THE EVOLUTIONARY STABILITY OF MORAL FOUNDATIONS*

MICHELLE AVATANEO
THOMAS NORMAN
NICOLA PERSICO

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 payoffs in a 2×2 game. We find that the five foundations are distinguishable from each other and evolutionarily stable. However, they are not a minimal set: strict subsets of the foundations suffice to describe all preferences that are evolutionarily stable. Not all evolutionarily stable foundations deliver social fitness improvement over the Nash equilibrium in the fitness game: we characterize which do. Finally, we study moral overdrive, that is, the situation in which the moral component of preferences totally dominates fitness payoffs 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 social fitness improvement. These results are partially extended to $n \times n$ games. We derive two testable implications from the theory and find empirical support for them. *JEL codes:* C73, D91.

I. INTRODUCTION

There are reasons to believe that human morality has been shaped 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,

* The article benefited from discussions with Jeff Ely and Tim Feddersen. We thank Francesco Crispano for valuable research assistance.

1. For example, babies as young as 34 hours old cry reflexively when exposed to crying sounds, 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 anti-

The Quarterly Journal of Economics (2025), 2459–2506. https://doi.org/10.1093/qje/qjaf019.
Advance Access publication on April 11, 2025.

twin studies find that several distinct moral principles are inherited.[2] This article asks: What kind of moral principles emerge from and survive an evolutionary process?

To make progress, we build on moral foundations theory (MFT), an influential empirical framework developed by Jonathan Haidt and colleagues.[3] MFT seeks to characterize the bases for social cooperation; it argues that all moral systems[4] are based on the following five "moral foundations" that regulate selfishness.[5]

(i) **Care**: This foundation is based on the idea of empathy and compassion. It emphasizes avoiding causing harm to others and the desire to care for those in need.
(ii) **Fairness**: This foundation involves concepts of justice, fairness, and reciprocity. It includes principles such as equality, proportionality, and the notion of treating others as you would like to be treated.
(iii) **Loyalty**: This foundation focuses on loyalty, group cohesion, and the valuing of one's in-group (such as family, community, or nation). It involves behaviors such as loyalty, patriotism, and group solidarity.
(iv) **Authority**: This foundation emphasizes respect for authority, tradition, and social hierarchy. It includes values such as obedience, deference to authority figures, and respect for customs and traditions.
(v) **Sanctity**: This foundation is concerned with issues of cleanliness, purity, and sanctity. It involves avoiding behaviors or actions that are seen as impure or contaminat-

social characters (who harm hinderers) over inappropriately prosocial ones (who help hinderers), actions that 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. See, in particular, Graham et al. (2011) and Graham et al. (2013).

4. According to MFT, "Moral systems are interlocking sets of values, virtues, norms, practices, identities, institutions, technologies, and evolved psychological mechanisms that work together to suppress or regulate selfishness and make social life possible." Graham et al. (2011, 368).

5. The explanatory verbiage that follows was generated by ChatGPT in response to the prompt "what is moral foundation theory?" (see screenshot and stable link in Online Appendix I), but similar (though less concise) explanatory verbiage appears in Haidt (2012), chapter 7, and Haidt and Graham (2007).

ing and upholding moral standards related to bodily integrity, chastity, and sacredness.

Analytically, these foundations are derived from factor analysis of responses to the MF questionnaire.[6] Individuals turn out to vary in the emphasis they place on each foundation.[7] Some scholars argue that this cross-person variation can be reduced to just two principal components rather than five separate foundations: a so-called individualizing factor (overlapping with Care and Fairness) and a so-called binding factor (overlapping with Loyalty, Authority, and Sanctity).[8] Regardless, there is agreement that empirically, individuals vary in their moral makeup along at least two dimensions. The fact that people differ in their emphasis on different moral foundations is referred to as "moral pluralism."

MFT has been extremely influential. Around the time of this writing, a Google Scholar search for moral foundations theory returned more than four million entries. The MF questionnaire (see Online Appendix I) is part of several institutional surveys, including one wave of the European Values Study. Recently, work that leverages MFT has been prominently published in economics (see Enke 2019, 2020, 2024). For this article, we regard MFT as an adequate empirical description of human moral judgment.

Although MFT's five foundations (or, depending on one's read of the literature, two principal components) are said to be selected by evolution among a broader set of theoretically possible moral principles,[9] no rigorous theoretical justification has been offered

6. The seminal paper performing factor analysis is Graham et al. (2011). The MF questionnaire is reproduced in Online Appendix I. Related survey instruments are available at https://moralfoundations.org/questionnaires/. Graham et al. (2013) provide a vignette elicitation instrument.

7. Intriguingly, the degree to which people emphasize certain principles has been found to correlate with political leanings, with liberals (in the U.S. sense of the term) most focused on Care and Fairness while conservatives valuing each foundation more evenly (Graham, Haidt, and Nosek 2009).

8. The "number of dimensions" literature is discussed later. The etymology of the words *individualizing* and *binding* is explained by Graham et al. (2011, 369), who write: "the 'individualizing foundations' . . . are all that are needed to support the individual-focused contractual approaches to society often used in enlightenment ethics, whereas . . . the 'binding foundations' are about binding people together into larger groups and institutions."

9. Graham et al. (2013, 60) state that the five foundations are "the concerns, perceptions, and emotional reactions that consistently turn up in moral codes

for this claim. Instead, and at odds with this multidimensional moral landscape, the existing game-theoretic literature based on individual selection identifies a single dimension, that is, a single evolutionarily stable moral principle that all agents have in the same degree. This literature, therefore, is so far mono-factor (people are "less or more moral" along a single dimension) and further predicts that every person is moral in the same degree.[10]

We study a game-theoretic setting in which agents play "fitness games" Π with randomly selected opponents and respond to psychological motives that induce potentially non–fitness maximizing best responses. These motives are captured by abstract functions $m(\Pi)$, which we call "moral principles." These moral principles are subject to evolutionary pressure. We focus on five special moral principles that represent MFT's foundations mathematically and find that these principles suffice to capture all evolutionarily stable behavior in every game type, consistent with MFT's empirical claim. We also find that several of these representations can coexist in any given game type, thus providing support for moral pluralism (meaning that different people can hold different moral principles in the same game type).

The model is as follows. Players repeatedly play a two-by-two symmetric fitness 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, a player does not maximize Π but $P = \Pi + m(\Pi)$. The role that $m(\cdot)$ plays in the analysis is to change a player's best-response function. The matrix-valued function $m(\cdot)$ will be called a moral principle. Moral principles are assumed to be observable.[11]

The space of all moral principles is large, and not all moral principles capture a sensible morality. For example, a function $m(\cdot)$ that given any fitness game Π, adds rewards (positive utils) to the outcomes in which players have more similar payoffs, is a representation of the sensible principle of Fairness. However, a function $\widetilde{m}(\cdot)$ that adds utils to unequal outcomes even when the inequality comes at the player's own expense encodes a weird

around the world, and for which there are already-existing evolutionary explanations." Footnote 4 quotes similar language from Graham et al. (2011).

10. The literature relying on group selection, as opposed to individual selection, has proved capable of generating a multifactor theory of morality; we discuss both literatures.

11. We relax this assumption in Section IV.A.

"anti-Fairness" morality—and yet, according to our definition, it too is a moral principle that evolution could potentially select.[12]

The evolutionary-stability concept operates on the space of moral principles. It is the same as in Dekel, Ely, and Yilankaya (2007): we say that "*m is evolutionarily stable for* Π" if adopting $\Pi + m(\Pi)$ guarantees a player an average fitness payoff that is no lower than any other player in the population—even if a small fraction of the population adopts some arbitrary mutant moral principle $\tilde{m}(\Pi)$.[13] Different moral principles can be stable in different game types.[14] If two different moral principles are stable in the same game type Π, we say that Π gives rise to moral pluralism.[15]

To formally embed MFT in a mathematical model of evolutionary selection, we define five special 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 mathematical functions collectively as the H5's (H for Haidt). Eqipped with these mathematical counterparts to the five foundations, we ask:

(i) (**Distinguishability**) Are the H5's distinct from each other, that is, do different elements m in the H5's give rise to strategically different matrices $\Pi + m(\Pi)$? This is nonobvious because it goes against the counting intuition that there are many H5's (to wit, five) compared to the low complexity of behavior that can be encoded in a 2×2 matrix $m(\Pi)$.

(ii) (**Spanning**) For every fitness matrix Π, is every evolutionarily stable matrix $m(\Pi)$ produced by a moral principle m in the H5? If so, our theory would validate MFT by pointing to the H5 (a very specific and empirically inspired subset of all m's) as a complete description of all evolutionarily stable behavior. Note that this is not a foregone conclu-

12. "Anti-fairness" is defined mathematically by replacing the word "equal" with "unequal" in Table I.

13. Note that evolutionary fitness is based on Π only, not on $\Pi + m(\Pi)$.

14. For example, we will show that Fairness is stable in all game types Π, but Care (to be defined later) is not.

15. If m and m' happen to be stable for a given Π, then we will show that any combination of m and m' is stable also. Hence, under moral pluralism, m and m' coexist in the sense that they will play against each other with some frequency.

sion: indeed, some matrices $m(\Pi)$ cannot be produced by any moral principle in the H5's (Counterexample 1: these matrices happen to be nonstable). For this reason, if the spanning question can be answered in the positive, an intriguing mathematical connection is uncovered between the H5's and evolutionary stability.

(iii) (**Minimal spanning**) Is the set of the H5's minimal, or are there strict subsets of the H5's that produce all evolutionarily stable matrices $m(\Pi)$? If the latter, could the subsets be interpreted as comprising an individualizing and a binding factor, as suggested by some of the empirical literature?

(iv) (**Moral code design**) Not all evolutionarily stable matrices $m(\Pi)$ necessarily lead to social fitness improvement in equilibrium play. Suppose a planner sought to design a stable minimal moral code that guaranteed all possible social fitness improvements: which subset of the H5's would the planner need to include in the code?

(v) (**Moral overdrive**) Can a 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 evolutionarily stable? Is such "moral overdrive" compatible with social fitness improvement?

The answer to question (i) is yes: for any pair of principles m and m' in the H5'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 H5's are pairwise distinguishable (Theorem 1). Distinguishability is a sanity check for our theory because if any two elements of the H5's were not distinguishable, our 2×2 setting would be too low-dimensional to allow for five different moral principles.

The answer to question (ii) is yes as well; this means that the H5's are, generically, "rich enough" to produce all the $m(\Pi)$'s that are evolutionarily stable for any given Π (Theorem 2). Because there are some (nonstable) matrices $m(\Pi)$ that cannot be produced by any moral principle in the H5's (Counterexample 1), Theorem 2 uncovers a nonobvious connection between the H5's and evolutionary stability.

But the H5's are not the smallest set that is rich enough: Theorem 3 identifies two proper subsets of the H5's, each of which is rich enough to produce all $m(\Pi)$'s that are evolutionarily sta-

ble for any given Π. Therefore, the answer to question (iii) is no: the H5's are not the minimal set that produces all evolutionarily stable $m(\Pi)$'s. Intriguingly, each of these two subsets happens to be constituted of one individualizing and one binding foundation, arguably consistent with the two-factor reading of the empirical evidence. Incidentally, Theorem 3 proves that moral pluralism is ubiquitous: every Π has at least two stable moral principles.

Question (iv) is answered in Theorem 4: a designer who seeks to maximize social fitness by designing a moral code based on the fewest number of principles in the H5's will only need to include Fairness or Loyalty in the moral code.

The answer to question (v) is yes: in some games, moral principles can be "blown out of proportion" and still be evolutionarily stable (Lemma 1). However, when this is the case, these moral principles do not improve social fitness relative to equilibrium play with Π (Theorem 5). In this sense, "moral overdrive" is incompatible with social fitness improvement. In particular, a moral code which, as in question (iv), was created to maximize social fitness would not need to rely on moral overdrive.

These findings are extended to settings where moral principles are not perfectly observable and, albeit partially, to $n \times n$ games. We conclude the article by deriving two comparative static results: in more market-oriented societies, there should be more Sanctity and less Care—in a sense made formal in Section V.C. We find empirical support for these implications in a cross-country analysis based on the World Values Survey.

This article contributes to the theoretical literature on the evolutionary stability of morality by providing theoretical support for three commonly made claims: that different moral principles apply in different settings (e.g., free-riding versus coordination game) and in the same setting; that people in the same setting can vary in the emphasis they place on different principles; and that these principles survive evolutionary selection. We provide a rigorous theoretical model in which these three claims hold simultaneously.

This study builds on Dekel, Ely, and Yilankaya (2007), but it innovates as follows. First, Dekel, Ely, and Yilankaya (2007) do not concern themselves with the function $m(\cdot)$ but characterize which matrices P are stable for every Π. But these stable matrices P need not be interpretable as arising from any natural morality, that is, the matrix $m(\Pi) = P - \Pi$ need not represent any natural moral prescription (many functions m do not:

for example, "anti-Fairness" as defined in footnote 12). We introduce restrictions that define which functions $m(\cdot)$ are "natural" in the sense that they have an empirical grounding in MFT's moral foundations—and ask whether all stable behavior can be generated by a "natural" $m(\cdot)$. Second, the characterization of stable preferences provided by Dekel, Ely, and Yilankaya (2007), while helpful, does not tell us which functions $m(\cdot)$ are stable because there is no one-to-one correspondence between stable preferences P and stable moral principles m: a given stable preference may be generated by several different moral principles, and the same moral principle can generate two different stable preferences (refer to Section II.D for a discussion). Essentially, whereas Dekel, Ely, and Yilankaya (2007) characterize non–morally interpretable stable best responses (implied by P), we characterize morally natural justifications $m(\cdot)$ for stable best responses—and these two objects are not in one-to-one correspondence. In particular, one cannot directly deduce from Dekel, Ely, and Yilankaya (2007) the existence of moral pluralism (as opposed to stable-preference pluralism) in any game Π. In sum, our contribution is to explore whether morality emerges from the particular notion of evolutionary stability put forth by Dekel, Ely, and Yilankaya (2007). In so doing, we are able to provide the first (to our knowledge) evolution-theoretic foundation for moral pluralism.

This article belongs to the indirect evolution of morality literature, which assumes that individuals play rationally for given preferences and that these preferences are subject to evolutionary selection according to the individual fitness of the behavior that they induce; see Alger and Weibull (2013, 2016, 2017) and Alger, Weibull, and Lehmann (2020). Unlike our work, this literature finds that morality is mono-factor in the sense that people are "less or more moral" along a single dimension.[16] The literature in which evolutionary selection operates at the group level has proved capable of generating heterogeneous non–fitness maximizing behaviors, either through cultural norms that vary across different groups (as argued informally by Henrich 2004) or through the long-run coexistence of different behavioral types in the same group (as in Boyd et al. 2003; Bowles and Gintis 2004, 2011). But this "group selection" literature relies on

16. Bester and Güth (1998) directly assume that morality is one-dimensional.

some friction or deviation from the harsh discipline of individual selection.

The literature on the indirect evolution of preferences (see Alger 2023 for a review) has shown that individuals can depart from fitness maximization under complete information (Güth and Yaari 1992; Güth 1995; Dekel, Ely, and Yilankaya 2007; Heifetz, Shannon, and Spiegel 2007a, 2007b). But this result breaks down if preferences are entirely unobservable (Ely and Yilankaya 2001; Güth and Peleg 2001; Ok and Vega-Redondo 2001; Dekel, Ely, and Yilankaya 2007), at least if the standard assumption of random matching is maintained. By contrast, if assortative matching is assumed, then behavior departs from fitness maximization once more, in the direction of "Kantian" behavior: players place some weight on the action that, if played by both players, would result in maximal fitness. The idea that evolution can select specifically for Kantian behavior first appeared in Bergstrom (1995)'s complete-information sibling analysis. This finding spawned a literature exploring what moral preferences are stable with incomplete information and assortative matching (Alger and Weibull 2013, 2016, 2017; Alger, Weibull, and Lehmann 2020), but its operation is quite different from the complete (or almost-complete) information analysis of the current article, and as a result, these papers yield a mono-factor theory of morality. Alger and Weibull's incomplete-information model has received some criticism, for instance, from Newton (2017), who shows that their results are not robust to an evolving degree of assortativity. The stability of Kantian behavior has also been criticized on its own terms by Akdeniz, Graser, and van Veelen (2020), who highlight a problem with Alger and Weibull's original result and model relaxations that raise further difficulties.[17]

MFT has spawned a "number of dimensions" debate concerning how many independent moral principles there really are empirically. The creators of MFT hold that there are five distinct foundations,[18] but others hold that the foundations can be subsumed under two high-level factors: an individualizing factor

17. For instance, if mixed equilibria are allowed, they show that altruism either replicates or outperforms Kantian morality.

18. Graham et al. (2011, 372) argue in favor of the five-factor model, as do Nilsson and Erlandsson (2015), Métayer and Pahlavan (2014), and Doğruyol, Alper, and Yilmaz (2019).

(overlapping with Care and Fairness) and a binding one (overlapping with Loyalty, Authority, and Sanctity).[19] Our findings are nuanced: while all five foundations postulated by MFT withstand the test of an evolutionary process, there are two minimal subsets of the foundations that are sufficient to generate all stable best responses in all games, and each minimal subset happens to contain one individualizing and one binding foundation, arguably consistent with the two-factor reading of the empirical evidence.

The MF questionnaire has recently been used in economics. Enke (2020) uses it to construct a one-dimensional index of "universalism vs communitarianism."[20] Enke (2020) does not seek to map the dimensions of statistical variation in moral attitudes across people; instead, it projects this variation onto a single dimension. In a similar spirit, the "moral kernel" in Enke (2019) is a single-dimensional projection (technically, the first principal component) of five different indices that proxy for moral values and beliefs that, in Enke's estimation, are prevalent in societies that emphasize communitarianism over universalism.[21] Thus, the aim of these papers is to project a multidimensional object (the moral makeup of people) onto a single dimension (a one-dimensional scale spanning communitarianism at one end and universalism at the other). In contrast, the goal of this article (and MFT) is to describe and investigate the number of dimensions of independent variation in the data, that is, the full dimensionality of moral variation across people.

II. The Model

Members of a large population are repeatedly matched at random to play a two-player, two-action normal-form game G

19. Moreira, Souza, and Guerra (2019) and Smith et al. (2017) argue for the two-factor model, and many studies implicitly ignore the five-factor model and skip directly to the two-factor model. See Barnett, Öz, and Marsden (2018), Malka et al. (2016), Napier and Luguri (2013), Stolerman and Lagnado (2020), Süssenbach, Rees, and Gollwitzer (2019), and Yilmaz and Saribay (2017). Enke (2020, 3690) discusses the "dimensionality debate."

20. See Enke (2024) for a recent review.

21. Of the five indices, two are single-dimensional constructs based on Haidt's MFT.

with action set {*cooperate*, *defect*}.[22] The players' *fitness payoffs* (though not necessarily their preferences) in game G are:

(1)

	cooperate	*defect*
cooperate	*a,a*	*b,c*
defect	*c,b*	*d,d*

We think of these payoffs as capturing the players' individual fitness or reproductive success, separate from their moral principles. A particular ordering of $\{a, b, c, d\}$ induces a specific *fitness game* or, with some abuse of notation, a *game*; for example, if $c > a > d > b$, we refer to (1) as a prisoner's dilemma game.

A fitness game is summarized by a *fitness matrix* Π, which captures the row player's fitness. So,

(2) $$\Pi = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Every agent in the population has the same Π. Let $\mathbb{M}$ denote the space of all 2×2 matrices. We examine all configurations of $\Pi \in \mathbb{M}$ subject to $a \geqslant d$; this restriction is without loss of generality and justifies calling "cooperation" the action that, if taken by both agents, leads to a fitness payoff of a for both players.

A mixed action in game G is denoted by the (2×1) column vector σ, and Δ denotes the set of all mixed actions. If a player with fitness matrix Π plays the mixed action σ against an opponent playing σ', then she receives expected fitness $\mathcal{A}(\sigma, \sigma') = \sigma^\top \Pi \sigma'$.

The players do not care directly about their fitness; instead, their preferences are represented by a *preference matrix* $P \in \mathbb{M}$. If a player with preferences P plays the mixed action σ against an opponent playing σ', then she receives expected utility $\mathcal{P}_P(\sigma, \sigma') = \sigma^\top P \sigma'$. 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 $\mu \in \mathcal{F}(\mathbb{M})$, where $\mathcal{F}(\mathbb{M})$ is the set of all finite-support distributions. We assume that matched players observe one another's preference matrices (this assumption is relaxed in Section IV.A). A *strategy* for a player with preference

22. The population is large to abstract away from repeated-game concerns.

$$\Pi \quad + \quad m(x, y; \Pi) \quad = \quad P$$

fitness moral principle preferences

FIGURE I

$m(x, y; \cdot)$ Generates P from Π

matrix P is a function $\beta_P : \mathbb{M} \to \Delta$ that specifies a mixed action conditional on the preference matrix of the matched opponent. A *Nash equilibrium profile* β^* is a strategy profile $\{\beta^*_P(\cdot)\}_{P \in \mathbb{M}}$ that, for all P's, is a best response to the opponent's equilibrium strategy conditional on observing the opponent's preferences P', that is,

$$\beta^*_P(P') \in \arg\max_{\sigma \in \Delta} \mathcal{P}_P(\sigma, \beta^*_{P'}(P)) \text{ for all } P, P' \in \mathbb{M}. \tag{3}$$

We restrict attention to symmetric equilibria, that is, equilibria in which players with the same P play the same equilibrium strategy.

Given a population distribution μ and an equilibrium strategy profile β^*, the average fitness of a player with preference matrix P is

$$\bar{\Pi}_P(\mu \mid \beta^*) = \sum_{P' \in \text{supp}(\mu)} \mathcal{A}(\beta^*_P(P'), \beta^*_{P'}(P)) \cdot \mu(P').$$

II.A. Moral Principles

Our object of interest is the difference between P and Π, which we interpret as reflecting a moral principle. Formally, we define moral principles as rules that take as input any fitness matrix Π and output a matrix which, when added to Π, produces P.

DEFINITION 1 (MORAL PRINCIPLE). A moral principle with weights x and y is a function $m(x, y; \cdot)$ that maps a fitness matrix Π into a 2×2 matrix with nonnegative entries $\{0, x, y\}$ that has at least one zero in each column and, by convention, x in the left column.

A moral principle is a single rule that applies across all Π's and may modify the player's best-response correspondence. That is, because the moral principle $m(x, y; \Pi)$ is added to the player's fitness matrix Π to obtain P (see Figure I), it may cause a player

to behave differently than would be dictated by Π alone. There is no loss of generality (and considerable economy of parameters) in requiring that two entries of the matrix in Definition 1 equal zero. Indeed, the sole function of moral principles is to change a player's best response—and two judiciously placed numbers x and y suffice to achieve any best response. Note that since Definition 1 states that x and y are nonnegative, moral principles represent rewards, not punishments.

Figure I illustrates the relationship between the fitness matrix Π, a moral principle $m(x, y; \cdot)$, and the preferences P generated by m. The fitness matrix Π induces a symmetric fitness game whose payoffs drive reproductive success. However, instead of maximizing Π the player maximizes P, which combines individual fitness and the moral principle.

We are interested in which moral principles m survive the evolutionary process in the fitness game induced by a given Π. For example, we are interested in whether the moral principle $m = Care$ survives the evolutionary process when Π induces a prisoner's dilemma fitness game.

II.B. Some Special Moral Principles: The H5's

Moral principles could be very complicated functions of x, y, and Π. Next we describe five principles that were empirically identified by MFT: Care, Fairness, Authority, Loyalty, and Sanctity. We call these principles the H5's.

To operationalize the H5's as functions $m(x, y; \Pi)$, we need to introduce the concept of Kantian action.

DEFINITION 2 (KANTIAN ACTION). The Kantian action is the pure-strategy action which, when taken by both players, yields the highest social fitness.

Whenever $a > d$, the Kantian action is cooperate.[23] Table I illustrates how we operationalize the H5 moral principles. Online Appendix I provides mathematical definitions of the H5's.

A few comments on how we operationalize the moral foundations. In Table I, Care is operationalized as a concern for the material well-being (i.e., individual fitness) of the other player, and Fairness is operationalized as a preference for more equal outcomes. We regard these operationalizations as consistent with

23. The case $a = d$ is nongeneric and will be ignored henceforth.

TABLE I
OPERATIONALIZING THE H5 MORAL PRINCIPLES

Moral principle	Operational rule: add utils x or y to the entry of Π which . . .
Care $(x, y; \Pi)$	for each given opponent's action makes the opponent best off fitness-wise.
Fairness $(x, y; \Pi)$	for each given opponent's action has the most equal fitness payoffs.
Authority $(x, y; \Pi)$	for each given opponent's action maximizes the row player's fitness relative to the opponent's.
Loyalty $(x, y; \Pi)$	maximizes the opponent's fitness when the opponent chooses the Kantian action and minimizes the opponent's fitness when the opponent does not.
Sanctity $(x, y; \Pi)$	corresponds to the Kantian action regardless of the opponent's action.

Note. Π represents the row player's fitness matrix, "the opponent" is the column player, and "the opponent's action" refers to either column of Π.

the literature and relatively uncontroversial.[24] The other three foundations are not operationalized in the existing literature, so our approach requires some explanation.

Our definition of Loyalty entails conditionality: the obligation runs out (and even reverses) if the object of loyalty behaves unworthily—in our setting, if the other player fails to take the Kantian action. This kind of conditionality separates the philosophical concept of "good" loyalty from the "bad" or unvirtuous loyalty that is blindly bestowed even on antisocial objects.[25] That (good) loyalty should be conditional makes sense: an employee would not be considered disloyal for whistle-

24. For example, Bester and Güth (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 Online Appendix II for a proof of these statements.

25. The classic example of unvirtuous loyalty is that of the loyal Nazi. The distinction between good and bad loyalty is discussed by Ewin (1992), who regards unlimited or uncritical loyalty as "bad loyalty." Similarly, the Stanford Encyclopedia of Philosophy's (SEP) entry on Loyalty has a section titled "Limiting Loyalty" in which one reads: "Loyalty to a particular object is forfeited—that is, its claims for the protection and reinforcement of associative identity and commitment run out—when the object shows itself to be no longer worthy or capable of being a source of associational satisfaction or identity-giving significance. That is, the claims run out for the once-loyal associate" (Kleinig 2022). This quote indicates that the SEP defines loyalty as inherently limited, that is, conditional, and thus virtuous, as we also do.

blowing on an employer who behaves unlawfully.[26] Our definition of Loyalty thus captures good (or virtuous) loyalty, arguably consistent with the ChatGPT blurb, which also captures "good loyalty" because it highlights prosociality (the well-being of the community or the nation) and not blind or unconditional loyalty.

Our definition of Authority captures a preference for high relative fitness, that is, for favorable inequality.[27] That humans have a preference for high relative outcomes is fairly uncontroversial.[28] We simply note that in a primitive society, favorable inequality, that is, priority over access to food and reproductive opportunities coincides with high status. The perquisites of high status can be summarized abstractly as being respected, deferred to, and obeyed. A preference for these perquisites is therefore implicitly a preference for favorable inequality. In this sense, our definition of Authority is consistent with the ChatGPT blurb.

Last, we define Sanctity as supreme selflessness: doing the right thing for the community regardless of personal cost. In our definition, it means fully internalizing the externalities of one's actions on the community. In an evolutionary context, some of the most critical externalities for community survival were related to communicable diseases, requiring personal hygiene and proper waste disposal. Indeed, according to MFT, Sanctity evolved to mitigate the private incentive to underinvest in preventing the spread of contagious microbes and parasites among people living in close proximity and sharing food; the emotion associated with a violation of sanctity is disgust (Graham et al. 2013, 68). Then, "once human beings developed the emotion of disgust and its cognitive component of contagion sensitivity, they began to apply the emotion to other people and groups for social and symbolic reasons that sometimes had a close connection to health concerns"

26. This case is examined by Varelius (2009).

27. Note that Authority is not the opposite of Fairness: the opposite of Fairness is a preference for inequality, whereas Authority is a preference for favorable inequality.

28. In a classic study (Solnick and Hemenway 1998), survey respondents were asked to choose between a world where they had more income than others and one where everyone's income was higher but the respondent had less than others. Half of the respondents preferred to have 50% less real income but high relative income. Bault, Coricelli, and Rustichini (2008) report similar findings. For a review of the empirical literature on status, see Anderson, Hildreth, and Howland (2015).

(Haidt and Graham 2009, 382). In this view, the moral virtue we call sanctity evolved in response to a health risk caused by selfish community members. Then its application expanded so that in modern times, we say that a saint is someone who is entirely selfless. This is why we define sanctity as doing the right thing for the community, that is, taking the Kantian action irrespective of one's material incentives.

We acknowledge that because our model is so stylized, some of our operationalizations can be questioned. For example, as concerns Loyalty, a natural definition in many contexts would be a preference for behaving kindly toward those whom one considers part of an in-group and spitefully toward those whom one considers part of an out-group. Lacking a definition of in- and out-groups, we approximate this notion as behaving kindly toward players whose actions deserve it and spitefully toward players who do not. Also, our definition of Authority only captures a part of what goes on in a hierarchical or power relationship; the element of respect for hierarchy, or deference to power, is missing. However, the model is too stylized to embed a hierarchical relationship, so our definition is, of necessity, limited. Finally, concerning Sanctity, some ethnographic literature questions whether the actions that are associated with sanctity in a given culture are in fact prosocial or, instead, they emerge arbitrarily out of cultural accident. Our operationalization of Sanctity focuses on the prosocial view. In general, some readers may feel that formalizing complex constructs such as loyalty, authority, and sanctity within the confines of a 2×2 simultaneous game with perfect information is a stretch (though the extensions to partial observability and $n \times n$ games in Section IV should reassure these readers). We agree, but we also see a virtue in the simplicity of this approach.

The next example illustrates the H5's in the context of a coordination game.

EXAMPLE 1 (H5 MORAL PRINCIPLES IN A COORDINATION GAME). The fitness matrix

$$\Pi = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix}$$

represents the row player's fitness payoff in the following coordination game:

	cooperate	*defect*
cooperate	$2, 2$	$-1, 1$
defect	$1, -1$	$0, 0$

.

The moral principle Care with weights x, y gives the row player extra utils in the entry which, for each given opponent's action (i.e., column) makes the opponent best off fitness-wise. Therefore,

$$\textit{Care}\,(x, y; \Pi) = \begin{bmatrix} x & y \\ 0 & 0 \end{bmatrix}.$$

The moral principle Fairness gives extra utils to the entry which, for each given opponent's action, has the most equal fitness payoffs. Therefore,

$$\textit{Fairness}\,(x, y; \Pi) = \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix}.$$

The moral principle Authority gives extra utils to the entry which, for each given opponent's action, makes the row player best off fitness-wise relative to the opponent (i.e., maximizes the row player's fitness advantage). Therefore,

$$\textit{Authority}\,(x, y; \Pi) = \begin{bmatrix} 0 & 0 \\ x & y \end{bmatrix}.$$

Applying the moral principle Loyalty requires identifying the Kantian action, which is cooperate. Loyalty requires boosting the utils of the row player who (i) conditional on the opponent playing cooperate, picks the action that makes the opponent best off fitness-wise; and (ii) conditional on the opponent playing defect, picks the action that makes the opponent worse off fitness-wise. Therefore,

$$\textit{Loyalty}\,(x, y; \Pi) = \textit{Fairness}\,(x, y; \Pi).$$

Finally, the moral principle Sanctity requires boosting the utils of the row player who plays the Kantian action, so:

$$\textit{Sanctity}\,(x, y; \Pi) = \textit{Care}\,(x, y; \Pi).$$

Example 1 reveals that several distinct moral principles can produce the same matrix $m(x, y; \Pi)$: for example, Fairness and Loyalty happen to produce the same matrix. Therefore, for this

Π, Fairness and Loyalty are not distinguishable. We pick up the issue of distinguishability in Section III.A.

Formally, we denote by **H5** the set that comprises all the moral principles listed in Table I, that is,

$$\mathbf{H5} = \{m(x, y; \cdot) \mid m \in \{Care, Fairness, Loyalty, Authority, Sanctity\}; \\ x, y \geqslant 0\}.$$

Online Appendix I provides mathematical definitions of the H5's. Online Appendix II highlights previous publications in which our operationalizations of different moral principles have previously surfaced, showing that our operationalizations are not entirely idiosyncratic.

II.C. Stability and Moral Pluralism

Dekel, Ely, and Yilankaya (2007) introduced a definition of stability for a given Π; we adopt their definition and lay it out using our notation. Recall that β^* denotes an equilibrium strategy profile, and μ is a population preference distribution. Throughout the article, agents with preferences in $supp(\mu)$ are referred to as incumbents, and agents with preferences not in $supp(\mu)$ are referred to as mutants: we also refer, equivalently, to mutant preferences and mutant moral principles. A pair (μ, β^*) is referred to as a *configuration*.

The first feature of a stable configuration (μ, β^*) is that all incumbents receive the same individual fitness—a property referred to as *balance*.

DEFINITION 3 (BALANCED CONFIGURATIONS). A configuration (μ, β^*) is balanced if

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

A configuration (μ, β^*) is balanced if the average fitness of all incumbents is the same. Intuitively, balance rules out (inherently unstable) scenarios where higher-fitness incumbents would spread at the expense of lower-fitness ones.

The second feature of a stable configuration pertains to how incumbents react to a small injection of mutants. Intuitively, a stable configuration is such that an injection of mutants does not change the equilibrium play among incumbents but in any equilibrium against any mutant, induces behavior that leaves the incumbents better off than the mutant. To formalize this statement, we need the following definition.

DEFINITION 4 (FOCAL STRATEGY PROFILES). Fix an equilibrium strategy profile β^* and some $\mu \in \mathcal{F}(\mathbb{M})$. The equilibrium strategy profile $\tilde{\beta}$ is focal on (μ, β^*) if $\tilde{\beta}_P(P') = \beta^*_P(P')$ for all $P, P' \in \text{supp}(\mu)$.

In this definition, intuitively, the pair (μ, β^*) denotes equilibrium play among incumbents, and $\tilde{\beta}$ denotes equilibrium play after an injection of mutants. The profile $\tilde{\beta}$ is focal on (μ, β^*) if it prescribes the same behavior as β^* whenever the player and her opponent are incumbents, but it may differ when the player or her opponent are mutants.

Denote by $B(\mu, \beta^*)$ the set of all equilibrium strategy profiles that are focal on (μ, β^*). Given any μ and mutant preference $\tilde{P}$, let $N_\varepsilon(\mu, \tilde{P}) = \{\mu' : \mu' = (1 - \varepsilon')\mu + \varepsilon'\tilde{P}, \varepsilon' < \varepsilon\}$ denote the set of post-mutation preference distributions resulting from the entry of at most ε mutants with preference $\tilde{P}$.

DEFINITION 5 (STABLE CONFIGURATIONS). A configuration (μ, β^*) is stable if it is balanced and if there exists $\varepsilon > 0$ such that, for every $\tilde{P} \in \mathbb{M}$ and $\tilde{\mu} \in N_\varepsilon(\mu, \tilde{P})$,

$$\bar{\Pi}_P(\tilde{\mu} \mid \tilde{\beta}) \geqslant \bar{\Pi}_{\tilde{P}}(\tilde{\mu} \mid \tilde{\beta}) \quad \text{for all } \tilde{\beta} \in B(\mu, \beta^*) \text{ and } P \in \text{supp}(\mu).$$

At a stable configuration, incumbents must all do equally well (balance) and weakly better than any mutant that might enter with preferences $\tilde{P}$, provided that mutant entrants are fewer than ε and do not disturb equilibrium play among incumbents (i.e., the post-entry equilibrium profile is focal on the pre-entry one). Stable configurations need not be unique.

Now we are ready to define stable preferences and stable moral principles.

DEFINITION 6 (STABLE PREFERENCES AND MORAL PRINCIPLES).

(i) A preference P is stable for Π if $P \in \text{supp}(\mu)$ for some stable configuration (μ, β).
(ii) A moral principle $m(x, y; \cdot)$ is stable for Π if the matrix $\Pi + m(x, y; \Pi)$ is stable for Π.

This stability concept is static and similar, in spirit, to neutral stability (Maynard Smith 1982).[29] For any given Π, Dekel, Ely, and Yilankaya (2007) characterize all stable preferences P;

29. Specifically: like neutral stability, Definition 6 makes it relatively easy for preferences to be stable because it only requires that mutants not be strictly (as

we list them in Online Appendix V. In Online Appendix V we also report, for each Π, the H5 moral principles that generate each stable preference.

Whenever two different moral principles in the H5 are stable for the same Π, we have moral pluralism.

DEFINITION 7 (MORAL PLURALISM). Π gives rise to moral pluralism if there is a stable distribution of preferences μ whose support includes more than one moral principle in the H5's.

Moral pluralism implies that there are at least two moral principles $m, m' \in \mu$. In our setting, m and m' will play against each other with some frequency in game Π, showing that several moral principles can coexist within the same fitness game. In our setting, moral pluralism arises for two reasons. First, certain game types Π admit several strategically nonequivalent stable preferences P, each inducing a different best response; second, a stable preference P may be generated by several different functions m, all of which give rise to the same best response. Both phenomena are illustrated in the next section.

II.D. Modeling Moral Principles Versus Modeling Preferences

A key innovation relative to Dekel, Ely, and Yilankaya (2007) is that we introduce several distinct moral principles, which are (empirically motivated from MFT) functions that map Π into a single element of $\mathbb{M}$. By contrast, Dekel, Ely, and Yilankaya (2007) characterize a single multivalued correspondence that maps each Π into the set of stable preferences for that Π. This abstract correspondence is not empirically motivated and is not automatically interpretable as describing "morality"—if nothing else, because interpretational problems arise when the correspondence is multivalued, that is, when it prescribes both cooperate and defect as strict best responses against some action (as in Example 1, where P and P' are stable and not strategically equivalent).

To further draw the distinction between moral principles and preferences, observe that even if we restrict attention to stable preferences, there is no one-to-one correspondence between stable H5 moral principles and stable preferences: for a given Π, a given stable preference P may be generated by several different

opposed to weakly, as in the stability concept of Maynard Smith and Price (1973) and Maynard Smith (1974)) better off than incumbents.

moral principles in the H5's. Vice versa, a given moral principle $m(x, y; \cdot)$ in the H5's can generate several different stable preferences depending on whether x and y are each large or small.

For instance, for the Π of Example 1, the two preferences

$$P = \begin{bmatrix} 3 & 3 \\ 1 & 0 \end{bmatrix} \text{ and } P' = \begin{bmatrix} 4 & -1 \\ 1 & 0 \end{bmatrix}$$

are stable and not strategically equivalent.[30] Note that P can be generated by $Care(1, 4; \Pi)$ or by $Sanctity(1, 4; \Pi)$: hence, these two moral principles generate the same preferences in this game. Vice versa, a single moral principle Care can also generate P and P' (the latter from $Care(2, 0; \Pi)$). Thus, for each given fitness game Π, different moral principles can lead an agent to the same best response, and the same moral principle can be used to justify different best responses depending on the amount of moral satisfaction (number of utils) that the moral principle attaches to different outcomes.

Essentially, Dekel, Ely, and Yilankaya (2007) characterize stable best responses, whereas we are interested in the moral justification for stable best responses. Since the two are not in one-to-one correspondence, the characterization of stable preferences provided by Dekel, Ely, and Yilankaya (2007), though useful, is insufficient to investigate morality.

III. Results

III.A. Distinguishability of the H5's

According to MFT, the H5's are empirically distinguishable from each other; but are they, theoretically, within our model? That is, do they generate different behavior? For a given Π, the answer is no: Example 1 shows that several different H5 moral principles generate the same preferences. This makes sense: in any given 2×2 game, we cannot expect five different best responses. In what sense, then, can the H5's be said to generate different behavior? Theorem 1 provides the answer.

To state Theorem 1, we need two definitions.

30. We know this from Dekel, Ely, and Yilankaya (2007), see also Stability Table 1 in Online Appendix V.

DEFINITION 8 (STRATEGIC EQUIVALENCE). Two preferences P and P' are strategically equivalent if their best-response correspondences are the same.

Strategic equivalence means that for every action taken by the opponent, including mixed actions, the player's best response(s) under P and P' are the same.

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 $x, y \geqslant 0$.

Intuitively, the moral principle m spans P if an agent with preference $\Pi + m(\Pi)$ best-responds exactly like an agent with preference P would against any possible opponent. We are now ready to state Theorem 1.

THEOREM 1 (PAIRWISE DISTINGUISHABILITY OF THE H5'S MORAL PRINCIPLES). For any pair m and m' of H5 moral principles, there exists a fitness matrix Π and a pair of preferences P and P' that are stable for it and are not strategically equivalent, such that m spans P but not P' and, conversely, m' spans P' but not P.

Proof. See Online Appendix II. □

Theorem 1 says that any two H5 principles generate different stable behavior in some game—even though they cannot generate different stable behavior in all games. This is a good sanity check because if any two elements of the H5's were not distinguishable, our 2×2 setting would be too low-dimensional to allow for five different moral principles.

Theorem 1 is not trivial because in any given game, many moral principles generate the same preferences. This can be seen in Example 1. 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 moral principle (or possibly an incumbent one) against which m prescribes a different equilibrium play than m'.

III.B. The H5's Span the Set of Stable Preferences

MFT asserts that empirically, the H5's account for all the moral behavior observed "in the wild." Analogously, in our theoretical model, we hope to generate all evolutionarily stable pref-

erences using the H5's. It is not obvious that this should be possible—after all, the H5's are just a few among all possible moral principles. Indeed, the next counterexample shows that the H5's are not rich enough to generate all possible (including non-stable) preferences.

COUNTEREXAMPLE 1 (PREFERENCE MATRICES GENERATED BY THE H5'S DO NOT SPAN THE SPACE OF ALL MATRICES). There exists a pair Π, P such that no moral principle in the H5's spans P.

Proof. See Online Appendix II. □

Counterexample 1 says that, for the purpose of generating preferences P, there is loss of generality in restricting attention to the H5's. Hence, it is not obvious that the H5's could 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, that is, for almost any fitness matrix Π, the H5'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.[31]

DEFINITION 10 (NONGENERICITY). Given two sets A, B belonging to a vector space V with $A \cap B \neq \emptyset$, A is nongeneric in B if it has lower dimension than B. If B is the parent space to which A belongs, then, for short, A is said to be nongeneric.

A property holds nongenerically in a set if it holds only on a subset of lower dimension than the whole set.[32] For example, a point is nongeneric in the real line, but an interval $[a, b]$ is not; the real line is nongeneric 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[33] is generic because it has the same dimension (four) as $\mathbb{M}$, the following game has dimension three and is thus nongeneric.

31. A similar notion of genericity is used by Govindan and Wilson (2012).

32. 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.

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

EXAMPLE 2 (PURELY COMMON-VALUE HAWK-DOVE GAME IS NONGENERIC). The fitness matrix where $b = c$

$$\begin{bmatrix} a & b \\ b & d \end{bmatrix}$$

and $b > a \geqslant d$, induces a pure common-value hawk-dove fitness game. The set of all such matrices is nongeneric in the space $\mathbb{M}$ of 2×2 matrices.

Rather than applying the concept of genericity to sets of matrices in $\mathbb{M}$, it will be expositionally convenient to apply this notion to the graphs in $\mathbb{M}^2$ that map fitness matrices into preference matrices.[34] We use this graph as a convenient litmus test for nongenericities in the space of fitness matrices and in the space of preference matrices. This is because a graph is nongeneric in $\mathbb{M}^2$ if either its domain or range are nongeneric in $\mathbb{M}$; in other words, a graph "inherits" any nongenericities in its domain and range.[35] 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 $\mathcal{H}5$ be the graph of the set of preferences spanned by the H5's, that is,

$$\mathcal{H}5 = \{(\Pi, P) \in \mathbb{M}^2 \mid P = \Pi + m(x, y; \Pi),\ m \in \mathbf{H5},\ x, y \geqslant 0\}.$$

We are now ready to state the main result of this subsection.

THEOREM 2 (THE H5'S GENERICALLY SPAN THE STABLE PREFERENCES). The graph $\mathcal{ES} \backslash \mathcal{H}5$ is nongeneric in $\mathcal{ES}$.

Proof. See Online Appendix II. □

The interpretation of Theorem 2 is that the H5's are "rich enough" to transform almost all fitness matrices Π into *almost all* their associated stable preferences P. Put differently, the H5's are capable of spanning almost all stable preference matrices for almost all Π's. This result is nonobvious because the H5's are not capable of spanning all (including nonstable) preference matrices.[36]

34. 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 $\mathbb{M}$.

35. 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 nongeneric.

36. See Counterexample 1. The contrast between Counterexample 1 (the H5's can't span) and Theorem 2 (the H5's span) is substantive—it is not a technical ar-

Therefore, Theorem 2 uncovers a nonobvious connection between the H5's and evolutionary stability.[37]

Next we develop intuition for why just a few moral principles, namely, the H5's, suffice to span all stable preferences (Theorem 2). Articulating this intuition requires stating the following definition.

DEFINITION 11 (SOCIAL FITNESS–MAXIMIZING PREFERENCES). A preference matrix P is social 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 individual fitness payoffs.

Intuitively, this definition says that a preference matrix is social fitness–maximizing if its Nash equilibrium implements the "socially fittest" mixed action. Note that social fitness–maximization may differ from implementing the Kantian action (Definition 2) because the latter is defined as a pure strategy.

From Dekel, Ely, and Yilankaya (2007), we know that a preference can only be stable for Π if it is social fitness–maximizing for Π in the sense of Definition 11. Moreover, for almost all Π's that have a stable preference, the social fitness–maximizing

tifact of the fact that Theorem 2 only holds generically, whereas Counterexample 1 is stated uniformly. Indeed, the nonspanning property in Counterexample 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 H5's can span any preference with $c > a$ and $b > d$, which are also generic in the space of 2×2 matrices.

37. What (nongeneric) 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 H5's. This is the pure common-value hawk-dove game of Example 2, which, of course, is nongeneric 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 H5's. The nonspanned preferences all have the form

$$\begin{bmatrix} e & f \\ e & h \end{bmatrix},$$

which is, of course, nongeneric in $\mathbb{M}$ (depending on the values of f and h, these preferences specialize to, in the notation of Dekel, Ely, and Yilankaya (2007), $\mathcal{BA}_1$, $\mathcal{AB}_1$, and θ_0). Furthermore, for those Π's, the nonspanned preferences are even nongeneric in $\mathcal{ES}(\Pi)$, meaning intuitively that the quasi-totality of the stable preferences are in fact spanned by the H5's.

mixed action happens to be the pure strategy cooperate,[38] and every preference matrix whose equilibrium set includes cooperate is strategically equivalent to:

$$\begin{bmatrix} e & f \\ 0 & 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} e & 0 \\ 0 & f \end{bmatrix} \tag{4}$$

for some $e, f \geqslant 0$. So for almost all Π's that have a stable preference, every stable preference must look like one of the two matrices in expression (4). The H5's are capable of spanning almost all these matrices for any Π. Therefore, the H5's are capable of spanning almost all stable matrices for almost all Π's. The salient feature of expression (4) is that the left matrix rewards the row player's cooperation unconditionally; the right matrix rewards it conditionally. Theorem 2 implies that these two reward patterns suffice to generate all stable preferences. Section IV.B shows that this intuition carries over beyond the 2×2 case to the $n \times n$ case.

We conclude with a housekeeping observation.

REMARK 1 (EACH H5 IS STABLE IN SOME GAME). Every element of the H5 is stable in at least one game Π.

Proof. Every element of the H5 is stable in all coordination games, that is, all games in which $a > c$ and $d > b$: see Stability Table 1 in Online Appendix V. □

III.C. Minimal Spanning Moral Principles

Theorem 2 shows roughly that the H5's are sufficient to span the set of almost all stable preferences for almost all fitness matrices. But it does not say that the H5's constitute a minimal set: a smaller set could, in principle, suffice to span almost all 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:

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

similarly, let $\mathcal{CL}$ denote the set generated by $\{Care, Loyalty\}$. Finally, let $\mathcal{D}$ denote the graph of the set of preferences generated

38. The only Π that has a stable preference for which the social fitness–maximizing mixed action is not the pure strategy cooperate is the common-value hawk-dove game of Example 2, which happens to be nongeneric.

by some subset D of the H5's, that is:

$$\mathcal{D} = \{(\Pi, P) \in \mathbb{M}^2 \mid P = \Pi + m(x, y; \Pi),\ m \in D \subseteq \mathbf{H5},\ x, y \geqslant 0\}.$$

The next theorem shows that there are two subsets D of the H5's, each of which is sufficient to span almost all the evolutionarily 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 this sense, these two subsets are minimal.

THEOREM 3 (TWO MINIMAL SPANNING PAIRS). If $\mathcal{D} \in \{\mathcal{FS}, \mathcal{CL}\}$, the graph $\mathcal{ES}\backslash\mathcal{D}$ is nongeneric in $\mathcal{ES}$. Moreover, if the graph $\mathcal{ES}\backslash\mathcal{D}$ is nongeneric in $\mathcal{ES}$, then $\mathcal{D}$ must contain $\mathcal{FS}$ or $\mathcal{CL}$.

Proof. See Online Appendix II. □

The theorem says that we can rationalize all evolutionarily stable moral behavior in almost all fitness games by appealing to fewer than five moral principles: either Sanctity and Fairness or Care and Loyalty. Intriguingly, these subsets happen to contain at least one individualizing component and one binding component, consistent with the two-factor version of MFT.[39]

Next we provide an 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,[40] just like the right matrix in expression (4); Sanctity boosts the top row of Π, just like the left matrix in expression (4). Therefore, intuitively, Fairness and Sanctity "morph" the matrix Π toward the matrices in expression (4), which we know are the only stable preferences for all Π. The mechanics are 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 expression (4) and Loyalty morphs

39. Recall that MFT classifies Care and Fairness as individualizing, whereas Loyalty and Sanctity are classified as binding foundations. It is encouraging that our mathematical definitions of these moral principles happen to be consistent with MFT's intuition behind this classification, which is that binding foundations promote the welfare of a larger group, whereas individualizing foundations do not (refer to footnote 8). Indeed, per Table I, our mathematical definitions of Loyalty and Sanctity are based on a concept of prosociality (the Kantian action), whereas our definitions of Care and Fairness are not. In this sense, our mathematical definitions are consistent with MFT's classification.

40. The exception being the common-value hawk-dove game of Example 2, which is nongeneric.

Π toward the other matrix. Because being able to span both matrices in expression (4) guarantees that almost all stable preferences can be spanned (refer to the discussion following expression (4), we have given the intuition we set out to provide.

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 nongeneric stable preferences, and Theorem 3 does not cover these nongeneric preferences. In this sense, Authority plays a less important role in our theory.

Theorem 3 immediately implies that moral pluralism (in the precise sense of Definition 7) is a feature of all fitness games Π. We record this observation next.

COROLLARY 1 (MORAL PLURALISM IS UBIQUITOUS). Generically, a fitness game Π gives rise to moral pluralism.

Proof. By Theorem 3, generically, game Π has at least a stable moral principle in the set $\{Fairness, Sanctity\}$ and another in the set $\{Care, Loyalty\}$. □

III.D. The Design of Minimal Moral Codes

We consider the problem of a hypothetical designer who has the power to instill certain H5 moral principles early on in people's lives. These principles, we posit, are instilled through early education and are Π-specific. For example, all children might be taught to apply Care when playing a prisoner's dilemma game and Fairness when playing a hawk-dove game. Following the early-education phase, these moral principles become subject to evolutionary pressures throughout a person's life, including interacting with mutant moral principles that the designer cannot control.[41]

We imagine that the designer follows two criteria when picking which moral principles to instill. First, intuitively, the chosen moral principles should be "evolution-proof," meaning that whatever moral principles are instilled in young children should not be disadvantageous to their individual fitness during a lifetime of social interaction. Second, intuitively, the designer seeks to econ-

41. In this section, evolutionary selection is interpreted as an adaptive learning process that takes place throughout a person's life; refer to Section V.A for a discussion of how to interpret evolution in our model.

omize on the number of distinct moral principles that are taught in school.

Formally, the first criterion identifies the correspondence $\mathcal{ES}(\Pi)$ that associates to each Π all the evolutionarily stable preferences, if they exist. For any given Π, there can be several stable preferences, so $\mathcal{ES}(\cdot)$ is a correspondence, not a function. Then, the designer identifies the selection(s) from the correspondence $\mathcal{ES}(\cdot)$ that can be generated by the fewest distinct moral principles in the H5's. In this second step, the designer is allowed to ignore a nongeneric set of Π's. The set of moral principles required to generate this selection is a minimal moral code.

What does a minimal moral code look like? For each (generic) Π that has a stable preference, the moral code must include at least one stable moral principle in the H5. In the hawk-dove game, it turns out that the only stable moral principles are Fairness and Loyalty, so any moral code must include at least one of these principles. Furthermore, the proof of Theorem 4 shows that Fairness and Loyalty happen to be stable in all games. Therefore, there are only two minimal moral codes: one based on Fairness and the other based on Loyalty. This observation is recorded in the following theorem.

THEOREM 4 (MINIMAL MORAL CODES). There are only two minimal moral codes: one based on Fairness, the other based on Loyalty.

Proof. See Online Appendix II. □

Theorem 4 differs from Theorem 3 in that the former is minimalistic, whereas the latter is comprehensive: Theorem 3 seeks to span all stable preferences for every Π, whereas Theorem 4 limits itself to the *minimal number* of moral principles that span at least one stable preference for every Π. A real-world interpretation of Theorem 4 is offered next.

MFT empirically asserts that individuals vary in their moral perceptions along at least two dimensions, and our theory shows that the evolutionary process admits at least two stable dimensions of independent moral variation (Theorem 3). In other words, if left undisturbed, evolution is consistent with moral pluralism. Theorem 4, in contrast, asserts that if a designer can control the number of moral principles present in a society, they can deliver all available fitness improvements with a single stable moral principle. In this sense, the designer problem admits a morally

totalitarian solution, which can load either exclusively on Fairness or exclusively on Loyalty (Theorem 4). Intriguingly, survey respondents who place more emphasis on the former (resp., latter) are more likely to be politically liberal (resp., conservative).[42] In this sense, Theorem 4 can be interpreted as saying that a designer can deliver all the available fitness improvements by relying either on a purely liberal or a purely conservative moral code—moral pluralism is not needed.

It is also worth contrasting Theorem 4 with the results of Alger and Weibull (2013). In their paper, the only moral code that is stable in all fitness games is referred to as *homo moralis*, corresponding, in our terminology, to Sanctity.[43] Of note, Sanctity is not featured in Theorem 4. This difference is not unexpected, of course, because our settings are different.

III.E. Intensity of Moral Principles and Moral Overdrive

One can ask whether, given a particular game Π, the stable matrices $m(x, y; \Pi)$ are "large," "small," or must be "just right" relative to Π. The case of interest is when these stable matrices 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, compatibility with moral overdrive means that it is evolutionarily stable for the moral component of preferences to swamp individual fitness as a decision-making criterion, and to totally drive decision making in game Π. Next we show that every moral principle in the $H5$ is compatible with moral overdrive in at least one game Π.

42. "Haidt and Graham (2007) . . . make the simple prediction that liberals would show greater reliance than conservatives upon the Care and Fairness foundations . . . whereas conservatives would show greater reliance upon the Loyalty and Authority foundations . . . and the Sanctity foundation. . . . They found support for their prediction, and this basic pattern has been found in many subsequent studies, using many different methods." Cited from Graham et al. (2013, 75).

43. In Online Appendix II we trace the formal connection between Sanctity and *homo moralis*.

LEMMA 1. Every moral principle in the H5's is compatible with moral overdrive in at least one game Π.

Proof. See Online Appendix II. □

Even though all the H5's are compatible with moral overdrive in some game, there is an interesting class of fitness games in which no stable moral principle is compatible with moral overdrive. This is the class of fitness games where stable moral principles actually improve social fitness. Social fitness–improving moral principles are defined next.

DEFINITION 13 (STRICTLY SOCIAL FITNESS–IMPROVING MORAL PRINCIPLES). A moral principle m is strictly social fitness–improving for Π if, for some $x, y \geqslant 0$, $\Pi + m(x, y; \Pi)$ is social fitness–maximizing for Π whereas Π is not social fitness–maximizing for itself.

Intuitively, a moral principle is strictly social fitness–improving if the symmetric Nash equilibrium among two players who adopt the principle generates strictly more social fitness than any symmetric Nash equilibrium among two players who play according to Π. Not many game types Π have strictly social fitness–improving moral principles for the simple reason that often, the best Nash equilibrium of Π is already social fitness–maximizing and thus not susceptible to improvement.[44] Only in the prisoner's dilemma and hawk-dove games is there a mixed action σ that, when played by both players, delivers a strictly higher social fitness than the best symmetric Nash equilibrium of Π. If we restrict attention to these games, it turns out that no stable moral principle (including those not in the H5's), is compatible with moral overdrive. This is what the next theorem says.

THEOREM 5 (MORAL OVERDRIVE IS INCOMPATIBLE WITH SOCIAL FITNESS IMPROVEMENT). There is no game Π in which a strictly social fitness–improving moral principle is compatible with moral overdrive.

Proof. See Online Appendix II. □

44. This observation highlights that moral principles are not stable because they improve social fitness. Instead, what makes a moral principle evolutionarily stable is that it is not a liability in the competition against mutants.

Theorem 5 says that if a moral principle is stable and fitness-improving, it cannot be an overriding principle, meaning x and y cannot be too large relative to Π. The reason is most transparent in the prisoner's dilemma. In this fitness game, all stable moral principles m in the H5 place x in the top-left entry of the matrix $m(x, y; \Pi)$.[45] Any stable moral principle must have two features. First, incumbents must cooperate when they meet another incumbent; because all incumbents cooperate in a stable configuration, this says that $x \geqslant c - a$. Second, incumbents must be willing to defect against any mutant who cooperates with a probability less than one, so it must be that $x \leqslant c - a$. These requirements together pin down the condition $x = c - a$, which is incompatible with moral overdrive. Intuitively, an evolutionarily stable morality must direct people to cooperate with others who share the stable morality, but to avoid being taken advantage of by mutants who behave amorally. To satisfy both requirements, moral principles must be commensurate with Π. To put it more colorfully, "turn the other cheek" can be a stable moral principle only up to a point; a society in which meekness is an overriding moral value will be taken over by mutants who behave amorally.[46]

IV. Extensions

IV.A. Relaxing the Observability Assumption

An important assumption in our theory is that moral principles are perfectly observable to the opponent.[47] Is it reasonable to assume that players can condition their strategy on (an accurate guess of) their opponent's moral principles? We think so. There is a literature in evolutionary psychology showing that humans are surprisingly adept at inferring the moral disposition even of relative strangers from cheap talk, visual cues, or involuntary tells. For example, experimental subjects who did not previously know each other were substantially more accu-

45. In a prisoner's dilemma, all moral principles in the H5 except Authority are stable.

46. Note that the statement of Theorem 5 is not restricted to moral principles in the H5's.

47. This is an important difference from Alger and Weibull (2013), who do not make this assumption. Instead, they assume assortative matching along similar "moral types."

rate in predicting their opponents' action (cooperate or defect) in a one-shot prisoner's dilemma if they were allowed to interact for 30 or even just 10 minutes before playing with each other (Frank, Gilovich, and Regan 1993). Similarly, pictures of a player's face taken at the time of decision (cooperate or defect) in a prisoner's dilemma game allowed external observers to predict the player's action (Vertplaetse, Vanneste, and Braeckman 2007). Finally, people were able to distinguish altruistic from non-altruistic strangers based solely on how the stranger told a fable (Little Red Riding Hood) on video; the stranger's facial expressions, including involuntary ones, seemed to hold important clues.[48]

Our argument is that humans have access to cues about their opponent's moral dispositions before playing a game with them. This is even more plausible in the conditions under which the evolutionary process took place, that is, in small communities. Indeed, Helzer et al. (2014) show that one's moral makeup is assessed quite accurately by those whom one interacts with more than occasionally: the authors document that experimental subjects assessed their own moral character in a way that aligned with the assessment of the subject's friends, family members, and acquaintances.

With this being said, perfect observability is admittedly an assumption of convenience: it allows us to leverage the machinery developed by Dekel, Ely, and Yilankaya (2007). It seems reasonable to assume partial observability in the evolutionary process, meaning that usually, but not always, humans interacted with others whose moral makeup was known to them. In what follows, we discuss which of our results survive and which need modification if we allow for high—but not perfect—observability.

The answer depends on how we implement partial observability. We study two scenarios. In Scenario A, the event that an agent observes her opponent's preferences is independent of the event that her own preferences are observed by the opponent. In Scenario B, these two events are perfectly correlated: whenever an agent observes her opponent's preference, the converse is true also. Overall, Theorems 1–3 remain unchanged, while Theorems 4 and 5 require adjustment depending on how imperfect observ-

48. See Brown, Palameta, and Moore (2003). Altruism was both self-reported and manipulated. Altruists were perceived as more "helpful," "concerned," and "attentive."

ability is modeled. The adjustments reflect the fact that with less-than-perfect observability, the hawk-dove game and, in Scenario A, the prisoner's dilemma do not admit any stable preference. This makes the statements in Theorems 4 and 5 easier to satisfy, but in Scenario A, trivially so. See Online Appendix III for details.

IV.B. $n \times n$ Games

Up to this point, in common with much of the literature on indirect preference evolution, we have limited our analysis to 2×2 games. It is clearly of interest to generalize our approach to larger games. This exercise raises a number of technical difficulties when there are more than two actions and in the case of many players, discussed by Fehr and Fischbacher (2003). We manage to make progress on the many-action case.

Fully extending the analysis to $n \times n$ fitness payoffs is challenging when $n > 2$, partly because the space of all $n \times n$ fitness matrices is not well-ordered with respect to cooperation/conflict—in a sense that will be made precise below. In this section, we make some progress by restricting attention to fitness matrices Π that are well-ordered with respect to conflict—a set that, we argue, is rich enough to be of economic interest. In this subset, we extend Theorems 1–5, focusing on (suitable generalizations of) the two principles of Care and Loyalty.

First, some notation. Let Π be an $n \times n$ matrix with generic element π_{ij}. Without loss of generality, let the rows of Π (the actions, in our context) be ordered so that the diagonal entries of Π decrease from top left to bottom right; this implies that choosing top rows is "more Kantian" than choosing bottom rows. Let $\mathbb{D}_n$ denote the space of all such matrices. Preferences P are now $n \times n$ matrices. The concepts of evolutionary stability, moral overdrive, and strict fitness improvement extend verbatim from Definitions 6, 12, and 13, respectively.

Next we introduce a notion of "well-orderedness" on the space of fitness matrices. Our result (Theorem 6) will concern well-ordered fitness matrices only.

DEFINITION 14 (WELL-ORDERED FITNESS MATRIX). We say that a fitness matrix Π is well-ordered with respect to conflict if, for all (i, j), $\pi_{iq} - \pi_{jq}$ has the same sign as $\pi_{qi} - \pi_{qj}$ if, and only if, $q > \overline{q}$, for some $\overline{q} \in \{0, ..., n\}$.

Coordination game

$$\begin{bmatrix} 9 & 4 & 2 \\ 6 & 8 & 1 \\ 5 & 3 & 7 \end{bmatrix}$$

Prisoners' Dilemma

$$\begin{bmatrix} 7 & 4 & 1 \\ 8 & 5 & 2 \\ 9 & 6 & 3 \end{bmatrix}$$

Hawk-Dove

$$\begin{bmatrix} 7 & 6 & 3 \\ 8 & 5 & 2 \\ 9 & 4 & 1 \end{bmatrix}$$

FIGURE II

Fitness Matrices That Are Well-Ordered with Respect to Conflict

In the definition, we adopt the convention that if $\pi_{iq} = \pi_{jq}$, then well-orderedness requires that $\pi_{qi} = \pi_{qj}$ regardless of q. To understand Definition 14, consider first a column with index $q > \overline{q}$. Intuitively, in a well-ordered matrix, whatever permutation happens to sort the entries of the qth column vector $\pi_{\cdot q}$ from highest to lowest also sorts the qth row vector $\pi_{q\cdot}$ in the same order. In other words, if either player plays action q (say, the row player), both players have the same fitness ranking over the other player's (the column player's) action choice—sort of a pure common-values assumption. Conversely, if $q \leqslant \overline{q}$ well-orderedness requires that, conditional on either player playing action q, both players have a total conflict of interest over the other player's action choice. Intuitively, then, a fitness matrix is well-ordered with respect to conflict if it induces a fitness game that has conflict of interest around the q most-Kantian actions, and common interest elsewhere.

Figure II shows three fitness matrices that are well-ordered with respect to conflict. The coordination game induced by the left-most matrix happens to have pure common values everywhere, corresponding to $\overline{q} = 0$ in Definition 14. The prisoner's dilemma game induced by the middle matrix in Figure II happens to have conflict of interest everywhere, corresponding to $\overline{q} = n$. The hawk-dove game induced by the right-most matrix happens to have conflict of interest only for the top row action, corresponding to $\overline{q} = 1$. In general, not every fitness matrix is well-ordered with respect to conflict.[49] However, Figure II suggests that well-ordered matrices are rich enough to cover many games of economic interest.

49. Even some 2×2 fitness matrices are not well-ordered with respect to conflict. For example, $\begin{bmatrix} 2 & 1 \\ -1 & 0 \end{bmatrix}$ is not well-ordered with respect to conflict because it induces common interest regarding the row player's action if and only if the column player chooses the Kantian action.

Next we generalize the two moral principles Care and Loyalty to the $n > 2$ environment.

DEFINITION 15 (CARE AND LOYALTY IN THE N × N SETTING). Take a vector $\mathbf{h} = [h_1, \ldots, h_n]$ of positive real functions of which $h_1, \ldots, h_\ell$ are nondecreasing and $h_{\ell+1}, \ldots, h_n$ nonincreasing. If $1 \leqslant \ell < n$, the operator $\mathbb{D}_n \to \mathbb{D}_n$ that, entry by entry, maps π_{ij} into $h_j(\pi_{ji})$ is denoted by LOYALTY($\mathbf{h}$; ·). If $\ell = n$, meaning that all the functions h are nondecreasing, the same operator is denoted by CARE($\mathbf{h}$; ·).

To understand Definition 15, consider the operator CARE. The $n \times n$ matrix [$\text{CARE}_{ij}(\mathbf{g}; \Pi)$] is a nonnegative matrix that, within each column, has higher values in those rows where i's opponent has a higher fitness level. Intuitively, the matrix [$\text{CARE}_{ij}(\mathbf{g}; \Pi)$] rewards player i for increasing player j's fitness. When $n = \ell = 2$, this definition of CARE specializes to the definition provided in Table I.

Analogously, the $n \times n$ matrix [$\text{LOYALTY}_{ij}(\mathbf{h}; \Pi)$] is a nonnegative matrix that in its first ℓ columns (corresponding to the opponent selecting one of the ℓ most Kantian actions), rewards player i for choosing rows (i.e., actions) that produce a higher fitness for i's opponent, and in its other columns penalizes player i for the same choices. Intuitively, LOYALTY can be interpreted as "conditional CARE." When $n = 2$ and $\ell = 1$, this definition of Loyalty specializes to the definition provided in Table I.

The reason we choose to restrict attention to Care and Loyalty among the H5's is that they constitute a minimal spanning pair in Theorem 3; also, conveniently, their definitions are relatively straightforward to generalize from the (2×2) to the $(n \times n)$ setting. By contrast, generalizing the definition of Fairness to the $(n \times n)$ setting is more complicated because it involves comparing player i's and j's payoffs.

We relax the notion of what it means for a moral principle to generate a preference equivalent to a stable preference P. In the previous sections, equivalence meant strategic equivalence (recall that spanning is based on strategic equivalence; see Definition 9). But when $n > 2$, this notion is too demanding: CARE and LOYALTY, as defined in Definition 15, are not adequate to span all stable preferences. However, CARE and LOYALTY can generate preferences equivalent to stable preferences in the following weaker sense.

DEFINITION 16 (PURE-STRATEGY EQUIVALENCE). Two matrices P and P' are pure-strategy equivalent if, for any given column player's pure action, they give rise to the same best-response set by the row player.

Pure-strategy equivalence should be taken to mean that two preference matrices are best-response similar. More precisely, if P and P' are not pure-strategy equivalent, they are very different strategically (which makes Theorem 6, part (i) a strong result). But if P and P' *are* pure-strategy equivalent, they are not "the same" strategically—even though pure-strategy equilibria against all opponents are the same. The problem is with mixed equilibria: even if P and P' are pure-strategy equivalent, given a column player's mixed action, the row player may be willing to play a specific mixed strategy under P but not under P'—which means that mixed-strategy equilibrium play need not be the same against all opponents. Therefore, if P is stable, P' need not be stable even if P and P' are pure-strategy equivalent. Thus, in general, pure-strategy equivalence need not preserve evolutionary stability; fortunately, for the purpose of Theorem 6, we show that it does (Theorem 6, part (ii)).

THEOREM 6 (GENERALIZATION OF THEOREMS 1–5). Let $n > 2$. Let $\Pi \in \mathbb{D}_n$ be a generic fitness matrix that is well-ordered with respect to conflict. Then,

(i) [generalizing Theorem 1] There exists an $n \times n$ fitness matrix $\Pi' \in \mathbb{D}_n$ that is well-ordered with respect to conflict, and vectors $\mathbf{h}, \mathbf{h}'$ such that LOYALTY($\mathbf{h}$; Π') and CARE($\mathbf{h}'$; Π') are stable and are not pure-strategy equivalent.
(ii) [generalizing Theorems 2 and 3] For any generic preference P that is evolutionarily stable for Π, a pure-strategy equivalent matrix P' exists that is evolutionarily stable and such that $P' = \Pi + \text{LOYALTY}(\mathbf{h}; \Pi)$ or $P' = \Pi + \text{CARE}(\mathbf{h}; \Pi)$ for suitable choice of $\mathbf{h}$.
(iii) [generalizing Theorem 4] The minimal moral code is a subset of {CARE, LOYALTY}.
(iv) [generalizing Theorem 5] Moral overdrive is not compatible with strict fitness improvement.

Proof. See Online Appendix IV. □

Theorem 6 generalizes Theorems 1–5, but with some notable restrictions. First, it restricts attention to fitness matrices that are well-ordered with respect to conflict. With this proviso, part (i) is the same as Theorem 1. Part (ii) should be interpreted roughly as "CARE and LOYALTY generate stable preferences P' that have the same pure-strategy equilibria as any stable P." Notably, part (ii) only covers two of the H5's: extending the analysis to all the H5 may be possible in the future. But in our favor, it is encouraging that CARE and LOYALTY together capture both an individualizing and a binding factor. Finally, part (iii) is admittedly unsatisfactory because it is essentially a restatement of part (ii). We believe that a sharper statement is true: that LOYALTY alone constitutes a minimal moral code—but we could not prove this statement. In effect, then, part (iii) is merely a placeholder for a tighter result that we hope to prove in the future. Part (iv) is a verbatim generalization of Theorem 5.

V. INTERPRETATION AND EMPIRICAL EVIDENCE

In this section, we discuss several interpretive points of our theory and provide some empirical evidence for it.

V.A. Moral Principles Can Evolve Through Genetic Transmission, Individual Development, or Interpersonal Transmission

The concept of evolutionary stability laid out in Definition 6 is abstract enough to encompass several possible channels of evolutionary transmission. First, and conceptually more straightforward, is reproduction: individuals who, in some game Π, adopt a moral principle that is not maximally fit for Π will have fewer offspring than those who, in that game, adopt the fittest moral principle—and will be weeded out over time. This genetic transmission mechanism can account for the heritability of moral principles documented by Zakharin and Bates (2023). Second, moral principles could be selected through individual development by reinforcement, from youth through adulthood, of those moral principles that 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 III.D. Finally, the moral principles of successful (in individual 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.

V.B. *Difference with Theories of Internalized Norms and Group Selection*

We define internalized norms as incentives that modify a player's individual fitness-maximizing behavior and exist only in the player's mind, not the physical world. According to Coleman (1994),[50] 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 agents who are affected by agent i's behavior. This pressure may come from formal laws and/or 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 behavior in ways that don't necessarily maximize individual fitness, and they 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 themselves in the players' minds. In our theory, moral principles can be evolutionarily stable even if they do not improve the opponents' (or society's) 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 set of internalizable norms is taken to be "all the equilibrium strategies that can be sustained in a repeated game setting" (as in Voss 2001), then this set is vast because of the multiplicity of self-enforcing strategies in a repeated game. By contrast, in our theory, the evolutionary selection mechanism limits the set of stable moral principles for any given Π.

Finally, as mentioned earlier, our theory is based on individual selection as opposed to group selection à la Bowles and Gintis (2011), Henrich (2004), Boyd et al. (2003), or Bowles and Gintis (2004). In this sense, ours is a model of the individualistic part of morality.

50. Our discussion of internalized norms follows Coleman (1994, 292ff).

	cooperate	*defect*
cooperate	*3,3*	*0,4*
defect	*4,0*	*1,1*

(A) Prisoner's Dilemma game

	trade	*don't trade*
trade	*3,3*	*3,0*
don't trade	*0,3*	*0,0*

(B) Market Game

FIGURE III

Prisoner's Dilemma Game and Market Game

V.C. Using the World Values Survey to Test Some Implications

We leverage the World Values Survey (WVS) to provide some empirical counterpart to the theory developed above. Although the WVS does not, to our knowledge, contain items that capture the H5 moral principles, we were able to locate suitable empirical counterparts for Sanctity and Care. Sanctity is captured by the postmaterialism index, which is larger when respondents rank abstract universalistic values that are good for society (democracy, freedom of speech/expression) above material considerations (low prices, law and order). Care is captured by a question that probes whether the respondent feels that "it is important to help people living nearby and to care for their needs."[51]

To develop testable implications from the theory, we focus on the degree to which an economy is more or less market-oriented. Formally, we contrast the two fitness games in Figure III. The game in Panel A captures a situation in which bilateral cooperation is necessary for value creation, such as two neighboring ranchers dealing with the problem of straying cattle. We expect this sort of interaction to be more typical of environments that are less market-oriented and more relationship-oriented. The game in Panel B represents the problem of an atomistic price-taking agent in a large market. The price-taking assumption implies that the agent can ignore every other agent's behavior; this is captured by the feature in Panel B that the opponent's actions are payoff-irrelevant. We expect this sort of interaction to be more frequent in market-oriented environments.

Our theory predicts that there is less Sanctity in the prisoner's dilemma than in the market game, meaning specifically that Sanctity is compatible with moral overdrive in the market

51. The postmaterialism index was developed by Inglehart and Abramson (1999). The question we use for Care is part of a module that was only administered in waves 5 and 6 of the WVS. See Online Appendix VI for further detail on variables' construction.

game but not in the prisoner's dilemma.[52] The theory also predicts that there is more Care in the prisoner's dilemma than in the market game, meaning specifically that Care is stable in the prisoner's dilemma but not stable in the market game.[53] In our interpretation of the theory, morality develops as a function of the type of fitness games the agents play; accordingly, we predict that in more market-oriented countries there should be more Sanctity and less Care.

These two predictions are confirmed in the data. Both panels of Figure IV plot, on their horizontal axes, an index capturing how market-oriented a country is,[54] and, on the vertical axes, our proxies for Sanctity (Panel A) and Care (Panel B).[55] The figure shows that there is more Sanctity and less Care in more market-oriented countries, as predicted by the theory. We view Figure IV as encouraging regarding the empirical content of our theory.

VI. Conclusion

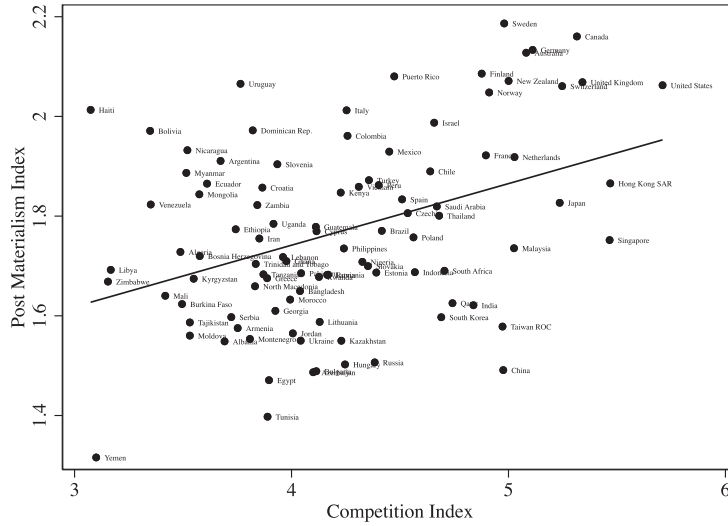
It is commonly argued that human morality has been shaped, at least in part, by evolutionary pressures. This article asks: what kind of moral principles emerge from and survive an evolutionary process? Can more than one moral principle be stable in a given game—that is, can moral pluralism be evolutionarily stable?

52. For the prisoner's dilemma, this statement was discussed following Theorem 5. For the market game, any moral principle—no matter how overwhelming relative to Π—that reinforces an agent's incentives to choose "trade" is stable. Sanctity is one such principle. Therefore, Sanctity is compatible with moral overdrive.

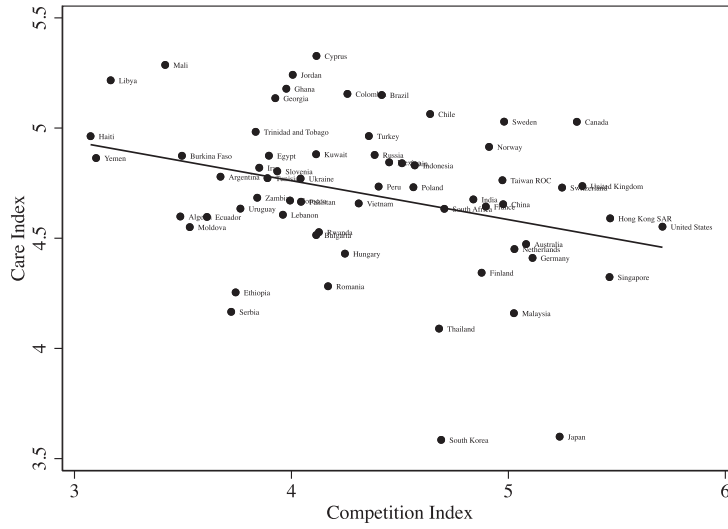
53. To be precise, Care is not applicable to the market game: in Panel B, consistent with the atomistic assumption, no player has an action that can improve the other player's fitness.

54. We compute the market orientation index as the average of four country-level scores given by the World Economic Forum. The scores are for goods market efficiency, labor market efficiency, financial market development, and market size. See Online Appendix VI for details about the index construction.

55. We compute the Sanctity index as the country average of the postmaterialism index of each WVS respondent. We compute the Care index as the country average of the responses ranging from 1 (lowest Care) to 6 (highest Care). See Online Appendix VI for details about the index construction. The same findings emerge from an individual-level regression which shows that living in a country that is more market-oriented than average is correlated with a response that is higher than average on Sanctity and less than average on Care (analysis available on request).



(A) More *Sanctity* in more market-oriented countries.



(B) Less *Care* in more market-oriented countries.

FIGURE IV

More Sanctity and Less Care in More Market-Oriented Countries

Elaboration based on the WVS and the World Economic Forum data. See Online Appendix VI.

We started from a theory in which, within each distinct fitness game Π that agents play, agents respond to psychological motives that give rise to potentially non–fitness maximizing best responses. These motives are captured by abstract functions $m(\Pi)$, which we have called moral principles. Among all mathematically possible such functions, we defined five to match the five moral foundations empirically identified by MFT: Care, Fairness, Authority, Loyalty, and Sanctity. We then asked whether our theory of evolutionary stability singles out these special moral principles.

We showed that any two moral principles in the H5 are distinguishable from each other in that they shape behavior differently; this is a sanity check of our theory. Moreover, we have shown that, together, the H5's are sufficient to generate all (generic) best-response behavior that can ever be evolutionarily stable in any (generic) fitness game. In this sense, our theory and MFT validate each other: MFT's empirical description of moral behavior connects to evolutionary stability, a purely mathematical concept, through the H5's.

However, we have also shown that these moral principles are somewhat redundant, in that two proper subsets of them are each rich enough to generate all evolutionarily stable best-response behavior. Intriguingly, these subsets happens to contain one individualizing and one binding foundation, consistent with the view—held by some empiricists—that most of the interpersonal variation captured by MFT can be reduced to just two principal components rather than five separate foundations: a so-called individualizing component (overlapping with Care and Fairness) and a so-called binding component (overlapping with Loyalty, Authority, and Sanctity). This redundancy result incidentally establishes that moral pluralism is evolutionarily stable in all games Π.

We have shown that a designer seeking to maximize social fitness by designing a moral code based on the fewest number of H5 principles will only need to include Fairness or Loyalty. Finally, we asked whether any moral principle can be blown out of proportion and still be evolutionarily stable. The answer is yes, but when this is the case, these moral principles do not improve social fitness relative to individual fitness-maximizing Nash equilibrium play.

Although the theory has been developed in 2×2 games with perfect observability, we have shown that many of these results generalize to some degree of imperfect observability and to $n \times$

n games. Still, some readers may feel that formalizing complex concepts such as loyalty, authority, and sanctity is a stretch in the context of a simultaneous move game with close to perfect information. We agree, but we see virtue in the simplicity of this approach as a building block for the analysis of more complex (and realistic) settings. Finally, we provide some empirical support for our theory based on the WVS.

Overall, this article provides theoretical support for three commonly made (but not rigorously justified) claims: that different moral principles apply in different settings (e.g., free-riding versus coordination game) and also in the same setting; that people in the same setting can vary in the emphasis they place on different moral principles; and that these principles survive evolutionary selection.

Supplementary Material

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online.

Data Availability

The data underlying this article are available in the Harvard Dataverse (Avataneo, Norman, and Persico 2025), https://doi.org/10.7910/DVN/ZPWIHU.

Northwestern University, United States
Magdalen College, Oxford, United Kingdom
Northwestern University, United States

References

Akdeniz, Aslihan, Christopher Graser, and Matthijs van Veelen, "Homo Moralis and Regular Altruists—Preference Evolution for When They Disagree," Discussion Paper, Tinbergen Institute, Amsterdam, 2020.

Alger, Ingela, "Evolutionarily Stable Preferences." *Philosophical Transactions of the Royal Society B*, 378 (2023), 20210505. https://doi.org/10.1098/rstb.2021.0505.

Alger, Ingela, and Jörgen W. Weibull, "Homo Moralis—Preference Evolution under Incomplete Information and Assortative Matching," *Econometrica*, 81 (2013), 2269–2302.

———, "Evolution and Kantian Morality," *Games and Economic Behavior*, 98 (2016), 56–67.

———, "Strategic Behavior of Moralists and Altruists," *Games*, 8 (2017), 38.

Alger, Ingela, Jörgen W. Weibull, and Laurent Lehmann, "Evolution of Preferences in Structured Populations: Genes, Guns, and Culture," *Journal of Economic Theory*, 185 (2020), 104951.

Anderson, Cameron, John Angus D. Hildreth, and Laura Howland, "Is the Desire for Status a Fundamental Human Motive? A Review of the Empirical Literature," *Psychological Bulletin*, 141 (2015), 574–601.

Avataneo, Michelle, Thomas Norman, and Nicola Persico, "Replication Data for: 'The Evolutionary Stability of Moral Foundations'," (2025), Harvard Dataverse. https://doi.org/10.7910/DVN/ZPWIHU.

Barnett, Michael D., Haluk C. M. Öz, and Arthur D. Marsden, "Economic and Social Political Ideology and Homophobia: The Mediating Role of Binding and Individualizing Moral Foundations," *Archives of Sexual Behavior*, 47 (2018), 1183–1194.

Bault, Nadege, Giorgio Coricelli, and Aldo Rustichini, "Interdependent Utilities: How Social Ranking Affects Choice Behavior," *PLoS One*, 3 (2008), e3477.

Bergstrom, Theodore C., "On the Evolution of Altruistic Ethical Rules for Siblings," *American Economic Review*, 85 (1995), 58–81.

Bester, Helmut, and Werner Güth, "Is Altruism Evolutionarily Stable?," *Journal of Economic Behavior & Organization*, 34 (1998), 193–209. https://doi.org/10.1016/S0167-2681(97)00060-7.

Bowles, Samuel, and Herbert Gintis, "The Evolution of Strong Reciprocity: Cooperation in Heterogeneous Populations," *Theoretical Population Biology*, 65 (2004), 17–28. https://doi.org/10.1016/j.tpb.2003.07.001.

———, *A Cooperative Species: Human Reciprocity and its Evolution*, (Princeton, NJ: Princeton University Press, 2011).

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter J. Richerson, "The Evolution of Altruistic Punishment," *Proceedings of the National Academy of Sciences*, 100 (2003), 3531–3535. https://doi.org/10.1073/pnas.0630443100.

Brown, William Michael, Boris Palameta, and Chris Moore, "Are There Nonverbal Cues to Commitment? An Exploratory Study Using the Zero-Acquaintance Video Presentation Paradigm," *Evolutionary Psychology*, 1 (2003), 147470490300100104. https://doi.org/10.1177/147470490300100104.

Coleman, James S., *Foundations of Social Theory*, (Cambridge, MA: Harvard University Press, 1994).

Dekel, Eddie, Jeffrey C. Ely, and Okan Yilankaya, "Evolution of Preferences," *Review of Economic Studies*, 74 (2007), 685–704.

Doğruyol, Burak, Sinan Alper, and Onurcan Yilmaz, "The Five-Factor Model of the Moral Foundations Theory Is Stable across WEIRD and Non-WEIRD Cultures," *Personality and Individual Differences*, 151 (2019), 109547. https://doi.org/10.1016/j.paid.2019.109547.

Ely, Jeffrey C., and Okan Yilankaya, "Nash Equilibrium and the Evolution of Preferences," *Journal of Economic Theory*, 97 (2001), 255–272. https://doi.org/10.1006/jeth.2000.2735.

Enke, Benjamin, "Kinship, Cooperation, and the Evolution of Moral Systems," *Quarterly Journal of Economics*, 134 (2019), 953–1019. https://doi.org/10.1093/qje/qjz001.

———, "Moral Values and Voting," *Journal of Political Economy*, 128 (2020), 3679–3729. https://doi.org/10.1086/708857.

———, "Moral Boundaries," *Annual Review of Economics*, 16 (2024), 133–157. https://doi.org/10.1146/annurev-economics-091223-093730.

Ewin, Robert E., "Loyalty and Virtues," *Philosophical Quarterly*, 42 (1992), 403–419. https://doi.org/10.2307/2220283.

Fehr, Ernst, and Urs Fischbacher, "The Nature of Human Altruism," *Nature*, 425 (2003), 785–791. https://doi.org/10.1038/nature02043.

Fehr, Ernst, and Klaus M. Schmidt, "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114 (1999), 817–868. https://doi.org/10.1162/003355399556151.

Frank, Robert H., Thomas Gilovich, and Dennis T. Regan, "The Evolution of One-Shot Cooperation: An Experiment," *Ethology and Sociobiology*, 14 (1993), 247–256. https://doi.org/10.1016/0162-3095(93)90020-I.

Govindan, Srihari, and Robert Wilson, "Axiomatic Equilibrium Selection for Generic Two-Player Games," *Econometrica*, 80 (2012), 1639–1699. https://doi.org/10.3982/ECTA9579.

Graham, Jesse, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto, " Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism," in *Advances in Experimental Social Psychology*, vol. 47, Patricia Devine and Ashby Plant, eds. (San Diego, CA: Academic Press, 2013), 55–130. https://doi.org/10.1016/B978-0-12-407236-7.00002-4.

Graham, Jesse, Jonathan Haidt, and Brian A. Nosek, "Liberals and Conservatives Rely on Different Sets of Moral Foundations," *Journal of Personality and Social Psychology*, 96 (2009), 1029–1046. https://doi.org/10.1037/a0015141.

Graham, Jesse, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto, "Mapping the Moral Domain," *Journal of Personality and Social Psychology*, 101 (2011), 366–385. https://doi.org/10.1037/a0021847.

Güth, Werner, "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives," *International Journal of Game Theory*, 24 (1995), 323–344. https://doi.org/10.1007/BF01243036.

Güth, Werner, and Bezalel Peleg, "When Will Payoff Maximization Survive? An Indirect Evolutionary Analysis: An Indirect Evolutionary Analysis," *Journal of Evolutionary Economics*, 11 (2001), 479–499.

Güth, Werner, and Menahem Yaari, "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game," in *Explaining Process and Change: Approaches to Evolutionary Economics*, Ulrich Witt, ed. (Ann Arbor: University of Michigan Press, 1992), 23–34.

Haidt, Jonathan, *The Righteous Mind: Why Good People Are Divided by Politics and Religion*, (New York: Vintage, 2012).

Haidt, Jonathan, and Jesse Graham, "When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals May Not Recognize," *Social Justice Research*, 20 (2007), 98–116. https://doi.org/10.1007/s11211-007-0034-z.

———, "Planet of the Durkheimians, Where Community, Authority, and Sacredness Are Foundations of Morality," in *Social and Psychological Bases of Ideology and System Justification*, John T. Jost, Aaron C. Kay, and Hulda Thurisdottir, eds. (New York: Oxford University Press, 2009), 371–401.

Hamlin, J. Kiley, "Moral Judgment and Action in Preverbal Infants and Toddlers: Evidence for an Innate Moral Core," *Current Directions in Psychological Science*, 22 (2013), 186–193. https://doi.org/10.1177/0963721412470687.

Heifetz, Aviad, Chris Shannon, and Yossi Spiegel, "The Dynamic Evolution of Preferences," *Economic Theory*, 32 (2007a), 251–286. https://doi.org/10.1007/s00199-006-0121-7.

———, "What to Maximize If You Must," *Journal of Economic Theory*, 133 (2007b), 31–57.

Helzer, Erik G., R. Michael Furr, Ashley Hawkins, Maxwell Barranti, Laura E. R. Blackie, and William Fleeson, "Agreement on the Perception of Moral Character," *Personality and Social Psychology Bulletin*, 40 (2014), 1698–1710.

Henrich, Joseph, "Cultural Group Selection, Coevolutionary Processes and Large-Scale Cooperation," *Journal of Economic Behavior & Organization*, 53 (2004), 3–35.

Inglehart, Ronald, and Paul R. Abramson, "Measuring Postmaterialism," *American Political Science Review*, 93 (1999), 665–677.

Kleinig, John, " Loyalty," in *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, ed. (Stanford, CA: Metaphysics Research Lab, Stanford University, 2022).

Malka, Ariel, Danny Osborne, Christopher J. Soto, Lara M. Greaves, Chris G. Sibley, and Yphtach Lelkes, "Binding Moral Foundations and the Narrowing

of Ideological Conflict to the Traditional Morality Domain," *Personality and Social Psychology Bulletin*, 42 (2016), 1243–1257.
Maynard Smith, John, "The Theory of Games and the Evolution of Animal Conflicts," *Journal of Theoretical Biology*, 47 (1974), 209–221.
———, *Evolution and the Theory of Games*(Cambridge: Cambridge University Press, 1982).
Maynard Smith, John, and George R. Price, "The Logic of Animal Conflict," *Nature*, 246 (1973), 15–18.
Métayer, Sébastien, and Farzaneh Pahlavan, "Validation of the Moral Foundations Questionnaire in French," *Revue internationale de psychologie sociale*, 27 (2014), 79–107.
Miettinen, Topi, Michael Kosfeld, Ernst Fehr, and Jörgen Weibull, "Revealed Preferences in a Sequential Prisoners' Dilemma: A Horse-Race between Six Utility Functions," *Journal of Economic Behavior & Organization*, 173 (2020), 1–25. https://doi.org/10.1016/j.jebo.2020.02.018.
Moreira Luana, Vianez, Mariane Lima de Souza, and Martins Guerra Valeschka, "Validity Evidence of a Brazilian Version of the Moral Foundations Questionnaire," *Psicologia: Teoria e Pesquisa*, 35 (2019), e35513. https://doi.org/10.1590/0102.3772e35513.
Napier, Jaime L., and Jamie B. Luguri, "Moral Mind-sets: Abstract Thinking Increases a Preference for 'Individualizing' over 'Binding' Moral Foundations," *Social Psychological and Personality Science*, 4 (2013), 754–759. https://doi.org/10.1177/1948550612473783.
Newton, Jonathan, "The Preferences of Homo Moralis Are Unstable under Evolving Assortativity," *International Journal of Game Theory*, 46 (2017), 583–589. https://doi.org/10.1007/s00182-016-0548-4.
Nilsson, Artur, and Arvid Erlandsson, "The Moral Foundations Taxonomy: Structural Validity and Relation to Political Ideology in Sweden," *Personality and Individual Differences*, 76 (2015), 28–32. https://doi.org/10.1016/j.paid.2014.11.049.
Ok, Efe A., and Fernando Vega-Redondo, "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario," *Journal of Economic Theory*, 97 (2001), 231–254. https://doi.org/10.1006/jeth.2000.2668.
Smith, Kevin B., John R. Alford, John R. Hibbing, Nicholas G. Martin, and Peter K. Hatemi, "Intuitive Ethics and Political Orientations: Testing Moral Foundations as a Theory of Political Ideology," *American Journal of Political Science*, 61 (2017), 424–437. https://doi.org/10.1111/ajps.12255.
Solnick, Sara J., and David Hemenway, "Is More Always Better?: A Survey on Positional Concerns," *Journal of Economic Behavior & Organization*, 37 (1998), 373–383. https://doi.org/10.1016/S0167-2681(98)00089-4.
Stolerman, Dominic, and David Lagnado, "The Moral Foundations of Human Rights Attitudes," *Political Psychology*, 41 (2020), 439–459. https://doi.org/10.1111/pops.12539.
Süssenbach, Philipp, Jonas Rees, and Mario Gollwitzer, "When the Going Gets Tough, Individualizers Get Going: On the Relationship between Moral Foundations and Prosociality," *Personality and Individual Differences*, 136 (2019), 122–131. https://doi.org/10.1016/j.paid.2018.01.019.
Varelius, Jukka, "Is Whistle-Blowing Compatible with Employee Loyalty?," *Journal of Business Ethics*, 85 (2009), 263–275. https://doi.org/10.1007/s10551-008-9769-1.
Verplaetse, Jan, Sven Vanneste, and Johan Braeckman, "You Can Judge a Book by Its Cover: The Sequel: A Kernel of Truth in Predictive Cheating Detection," *Evolution and Human Behavior*, 28 (2007), 260–271. https://doi.org/10.1016/j.evolhumbehav.2007.04.006.
Voss, Thomas, "Game Theoretical Perspectives on the Emergence of Social Norms," in *Social Norms*, Michael Hechter and Karl-Dieter Voss, eds. (New York: Russell Sage Foundation, 2001).

Yilmaz, Onurcan, and S. Adil Saribay, "Activating Analytic Thinking Enhances the Value Given to Individualizing Moral Foundations," *Cognition*, 165 (2017), 88–96. https://doi.org/10.1016/j.cognition.2017.05.009.
Zakharin, Michael, and Timothy C. Bates, "Testing Heritability of Moral Foundations: Common Pathway Models Support Strong Heritability for the Five Moral Foundations," *European Journal of Personality*, 37 (2023), 485–497. https://doi.org/10.1177/08902070221103957.