > c$. From Stability Table C.1 we know that, for this game, *Care* and *Fairness* span

$$P_1 = \Pi + M_1 = \begin{array}{|c|c|}\hline 2 & 0 \\ \hline -1 & 1 \\ \hline \end{array} + \begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array} = \begin{array}{|c|c|}\hline 2+x & 0 \\ \hline -1 & 1+y \\ \hline \end{array}$$

and *Authority*, *Sanctity* and *Loyalty* span

$$P_2 = \Pi + M_2 = \begin{array}{|c|c|}\hline 2 & 0 \\ \hline -1 & 1 \\ \hline \end{array} + \begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array} = \begin{array}{|c|c|}\hline 2+x & y \\ \hline -1 & 1 \\ \hline \end{array}.$$

For no values of $x, y \geq 0$ can P_1 or P_2 be strategically equivalent to

$$P' = \begin{array}{|c|c|}\hline 0 & y \\ \hline x & 0 \\ \hline \end{array}.$$

□

Proof of Theorem 2 From Stability Tables C.1 and C.2, we know that for all games in these tables, all evolutionarily stable preferences are strategically equivalent to an element of the set

$$\mathbf{ES} = \{\mathcal{AA}, \mathcal{AB}_\alpha \text{ with } \alpha \in [0,1], \mathcal{BA}_1, \theta_0\}.$$

The lowest-dimensional vector space containing **ES** has dimension 4, as one needs 4 parameters to describe some preferences. From Stability Tables C.1 and C.2 we also know that the only preferences that can't be spanned by any of the principles in the $5m$ are preferences strategically equivalent to one element of the set

$$\mathbf{NG} = \{\mathcal{AB}_1, \mathcal{BA}_1, \theta_0\}.$$

The lowest-dimensional vector space containing **NG** has dimension 3, as one needs at most 3 parameters to describe all preferences in the set. Hence, the set of preferences not spanned by the $5m$, **NG**, is non-generic in the set of evolutionarily stable preferences ES.

From Stability Tables C.3, C.4 and C.5, we know that for all fitness games represented on those tables, the $5m$ span the entire set of evolutionarily stable preferences.

The remaining fitness games are either the pure-common value Hawk-Dove game, which is non-generic in the space of 2x2 matrices, or fitness games which do not have any evolutionarily stable preferences.

Hence, since graphs inherit the genericity properties of both their domain and the range, $\mathcal{ES}\backslash 5\mathcal{M}$ is non-generic in $\mathcal{ES}$. □

D.4 Proofs for Section 3.3

Proof of Theorem 3 From Stability Tables C.1 and C.2 and from the proof of Theorem 2 we know that, for the fitness games represented in those tables, the set of generic evolutionarily stable preferences are all preferences strategically equivalent to one element of

$$ES\backslash NG = \{\mathcal{AA}, \mathcal{AB}_\alpha \text{ with } \alpha \in (0,1)\}.$$

Sufficiency

To prove that $\mathcal{ES}\backslash\mathcal{D}$ is non-generic in $\mathcal{ES}$, we must show that both

$\{Sanctity, Fairness\}$ and $\{Care, Loyalty\}$ are sufficient to span all generic evolutionarily stable preferences for all generic fitness games.
As can be seen in Stability Tables C.1 and C.2, for all fitness games represented in those tables, *Sanctity* spans $\mathcal{AA}$ and *Fairness* spans $\mathcal{AB}_\alpha$ with $\alpha \in (0,1)$. Furthermore, either *Care* spans $\mathcal{AA}$ and *Loyalty* spans $\mathcal{AB}_\alpha$ with $\alpha \in (0,1)$, or *Loyalty* spans $\mathcal{AA}$ and *Care* spans $\mathcal{AB}_\alpha$ with $\alpha \in (0,1)$. For the fitness games represented in Stability Tables C.3, C.4 and C.5, *Fairness* spans all stable preferences, and whenever *Care* can't span some stable preference, *Loyalty* can. Since all other fitness games are either non-generic in the space of 2×2 matrices or do not have any stable preferences, both $\{Fairness, Sanctity\}$ and $\{Care, Loyalty\}$ are sufficient to span all generic evolutionarily stable preferences for all generic fitness games.

Necessity
To prove that if $\mathcal{ES}\backslash\mathcal{D}$ is non-generic in $\mathcal{ES}$, $\mathcal{D}$ contains either $\mathcal{FS}$ or $\mathcal{CL}$, we must show that either $\{Sanctity, Fairness\}$ or $\{Care, Loyalty\}$ are necessary to span all generic evolutionarily stable preferences for all generic fitness games.
For fitness games represented in table C.5, the set of generic evolutionarily stable preferences is the set of all preferences that are strategically equivalent to $\mathcal{AB}_1$. As can be seen in the stability table, either *Fairness* or *Loyalty* is necessary to span $\mathcal{AB}_1$. For fitness game sub-type $a > c > d > b$, represented in the third column of table C.1, either *Sanctity* or *Care* are necessary to span $\mathcal{AA}$. Hence, one element of $\{Fairness, Loyalty\}$ and one of $\{Care, Sanctity\}$ are needed to span all generic evolutionarily stable preferences for all generic fitness games. No element of $\{Loyalty, Sanctity\}$ can span $\mathcal{AB}_\alpha$ with $\alpha \in (0,1)$ for fitness games $a > b > d > c$ represented in the second column of Stability Table C.2, and no element of $\{Care, Fairness\}$ can span $\mathcal{AA}$ for fitness games $a > d > b > c$ represented in the second column of Stability Table C.1. Hence, either $\{Fairness, Sanctity\}$ or $\{Care, Loyalty\}$ are necessary to span all generic evolutionarily stable preferences in all generic fitness games. □

D.5 Proofs for Section 3.4

Proof of Theorem 4 From Stability Table C.5, we know that for the Hawk-Dove fitness games represented in the table, only *Fairness* or *Loyalty* are capable of spanning the generic evolutionarily stable preferences. Furthermore, from tables

C.1, C.2, C.3, C.4 we know that both *Fairness* and *Loyalty* are capable of spanning some evolutionarily stable preferences. All other fitness games are either non-generic or do not have any stable preferences. Hence, both *Fairness* and *Loyalty* can be a minimal moral code. □

D.6 Proofs for Section 3.5

Proof of Lemma 2 From stability stable C.1's column 3, we know that for fitness games with $a > c > d > b$, $Sanctity(x, y; .)$ and $Care(x, y; .)$ with $x \geq 0$ and $y > d - b$ span those evolutionarily stable preferences that are strategically equivalent to $\mathcal{AA}$. Since $x \geq 0$ and $y > d - b$ implies $kx \geq 0$ and $ky > d - b$ for all $k \geq 1$, *Sanctity* and *Care* are compatible with moral overdrive in those fitness games.

Similarly, from Stability Table C.1's column 4, we know that for fitness games with $a > d > b > c$, $Authority(x, y; .)$ and $Loyalty(x, y; .)$ with $x \geq 0$ and $y > d - b$ span those evolutionarily stable preferences that are strategically equivalent to $\mathcal{AA}$. Since $x \geq 0$ and $y > d-b$ implies $kx \geq 0$ and $ky > d-b$ for all $k > 1$, *Authority* and *Loyalty* are compatible with moral overdrive in those fitness games.

In all generic fitness games games, *Fairness* generates moral proclivity matrix $\begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array}$,which means *Fairness* spans preference $P = \begin{array}{|c|c|}\hline a+x & b \\ \hline c & d+y \\ \hline \end{array}$. In fitness games with $a > c$ and $d > b$, which are represented in Stability Table C.1, for any $x, y \geq 0$, P is strategically equivalent to $\mathcal{AB}_\alpha$ for some $\alpha \in [0, 1]$, which, for those fitness games, is evolutionarily stable. Since $x \geq 0$ and $y \geq 0$ implies $kx \geq 0$ and $ky \geq 0$ for all $k \geq 1$, *Fairness* is compatible with moral overdrive in those fitness games. □

Proof of Theorem 5 For fitness games represented in Stability Tables C.1, C.2, C.3, the fittest symmetric strategy profile is (A, A). Furthermore, for all those fitness games, (A, A) is already a Nash equilibrium. Hence, there does not exist a moral principle m which can strictly improve social fitness. For fitness games represented in Stability Table C.6, the fittest symmetric strategy profile is the mixed strategy in which both agents play A with probability $\frac{b-d}{b-d+c-a}$, but that mixed strategy profile is already a Nash equilibrium. Hence, there does not exist a moral principle m which can strictly improve social fitness. For fitness games represented in Stability Tables C.4 and C.5, which are the Prisoners' Dilemma and Hawk-Dove games in which (A, A) is the fitness-maximizing strategy profile, (A, A) is not a Nash equilibrium. Hence, there are some moral principles m that could potentially strictly improve social fitness.

However for those games, as can be seen in the stability tables, 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. All other fitness games do not have evolutionarily stable preferences. □

D.7 Proofs for Section 4

Lemma 4. *Any moral principle m that is compatible with moral overdrive for fitness game Π, is also evolutionarily stable for fitness game Π for any $x, y \geq 0$.*

Proof of Lemma 4 For fitness games represented in Stability Tables C.1 and C.2, the set of evolutionarily stable preferences is the set of preferences which are strategically equivalent to some preferences in $\{\mathcal{AA}, \mathcal{AB}_\alpha \text{ with } \alpha \in [0,1], \mathcal{BA}_1, \theta_0\}$. For those fitness games, only moral principles which generate either moral proclivity matrices $\begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array}$ or $\begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array}$ are compatible with moral overdrive because:

- If $x \geq 0$ and $y > max\{0, d-b\}$, $\Pi + \begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array} = \begin{array}{|c|c|}\hline a+x & b+y \\ \hline c & d \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AA}$. Since $x \geq 0$ and $y > max\{0, d-b\}$ implies $kx \geq 0$ and $ky > max\{0, d-b\}$ for all $k \geq 1$, $\Pi + \begin{array}{|c|c|}\hline kx & ky \\ \hline 0 & 0 \\ \hline \end{array}$ is also strategically equivalent to $\mathcal{AA}$ for all $k \geq 1$.

- If $x \geq 0$ and $y > max\{0, b-d\}$, $\Pi + \begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array} = \begin{array}{|c|c|}\hline a+x & b \\ \hline c & d+y \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AB}_\alpha$ for some $\alpha \in [0,1]$. Since $x \geq 0$ and $y > max\{0, b-d\}$ implies $kx \geq 0$ and $ky > max\{0, b-d\}$ for all $k \geq 1$, $\Pi + \begin{array}{|c|c|}\hline kx & 0 \\ \hline 0 & ky \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AB}_\beta$ for some $\beta \in [0,1]$ for all $k \geq 1$.

- As can be seen in Stability Tables C.1 and C.2, for moral proclivity matrices $\begin{array}{|c|c|}\hline 0 & 0 \\ \hline x & y \\ \hline \end{array}$ and $\begin{array}{|c|c|}\hline 0 & y \\ \hline x & 0 \\ \hline \end{array}$ to be stable, x and/or y have to be either small enough or take a specific value. However, it is not possible for kx and ky to continue being small or continue taking the same value as x and y for all $k \geq 1$.

Hence, for fitness games represented in Stability Tables C.1 and C.2, only moral principles that generate moral proclivity matrices $\begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array}$ or $\begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array}$ are compatible with moral overdrive. Moreover, those moral proclivity matrices are evolutionarily stable for all $x, y \geq 0$ because:

- For fitness games represented in Stability Table C.1, if $y \leq d-b$, $\Pi + \begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array} = \begin{array}{|c|c|}\hline a+x & b+y \\ \hline c & d \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AB}_\alpha$ for some $\alpha \in [0,1]$, and if $y > d-b$, $\Pi + \begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array} = \begin{array}{|c|c|}\hline a+x & b+y \\ \hline c & d \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AA}$. Furthermore, for $y \geq 0$, $\Pi + \begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array} = \begin{array}{|c|c|}\hline a+x & b \\ \hline c & d+y \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AB}_\alpha$ for some $\alpha \in [0,1]$.

- For fitness games represented in Stability Table C.2, if $y < b-d$, $\Pi + \begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array} = \begin{array}{|c|c|}\hline a+x & b \\ \hline c & d+y \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AA}$, and if $y \geq b-d$, $\Pi + \begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array} = \begin{array}{|c|c|}\hline a+x & b \\ \hline c & d+y \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AB}_\alpha$ for some $\alpha \in [0,1]$. Furthermore, for $y \geq 0$, $\Pi + \begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array} = \begin{array}{|c|c|}\hline a+x & b+y \\ \hline c & d \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AA}$.

Hence, for fitness games represented in C.1 and C.2, any moral principle compatible with moral overdrive is evolutionarily stable for all $x, y \geq 0$.

For fitness games represented in Stability Table C.3, the set of evolutionarily stable preferences is the set of preferences which is strategically equivalent to some preferences in $\{\mathcal{AA}, \mathcal{AB}_\alpha \text{ with } \alpha \in [0, \frac{a-d}{b-d})\}$. For those fitness games only moral principles which generate moral proclivity matrix $\begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array}$ are compatible with moral overdrive because:

- If $x \geq 0$ and $y \geq 0$, $\Pi + \begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array} = \begin{array}{|c|c|}\hline a+x & b+y \\ \hline c & d \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AA}$. Since $x \geq 0$ and $y \geq 0$ implies $kx \geq 0$ and $ky > 0$ for all $k \geq 1$, $\Pi + \begin{array}{|c|c|}\hline kx & ky \\ \hline 0 & 0 \\ \hline \end{array}$ is also strategically equivalent to $\mathcal{AA}$.

- From Stability Table C.3, we know moral proclivity $\begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array}$ with $x \geq 0$ and $y = \frac{\alpha}{1-\alpha}(a-c+x) + (b-d)$ spans $\mathcal{AB}_\alpha$. So if $x = 0$ and $y \in [b-d, b-d+\frac{(a-d)}{(b-a)}(a-c))$, $\Pi + \begin{array}{|c|c|}\hline x & 0 \\ \hline 0 & y \\ \hline \end{array} = \begin{array}{|c|c|}\hline a+x & b \\ \hline c & d+y \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AB}_\alpha$ for some $\alpha \in [0, \frac{a-d}{b-d})$. However, $ky > b-d+\frac{(a-d)}{(b-a)}(a-c)$ for a large k, so $\Pi + \begin{array}{|c|c|}\hline kx & 0 \\ \hline 0 & ky \\ \hline \end{array} = \begin{array}{|c|c|}\hline a+kx & b \\ \hline c & d+ky \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AB}_\alpha$ with $\alpha > \frac{a-d}{b-d}$ for some k, which is not stable.

- From Stability Table C.3, we also know that for moral proclivity matrices $\begin{array}{|c|c|}\hline 0 & 0 \\ \hline x & y \\ \hline \end{array}$ and $\begin{array}{|c|c|}\hline 0 & y \\ \hline x & 0 \\ \hline \end{array}$ to be stable, x and/or y have to be either small enough or take a

specific value. However, it is not possible for kx and ky to continue being small or continue taking the same value as x and y for all $k \geq 1$.

Hence, for fitness games represented in Stability Table C.3, only those moral principles that generate moral proclivity matrix $\begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array}$ are compatible with moral overdrive. Since for $x \geq 0$ and $y \geq 0$, $\Pi + \begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline \end{array} = \begin{array}{|c|c|}\hline a+x & b+y \\ \hline c & d \\ \hline \end{array}$ is strategically equivalent to $\mathcal{AA}$, any moral principle compatible with moral overdrive is evolutionarily stable for all $x, y \geq 0$.

For fitness games represented in Stability Tables C.4, C.5 and C.6, all stable moral proclivity matrices require x to take a specific value. Since kx can't take the same value as x for any $k \neq 1$, there aren't any moral principles which are compatible with moral overdrive for those fitness games. □

References

Alger, I. Evolutionarily Stable Preferences. *Philosophical Transactions of the Royal Society B*, 378:20210505, 2023.

Alger, I. and Weibull, J. W. Homo Moralis—Preference Evolution Under Incomplete Information and Assortative Matching. *Econometrica*, 81:2269–2302, 2013.

Alger, I. and Weibull, J. W. Evolution and Kantian Morality. *Games and Economic Behavior*, 98:56–67, 2016.

Alger, I. and Weibull, J. W. Strategic Behavior of Moralists and Altruists. *Games*, 8: 38, 2017.

Alger, I., Weibull, J. W., and Lehmann, L. Evolution of Preferences in Group-Structured Populations: Genes, Guns, and Culture. *Journal of Economic Theory*, 185:104951, 2020.

Atari, M., Haidt, J., Graham, J., Koleva, S., Stevens, S.T. and Dehghani, M. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 2023.

Bester, H., and Guth, W. Is Altruism Stable? *Journal of Economic Behavior and Organization*, 34:193–209, 1998.

Binmore, K. *Playing for Real: A Text on Game Theory.* Oxford University Press, Oxford, 2007.

Bowles, S., and Gintis, H. *A Cooperative Species: Human Reciprocity and its Evolution.* Princeton University Press, 2011.

Charness, G. and Rabin, M. Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, 117(3): 817–869, 2002.

Coleman, J.S. *Foundations of social theory.* Harvard University Press, 1994.

Corning, P. *The Fair Society: The Science of Human Nature and the Pursuit of Social Justice.* University of Chicago Press, 2011.

Cressman, R. *The Stability Concept of Evolutionary Game Theory.* Springer Verlag, Berlin, 1992.

Dekel, E., Ely, J. C., and Yilankaya, O. Evolution of Preferences. *Review of Economic Studies*, 74:685–704, 2007.

Ely, J. C. and Yilankaya, O. Nash Equilibrium and the Evolution of Preferences. *Journal of Economic Theory*, 97:255–272, 2001.

Fehr, E. and Schmidt, K. M. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114:817–68, 1999.

, Govindan, S. and Wilson, R. Axiomatic Equilibrium Selection for Generic Two-Player Games. *Econometrica*, 80:1639-1699, 2012.

Graham, J., Haidt, J., and Nosek, B. A. Liberals and Conservatives Rely on Different Sets of Moral Foundations. *Journal of Personality and Social Psychology*, 96:1029–1046, 2009.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., and Ditto, P. H. Mapping the Moral Domain. *Journal of Personality and Social Psychology*, 101:366–385, 2011.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in Experimental Social Psychology* (Vol. 47, pp. 55–130). Academic Press, 2012.

Güth, W. An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives. *International Journal of Game Theory*, 24:323–344, 1995.

Güth, W. and Peleg, B. When Will Payoff Maximization Survive? An Indirect Evolutionary Analysis. *Journal of Evolutionary Economics*, 11:479–499, 2001.

Güth, W. and Yaari, M. E. Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach. In Witt, U., editor, *Explaining Process and Change*. University of Michigan Press, Ann Arbor, Michigan, 1992.

Haidt, J. *The Righteous Mind*. Pantheon Books, New York, 2012.

Hamlin, J. K. Moral Judgment and Action in Preverbal Infants and Toddlers: Evidence for an Innate Moral Core. *Current Directions in Psychological Science*, 22:186–193, 2013.

Heifetz, A., Shannon, C., and Spiegel, Y. What to Maximize if You Must. *Journal of Economic Theory*, 133:31–57, 2007.

Helzer, E. G., Furr, R. M., Hawkins, A., Barranti, M., Blackie, L. E., and Fleeson, W. Agreement on the perception of moral character. Personality and Social Psychology Bulletin, 40:1698–1710, 2014.

Maynard Smith, J. The Theory of Games and the Evolution of Animal Conflicts. *Journal of Theoretical Biology*, 47:209–221, 1974.

Maynard Smith, J. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, 1982.

Maynard Smith, J. and Price, G. R. The Logic of Animal Conflict. *Nature*, 246:15–18, 1973.

Messick, D. M., and Sentis, K. P. Estimating Social and Nonsocial Utility Functions from Ordinal Data. *European Journal of Social Psychology*, 15(4): 389–399, 1985.

Miettinen, T., Kosfeld, M., Fehr, E., and Weibull, J. Revealed Preferences in a Sequential Prisoners' Dilemma: A Horse-Race between Six Utility Functions. *Journal of Economic Behavior and Organization*, 173:1–25, 2020.

Nachbar, J. H. 'Evolutionary' Selection Dynamics in Games: Convergence and Limit Properties. *International Journal of Game Theory*, 19:59–89, 1990.

Norman, T. W. L. Equilibrium Selection and the Dynamic Evolution of Preferences. *Games and Economic Behavior*, 74:311–320, 2012.

Ordóñez, L. D., Connolly, T., and Coughlan, R. Multiple Reference Points in Satisfaction and Fairness Assessment. *Journal of Behavioral Decision Making*, 13(3): 329–344, 2000.

Ok, E. A. and Vega-Redondo, F. On the Evolution of Individualistic Preferences: An Incomplete Information Scenario. *Journal of Economic Theory*, 97:231–254, 2001.

Robson, A.J. Group Selection: A Review Essay on Does Altruism Exist? by David Sloan Wilson. *Journal of Economic Literature*, 55(4):1570-1582, 2017.

Robson, A. J. and Samuelson, L. The Evolutionary Foundations of Preferences. In Benhabib, J., Bisin, A., and Jackson, M., editors, *The Social Economics Handbook*, pp. 221–310. North Holland, 2011.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., and Zamir, S. Bargaining and Market Behavior in Jerusalem, Llubljana, Pittsburgh and Tokyo: An Experimental Study. *American Economic Review*, 81:1068–1095, 1991.

Samuelson, L. Stochastic Stability in Games with Alternative Best Replies. *Journal of Economic Theory*, 64:35–65, 1994.

Samuelson, L., and Zhang, J. Evolutionary Stability in Asymmetric Games. *Journal of Economic Theory*, 57:363–391, 1992.

Sandholm, W. H., Dokumaci, E., and Franchetti, F. Dynamo: Diagrams for Evolutionary Game Dynamics. http://www.ssc.wisc.edu/∼whs/dynamo, 2012.

Schein, C. and Gray, K. The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. *Personality and Social Psychology Review*, 22:32–70, 2018.

Segal, U., and Sobel, J. Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings. *Journal of Economic Theory*, 136(1):197–216, 2007.

Selten, R. A Note on Evolutionarily Stable Strategies in Asymmetric Animal Contests. *Journal of Theoretical Biology*, 84:93–101, 1980.

Suhler, C. L., and Churchland, P. Can Innate Moral 'Foundations' Explain Morality? Challenges for Haidt's Moral Foundations Theory. *Journal of Cognitive Neuroscience*, 23:2103–2116, 2011.

Taylor, P. D. Evolutionary Stable Strategies with Two Types of Players. *Journal of Applied Probability*, 16:76–83, 1979.

Taylor, P. D., and Jonker, L. B. Evolutionary Stable Strategies and Game Dynamics. *Mathematical Biosciences*, 40:145–156, 1978.

Turiel, E. *The Development of Social Knowledge: Morality and Convention*. Cambridge, 1983.

Voss, T. *Game-theoretical perspectives on the emergence of social norms*. na, 2001.

Weibull, J. W. *Evolutionary Game Theory*. The MIT Press, Cambridge, MA, 1995.

Zakharin, M., and Bates, T. C. Remapping the Foundations of Morality: Well-Fitting Structural Model of the Moral Foundations Questionnaire. *PloS one* 16(10), 2021.

Zakharin, M., and Bates, T. C. Testing heritability of moral foundations: Common pathway models support strong heritability for the five moral foundations. *European Journal of Personality* 37:485–497, 2023.