Magical thinking: A representation result

Brendan Daley
The Fuqua School of Business, Duke University

Philipp Sadowski
Department of Economics, Duke University

This paper suggests a novel way to import the approach of axiomatic theories of individual choice into strategic settings and demonstrates the benefits of this approach. We propose both a tractable behavioral model as well as axioms applied to the behavior of the collection of players, focusing first on prisoners' dilemma games. A representation theorem establishes these axioms as the precise behavioral content of the model, and that the model's parameters are (essentially) uniquely identified from behavior. The behavioral model features *magical thinking*: players behave as if their expectations about their opponents' behavior vary with their own choices. The model provides a unified view of documented behavior in a range of often studied games, such as the prisoners' dilemma, the battle of the sexes, hawk–dove, and the stag hunt, and also generates novel predictions across games.

Keywords. Magical thinking, axioms/representation theorem, prisoners' dilemma, coordination games.

JEL classification. C7, D8.

1. Introduction

This paper suggests a novel way to import the approach of axiomatic theories of individual choice into game-theoretic settings. We propose a behavioral model of play in symmetric 2×2 games, which features *magical thinking*: players behave as if they expect that choosing an action a increases the likelihood that their opponents also select action a. We then provide axioms and a representation result that establishes the equivalence between the axioms and the equilibrium play of the behavioral model, focusing

Brendan Daley: bd28@duke.edu
Philipp Sadowski: p.sadowski@duke.edu

Earlier working versions circulated under the titles "Foundations for cooperation in the prisoners' dilemma" and "A strategic model of magical thinking: Axioms and analysis." We are grateful to a co-editor and three anonymous referees for helpful comments and suggestions. We also thank David Ahn, Attila Ambrus, Pierpaolo Battigali, Alexandre Belloni, Roland Benabou, Haluk Ergin, Drew Fudenberg, Benjamin Golub, Rachel Kranton, Peter Landry, David McAdams, Wolfgang Pesendorfer, David Rand, Luca Rigotti, Jacob Sagi, Kota Saito, Larry Samuelson, Todd Sarver, Eldar Shafir, Ran Spiegler, Tomasz Strzalecki, and Curtis Taylor for useful discussions, and NSF Grant SES 1461502 for financial support.

DOI: 10.3982/TE2099

first on behavior in prisoners' dilemma (PD) games. Further, the model's parameters are (essentially) uniquely identified from behavior.

The novelty lies in the behavioral data to which our axioms apply. The axioms concern players' preferences over *actions* contingent on the payoffs of the (one-shot) game, rather than preferences over outcomes. In addition, they restrict not only individual behavior, but also place a joint restriction on the behavior of a finite collection of players. We motivate our axioms as simple and intuitive behavioral regularities across games and individuals, without reference to any particular strategic model.

The contribution of the paper is therefore threefold. First, we provide a tractable and empirically plausible theory of magical thinking, a phenomenon that has received attention in psychology and philosophy (discussed below), applied to strategic games. Most importantly here, we demonstrate that our model provides a unified view of observed behavior in a range of often studied games including the battle of the sexes, hawk–dove (also known as chicken), and the stag hunt, in addition to the PD.

Second, distinct from typical work in applied or behavioral game theory, we present a representation result that establishes equivalence between the model's predictions and a set of empirically plausible axioms. This result allows for the evaluation and empirical testing of the model, and facilitates its comparison to alternative theories. Further, the model's parameters can be identified from behavior, which is both useful for comparative statics and allows the analyst to traverse between the model and the axioms whenever convenient. For example, observed behavior satisfying the axioms on PD games can be used to identify the parameters of the model, which can be used in turn to generate predictions for (not yet observed) behavior in a different set of games. All of this is important for applied work.

Third, a key component of our approach is that the axioms apply to players' preferences over actions (rather than outcomes). Axiomatizing this type of data has the following benefits, numbered B1–B3. (B1) The primitive of our axiomatic analysis is exactly the type of data we aim to address, namely players' preferences over their own actions, across games and across players. (B2) This type of data is straightforward and common to collect in experiments. (B3) We do not have to rely on auxiliary assumptions about an equilibrium concept or on commonality of beliefs. Instead, as we discuss below, we can *derive* these, as well as individual value functions, from the data. We hope that our approach will prove useful in future research beyond this one application.

The domain of games

For several reasons, we begin our analysis on the set of PD games. PD games constitute perhaps the most important class of games in applications, and cooperation in the (one-shot) PD is a much discussed behavioral puzzle.[1] We demonstrate that our model

[1] Of course, cooperation is easier to explain in the repeated PD, provided players are patient enough. For finitely repeated versions, reputation models starting with Kreps et al. (1982) offer a potential explanation. For infinitely repeated versions, cooperation is part of some subgame perfect Nash equilibria. Interestingly, in an infinitely repeated, noisy version, Fudenberg et al. (2012) find substantial levels of cooperation (over 30% across all rounds) under parameters for which the *unique* equilibrium strategy is always defect.

makes behavioral predictions distinct from other explanations of cooperative behavior in PD games. Further, focusing on PD games helps build intuition for the workings of the model. Most importantly though, we demonstrate that behavior in PD games provides sufficient data to precisely characterize the behavioral model via axioms and to identify its parameters. Isolating such a small, yet economically interesting, domain for both the representation and identification results has the same advantages as it does in theories of individual choice.[2]

We then apply the behavioral model to all symmetric 2×2 games, where its predictions continue to align with experimental evidence.[3] Hence, the model's ability to explain behavior is not tailored to PD games at the expense of descriptive accuracy in other games in the class, but instead it provides a single account of observed play. Correspondingly, the model generates novel predictions for how behavioral patterns should correlate across games. Finally, we extend the model to allow for larger action sets, and investigate the manner in which the connection between magical thinking and cooperative behavior likewise extends.

Further extensions of the model are possible, but would require additional modeling choices. At a very general level, the two key components are that players believe their action choices have stochastic influence over the decisions of others and that equilibrium beliefs are biased as a result (evidence for each is discussed in Section 4). In principle, players could have arbitrary (magical) beliefs about how their choices affect others. However, given the formulation of magical thinking (and related concepts) in psychology and philosophy, we believe a natural starting point is for players to believe they influence others to select the same action as they do. Although it is possible that real-world context could imbue meaning into strategically irrelevant action labels, symmetric games provide a setting in which "the same action" is meaningful strategically. Next, in games with more players, there would be the added modeling choice of *which* other players i believes he is influencing and whether there is correlation in his perceived influencing.[4] These modeling choices will likely need to be tailored to the setting at hand, but the two key features would remain.

Summary of results

Within the model, each player i in the collection of players, I, is endowed a type, α_i, and there is a cumulative distribution function (CDF) over types, F, from which players perceive types to be independent and identically distributed (i.i.d.) draws. Given a game,

[2] First, it distills the behavioral implications of the psychological phenomenon (here magical thinking) by abstracting from as many complications as possible. Second, the less data needed for identification of the parameters, the better. Third, the model can be assessed by testing the individual axioms on the small domain of interest (for example, see our comparison to alternative explanations of cooperation in PD games in Section S.1 of the Supplement, available in a supplementary file on the journal website, http://econtheory.org/supp/2099/supplement.pdf). In contrast, an axiomatization of the same model on a larger domain might involve axioms that have less bite when restricted to the small domain, and can therefore not serve as a checklist for testing the model on that domain.

[3] Extension of the axiomatic analysis is found in Section S.3 of the Supplement.

[4] For example, in a two-party voting game, player i might believe that turning out independently increases the probability that others from his own party turn out, while not affecting the turnout of the opposing party.

player i forms the following nonstandard beliefs. He assigns probability α_i that the action of his anonymous opponent, j, will correlate perfectly with his own, and probability $(1 - \alpha_i)$ that j's action will be determined independently. In the latter case, i's belief about j's behavior is consistent with j's equilibrium strategy. We refer to α_i as i's degree of magical thinking. A player with $\alpha_i = 0$ corresponds to a standard game-theoretic agent—though, one who recognizes that he may be playing against a nonstandard opponent. We characterize the equilibria of the model, and establish a necessary and sufficient condition on F for the equilibrium to be unique in all PD games.

Turning to the axioms, as one would expect, some of them describe plausible regularities of individual behavior. Specifically, we posit *Monotonicity*, which requires that a player who is willing to defect in one PD does not prefer to cooperate in another PD with greater payoffs from defection, as well as appropriate notions of Continuity, Convexity, and Invariance to Positive Affine Payoff Transformations. In addition, we posit a novel *Interplayer Sensitivity Comparison* axiom. Roughly, the idea behind the axiom is that the behavior of a player who is more prone to defection is also more sensitive to changes in the gains from defecting on a cooperating opponent. We will see that this pattern is consistent with a player's willingness to cooperate being responsive to the true cost of doing so. In surveying the experimental literature, we find that our axioms are broadly consistent with the available evidence and also offer new testable implications for future studies.

Our representation theorem establishes that the axioms are equivalent to the behavioral model *with* the condition on F that is necessary and sufficient for uniqueness of the equilibrium in all PD games. Further, the collection of types $(\alpha_i)_{i\in I}$ and the quantiles $(F(\alpha_i))_{i\in I}$ are uniquely identified from behavior, which allows us to provide stronger comparative statics in terms of those parameters.[5] Finally, note that in the representation, F is the common belief among players regarding the distribution that types are drawn from. In the Supplement, Section S.2, we provide an axiomatic characterization of this belief being *empirically valid* when the collection of players is arbitrarily large.

In addition to generating a positive degree of cooperation in PD games that decreases monotonically with the incentives for defection, the model comports with observed behavior in other well known games. In hawk–dove games, our model predicts that players will choose *dove* more often than is predicted by the symmetric (mixed-strategy) Nash equilibrium of the standard model, in line with experimental evidence. In battle of the sexes games, the prediction of our model matches the symmetric (mixed-strategy) Nash equilibrium of the standard model, which also aligns with experimental findings. Consider next coordination games with multiple symmetric Nash equilibria that are Pareto ranked (e.g., the stag hunt game). Our model uniquely predicts coordination on the payoff-dominant Nash equilibrium only if it is also not "too risky," in a sense similar to the concept of *risk dominance* (Harsanyi and Selten 1988), and in line with evidence. However, the prediction is more nuanced than risk dominance in that

[5] That is, because of identification, our comparative statics (Section 3) describe not only the implication of changes in parameters for changes in behavior (as is common in applied game theory), but can establish equivalence between them (as is standard in decision theory).

whether the payoff-dominant Nash equilibrium is too risky depends on the (perceived) distribution of types, F.

Note that the model's ability to capture all of these findings does *not* owe to any flexibility across different games. Our results show that play in PD games alone (essentially) pins down the model, leaving no additional flexibility. Hence, the model also makes predictions across classes of games that are often studied independently. For example, collections with higher rates of cooperation in PD games also have a larger set of coordination games in which the payoff-dominant Nash equilibrium is uniquely selected in our model.

Of course, alternative explanations of nonstandard behavior in games—most notably models based on other-regarding preferences—have been studied and shown to align with important experimental findings. However, in both PD games as well as other prominent games in our domain, there remains significant evidence of nonstandard behavior that is not explained by these theories, but is consistent with our model of magical thinking, as we discuss in Sections 4, 5.1, and S.1 in the Supplement.

Magical thinking

Psychologists have collected evidence that is consistent with individuals exhibiting magical thinking. Starting first with inanimate "opponents," the term *illusion of control* was coined by Langer (1975) to describe subjects who acted as if their choices had influence over physical outcomes. For example, subjects placed higher bets on a coin about to be flipped than on a coin already flipped, but whose outcome was still unknown.[6]

Section 4 discusses evidence suggestive of magical thinking in strategic settings. Presenting one example here may be useful. Shafir and Tversky (1992) had subjects play a standard PD with the twist that in some treatments the game was played sequentially, such that one player knew the other's action before choosing his own. They observed that second-movers cooperate significantly less often in the sequential PD—even following cooperation by the first-mover—than in the standard, simultaneous-move version of the game. This finding is inconsistent with standard forms of other-regarding preferences (such as reciprocity), but can be explained by players believing that their actions directly influence their opponents' not-yet-chosen action, but cannot influence those that have already been taken.

Throughout, we refer to magical thinking as the belief that one's action choice influences one's opponent to choose the same action. A related notion is found in a normative debate in philosophy that concerns Newcomb's paradox (Nozick 1969) and extends to the PD if one presumes a notion of *self-similarity*.[7] *Evidentiary* decision theorists argue that one's opponent is probably similar to one's self, and hence one *should* believe that the other player will go through the same deliberations and come to a similar conclusion as one's self (Lewis 1979, Jeffrey 1983). They conclude from this that cooperation

[6] The interpretation that a decision-maker's beliefs about random states of nature vary with his own choice is also common in the theory of ambiguity aversion (see, for example, Gilboa and Schmeidler 1989).

[7] Such as described by Rubinstein and Salant (2016) (and citations therein) as the belief that others are likely to make similar judgements and choices as one's self.

is the optimal choice. Hence, while their psychological mechanism is slightly different, evidentiary decision theorists advocate for a player to behave as if his choice influences his opponent's choice, and the notion is observationally equivalent to magically thinking on our domain. In contrast, *causal* decision theorists argue that one *should not* believe that one's own action affects the other player's action, as the simultaneous-move game leaves no room for a causal explanation (Joyce 1999).

We mention this debate not because we will participate in it—the nature of the behavioral data we consider presupposes that magical thinking is a cognitive error—but to highlight that a number of intelligent, serious individuals have reasoned in such a manner.[8] Finally, a similar idea is apparent in common casual reasoning, such as, "I contribute/recycle/volunteer because if I did not, then how could I *believe* that others are doing it?"

The remainder of the paper is organized as follows. For PD games, Section 2 presents our model, axioms, and representation theorem. Section 3 presents comparative statics, and Section 4 compares our theory to experimental evidence and alternative theories of play. Section 5 first applies the model to all symmetric 2×2 games and then extends it to allow for larger action sets. Section 6 provides extended discussion including a comparison of our axiomatic methodology to alternative approaches. Proofs are given in the Appendix. The Supplement comprises Sections S.1– S.3, which contain extended formal results.

2. A theory of magical thinking

We begin with the class of prisoners' dilemma games as shown in Figure 1, where $r > p$ and $x, y > 0$, which we refer to as PD^0 (the reason for the superscript will become apparent shortly).[9] In each game, two players, i and j, can each choose to defect (d) or to cooperate (c). Often $r + x$ is denoted as t and $p - y$ is denoted as s, but the above parametrization will be more convenient for our purposes. Note that x captures the benefit from defecting on a cooperating opponent, while y is the benefit from defecting on a defector. We refer to an arbitrary game as $g \in \mathrm{PD}^0$ or, if it is useful to be more explicit about its payoffs, as (r, p, x, y). We consider a finite collection of players, indexed by $I := \{1, \dots, n\}$, and each player i's preferred action for each possible game in PD^0 when played as a one-shot game against an anonymous opponent, as is typical in experimental settings.

We present the behavioral model, or representation, first and then present the axioms in Section 2.2. Compared to axiomatic theories of individual choice, the most notable procedural difference is the necessity to conduct equilibrium analysis (Section 2.1.1) so as to apply our representation.[10]

[8] Experimental evidence suggestive of evidentiary reasoning is found in Quattrone and Tversky (1984).

[9] Throughout, we interpret game payoffs in monetary terms to facilitate comparison with experimental findings. However, there is no formal sense in which our theory relies on this interpretation rather than the interpretation of game payoffs as von Neumann–Morgenstern (vNM) utilities, as is customary in game theory. (See the discussion of methodology in Section 6 for more.)

[10] This can be viewed as a generalization of the single-agent exercise. There the prototypical result is the equivalence between axioms and a decision-maker acting as if he maximizes a certain utility function.

		Player j	
		c	d
Player i	c	r, r	$p-y, r+x$
	d	$r+x, p-y$	p, p

FIGURE 1. An arbitrary prisoners' dilemma in PD^0.

2.1 *The behavioral model*

For the set of atomless probability distributions each with support $[0, 1]$ and differentiable CDF, let $\mathcal{F}$ be the corresponding set of CDFs. In the behavioral model, each player $i \in I$ is privately endowed with a type $\alpha_i \in [0, 1]$. In addition, there is a common prior that types are drawn i.i.d. from a distribution with CDF $F \in \mathcal{F}$. For each $g \in \text{PD}^0$, player i evaluates the expected payoff of action $a_i \in \{c, d\}$ as

$$\begin{aligned} V_i(c) &= \alpha_i \cdot r + (1-\alpha_i)\big[P_i \cdot (p-y) + (1-P_i) \cdot r\big], \\ V_i(d) &= \alpha_i \cdot p + (1-\alpha_i)\big[P_i \cdot p + (1-P_i)(r+x)\big], \end{aligned} \tag{1}$$

where P_i is the probability i assigns to being defected on in game g, *conditional* on a_i and a_j being determined independently. That is, i evaluates options as if he thinks that there is probability α_i that his opponent will match whatever action choice i makes, and probability $1-\alpha_i$ that his opponent determines a_j uninfluenced by a_i. This is the sense in which player i exhibits magical thinking, and the degree to which he does so is measured by α_i.

Given a game $g \in \text{PD}^0$, a strategy for player i (denoted σ_i) is completely characterized by the probability with which he selects $a \in \{c, d\}$ if his type is α_i (denoted $\sigma_i(a|\alpha_i) \in [0, 1]$), and his interim expected payoff from strategy σ_i is $\sigma_i(c|\alpha_i)V_i(c) + \sigma_i(d|\alpha_i)V_i(d)$.[11] Throughout, we consider only symmetric equilibria, defined as follows.

DEFINITION 1. Fix any CDF F and $g \in \text{PD}^0$. An *equilibrium* is a pair (σ, P), such that, with V_i as given by (1), the following statements hold:

(i) For all $i \in I$, $\sigma_i = \sigma$.

However, each choice problem can be interpreted as a single-player game, with the notions of optimization and equilibrium coinciding. Therefore, the standard result is identical to showing that the axioms are equivalent to the decision-maker playing an *equilibrium* in every (single-player) game where payoffs are defined by the utility representation.

[11] There can be measurability issues for mixed strategies with uncountable type spaces (Aumann 1964). We use a convenient formulation that handles those issues. A strategy is a function $\sigma_i : \mathcal{A} \times [0, 1] \to [0, 1]$, where $\mathcal{A}$ is the collection of all subsets of $\{c, d\}$, that satisfies two properties: (i) for every $B \in \mathcal{A}$, the function $\sigma_i(B|\cdot) : [0, 1] \to [0, 1]$ is measurable, and (ii) for every $\alpha_i \in [0, 1]$, the function $\sigma_i(\cdot|\alpha_i) : \mathcal{A} \to [0, 1]$ is a probability measure. In a slight abuse of notation, then, we write $\sigma_i(\{a\}|\alpha_i)$ as $\sigma_i(a|\alpha_i)$, and if $\sigma_i(a|\alpha_i) = 1$, we say that player i chooses/selects/plays action a when his type is α_i. See Milgrom and Weber (1985) for further details and equivalence between this and other notions of mixing with uncountable type spaces. Finally, while the formula for interim expected payoff is standard (taking (1) as given), it implies that the bias in a player's beliefs depends only on his type and ultimate action choice, and *not* on $\sigma_i(\cdot|\alpha_i)$ directly.

(ii) For all $i \in I$ and $a, a' \in \{c, d\}$, $\sigma(a|\alpha_i) > 0 \implies V_i(a) \geq V_i(a')$.

(iii) For all $i \in I$, $P_i = P = \int_0^1 \sigma(d|\alpha)\, dF(\alpha)$.

The first two requirements are standard: the first is the symmetry condition; the second states that the strategy assigns positive probability only to actions that yield the highest expected payoff, given a player's type and beliefs. The third requires that any player's belief conditional on *not* influencing his opponent is consistent with his opponent's equilibrium strategy. If $\alpha_i = 0$, player i corresponds to a standard game-theoretic agent in that he assigns probability zero to directly influencing his opponent, and his belief about his opponent's behavior is consistent with his opponent's equilibrium strategy. If $\alpha_i > 0$, player i's belief is a convex combination of this belief and the belief that i's opponent will match the action played by i.[12]

2.1.1 *Equilibrium analysis* We now characterize the equilibrium properties of the behavioral model. First, we observe that the set of equilibria is invariant to positive affine transformations of the payoffs.

Lemma 1. *If (σ, P) is an equilibrium of the game $(r, p, x, y) \in \mathrm{PD}^0$, then it is also an equilibrium of the game $\kappa(r + \xi, p + \xi, x, y) \in \mathrm{PD}^0$ for all $\kappa > 0$ and $\xi \in \mathbb{R}$.*

All proofs are located in the Appendix. From the lemma, the set of equilibria is identical in games (r, p, x, y) and $(1, 0, \frac{x}{r-p}, \frac{y}{r-p})$, the latter being the positive affine transformation of the former with $\kappa = \frac{1}{r-p} > 0$ and $\xi = -p$. Let $\mathrm{PD} \subset \mathrm{PD}^0$ denote the subset of games in which r and p are normalized to 1 and 0, respectively, with $(x, y) \in \mathrm{PD}$ being an arbitrary element. Given Lemma 1, it is sufficient to characterize equilibrium behavior for games in PD, which we focus on for the remainder of Section 2.1.

Definition 2. An equilibrium (σ, P) is a *cutoff equilibrium* if σ is of the form $\sigma(d|\alpha) = 1$ if $\alpha < \alpha^*$ and $\sigma(d|\alpha) = 0$ if $\alpha > \alpha^*$, for some $\alpha^* \in [0, 1]$.

Proposition 1. *For any $F \in \mathcal{F}$ and $(x, y) \in \mathrm{PD}$, (i) any equilibrium is a cutoff equilibrium with $\alpha^* \in (0, 1)$, (ii) α^* is an equilibrium cutoff if and only if it is a solution to (2) below, and (iii) an equilibrium exists.*

Fixing any $(x, y) \in \mathrm{PD}$, the cutoff nature of the equilibrium is immediate: for any (common) equilibrium belief $P_i = P$, $V_i(c) - V_i(d)$ is strictly increasing in α_i. Then, in equilibrium, $P_i = P = F(\alpha^*)$, and the cutoff type, α^*, is indifferent between c and d. So

[12] Because players in the model seek to maximize their expected payoff (albeit, with nonstandard beliefs), one could obviously employ an alternative, reduced-form assumption that a player simply receives a direct utility gain from selecting c. In Section 4 we discuss how this modeling choice would require a counterintuitive form of dependence on the payoff parameters to emulate our model (which is only exacerbated when we extend to games beyond PD^0) and be at odds with additional experimental evidence.

the set of equilibria is identical to the set of solutions to the equation[13]

$$\begin{aligned} V_i(c|\alpha_i=\alpha^*) &= \alpha^*\cdot 1+(1-\alpha^*)\big[F(\alpha^*)\cdot(-y)+(1-F(\alpha^*))\cdot 1\big] \\ &= \alpha^*\cdot 0+(1-\alpha^*)\big[F(\alpha^*)\cdot 0+(1-F(\alpha^*))\cdot(1+x)\big]=V_i(d|\alpha_i=\alpha^*). \end{aligned} \tag{2}$$

Noting that for $\alpha_i = 0$, $V_i(c|\alpha_i = \alpha^*) < V_i(d|\alpha_i = \alpha^*)$ and for $\alpha_i = 1$, $V_i(c|\alpha_i = \alpha^*) > V_i(d|\alpha_i = \alpha^*)$, all solutions to (2) are interior and existence is guaranteed by the continuity of both the left- and right-hand sides. This leaves only the question of uniqueness.

Proposition 2. *For any fixed $F \in \mathcal{F}$, there is a unique equilibrium cutoff in each $(x, y) \in$ PD if and only if $\frac{F'(\alpha)}{F(\alpha)} \leq \frac{1}{\alpha-\alpha^2}$ for all $\alpha \in (0, 1)$ (hereafter referred to as Condition S).*

Condition S restricts how steep F can be, by limiting its reverse hazard rate, in a manner that depends on α. For example, the CDF $F(\alpha) = \alpha^{1/k}$, $k \geq 1$, satisfies the condition, even though $\frac{F'(\alpha)}{F(\alpha)} \to \infty$ as $\alpha \to 0$. Note, then, that by taking k arbitrarily large, we can generate arbitrarily close approximations of the standard model (in which $F(\alpha) = 1$ for all $\alpha \in [0, 1]$), while continuing to satisfy Condition S.

To gain intuition for the potential multiplicity of equilibria, first note that for type α_i, defection carries the cost of $r - p = 1$ with perceived probability α_i, while the benefit of defection is $F(\alpha^*)y + (1 - F(\alpha^*))x$ with perceived probability $1 - \alpha_i$. If $x > y$, then the benefit of defection is decreasing in $F(\alpha^*)$ (i.e., the probability that one's opponent defects if his choice is made independently), and the indifference equation (2) has a unique solution.[14] But if $x < y$, then the benefit of defection is increasing in $F(\alpha^*)$. If F is steep on some range this means that (in expectation) there are many players making essentially the same calculation; so each is happy to cooperate if the equilibrium calls for all of them to do so, but each prefers to defect if the equilibrium calls for them all to do so. These types face a coordination problem. This problem is ameliorated if F is never too steep. Not surprisingly, the most difficult games in which to maintain uniqueness are those with the smallest x values, which are used to derive the tightness of Condition S for uniqueness (see the proof in the Appendix).

2.2 *The axioms*

We now present the axioms, doing so without reliance on the model. The data we consider are each player i's preferred action for each possible game in PD^0 when played as a one-shot game against an anonymous opponent. The behavior of player i partitions PD^0 into three sets: the set of games D_i^0 for which i strictly prefers d, the set of games C_i^0 for which i strictly prefers c, and the set of games $M_i^0 = \text{PD}^0 \setminus (D_i^0 \cup C_i^0)$ for which i is indifferent in his choice of d or c. We denote by $\overline{D}_i^0 = \text{PD}^0 \setminus C_i^0$ and $\overline{C}_i^0 = \text{PD}^0 \setminus D_i^0$ the sets of games for which i weakly prefers d or c, respectively. The primitive of our analysis

[13] Definition 2 does not specify the behavior of the cutoff type, who is indifferent between c and d. We do not always distinguish equilibria that have the same cutoff, but in which the cutoff type behaves differently since this type has measure zero and the distinction has no effect on payoffs.

[14] For $x = y$, (2) has a unique solution, which is independent of F: $\alpha^* = \frac{x}{1+x} = \frac{y}{1+y}$.

is the collection of pairs $(D_i^0, C_i^0)_{i\in I}$, which fully summarizes the behavior of all players in I.[15]

Our first four axioms consider individual behavior. It can be noted that a player who adheres to the standard prediction of always defecting, $D_i^0 = \text{PD}^0$, satisfies all of these axioms (and can never generate a violation of our fifth and final axiom).

Axiom 1 (Invariance to Positive Affine Transformations). *For all $i \in I$, if $(r, p, x, y) \in D_i^0$, then $\kappa(r + \xi, p + \xi, x, y) \in D_i^0$ for all $\kappa > 0$ and $\xi \in \mathbb{R}$, and analogously for C_i^0.*

The axiom states that positive affine transformations of all game payoffs have no effect on individual behavior. For the dollar stakes used in the laboratory, evidence seems to be consistent with the axiom, both in the prisoners' dilemma and also in many other games (see Section 4). The axiom has a flavor of risk neutrality (which we have already seen is part of the behavioral model). One interpretation is that subjects themselves treat strategic risk differently from environmental risk, focusing on the strategic aspects of their choice rather than their attitude toward risk.[16]

Axiom 1 implies that any player i behaves identically in games (r, p, x, y) and $(1, 0, \frac{x}{r-p}, \frac{y}{r-p})$. Hence, under Axiom 1, it is sufficient to characterize behavior on the subset $\text{PD} \subset \text{PD}^0$. We pose the remainder of our axioms on PD, meaning that, on their own, they are weaker than their obvious counterparts applying to PD^0. To do so, let $D_i = D_i^0 \cap \text{PD}$, and analogously for C_i, M_i, $\overline{D}_i$, and $\overline{C}_i$.

The remaining two payoff parameters, x and y, correspond to the two motives for defection: the exploitative motive of gaining at the expense of a cooperating opponent and reaping an extra payoff of x, and the defensive motive to avoid being the "sucker" and losing y. Our remaining axioms describe the effects of changing x and y on behavior.

Axiom 2 (Continuity). *For all $i \in I$, D_i and C_i are open.*

The axiom says that no individual has a jump from a strict preference for defection to a strict preference for cooperation as the motives for defection vary continuously.

Axiom 3 (Monotonicity). *For all $i \in I$, if $(x, y) \in \overline{D}_i$, $(x', y') \geq (x, y)$, and $(x', y') \neq (x, y)$, then $(x', y') \in D_i$.*

The axiom requires that strengthening the motives for defection (at least one of them strictly) will lead a player who initially weakly prefers to defect to strictly prefer defection.

[15] Our primitive differentiates the games where i strictly prefers d or c from those in which he is indifferent. This is analogous to the standard assumption in axiomatic decision theory that the primitive is a preference relation (not simply choice), which also distinguishes strict from weak preferences. Formally, for every $g \in \text{PD}^0$, i ranks the actions in $\{d, c\}$. Each ranking is a complete binary relation $\succsim_i^g$. Our primitive is $(D_i^0, C_i^0)_{i\in I}$, where D_i^0 and C_i^0 are the subsets of PD^0 for which $d \succ_i^g c$ and $c \succ_i^g d$, respectively.

[16] Of course, the axiom is also consistent with the alternative interpretation of game payoffs as vNM utilities as is customary in game theory. See the discussion of methodology in Section 6 for more.

Axiom 4 (Convexity). *For all $i \in I$, D_i and C_i are convex.*

The intuition behind the axiom is that a larger change in the motives for defection should have a weakly larger effect on behavior than does a proportionally smaller change. Suppose that player i strictly prefers to, say, defect in both (x, y) and (x', y'). The change from (x, y) to (x', y') can be interpreted as trading off the two motives at a *rate*, $\frac{y'-y}{x'-x}$, and a *scale*, normalized to 1. Comparatively, the change from (x, y) to $(\gamma x + (1-\gamma)x', \gamma y + (1-\gamma)y')$, where $\gamma \in (0, 1)$, is unambiguously smaller: it trades off the two motives at the same rate, but on a smaller scale. Axiom 4 states that if the larger change in payoffs does not alter i's strict preference for d (or for c), then neither should this smaller change in the payoffs.[17]

While we allow different players to behave differently in a given game, we now pose a new type of axiom that compares the behavior of any two players across games. Informally, the interplayer axiom says the following: Suppose that player i defects under lower incentives for defection than does j. Then, when i is at the cusp of flipping between d or c, his choice is more sensitive to changes in x (the exploitative motive) than is j's choice when j is likewise at the cusp.

It seems natural that the interpretation of an interplayer axiom would be contingent on at least some basic properties of individual behavior; in our case this will be Monotonicity (Axiom 3). Intuitively, if a player cooperates in a given prisoners' dilemma game, he does so at a cost to his own game payoff. This cost depends on his opponent's behavior: specifically, the more likely the opponent is to cooperate, the greater is the influence of x on this cost. Hence, if all players satisfy Axiom 3, then in games where player i (who defects under lower incentives for defection) is on the cusp of flipping his behavior it must be that the arbitrary opponent is more likely to be cooperating than in games where player j (who defects only under higher incentives for defection) is similarly on the cusp. If behavior is responsive to the true cost of cooperation, then player i's behavior should be more sensitive to changes in x than is player j's. We now present the formalisms.

Definition 3. For $H, H' \subset \text{PD}$ we write $H < H'$ if, for all $(x, y) \in H$ and $(x', y') \in H'$, $x < x'$ and $y < y'$.

Axiom 5 (Interplayer Sensitivity Comparison). *For all $\{i, j : i \neq j\} \subset I$ and $\varepsilon, \delta \in \mathbb{R}_{++}$, if (i) $\{(x, y), (x+\varepsilon, y-\delta)\} < \{(x', y'), (x'+\varepsilon, y'-\delta)\}$, (ii) $(x, y) \in \overline{D}_i$, (iii) $(x+\varepsilon, y-\delta) \in \overline{C}_i$, and (iv) $(x', y') \in \overline{C}_j$, then (v) $(x'+\varepsilon, y'-\delta) \in C_j$.*

The axiom is illustrated in Figure 2. To see that it captures the pattern described above, note first that, in the context of Axiom 3, (i), (ii), and (iv) imply that player i indeed defects under lower incentives for defection than does player j in the four games.[18]

[17] It may be useful to note that while reminiscent of the classic two-good consumer-preference diagram, in our context the choice objects are c and d, not (x, y) bundles; so M_i is *not* an indifference curve, and D_i and C_i are not better than/worse than sets, meaning Axiom 4 is not related to the standard convexity-of-consumer-preferences assumptions (for example, Mas-Colell et al. 1995, Chapter 3.B).

[18] Let H denote the set of four games. To see that i is more prone to defection than j in H, note that Axiom 3 implies that $\{(x', y'), (x'+\varepsilon, y'-\delta)\} \subset D_i$ and that $\{(x, y), (x+\varepsilon, y-\delta)\} \subset C_j$. Therefore, $\overline{D}_j \cap H \subsetneq D_i \cap H$ and $\overline{C}_i \cap H \subsetneq C_j \cap H$.

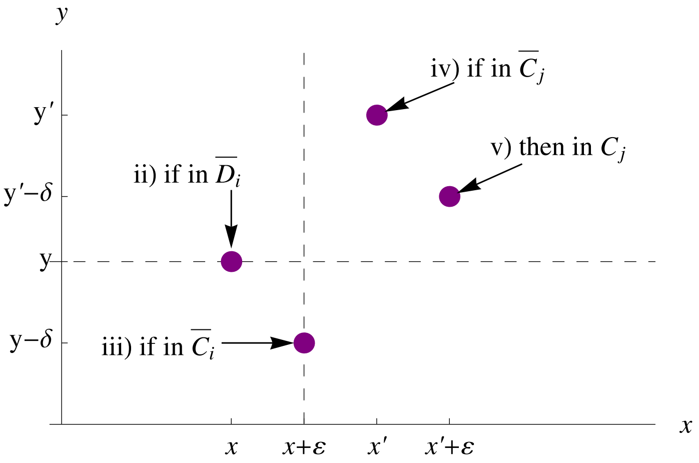

y
x
y′
y′−δ
y
y−δ
x
x+ε
x′
x′+ε
iv) if in $\overline{C}_j$
v) then in C_j
ii) if in $\overline{D}_i$
iii) if in $\overline{C}_i$

FIGURE 2. Depiction of Axiom 5. Notice that (i) holds, so (ii)–(iv) imply (v).

Second, (ii) and (iii) imply that i is willing to flip between choosing d or c when moving from (x, y) to $(x + \varepsilon, y - \delta)$. Third, (iv) says that j is willing to cooperate in (x', y'). Now, clearly, the movements from (x, y) to $(x + \varepsilon, y - \delta)$ and from (x', y') to $(x' + \varepsilon, y' - \delta)$ entail the same increase, ε, in the exploitative motive and the same reduction, δ, in the defensive motive. Hence, if, contrary to (v), j were willing to defect in $(x' + \varepsilon, y' - \delta)$, then j would have to be *more* sensitive to changes in x (relative to changes in y) than is i, which violates the pattern described at the outset. Hence, Axiom 5 requires that (i)–(iv) imply (v).

We note that insofar as one views both defection in more games and a greater responsiveness to the exploitative motive to be features of a more "aggressive disposition" on the part of player i, the axiom is consistent with the view, and the motivation based on objective incentives and Axiom 3 provides a microfoundation for this correlation.

Finally, as this type of interplayer axiom is novel to our approach, it may be worth previewing the role it plays in the representation result. The intuition provided for the axiom above refers to behavior being responsive to the true cost of cooperation. In the representation, player i's behavior is a response to the cost of cooperation as measured by his perception of the distribution F, call it F^i. The axiom, then, disciplines the heterogeneity in this perception. As we will see, it ensures that $F^i(\alpha_i) \leq F^j(\alpha_j)$ if and only if $\alpha_i \leq \alpha_j$, which must hold if all players perceive the same F.[19]

2.3 *The representation theorem*

Having studied the behavioral model and the axioms, we present the representation result.

[19] Conversely, if the behavioral model were expanded to accommodate heterogenous perceptions of F, and $F^i(\alpha_i) > F^j(\alpha_j)$ despite $\alpha_i \leq \alpha_j$, the implied behavior would violate Axiom 5.

Definition 4. For $I' \subset I$, the behavior of the players in I', $(D_i^0, C_i^0)_{i \in I'}$, *can be explained by* the behavioral model $[F, (\alpha_i)_{i \in I}]$ if for all $g \in \text{PD}^0$ there exists an equilibrium such that, with V_i as defined by (1), the following statements hold:

(i) For all $i \in I'$, $g \in C_i^0$ if and only if $\{c\} = \text{argmax}_{\{c,d\}}\{V_i(c), V_i(d)\}$.

(ii) For all $i \in I'$, $g \in D_i^0$ if and only if $\{d\} = \text{argmax}_{\{c,d\}}\{V_i(c), V_i(d)\}$.

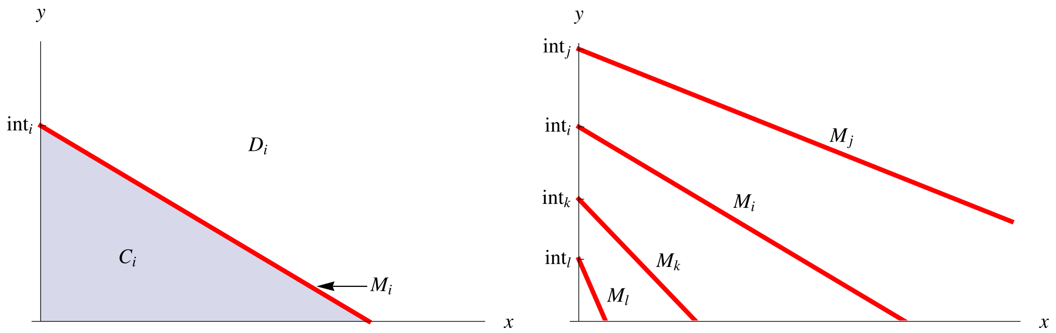
Theorem 1. *The primitive $(D_i^0, C_i^0)_{i \in I}$ satisfies Axioms 1–5 if and only if it can be explained by a behavioral model $[F, (\alpha_i)_{i \in I}]$, where $F \in \mathcal{F}$ satisfies Condition S. Furthermore, for all $i \in I$, α_i and $F(\alpha_i)$ are unique.*

Before sketching the proof, it is worth noting a few interesting features. First, a central concern in representation results is the degree to which the parameters in the representation, here F and $(\alpha_i)_{i \in I}$, are unique. Theorem 1 establishes that each player's α_i (the degree to which he exhibits magical thinking) is uniquely determined by the primitive and will, in fact, only depend on (D_i^0, C_i^0) as we sketch below. Further, the quantiles of F at all α_i in the collection are also unique.

Second is the interpretation of the magical-thinking component. Given the nature of our primitive, we have taken the position that this is an error, and the choices of each player are not directly influenced by the choices of any other player. In other words, *our* assumptions about the nature of human agency are the standard ones, but we allow that the players act as if they have nonstandard ones. There is also an important subtlety in understanding the F in the representation: (it is as if) F is the CDF of the distribution that all players perceive the α-types to be drawn from. This suggests an interpretation in which the players conceive of a grand population of which I is a random sample. In Section S.2, we provide an axiomatic characterization of this belief being empirically valid when the collection is large.

Third, a common concern in game-theoretic analysis is the issue of equilibrium multiplicity.[20] A reader might therefore object to the terminology that a model *can explain* behavioral data if the data are always consistent with one of the model's equilibria (Definition 4) as too permissive. The definition was chosen so that equilibrium uniqueness is not forced into the very notion of representation. Nevertheless, this objection is easily addressed. Notice that Theorem 1 includes the provision that F satisfies Condition S. Under this provision, Proposition 2 (with Lemma 1) guarantees that the equilibrium cutoff is unique for all $g \in \text{PD}^0$ (and all equilibria are cutoff equilibria (Proposition 1)). It is immediate, therefore, that the representation satisfies the more stringent definition of *can explain* attained if the requirements of Definition 4 must instead hold in *all* equilibria.

[20] In single-player games/decision problems, the agent may be indifferent between multiple payoff-maximizers, which can be interpreted as equilibrium multiplicity. However, in this scenario, the payoff to all agents is equivalent across all equilibria (by hypothesis). In general, the same statement does not hold for multiplayer games with multiple equilibria. This is one reason why the multiplicity issue is of perhaps greater concern in game theory than in decision theory.



(a) Individual Behavior in PD if $D_i, C_i \neq \emptyset$

(b) Behavior of four players in PD

FIGURE 3. (a) A player i's behavior in PD, for whom $D_i, C_i \neq \varnothing$. (b) The M-lines for four distinct players; note how they fan out.

A couple of notational definitions will simplify exposition for the remainder of the paper. Let $\mathcal{F}_S$ denote the set of CDFs in $\mathcal{F}$ that satisfy Condition S. Let α_{-i} denote an arbitrary assignment of types to players in $I \setminus \{i\}$ (i.e., $(\alpha_j)_{j \in I \setminus \{i\}}$).

Sketch of proof of Theorem 1 It is clear that Lemma 1 is the precise behavioral content of Axiom 1. Hence, we need only prove that Axioms 2–5 are equivalent to the behavioral model on PD.

As is typical, showing that the representation implies the axioms is the easier direction. First, extreme players, $\alpha_i = 0, 1$, either always defect or always cooperate, so trivially satisfy our axioms. Next, recall that in the behavioral model, the unique equilibrium of any game $(x, y) \in \text{PD}$ is of cutoff form, where the cutoff, α^*, is characterized by (2). To find the set of games in PD for which i is indifferent between c and d, fix $\alpha_i \in (0, 1)$ and solve (2) for y as a function of x to get

$$M_i = \left\{(x, y) \in \text{PD}\middle| y = \frac{\alpha_i}{(1-\alpha_i)F(\alpha_i)} - x\left(\frac{1 - F(\alpha_i)}{F(\alpha_i)}\right)\right\}.$$

Note that M_i is a downward sloping line in PD. The games D_i and C_i are the strict-upper- and strict-lower-contour sets of M_i, respectively (Figure 3(a)). Axioms 2–4 follow immediately.

In addition, observe that $\frac{\alpha_i}{(1-\alpha_i)F(\alpha_i)}$ is weakly increasing and $\frac{1-F(\alpha_i)}{F(\alpha_i)}$ is strictly decreasing in α_i; the former by Condition S, the latter by $F \in \mathcal{F}$. This implies that if $0 < \alpha_i < \alpha_j < 1$, then M_i and M_j do not intersect in PD and, further, that they "fan out" as x increases (Figure 3(b)). It is straightforward to verify that this property ensures Axiom 5.

The proof that the axioms imply the representation has two main parts. In the first part, we show that for any individual player i, if (D_i, C_i) satisfies Axioms 2–4, then there exists a pair $(\alpha_i, F_i) \in [0, 1]^2$ such that (D_i, C_i) can be explained by any model $[F, (\alpha_i, \alpha_{-i})]$ satisfying $F \in \mathcal{F}$ and $F(\alpha_i) = F_i$. Further, α_i and F_i are unique. In other words, the axioms on individual behavior are enough to establish that each individual is playing in accordance with our behavioral model—though not necessarily with agreement among individuals about F. The second part of the proof establishes that there is a common $F \in \mathcal{F}_S$ that can simultaneously explain all of $(D_i, C_i)_{i\in I}$. This relies on Axiom 5.

To begin the first part suppose that (D_i, C_i) satisfies Axioms 2–4: Continuity, Monotonicity, and Convexity. By Continuity, it is straightforward to show that either (i) $D_i = \mathrm{PD}$, (ii) $C_i = \mathrm{PD}$, or (iii) $M_i \neq \varnothing$. If (i), then $(\alpha_i, F_i) = (0, 0)$, and if (ii), then $(\alpha_i, F_i) = (1, 1)$. Suppose now that (iii) holds. Continuity and Monotonicity imply that there is a continuous, strictly decreasing function $\overline{y}$ such that $M_i = \{(x, y) \in \mathrm{PD} | y = \overline{y}(x)\}$, $C_i = \{(x, y) \in \mathrm{PD} | y < \overline{y}(x)\}$, and $D_i = \{(x, y) \in \mathrm{PD} | y > \overline{y}(x)\}$. Finally, Convexity of D_i and C_i means $\overline{y}$ is linear, so can be summarized by two scalars that we denote int_i and slp_i: $M_i = \{(x, y) \in \mathrm{PD} | y = \mathrm{int}_i - \mathrm{slp}_i \cdot x\}$.

Having established the linearity of M_i from behavioral data, recall from the argument above that in the behavioral model,

$$M_i = \left\{(x, y) \in \mathrm{PD} \middle| y = \frac{\alpha_i}{(1-\alpha_i)F(\alpha_i)} - x\left(\frac{1 - F(\alpha_i)}{F(\alpha_i)}\right)\right\}.$$

Inverting the bijection $(\mathrm{int}_i, \mathrm{slp}_i) = (\frac{\alpha_i}{(1-\alpha_i)F_i}, \frac{1-F_i}{F_i})$ establishes the first part of the proof.

For the second part, consider two players i and j, such that $M_i, M_j \neq \varnothing$ and who satisfy Axiom 5. This means $\mathrm{int}_i < \mathrm{int}_j$ implies $\mathrm{slp}_i > \mathrm{slp}_j$.[21] The translation of this condition under the bijection yields that $0 < \alpha_i < \alpha_j < 1$ implies $F_i < F_j \leq F_i \frac{\alpha_j(1-\alpha_i)}{\alpha_i(1-\alpha_j)}$. The first inequality means that there exists a strictly increasing CDF F that, together with $(\alpha_i)_{i\in I}$, can simultaneously explain the behavior of all players (the inclusion of the $\alpha_i = 0, 1$ players is trivial). The second inequality is a discretized version of Condition S. It is then straightforward, but cumbersome, to show that it is without loss of generality to take F to be differentiable and to satisfy Condition S.

Finally, a comment on the properties of F in the representation. As made clear from the sketch above, the axioms do not require F to have full support or to be differentiable, but merely allow for these properties. This is because the data of a finite number of players generate values for F at only a finite number of points (Section S.2 provides an analysis with a continuum of players). These features are chosen to be part of the representation because they are commonly assumed, appealing properties for applied models that facilitate a tractable analysis (recall Section 2.1). For example, they allow for a simple statement of Conditions S. It is not difficult to show that a larger class of behavioral models satisfies the axioms, and that any primitive that satisfies the axioms can be explained by another model $[F, (\alpha_i)_{i\in I}]$, where F lacks full support and/or is not everywhere differentiable. It is worth noting, however, that the unique identification of parameters in Theorem 1 continues to hold across this larger class of models since, as outlined above, these parameters are pinned down by individual behavior that satisfies Axioms 1–4.

[21] To see this, note that $\mathrm{int}_i < \mathrm{int}_j$ implies that there are games (x, y), $(x + \varepsilon, y - \delta)$, (x', y'), and $(x' + \varepsilon, y' - \delta)$ that satisfy (i) $\{(x, y), (x + \varepsilon, y - \delta)\} < \{(x', y'), (x' + \varepsilon, y' - \delta)\}$, (ii) $(x, y) \in M_i$, (iii) $(x + \varepsilon, y - \delta) \in M_i$, and (iv) $(x', y') \in M_j$. Axiom 5 then implies that $(x' + \varepsilon, y' - \delta) \in C_j$ and, consequently, $\mathrm{slp}_i > \mathrm{slp}_j$.

3. Comparative statics

In this section we illustrate how the predictions of the model vary with the parameters. In light of Axiom 1/Lemma 1, we do so on the smaller set of games, PD, without loss.

Definition 5. Let $|A|$ be the size of any finite set of players A. Consider two arbitrary sets of players A and $\widetilde{A}$ such that $|A| = |\widetilde{A}|$.

- We say that, in $H \subset \mathrm{PD}$, the set of players A *defects (weakly) more* than $\widetilde{A}$ if $|\{i \in A | (x, y) \in D_i\}| \geq |\{j \in \widetilde{A} | (x, y) \in D_j\}|$ for each $(x, y) \in H$.
- We say that the set of players A *is (weakly) more influenced by x relative to y* than is $\widetilde{A}$ if A defects more than $\widetilde{A}$ in $\{(x, y) | x \geq y\}$ and $\widetilde{A}$ defects more than A in $\{(x, y) | x \leq y\}$.

The notion of *defects more* is straightforward. For singletons $A = \{i\}$ and $\widetilde{A} = \{j\}$, it is simply that in $H \subset \mathrm{PD}$, player i defects (weakly) more than player j if $D_j \cap H \subset D_i \cap H$. When convenient, we use the term *cooperates (weakly) more* for the obvious analog. The notion of *more influenced by x* isolates the idea that players in set A are more driven to defection than players in $\widetilde{A}$ when x is relatively large but *without* being more prone to defection overall.

We begin with comparative static results that, as is typically done in applied work, investigate the effects of varying one parameter, *assuming* (rather than determining from behavior) that all other parameters stay fixed. The cutoff feature of equilibria (Proposition 1) immediately gives us our first comparative static: for fixed $F \in \mathcal{F}_S$, a player of type α cooperates more in PD than does a player of type $\widetilde{\alpha}$ if and only if $\alpha \geq \widetilde{\alpha}$. Intuitively, a player who believes he has more influence over his opponent's behavior cooperates in a larger set of games.

Proposition 3 below explores how predictions change as the *population* becomes more inclined toward magical thinking (in the sense of first-order stochastic dominance). It shows the equivalence between a first-order stochastically ranked pair of distributions and properties of both choice behavior in the observable domain (i.e., (b) and (d)) and their manifestations in the behavioral model (i.e., (c) and (e)). This may also serve to illustrate the usefulness of the equivalence between the axioms and the representation.

Proposition 3. *For any $F, \widetilde{F} \in \mathcal{F}_S$, let I and $\widetilde{I}$ be independently drawn collections of common size n from F and $\widetilde{F}$, respectively. For any $(x, y) \in \mathrm{PD}$, let $\alpha^*_{x,y}$ and $\widetilde{\alpha}^*_{x,y}$ be the unique equilibrium cutoffs for F and $\widetilde{F}$, and let the random variables $k_{x,y}$ and $\widetilde{k}_{x,y}$ be the number of players cooperating in their respective collections. The following statements are equivalent:*

(a) The CDF F first-order stochastically dominates (f.o.s.d.) $\widetilde{F}$ (i.e., $F(\alpha) \leq \widetilde{F}(\alpha)$ $\forall \alpha \in [0, 1]$).

(b) For all $(x, y) \in \mathrm{PD}$, the distribution of $k_{x,y}$ f.o.s.d. the distribution of $\widetilde{k}_{x,y}$.

(c) *For all* $(x, y) \in \mathrm{PD}$, $F(\alpha^*_{x,y}) \leq \widetilde{F}(\widetilde{\alpha}^*_{x,y})$.

(d) *For any* $\alpha \in [0, 1]$, *a player of type* α *is more influenced by* x *relative to* y *when facing* F *than when facing* $\widetilde{F}$.

(e) *For any* $(x, y) \in \mathrm{PD}$, $\alpha^*_{x,y} \leq \widetilde{\alpha}^*_{x,y}$ *if and only if* $x \leq y$.

Interpreting the proposition, (b) and (c) show specific manners in which greater degrees of population-wide magical thinking and of cooperation are synonymous. Notice that (b) is only useful if the analyst either assumes the empirical validity of F and $\widetilde{F}$ (see Section S.2), or if she is interested in understanding how much cooperation the players themselves predict as their common belief about the distribution of α-types changes—which does provide some useful intuition for the final two claims.

The final two statements are perhaps a bit more surprising. They can be interpreted as answering the question, "How does the behavior of the player with magical-thinking type α change if (the players believe that) the magical thinking of the population increases/decreases?" The answer depends on the relative magnitudes of the two motives for defection. From (b) and (c), F f.o.s.d. $\widetilde{F}$ means more cooperation from the F population than from the $\widetilde{F}$ population. As discussed following Proposition 2, when $x < y$, players want to cooperate if enough others are cooperating, which (d) and (e) reflect. However, when $x > y$, the gain from defecting on cooperators is relatively large, and the α-type takes advantage of increased cooperation in the populace by defecting in more games when facing F than when facing $\widetilde{F}$.

In axiomatic theories of individual choice, customarily, the aim of comparative statics results is to disentangle the behavioral content of different parameters, relying crucially on the separate identification of those parameters. Consider first the individual types $(\alpha_i)_{i\in I}$. If the analyst wishes to know if differences in the behaviors of two collections are at least partially due to differences in individual types, she can leverage the facts that in the model, equilibrium behavior is independent of F when $x = y$ (Section 2.1.1), and that any player's type can be identified from play in such games. Intuitively, when $x = y$ any player's incentive to defect is independent of what he believes about his opponent's decision. This is formalized in Proposition 4(a) below.

For the commonly believed distribution of types, F, part (b) of the proposition captures the exact behavioral content of keeping the actual types in the collection fixed and changing only these beliefs. Similar to Proposition 3(a) and (d), (the discretized analog of) a first-order stochastic shift in beliefs is equivalent to players becoming more influenced by x.

Proposition 4. *Consider two collections* I *and* $\widetilde{I}$ *such that* $|I| = |\widetilde{I}| = n$, *and whose behavior is described by* $[F, (\alpha_j)_{j\in I}]$ *and* $[\widetilde{F}, (\widetilde{\alpha}_j)_{j\in\widetilde{I}}]$, *respectively, with* $F, \widetilde{F} \in \mathcal{F}_S$ *and each collection ordered by increasing* α *values.*

(a) *In* $\{(x, y)|x = y\}$, *player* $i \in I$ *defects more than player* $j \in \widetilde{I}$ *if and only if* $\alpha_i \leq \widetilde{\alpha}_j$.

(b) *Collection* I *is more influenced by* x *relative to* y *than is* $\widetilde{I}$ *if and only if, for all* $i \leq n$, $\alpha_i = \widetilde{\alpha}_i$ *and* $F(\alpha_i) \leq \widetilde{F}(\widetilde{\alpha}_i)$.

4. Evidence and alternative theories

In this section, we first discuss how the available experimental evidence aligns with our axioms. We then discuss additional evidence, drawn from studies of manipulated variants of PD games, finding support for the magical-thinking interpretation of the behavioral model.

The rationale for discussing both types of evidence is as follows. The utility of our representation result is that it establishes (a) the (nonobvious) behavioral content of a model built on a documented psychological phenomenon (see the Introduction), applied to a domain of economic interest, and (b) that empirically plausible axioms on the domain of interest can be explained by a tractable model that is not obvious from mere inspection of those axioms. Hence, the first set of evidence presented speaks to the plausibility of the axioms as empirical regularities, while the second set speaks more to the relevance of the psychological decision-making process.

Starting with Rapoport and Chammah (1965), experimentalists have investigated how the payoffs in the prisoners' dilemma affect observed levels of cooperation.[22] For the stakes typically used in experiments, a positive affine transformation of the game payoffs seems to have little effect on the level of cooperation in the prisoners' dilemma (for example, Jones et al. 1968), or on play in games more generally (Camerer and Hogarth 1999, Kocher et al. 2008), consistent with Axiom 1. For very significant stakes, evidence from televised game shows where contestants play a one-shot prisoners' dilemma (of course, without anonymity) paints a similar picture (List 2006, Van de Assen et al. 2012). In fact, Axiom 1 is commonly assumed, and most experiments do not even test it. Also, as in more familiar contexts, continuity (Axiom 2) is hard to falsify empirically and should be thought of as a technically useful abstraction.

The main experimental finding for prisoners' dilemma games is that a substantial proportion of subjects choose to cooperate (see Dawes and Thaler 1988 for a survey), and that cooperation monotonically decreases with the motives to defect: x and y. For example, Charness et al. (2016) find that cooperation levels decrease monotonically from 60% to 23% when varying (x, y) on an increasing path from $(\frac{1}{4}, \frac{1}{4})$ to $(4, 1)$ (modulo positive affine transformations). Monotonicity has also been verified within subject (Ahn et al. 2001, Engel and Zhurakhovska 2016), giving strong support to Axiom 3. Any theory that aims to explain observed play in PD games should account for this evidence.

Axiom 4 is testable, but the available evidence on play in the PD is too incomplete to evaluate it directly. However, again starting with Rapoport and Chammah (1965), various unidimensional indices have been proposed (though with little theoretical foundation) to capture the magnitude of the incentive to defect, depending on the payoff parameters, and then used to forecast the level of cooperation across different prisoners' dilemma games. Empirically, the best validated of such indices are increasing in

[22] Ignoring the possible differences in behavior between the one-shot game and its finitely repeated version, early experimental works simply report aggregate behavior across rounds and subjects (for example, Rapoport and Chammah 1965, Steele and Tedeschi 1967, Jones et al. 1968). More recent studies of the one-shot game either randomly rematch subjects after every round of play (for example, Ahn et al. 2001), or use the "strategy method" (Engel and Zhurakhovska 2016), or truly have each subject play just a single game one time (Charness et al. 2016).

$\frac{r-p}{r-p+x+y}$ (see Steele and Tedeschi 1967, for example). This ratio is invariant to positive affine transformations of game payoffs, consistent with Axiom 1, and becomes $\frac{1}{1+x+y}$ in PD. Therefore, these indices predict that the level curves of constant aggregate cooperation will be thin, linear, and downward sloping, as they are in our model, owing to Axioms 1–4 and the fact that individual M_i lines do not cross, an implication of Axiom 5. The empirical support for these indices then provides indirect evidence in support of Axioms 1–4, but not of the differing slopes of level curves that are also implied by Axiom 5 (illustrated in Figure 3(b)), meaning our axioms/model provide a more nuanced prediction.[23]

Axiom 5 is a novel type of assumption that is central for our theory. It describes the correlation of behavior across players and games. This correlation has not been a focus of experimental investigation. Recall that if players are sensitive to the true cost of cooperating, Axiom 3 implies Axiom 5. The strong support in favor of Axiom 3, therefore, strengthens the empirical plausibility of Axiom 5. Ultimately, however, the validity of the axiom is an empirical question, and in that sense our theory suggests a fruitful avenue for future experiments.

Because the axioms distill the precise behavioral content of our theory, they facilitate comparison not only with the experimental evidence, but also with alternative models. In Section S.1, we formally demonstrate that canonical models with the three most common forms of other-regarding preferences—altruism (Ledyard 1995, Levine 1998), inequity aversion (Fehr and Schmidt 1999), and reciprocity (Rabin 1993)—violate our axioms, and hence make different predictions on our domain.[24] In McKelvey and Palfrey's (1995) notion of quantal-response equilibrium (QRE) each player chooses every available action with positive probability, which can be interpreted as random errors. Immediately then, QRE predicts a positive degree of cooperation in the prisoners' dilemma. Further, given the distribution of opponent play, the probability of selecting an action increases with the expected payoff from doing so, as is also true in our model. However, despite the many degrees of freedom afforded QRE, its implications for aggregate behavior differ from those of our model.[25] More importantly though, instead

[23] We are unaware of studies that provide detailed enough data to test the predictions of our model against the predictions based on these indices.

[24] A succinct intuition is that the most altruistic players in a population always fail Axiom 3 because, in games where they (correctly) predict their opponent will defect with probability 1, increasing their opponent's payoff from doing so *increases* the altruistic player's preference for cooperation. The models of inequity aversion and reciprocity have a coordination feature to them: players are willing to cooperate if and only if they believe cooperation by their opponent is sufficiently likely. This leads to equilibrium multiplicity: for every game, all players defecting is an equilibrium, but in some games cooperation by some players occurs in other equilibria. Further, because of this coordination component, the set of games that have equilibria with some cooperation end abruptly, as coordinated cooperation unravels due to a small increase in the incentive to defect, leading to abundant violations of Axiom 2.

[25] For example, even though the expected payoff from defection is always larger than from cooperation, (in expectation) the majority of individuals in our model will cooperate for small enough x and y, in line with evidence (Charness et al. 2016), but in contrast to QRE. Also, beyond PD games, there are games for which our model predicts that some actions are never played; for instance, any game where the socially optimal action is also dominant (Proposition 5). See also Proposition 10 on games with larger action sets.

of attributing differences in observed behavior to randomness, our axioms and model speak directly to heterogeneity in individual behavior.

We now discuss evidence suggestive of magical thinking from games that are in natural extensions of our domain (which for brevity we do not formalize here):

(i) Most immediately, players in our model would have completely standard preferences over the domain of final game-payoff vectors (unlike altruistic or inequity-averse players). Consistent with this, when the prisoners' dilemma is modified to have a passive opponent (so the unconstrained player is unilaterally selecting the payoff vector), higher rates of "defection" are found (Ellingsen et al. 2012).

(ii) Shafir and Tversky's (1992) observation that the level of cooperation by second-movers is significantly lower in the sequential prisoners' dilemma than in the standard, simultaneous-move version of the game—even if the first-mover cooperates—is highly suggestive of magical thinking, but inconsistent with standard forms of other-regarding preferences. Reciprocity, notably, predicts that second-movers should be more likely to cooperate following cooperation than in the simultaneous-move game.[26]

(iii) In a similar vein, Morris et al. (1998) find that the temporal order of moves affects cooperation even when the decision of the first-mover is *not* revealed. Consistent with magical thinking being the belief that one may directly influence the (yet unchosen) action of one's opponent, they find greater cooperation when players move first compared to second. Other-regarding preferences (as well as the evidentiary-reasoning interpretation of the beliefs in our model; see the Introduction) provide no rationale for this discrepancy, as play should be invariant to this strategically irrelevant difference in the games.

(iv) In a number of studies, experimental subjects played prisoners' dilemma games and were also asked to predict the behavior of their opponents. Subjects who defected were more likely to predict that their opponents would defect.[27] This feature is implied by the interpretation of our model, but absent from models with other-regarding preferences. While there may seem to be a sense in which it is consistent with reciprocity—players are more likely to cooperate when they expect cooperation from others—it is clearly inconsistent with standard notions of

[26] In the sequential-move game almost no second-movers cooperated after observing defection by their opponent. Perhaps more surprisingly, only about 15% of second-movers cooperated after observing cooperation. At the same time, and in line with other PD experiments (Dawes and Thaler 1988), 37% of subjects cooperated in the standard, simultaneous-move PD. In the study only a small subset of the games each subject played were prisoners' dilemma games. There is some evidence that repeated play of the one-shot, sequential prisoners' dilemma can reverse their observation (Clark and Sefton 2001), possibly because ethical considerations, like reciprocity, become more salient through frequent, uninterrupted repetition.

[27] See Dawes et al. (1977), Orbell and Dawes (1991), Engel and Zhurakhovska (2016), Rubinstein and Salant (2014). Rubinstein and Salant (2014) suggest that players' ex post reported beliefs will accurately reflect the beliefs their choices were based upon in prisoners' dilemma games (which feature a dominant strategy), but that this may not be the case in games such as hawk–dove (where either action is a best response to some belief).

equilibrium, even if players care about reciprocity. Either the cooperators are too optimistic or the defectors are too pessimistic about their opponents' behavior.

Finally, it is clear that magical thinking introduces a perceived benefit from cooperation. One could, of course, consider a reduced-form model in which each player may have a direct utility gain from choosing c over d. This gain might be interpreted as a form of "warm glow" (as introduced by Andreoni 1989 in the context of public-good games). As an alternative explanation for our data, such a model would have the following flaws.

First, to align with the expected-payoff calculations in our model on PD^0, this utility gain would have to be independent of x and y, but increase proportionally when simultaneously scaling r and p. That is, even though the warm glow a player would obtain by cooperating would have to vary across games, it would not depend on how strong were the motives for defection that he overcame—including them being arbitrarily small. The reduced-form model would give the analyst no intuition for this seemingly curious form of dependence. In contrast, our model provides a psychological mechanism, that of magical thinking, which generates it. Second, this model of warm glow would be at odds with the evidence in (i)–(iv) above. Third, in the next section we extend our model beyond the prisoners' dilemma to games in which it is unclear how to interpret as warm glow the utility gain a player would need to receive from selecting one action over the other so as to match the predictions of our model. For example, for any battle of the sexes game there would need to be no warm glow attached to either action choice, but there are two games, arbitrarily nearby, such that a player would need to receive a warm glow from selecting his preferred meeting event (instead of his opponent's) in one game but the reverse in the other game.

5. Beyond prisoners' dilemma games

We now extend our game-theoretic analysis beyond PD games. Section 5.1 takes the behavioral model characterized by Theorem 1 and investigates its predictions for all symmetric 2×2 games.[28] Section 5.2 extends the model to accommodate arbitrary finite action sets.

5.1 *Symmetric* 2×2 *games*

Let $S^0 := \{(r, p, x, y) | r \geq p\}$, with labels as in Figure 1, denote the set of all symmetric 2×2 games. As discussed in the Introduction, such games give strategic meaning to the notion that magical thinkers believe they influence others to select the same action as they do, without having to rely on arbitrary labels of actions. Therefore without loss, c (respectively, d) still corresponds to the action leading to the weakly superior (inferior) symmetric outcome, but outside the prisoners' dilemma we no longer refer to it as cooperate (defect).

We find that our model provides a unified explanation of the experimental evidence in several of the most often studied games: hawk–dove/chicken, the stag hunt, and the

[28] In Section S.3, we explore the extension of the axiomatic component to this domain.

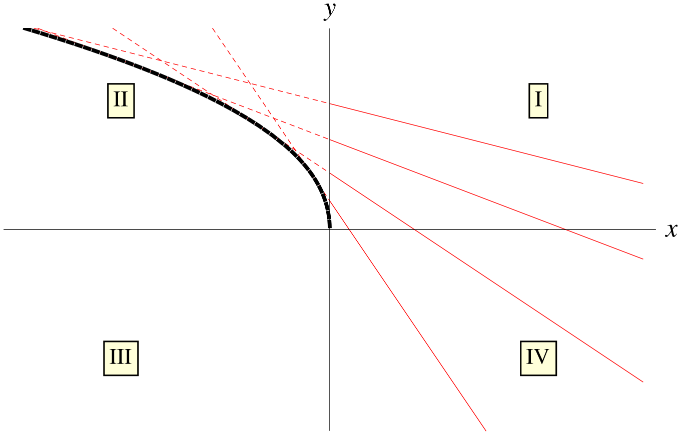


FIGURE 4. Depiction of the set of symmetric 2 × 2 games in which $r = 1$ and $p = 0$.

battle of the sexes (in addition to the prisoners' dilemma). We also compare the predictions of our model to Nash equilibrium in the standard model (hereafter, simply *Nash equilibrium*).[29] We consider first the generic case in which $r \neq p$, followed by the nongeneric complement.

5.1.1 *If symmetric outcomes are* not *payoff equivalent* Let S_G^0 be the (generic) set of games $\{(r, p, x, y) | r > p\}$, of which PD^0 is a subset. Lemma 1 remains valid, so we normalize $r = 1$ and $p = 0$, and again denote this normalized subset as S_G (represented as the plane in Figure 4). It remains true that for any $g \in S_G$, all equilibria are cutoff equilibria (Definition 2), and that a player of type α satisfies the indifference equation (2) if and only if

$$g \in \tilde{M}_\alpha := \left\{(x, y) \middle| y = \frac{\alpha}{(1-\alpha)F(\alpha)} - x\left(\frac{1-F(\alpha)}{F(\alpha)}\right)\right\}.$$

Let $B : \mathbb{R}_- \to \mathbb{R}$ be the lower envelop of $(\tilde{M}_\alpha)_{\alpha \in (0,1)}$ on the domain $x \leq 0$. Notice that (i) $B(0) \geq 0$, (ii) B is decreasing and concave, and (iii) $\lim_{x \to -\infty} B(x) = \infty$. Figure 4 depicts B (and four sample $\tilde{M}$-lines) for the case of $F(\alpha) = \sqrt{\alpha}$.

PROPOSITION 5. *For any $g \in S_G$, an equilibrium exists, all equilibria are cutoff, and the following statements hold:*

- *If $x > 0$, then the equilibrium cutoff α^* is unique, interior (i.e., $\alpha^* \in (0, 1)$), and characterized by (2).*
- *If $x \leq 0$, then $\alpha^* = 0$ is an equilibrium cutoff. It is unique if and only if $y < B(x)$.*

The labeled quadrants of Figure 4 serve as a useful taxonomy for our discussion of the games in S_G. Quadrant I corresponds to PD, which we have focused on up to

[29] We maintain our focus on symmetric equilibria (of both our model and the standard one). In a truly symmetric, anonymous, one-shot setting, asymmetric equilibria seem implausible as neither player would have any way of knowing if he were player 1 or player 2.

now. We proceed clockwise. For brevity, we focus the discussion on the interiors of each quadrant.

Quadrant IV. The defining feature of prisoners' dilemma games is that there are strict gains to a player for selecting d whether his opponent is playing d or c (i.e., $x, y > 0$). The games of quadrant IV retain the latter, meaning there are still gains from unilaterally deviating away from the better symmetric outcome (c, c). A particularly well know example of such games are hawk–dove (also known as chicken) games, where $y \in (-1, 0)$. Action c corresponds to *dove* and d to *hawk*.

Proposition 5 establishes that the equilibrium characterization results for PD (Section 2.1) extend unchanged to these games, and it is straightforward to show that Proposition 3 extends verbatim as well. In addition, we find the following result. For any $g \in S_G^0$ with $x > 0$, let π_g be the probability with which a player selects d in the unique symmetric Nash equilibrium of g. The corresponding probability in our behavioral model is $F(\alpha_g^*)$.

Proposition 6. *For any $g \in S_G^0$ with $x > 0$, $\pi_g > F(\alpha_g^*)$. In addition, if x and y are held fixed and $(r - p) \to 0$ (or, more generally, if $\frac{r-p}{x+|y|} \to 0$), then $(\pi_g - F(\alpha_g^*)) \to 0$.*

The result states that players are drawn to the action that produces the superior symmetric outcome more often than is predicted by the symmetric Nash equilibrium. This is consistent with experimental findings in the hawk–dove game (for example, Rubinstein and Salant 2014). However, as the difference between the symmetric outcomes disappears so too does the difference in the two models' predictions. Intuitively, as the difference between the symmetric outcomes disappears, the magical-thinking component has a vanishing impact on any player's *ranking* between c and d (even though players with different α-types still differ in their expectations over opponent play). Section 5.1.2 covers the limit case where $r = p$.

Quadrant III. In these games c is *both* the action leading to the better symmetric outcome and a dominant strategy (even without magical thinking). It seems natural that all players should then choose c—as they do in the unique equilibrium of our behavioral model by Proposition 5.

Quadrant II. Quadrant II consists of coordination games, such as the stag hunt, in which both symmetric outcomes constitute Nash equilibria, but (c, c) Pareto dominates *all* other outcomes. The choice of d in such games seems empirically implausible if the loss $x \leq 0$ of playing d rather than c against an opponent playing c is large, and the gain $y > 0$ of playing d rather than c against an opponent playing d is small. Players should find it natural to coordinate on c in such a game. At the same time, if the gain y of playing d against d is large compared to the loss $x \leq 0$ of playing d against c, then it becomes risky to rely on the opponent to play c, and d also becomes a plausible choice. These intuitions are supported by experimental evidence (for example, Straub 1995).

From Proposition 5, our behavioral model is consistent with all players selecting c, and it *uniquely* predicts this behavior for a subset of those games where coordinating on c is not "too risky" in the sense just described. This set is precisely characterized as the strict-lower-contour set of B. Hence, our behavioral model generates a unique equilibrium prediction in more games than does the standard model. More generally, the set of games for which our model makes a unique prediction is larger (in the sense of set inclusion) with the more magical thinking there is in the population (in the sense of a first-order stochastically dominant shift of F). For games in the upper-contour set of B, where the trade-off between the overall payoffs (higher under (c, c)) and riskiness is more pronounced, our model does not make a unique prediction and can accommodate a significant proportion of players selecting d.[30]

The intuition we gave for the implausibility of selecting d when $|x|$ is large compared to y is reminiscent of the motivation for the *risk dominance* criterion (Harsanyi and Selten 1988). It is easy to verify that, in the standard model, (c, c) is risk dominant when $|x| > y$ and (d, d) is risk dominant when $|x| < y$. Our boundary, B, is more nuanced than the fixed linear one implied by risk dominance, as it depends on the (perceived) distribution of α-types. Our model, therefore, provides flexibility, though within constraints, for explaining behavioral data in this quadrant of games by varying F, and at the same time connects behavior in this quadrant to behavior in other games. For example, collections with higher rates of cooperation in prisoners' dilemma games also have a larger set of quadrant-II coordination games in which the payoff-dominant Nash equilibrium is uniquely selected in our model.

5.1.2 *If symmetric outcomes are payoff equivalent* Consider now the (nongeneric) set of games $S_N^0 := \{(r, p, x, y)|r = p\}$. In such games our model of magical thinking is not behaviorally distinct from the standard model.

Proposition 7. *For any $g \in S_N^0$, an equilibrium exists.*

- *If (σ, P) is an equilibrium (of our model), then there exists a symmetric Nash equilibrium characterized by $\pi_g = P$.*
- *If π_g characterizes a symmetric Nash equilibrium, then there exists σ such that (σ, π_g) is an equilibrium (of our model).*[31]

In line with the limit property established in Proposition 6, when there is no payoff difference between the symmetric outcomes, magical thinking does not influence behavior in one direction or the other. Hence, the cutoff property is no longer a requirement for equilibrium, as there is no reason that players with higher α-types are more drawn to c.

Though not always labeled as a symmetric game, battle of the sexes games are a subset of S_N^0 in which $x > 0 > y$, $x \neq -y$. In our theory, action labels are only for the convenience of the analyst; it is the symmetry of the game that determines what "taking

[30] For (x, y), $y \geq B(x)$, $\alpha \in (0, 1)$ is an equilibrium cutoff if and only if $(x, y) \in \tilde{M}_\alpha$. Hence, the number of equilibria in (x, y) in which not all types select c is the number of $\tilde{M}_\alpha$-lines that pass through (x, y).

[31] More specifically, (i) if π_g characterizes a weak Nash equilibrium, then any σ such that $\int_0^1 \sigma(d|\alpha)\, dF(\alpha) = \pi_g$ constitutes an equilibrium; (ii) if π_g characterizes a strict Nash equilibrium, then in any equilibrium, $\sigma(d|\alpha) = \pi_g$ for all $\alpha \in [0, 1)$, but $\sigma(d|1)$ can be arbitrary since $\alpha = 1$ players are indifferent between c and d when $r = p$.

the same action" means. In a battle of the sexes game then, c and d do not correspond to "go to the ballet/boxing match," but to "go to my own/my opponent's preferred event" (with the labeling depending on the ranking of x and $-y$).

For a magical thinker i, therefore, both c and d are self-defeating: by being "selfish" and choosing his preferred event, i believes it more likely that his opponent j will likewise choose j's preferred event, but also analogously if i tries to be "accommodating" by choosing j's preferred event. The magical-thinking component then has no effect on preferences over actions, and equilibrium play is just as in the standard model.

Any battle of the sexes game has a unique symmetric Nash equilibrium, and hence our behavioral model predicts the same distribution of observed behavior. Notably, this common prediction is substantiated by the experimental studies of battle of the sexes games.[32] The ability to explain experimental findings across the well known games surveyed in this paper serves as another key distinction between our model and models of other-regarding preferences discussed in Section 4, each of which predict patterns of play in the battle of the sexes that differ from the prediction of the standard model.[33]

5.2 *Accommodating arbitrary finite action sets*

Allowing arbitrary finite action sets requires the following additional notation. Let $A := \{0, 1, \dots, K\}$ and let $v(k, k')$ be the (finite) game payoff a player receives from selecting action k when his opponent selects k'. Define $s(k) := v(k, k)$, and, without loss, order the actions such that $s(\cdot)$ is nondecreasing in k. To avoid technicalities, we consider the generic case in which $s(k) < s(k + 1)$ for all $k < K$.[34] Let Γ denote the set of such games.

In the extended behavioral model, each player $i \in I$ is still privately endowed with a type $\alpha_i \in [0, 1]$ and there is a common prior that types are drawn i.i.d. from a distribution with CDF $F \in \mathcal{F}$. For each game, player i evaluates the expected payoff of action $a_i = k \in A$ as

$$V_i(k) = \alpha_i s(k) + (1 - \alpha_i) \sum_{k' \in A} P_i(k') v(k, k'), \tag{3}$$

where $P_i(k')$ is the probability i assigns to $a_j = k'$, *conditional* on a_i and a_j being determined independently. His strategy, σ_i, is again characterized by the probability with

[32] Camerer (2003, Chapter 7.2) summarizes the evidence and concludes, "Even if the subjects are not deliberately randomizing, the data are consistent with the idea that, as a population, they are mixing in the [symmetric Nash] equilibrium proportions." Note that one could have imagined an alternative concept of magical thinking in the battle of the sexes: by choosing to go to the ballet, for example, a player believes it is more likely that his opponent will choose to go to the ballet as well. In addition to relying on strategically irrelevant action labels, this alternative model would predict that players select their preferred event more frequently than is found in the experimental evidence.

[33] This result is not difficult to demonstrate. We omit the analysis for the sake of brevity.

[34] If the ranking is not strict, our characterization (Proposition 8) holds only up to payoff equivalence: an equilibrium exists and for any equilibrium there exists a payoff-equivalent increasing equilibrium. Also, if $s(1) = s(K)$, then Proposition 7 extends and the predictions of our model are not behaviorally distinct from symmetric Nash equilibrium in the standard model.

which he selects each $k \in A$ if his type is α_i (denoted $\sigma_i(k|\alpha_i)$). The equilibrium notion is the immediate extension of Definition 1.[35]

When evaluating a potential action k, it is clear from (3) that the importance given to $s(k)$, the payoff generated by the symmetric profile (k, k), is increasing in α_i. This leads to the following increasing-in-type property of equilibrium.

Definition 6. A strategy σ is *increasing* if there exists $\alpha_0^* = 0 \leq \alpha_1^* \leq \cdots \leq \alpha_{K+1}^* = 1$ such that $\sigma(k|\alpha) = 1$ for all $\alpha \in (\alpha_k^*, \alpha_{k+1}^*)$.

Proposition 8. *For any $F \in \mathcal{F}$ and $g \in \Gamma$, an equilibrium (σ, P) exists, and in all equilibria σ is increasing.*

5.2.1 *Pure dilemma games* Given that the bulk of our analysis has focused on PD games, for brevity, we focus our remaining analysis on their natural extension, which we refer to as pure dilemma games.

Definition 7. A game $g \in \Gamma$ is a *pure dilemma* if $v(k, k') > v(k+1, k')$ for all k, k'.

In terms of game payoffs then, there is always a strict individual incentive to play a lower (indexed) action, but higher actions can be viewed as "more cooperative." In the standard model there is a unique Nash equilibrium: all players select $a_i = 0$. By comparison, in any equilibrium of our model, a positive measure of types select actions other than 0. Hence, just as in PD games, magical thinking generates a positive degree of cooperation, while the standard model predicts none.

As in Section 3, we can investigate whether there are connections between notions of more magical thinking and more cooperative behavior. Again, it is immediate from the increasing property of equilibrium (Proposition 8) that in any pure dilemma, a player's cooperation level is increasing in type. What about increases in population-wide magical thinking, as measured by a first-order shift in F? The following proposition identifies a sufficient condition on the underlying game under which this change leads to uniformly greater cooperation.

Definition 8. For pure dilemma game $g \in \Gamma$, an increasing strategy σ is *more cooperative* than increasing strategy $\hat{\sigma}$ if $\alpha_k^* \leq \hat{\alpha}_k^*$ for all k.

Proposition 9. *Let $g \in \Gamma$ be a supermodular pure dilemma game.*[36]

(i) For any $F \in \mathcal{F}$, there exists a most and a least cooperative equilibrium, denoted (σ_F^M, P_F^M) and (σ_F^L, P_F^L), respectively.

[35] With $P = (P(k))_{k \in A}$ and requirement (iii) of the definition generalizing to $P_i(k) = P(k) = \int_0^1 \sigma(k|\alpha)\, dF(\alpha)$ for all $i \in I$ and $k \in A$. This specification represents the literal extension of magical thinking as "players believe they influence others to select the same action as they do," as discussed in the Introduction. With more than two actions, one could envision more general notions of magical thinking that still capture its essence: players believe they influence the opponent to select an action more similar to their own action than the opponent otherwise would have. For simplicity, we consider only the literal extension.

[36] That is, for all $k' \geq k$ and $l' \geq l$, $v(k', l') - v(k, l') \geq v(k', l) - v(k, l)$.

(ii) If $F \in \mathcal{F}$ first-order-stochastically dominates $\widetilde{F} \in \mathcal{F}$, then σ_F^M is more cooperative than $\sigma_{\widetilde{F}}^M$ and σ_F^L is more cooperative than $\sigma_{\widetilde{F}}^L$.

First consider PD games. A PD game is supermodular if and only if $y \geq x$. Notice that such a PD game satisfies the very common assumption that (c, c) maximizes the sum of game payoffs: $2r > (r + x) + (p - y)$ or, equivalently, $y > x - (r - p)$. This feature generalizes: if $g \in \Gamma$ is supermodular, then $2s(k) > v(l, l') + v(l', l)$ for all $l \leq l' \leq k$ with $l < k$.

For a supermodular PD game, from Propositions 2 and 3, we know that if $F, \widetilde{F}$ satisfy Condition S, then each has a unique equilibrium, (σ_F, P_F) and $(\sigma_{\widetilde{F}}, P_{\widetilde{F}})$, and F f.o.s.d. $\widetilde{F}$ implies σ_F is more cooperative than $\sigma_{\widetilde{F}}$. Proposition 9 generalizes this comparative static along two dimensions: it does not assume Condition S—so there may be multiple equilibria even if $g \in \text{PD}^0$—and it applies to pure dilemma games beyond the PD. The intuition remains similar. Fix a game and imagine first that the equilibrium was unchanged following a shift from $\widetilde{F}$ to F. Then a given α-type would face a more cooperative (perceived) distribution of play. But just as Section 3 demonstrated for PD games, depending on payoff parameters, this change could make the α-type more or less cooperative (i.e., seeking to take advantage of the population's greater degree of cooperation in the latter case). Supermodularity of g implies the latter case never obtains.[37]

We conclude by identifying a class of games for which all of the analysis from preceding sections (both game-theoretic and axiomatic) applies.

Definition 9. A pure dilemma game $g \in \Gamma$ has *increasing returns to joint cooperation* if, for all k, k',

$$\frac{s(k+1) - s(k)}{v(k, k') - v(k+1, k')} \geq \frac{s(k) - s(k-1)}{v(k-1, k') - v(k, k')}.$$

Notice that this condition is neither weaker nor stronger than supermodularity. It requires that the benefit to increased joint cooperation increase at a rate greater than the increase in the individual benefit from lowering one's own action, independent of the action selected by the opponent. Consider, for example, a public-good game where both players can contribute any amount between 0 and K dollars to the public good, and the function Φ (with $\frac{1}{2} < \Phi' < 1$) measures the individual benefit derived from the amount of public good provided given the total contribution. That is, $v(k, k') = K - k + \Phi(k + k')$. If Φ is (weakly) convex, then the returns to joint cooperation are (weakly) increasing.

Proposition 10. *If a pure dilemma game $g \in \Gamma$ has increasing returns to joint cooperation, then in any equilibrium, $\alpha_1^* = \alpha_K^* \in (0, 1)$ and is equal to an equilibrium cutoff of the PD game generated by deleting actions $\{1, \ldots, K - 1\}$.*

[37] Notice that the notion of "more cooperative" in Definition 8 is a strong one: each α-type plays a (weakly) higher action. Alternatively, for $F, \widetilde{F} \in \mathcal{F}$ and corresponding equilibria $(\sigma, P), (\widetilde{\sigma}, \widetilde{P})$, one could view the first population's behavior as more cooperative if its F-measured distribution of play, P, f.o.s.d. the second population's $\widetilde{F}$-measured distribution of play, $\widetilde{P}$, even if fixed α-types play less cooperative actions under σ than under $\widetilde{\sigma}$. Under Condition S, Proposition 3 shows that a first-order shift in F implies more cooperation in this weaker sense for all PD games. It is not difficult to construct further examples outside PD games.

Under increasing returns to cooperation, (almost) all types play either 0 or K and intermediate actions play no (meaningful) role. Therefore, our equilibrium characterization and tight condition for uniqueness, our comparative statics, and even our axiomatization all apply to this class of games with the obvious additional axiom that intermediate actions are never chosen.

An example of an experimental study of linear two-player public-good games is Capraro et al. (2014), who investigate how the distribution of play responds to changes in the social benefit from cooperation, captured by the scalar ϕ, where $\Phi(k+k') = \phi(k+k')$. They find that, independent of the value of ϕ, about 20% of subjects contribute half their endowment ($a_i = \frac{K}{2}$), and argue that these subjects are likely following a simple heuristic. The predictions of our model are well aligned with behavior in the remaining 80% of subjects. The vast majority of these subjects (75% of the total population) choose an extreme action $a_i \in \{0, K\}$ (in line with Proposition 10), and the proportion of them choosing K increases with ϕ (as predicted by Axiom 3 (Monotonicity), in the aggregate).[38]

6. Discussion

Methodology

Our approach connects behavioral axioms on the observed play of a collection of players to a representation that suggests a procedural interpretation of individual behavior and an equilibrium concept. This is analogous to the standard axiomatic analysis of individual choice (see footnote 10). Throughout the paper we have stressed this analogy, as well as the differences that arise when leveraging our richer domain of group behavior.

At the outset, we discussed several benefits of this methodology. Here we compare it to related, alternative approaches. In (what we refer to as) the standard approach for connecting behavioral axioms to strategic, multi-agent environments, first, axioms characterize a specific utility representation of individual preferences regarding (lotteries over) physical allocations; then, second, physical games are described in terms of those utilities; finally, strategic analysis is performed according to an exogenously given solution concept (usually an equilibrium notion). The prototypical example of the standard approach assumes that players care only about their own physical payoffs and have risk preferences as axiomatized by Von Neumann and Morgenstern (1944). As another example, Rohde (2010) provides axioms for the inequity-averse utility function employed in the game-theoretic analysis of Fehr and Schmidt (1999). (See also Dillenberger and Sadowski 2012, Saito 2013.)

Relative to the benefits of our axiomatic approach that were listed as (B1)–(B3) in the Introduction, the standard approach has the following differences. In contrast to (B1), it relies on the assumption that behavior observed in the individual context is tightly connected to behavior in the strategic context. Clearly, it cannot achieve (B3), as it is not possible to derive that behavior corresponds to any particular equilibrium notion,

[38] In the experiment, $v(k, k') = K - k + wk'$, with $2 \leq w \leq 10$. So each game is a pure dilemma and strategically equivalent to our description of a linear public-good game with $\phi = \frac{w}{w+1}$.

or that prior beliefs are common, by axiomatizing the objective function of each player separately.

Notably, once preferences over physical allocations are accounted for, if a given physical game remains a PD when represented in terms of these utilities, then the standard approach *cannot* explain cooperative behavior in this game—as each player *should* like higher utility payoffs for himself and be indifferent toward the utility payoffs of others. That is, explanations of cooperative behavior via altruism or inequity-aversion merely establish that games that look like PDs in terms of physical game payoffs may not actually be PDs in terms of utility payoffs. In contrast, our theory is robust to this alternative interpretation of game payoffs as utilities. Insofar as cooperation in such games is plausible, the standard approach's inability to explain it could be due to a discrepancy between preferences in the individual and strategic contexts (related to (B1)), or to inappropriate assumptions about beliefs or the equilibrium notion (related to (B3)).

Segal and Sobel (2007), Segal and Sobel (2008) go beyond the standard approach by fixing a single game and using as an additional primitive individual preferences over own (mixed) strategies, which may depend on what the (mixed) strategy profile is "supposed to be," which is referred to as the "context." They axiomatize a representation that can, for example, accommodate reciprocal preferences (Rabin 1993), and then employ the natural extension of Nash equilibrium as the solution concept. Clearly, their model differs from (B3) in the same manner as does the standard approach. In addition, the elicitation of the primitive requires that, for a single game, the analyst uncover the player's preferences over his own strategies given each possible mixed-strategy profile. In contrast to (B2), this is not data that is commonly collected, and faces the potential difficultly that the analyst must meaningfully communicate to each subject (i.e., have them truly believe) that the opponents are *actually* using that profile. At the very least this is more involved than simply asking subjects how they would like to play.

Our approach also differs from well known axiomatic treatments in bargaining and cooperative game theory, most prominently in Nash bargaining, where axioms directly characterize the outcome rather than a model of play in the strategic setting coupled with a solution concept.[39] Given the difficulty of accurately capturing the nuances of, say, bilateral negotiations, that such analysis does not rely on explicit modeling of the strategic situation is often seen as a strength. However, in the simplified setting of simultaneous-move, one-shot games, characterizing play seems a more natural objective. In addition, our representation suggests an intuitive explanation of the individual

[39] See Thomson (2001) for a thorough review of this approach. In addition, while axiomatic approaches in cooperative game theory characterize solutions without a strategic model, axiomatic approaches in noncooperative game theory typically take as given the structure of the strategic model—by assuming that all players and the analyst view the game in the same way and that players are "rational" (i.e., they maximize excepted utility with respect to some (nonmagical) belief about opponents' play)—with the aim of characterizing particular solution concepts (e.g., rationalizability, Nash equilibrium, correlated equilibrium, etc.). Again, see Thomson (2001, Section 12.3), and Blonski et al. (2011) for an application to equilibrium selection in repeated games. Outside the axiomatic literature, Bergemann et al. (2017) consider behavior in games to identify *interdependent* preferences over outcomes. Because the aim is identification of preferences, they take as given both the structure of the model and the solution concept.

(x, y)	Approx. Periods Until Stabilization	Approx. Stable % of Cooperation	Study
(1.00, 3.00)	Study ended after 10	$\lesssim 7\%$	Bó et al. (2010)
(2.33, 2.33)	20 out of 200	10%	Bereby-Meyer and Roth (2006)
(1.67, 1.33)	0 out of 200	19%	Andreoni and Miller (1993)
(0.44, 0.78)	10 out of 20	22%	Cooper et al. (1996)
(0.33, 0.11)	20 out of 75	37%	Aoyagi and Fréchette (2009)

TABLE 1. Studies of the one-shot prisoners' dilemma with random, anonymous rematching.

decision-making process, which enables us to provide comparative statics in terms of the model's parameters.

For any fixed physical game, our behavioral model resembles a Bayesian game, in that each player is endowed with a type that affects how he evaluates the expected payoff of a potential strategy. The difference, of course, is that in standard Bayesian games, a player's type maps outcomes into payoffs, whereas in our model, type affects the player's expectations about what outcomes will obtain depending on his action choice. Following Savage (1954), the derivation of subjective beliefs is a central concern in the context of individual choice. Our model provides an example where beliefs (here, about both the opponent's type and action choice) are derived from behavior in a strategic setting.

Repetition

There are various experimental studies that report on the evolution of cooperation when the same one-shot prisoners' dilemma is played repeatedly, with opponents randomly and anonymously rematched after every round. Many of these studies find an initial decline in the incidence of cooperation before it stabilizes at a nonzero level. For a sample of studies Table 1 reports each of their featured one-shot games $(x, y) \in \mathrm{PD}$ (modulo positive affine transformations), the approximate number of periods after which stabilization was reached, and the approximate average levels of cooperation thereafter.[40] Note that the stable levels of cooperation summarized in the table give further support to Monotonicity (Axiom 3) in the aggregate.

As with most theories of behavior in one-shot settings, our theory does not formally provide any explanation for the dynamics before steady state is reached. The typical explanation for a pattern of initially varying behavior followed by stability is that subjects are initially learning about the game (e.g., how it works, how others play, etc.); see Camerer and Fehr (2003) for a discussion. The interesting feature in this particular instance is that initial play is systematically more cooperative than steady state. While a formal model along these lines is beyond the scope of this paper, one possible explanation for this pattern is that subjects (act as if they) revise their estimates of their

[40] Not all studies provided these numbers explicitly. In these cases, they are estimates based on the information the studies do provide.

own α-types based on play. Because they are not in fact magical, the updating will be systematically biased downward, leading subjects to cooperate less.[41]

Magical-thinking-like notions in other strategic models

In models of oligopolistic competition, the notion of *conjectural variation* (Bowley 1924, Pigou 1924) bears some resemblance to magical thinking. However, in this literature a firm's belief about how its rival will respond to its action is typically interpreted as capturing a sequential response. In Roemer's (2010, 2013) *Kantian equilibrium*, each player prefers the equilibrium to any strategy profile that features identical deviations by all players. Related features are found in Feddersen and Sandroni (2006), who introduce *rule-utilitarian* players into a model of voting (see also Coate and Conlin 2004, Ali and Lin 2013). As suggested by their names, the modeling of both Kantian equilibrium and rule-utilitarian players are motivated by ethical concepts, in contrast to our psychological interpretation of magical thinking. While these different motivations may have similar behavioral consequences in some settings, our motivation more naturally allows for heterogeneity among players that is absent from these models.[42] In addition, the interpretation of magical thinking is more in line with the evidence discussed in Section 4.[43] Of course, as we have stressed throughout, this paper is also—and most importantly—distinguished by providing a tight axiomatic characterization of our behavioral model.

We conclude by noting that magical thinking is likely not an appropriate description of behavior in all games for which the standard game-theoretic predictions are unsatisfying, be they inaccurate and/or weak due to multiplicity (impeding applied/policy research, argues Pakes 2008). An ideal axiomatization would alleviate both problems by avoiding false predictions and ruling out multiplicity where it is descriptively inappropriate. Our model alleviates the first concern in prisoners' dilemma games, and improves on the second in coordination games by eliminating equilibrium multiplicity in games where coordination on the better symmetric outcome is intuitive.[44] While our representation features an equilibrium concept, this need not be the case in other contexts. As in theories of individual choice, the goal should be to connect testable and plausible behavioral axioms to an intuitive, tractable, and identified representation that may, or may not, have the strategic flavor of equilibrium.

[41] This could perhaps be because the α_i in our behavioral model (of the single-iteration game) represents only the *expected* influence i believes he possesses, but his beliefs allow that his influence may vary across subject pools or other environmental features. That play stabilizes at nonzero levels of cooperation suggests that the lower bound on α_i is believed to be positive by some i.

[42] Specifically, because in these models all Kantians or rule utilitarians evaluate strategies in the same way, any heterogeneity in nonstandard behavior is driven completely by asymmetry in the physical aspects of the game (e.g., variations in the cost of voting). In contrast, even in symmetric games, our model captures heterogeneity in behavior (e.g., in the sets of PD games that different players choose to cooperate in).

[43] Additional models in which players' beliefs about opponent play may be biased include Orbell and Dawes (1991), Bernheim and Thomadsen (2005), Masel (2007), Capraro and Halpern (2015), al Nowaihi and Dhami (2015).

[44] In addition, our behavioral model introduces the possibility of equilibrium multiplicity even in the PD, depending on F. It is then the axioms that rule out models with multiplicity, again showing that the second concern can also be addressed axiomatically.

Appendix: Proofs

Proof of Lemma 1. Fix any $(r, p, x, y) \in \mathrm{PD}^0$, and suppose that (σ, P) is an equilibrium according to Definition 1. Then, for any player i,

$$V_i(c) - V_i(d) = \alpha_i\bigl[r - p + (1-P)x + Py\bigr] - \bigl[(1-P)x + Py\bigr], \tag{4}$$

and player i strictly prefers c, strictly prefers d, or is indifferent if (4) is positive, negative, or zero, respectively. Hence, it is sufficient to show that the sign of (4) is unchanged for all α_i when the payoffs are transformed to $\kappa(r + \xi, p + \xi, x, y)$, where $\kappa > 0$. Then

$$\begin{aligned} V_i(c) - V_i(d) &= \alpha_i\bigl[\kappa r + \kappa\xi - \kappa p - \kappa\xi + (1-P)\kappa x + P\kappa y\bigr] - \bigl[(1-P)\kappa x + P\kappa y\bigr] \\ &= \kappa\bigl(\alpha_i\bigl[r - p + (1-P)x + Py\bigr] - \bigl[(1-P)x + Py\bigr]\bigr). \end{aligned} \tag{5}$$

Because $\kappa > 0$, the signs of (4) and (5) are identical. □

Proof of Proposition 1. *Claim (i).* Fix any $(x, y) \in \mathrm{PD}$, and suppose that (σ, P) is an equilibrium according to Definition 1. Then

$$V_i(c) - V_i(d) = \alpha_i\bigl[1 + (1-P)x + Py\bigr] - \bigl[(1-P)x + Py\bigr]. \tag{6}$$

Player i strictly prefers c, strictly prefers d, or is indifferent if (6) is positive, negative, or zero, respectively. For any $P \in [0, 1]$, (i) if $\alpha_i = 1$, then (6) is positive, and (ii) (6) is linear in α_i. It follows that the equilibrium must be a cutoff equilibrium and that $\alpha^* < 1$. Suppose now that $\alpha^* = 0$. Then, by Definition 1, $P = 0$. But then $V_i(c|\alpha_i = 0) - V_i(d|\alpha_i = 0) = -x < 0$, which contradicts $\alpha^* = 0$, establishing the result.

Claims (ii) and (iii). That solutions to (2) and equilibrium cutoffs are identical follows immediately from the properties of (6) discussed in the proof of Claim (i). It is therefore sufficient to establish existence of a solution to (2). If $x = y$, (2) has a unique solution: $\alpha^* = \frac{x}{1+x} = \frac{y}{1+y}$. If $x \neq y$, any solution to (2) is (implicitly) characterized by

$$F(\alpha^*) = T(\alpha^*|x, y) := \frac{\alpha^* - (1-\alpha^*)x}{(1-\alpha^*)(y-x)}. \tag{7}$$

If $x > y$, then $\lim_{\alpha\to 0} T(\alpha|x, y) = \frac{x}{x-y} > 1$ and $\lim_{\alpha\to 1} T(\alpha|x, y) = -\infty$. Further, T is continuous and strictly decreasing in α. Hence, it must intersect F, a continuous CDF on $[0, 1]$, exactly once. If $x < y$, then $\lim_{\alpha\to 0} T(\alpha|x, y) = \frac{x}{x-y} < 0$ and $\lim_{\alpha\to 1} T(\alpha|x, y) = \infty$. Further, T is continuous and strictly increasing in α. Hence, it must intersect F, a continuous CDF on $[0, 1]$, at least once. □

Proof of Proposition 2. From Proposition 1, the number of equilibrium cutoffs is the number of solutions to (2), and existence is established. For $x \geq y$, the arguments given in the proof of Proposition 1 demonstrate uniqueness of the solution for any $F \in \mathcal{F}$.

Now fix arbitrary $x < y$ and suppose $F'(\alpha) \leq \frac{F(\alpha)}{\alpha-\alpha^2}$ for all $\alpha \in (0,1)$. Consider a solution $\alpha^* \in (0,1)$:

$$F'(\alpha^*) \leq \frac{F(\alpha^*)}{\alpha^* - (\alpha^*)^2} = \frac{T(\alpha^* | x, y)}{\alpha^* - (\alpha^*)^2} = \frac{\alpha^*(1+x) - x}{\alpha^*(1-\alpha^*)^2(y-x)}. \tag{8}$$

Further,

$$T'(\alpha^* | x, y) = \frac{1}{(1-\alpha^*)^2(y-x)}. \tag{9}$$

It is a matter of simple algebra to see that the rightmost term in (8) is strictly less than (9) for any x, y, α^* such that $0 < x < y$ and $\alpha^* \in (0,1)$. Hence, at any solution to (2), T intersects F from below. Because both functions are continuous they can intersect at most once.

To see that uniqueness fails if the condition is not satisfied, suppose there exists $\alpha_0 \in (0,1)$ such that $F'(\alpha_0) > \frac{F(\alpha_0)}{\alpha_0-\alpha_0^2}$. For any $(x,y) \in \text{PD}$ such that $y > x$, T is continuous, $\lim_{\alpha\to 0} T(\alpha|x,y) = \frac{x}{x-y} < 0$, and $\lim_{\alpha\to 1} T(\alpha|x,y) = \infty$. Hence, there must exist at least one solution in which T intersects F from below. Therefore, if for the same game there exists a solution in which T intersects F from above, then there are multiple solutions. Let $Y(x|\alpha, F(\alpha))$ be the function such that α solves (2) given $F(\alpha)$, x, and $y = Y(x|\alpha, F(\alpha))$; that is,

$$Y(x|\alpha, F(\alpha)) = \frac{\alpha - (1-\alpha)(1-F(\alpha))x}{(1-\alpha)F(\alpha)}.$$

Notice that given any $(\alpha, F(\alpha)) \in (0,1)^2$, for all $x < \frac{\alpha}{(1-\alpha)(1-F(\alpha))}$, $Y(x|\alpha, F(\alpha)) > 0$, meaning for such x, $(x, Y(x|\alpha, F(\alpha))) \in \text{PD}$. Finally, it is straightforward that

$$\lim_{x\to 0}\big(T'(\alpha|x, Y(x|\alpha_0, F(\alpha_0)))|_{\alpha=\alpha_0}\big) = \frac{F(\alpha_0)}{\alpha_0 - \alpha_0^2}.$$

By supposition, $F'(\alpha_0) > \frac{F(\alpha_0)}{\alpha_0-\alpha_0^2}$. Therefore, because T' is continuous in both x and y, there exists $x > 0$ small enough such that T intersects F from above at α_0 for the game $(x, Y(x|\alpha_0, F(\alpha_0)))$. □

Proof of Theorem 1. *Representation* $\Longrightarrow$ *Axioms*. Consider a collection I with primitive $(D_i^0, C_i^0)_{i\in I}$ that satisfies the representation. Because each game has a unique equilibrium cutoff, Lemma 1 immediately implies Axiom 1 is satisfied. To verify that the primitive satisfies the remaining axioms it is sufficient to focus only on PD and $(D_i, C_i)_{i\in I}$.

Propositions 1 and 2 immediately imply the following. First, if $\alpha_i = 0$, then $D_i = \text{PD}$, and if $\alpha_i = 1$, then $C_i = \text{PD}$. Second, if $\alpha_i \in (0,1)$, then $M_i = \{(x,y) \in \text{PD} | \alpha^*_{x,y} = \alpha_i\}$. Third, if $\alpha_i = \alpha_j$, then $(D_i, C_i) = (D_j, C_j)$.

Now fix arbitrary $\alpha_i \in (0,1)$ and solve (2) to get that

$$M_i = \{(x,y) \in \text{PD} | \alpha^*_{x,y} = \alpha_i\} = \left\{(x,y) \in \text{PD} \middle| y = \frac{\alpha_i}{(1-\alpha_i)F(\alpha_i)} - x\left(\frac{1-F(\alpha_i)}{F(\alpha_i)}\right)\right\}.$$

That is, M_i forms a line in PD. Define $\text{int}_i = \frac{\alpha_i}{(1-\alpha_i)F(\alpha_i)}$ and $\text{slp}_i = \frac{1-F(\alpha_i)}{F(\alpha_i)}$. It follows that if $0 < \alpha_j < \alpha_i < 1$, then $\text{int}_i \geq \text{int}_j$ and $\text{slp}_i < \text{slp}_j$. The latter is obvious since $\alpha_i > \alpha_j \implies F(\alpha_i) > F(\alpha_j)$ because $F \in \mathcal{F}$. To see the former,

$$\frac{d}{d\alpha}\left(\frac{\alpha}{(1-\alpha)F(\alpha)}\right) = \frac{F(\alpha) - \big(\alpha - \alpha^2\big)F'(\alpha)}{\big((1-\alpha)F(\alpha)\big)^2} \geq 0 \quad \forall \alpha \in (0,1) \qquad \Longleftrightarrow \qquad \text{Condition S.}$$

For arbitrary player $\alpha_i \in (0,1)$, let MU_i and ML_i be the strict-upper- and strict-lower-contour sets of M_i (within PD), respectively. Now, consider $(x,y) \in \text{MU}_i$. From Proposition 2, there exists unique $\alpha^*_{x,y}$, and it is distinct from α_i by $(x,y) \notin M_i$. From the argument above, whenever $\alpha_j \leq \alpha_i$ then $\text{ML}_j \subseteq \text{ML}_i$. Therefore, $\alpha^*_{x,y} > \alpha_i$. By the cutoff form of the equilibrium, $(x,y) \in D_i$. Therefore, $D_i = \text{MU}_i$. An analogous argument establishes $C_i = \text{ML}_i$.

Having completed the description of the data, $(D_i, C_i)_{i\in I}$, that the representation generates, we are ready to verify the axioms. That extreme players, $\alpha_i = 0, 1$, satisfy Axioms 2–4 is clear, so consider any player i such that $\alpha_i \in (0,1)$. Axiom 2 is satisfied since the sets $C_i = \{(x,y) \in \text{PD} | y < \text{int}_i - x \cdot \text{slp}_i\}$ and $D_i = \{(x,y) \in \text{PD} | y > \text{int}_i - x \cdot \text{slp}_i\}$ are open in PD. For Axiom 3, if $(x,y) \in \overline{D}_i$, then for any $(x',y') \geq (x,y)$ such that $(x',y') \neq (x,y)$ it follows that $(x',y') \in \text{MU}_i = D_i$. To verify Axiom 4, suppose both (x,y) and (x',y') are elements of D_i. Then

$$\begin{aligned} y > \text{int}_i - x \cdot \text{slp}_i \quad &\implies \quad \gamma y > \gamma(\text{int}_i - x \cdot \text{slp}_i), \\ y' > \text{int}_i - x' \cdot \text{slp}_i \quad &\implies \quad (1-\gamma)y' > (1-\gamma)\big(\text{int}_i - x' \cdot \text{slp}_i\big) \\ &\implies \quad \gamma y + (1-\gamma)y' > \gamma(\text{int}_i - x \cdot \text{slp}_i) + (1-\gamma)\big(\text{int}_i - x' \cdot \text{slp}_i\big) \\ &\implies \quad \gamma y + (1-\gamma)y' > \text{int}_i - \big(\gamma x + (1-\gamma)x'\big)\text{slp}_i. \end{aligned}$$

Hence, $(\gamma x + (1-\gamma)x', \gamma y + (1-\gamma)y') \in D_i$. A symmetric argument holds if $\{(x,y), (x',y')\} \subset C_i$.

Finally, Axiom 5. Suppose the hypotheses of the axiom are satisfied for two distinct players i and j. Then it must be that $0 < \alpha_i < \alpha_j$. If $\alpha_j = 1$, then $C_j = \text{PD}$ and the axiom is trivial. If $\alpha_j < 1$, then $0 < \alpha_i < \alpha_j$ implies $\text{slp}_j < \text{slp}_i$ (above). Further, by hypotheses (ii) and (iii), $\text{slp}_i \leq \frac{\delta}{\varepsilon}$. Finally, $\text{slp}_j < \frac{\delta}{\varepsilon}$ and $(x',y') \in \overline{C}_j$ imply $(x'+\varepsilon, y'-\delta) \in C_j$, completing the proof.

Axioms $\implies$ *Representation.* The majority of the proof concerns behavior in the set of games PD (that is $(D_i, C_i)_{i\in I}$). In a series of lemmas we establish that Axioms 2–5 imply the representation on this smaller domain. Lemmas A.1 and A.2 demonstrate that if (D_i, C_i) satisfies Axioms 2–4, then there is a unique value for α_i and a unique scalar F_i such that any behavioral model $[F, (\alpha_i, \alpha_{-i})]$ with $F \in \mathcal{F}$ and $F(\alpha_i) = F_i$ can explain the behavior of player i. Lemmas A.3–A.5 then show that there exists $F \in \mathcal{F}_S$ that simultaneously satisfies the required values for all $i \in I$. Therefore, by Proposition 2, for all $(x,y) \in \text{PD}$, under this F there is a unique equilibrium cutoff. This ensures that in each game there is an equilibrium consistent with the behavior of all players; hence, the behavioral model using this assignment of F and the mandated α_i-values can explain

$(D_i, C_i)_{i \in I}$ (Lemma A.6). It is then an immediate corollary that the addition of Axiom 1 implies the representation on the full domain, PD^0 (Lemma A.7). This completes the proof. □

Fact A.1. *Fix any player i. If (D_i, C_i) satisfies Axiom 2, and $C_i \neq \varnothing$ and $D_i \neq \varnothing$, then for any $(x, y) \in C_i$, $(x', y') \in D_i$, and continuous path $p : [0, 1] \to \text{PD}$ such that $p(0) = (x, y)$ and $p(1) = (x', y')$, there exists $t \in (0, 1)$ such that $p(t) \in M_i$.*

Proof. Let $\bar{t} = \sup\{t | p(t') \in C_i \ \forall t' \in [0, t]\}$. Because (x, y) is an arbitrary element of C_i, it is sufficient to show that $p(\bar{t}) \in M_i$. Suppose that $p(\bar{t}) \in C_i$. Then, by definition of $\bar{t}$, for any $\varepsilon > 0$ there exists $t \in (\bar{t}, \bar{t} + \varepsilon)$ such that $p(t) \notin C_i$. Because p is continuous, this contradicts C_i being open (and, hence, Axiom 2). Now, suppose that $p(\bar{t}) \in D_i$. By definition of $\bar{t}$, for all $\varepsilon > 0$ there exists $t \in (\bar{t} - \varepsilon, \bar{t})$ such that $p(t) \in C_i$, and therefore $p(t) \notin D_i$. Because p is continuous, this contradicts D_i being open (and, hence, Axiom 2). Hence, $p(\bar{t}) \in M_i$. □

Lemma A.1. *Fix any player i such that (D_i, C_i) satisfies Axioms 2–4. If $D_i \neq \varnothing$ and $C_i \neq \varnothing$, then there is a unique pair $(\text{int}_i, \text{slp}_i) \in (0, \infty)^2$ such that $D_i = \{(x, y) \in \text{PD} | y > \text{int}_i - \text{slp}_i \cdot x\}$ and $C_i = \{(x, y) \in \text{PD} | y < \text{int}_i - \text{slp}_i \cdot x\}$. If $D_i = \varnothing$, then $C_i = \text{PD}$, and if $C_i = \varnothing$, then $D_i = \text{PD}$.*

Proof. Consider the three possible cases.

Case 1: $D_i \neq \varnothing$ and $C_i \neq \varnothing$. Axiom 2 implies that not only is M_i nonempty, but is nonsingleton (see Fact A.1). Therefore, let $\{(x_1, y_1) \neq (x_2, y_2)\} \subset M_i$, with $x_1 \leq x_2$. By Axiom 3, $x_1 < x_2$ and $y_1 > y_2$. Again employing Axioms 2 and 3, we see that $M_i \cap \{(x, y) \in \text{PD} | x \in [x_1, x_2]\}$ must consist of a strictly decreasing function $\bar{y}$, where $\bar{y}(x_1) = y_1$ and $\bar{y}(x_2) = y_2$. For any $x \in [x_1, x_2]$, if $y > \bar{y}(x)$, then $(x, y) \in D_i$, and if $y \in (0, \bar{y}(x))$, then $(x, y) \in C_i$. Hence, Axiom 4 implies that $\bar{y}$ is linear. Let $\text{slp}_i := \frac{y_1 - y_2}{x_2 - x_1} \in (0, \infty)$ and $\text{int}_i := (y_1 + \text{slp}_i \cdot x_1) \in (0, \infty)$.

Since the above applies to any pair of games in M_i, all games in M_i must fall on the same line: $M_i \subseteq \{(x, y) \in \text{PD} | y = \text{int}_i - \text{slp}_i \cdot x\}$. But, if the inclusion were strict, Axiom 2 would be violated (again, see Fact A.1). Hence, $M_i = \{(x, y) \in \text{PD} | y = \text{int}_i - \text{slp}_i \cdot x\}$. The claimed structures of D_i and C_i follow from Axiom 3.

Case 2: $D_i = \varnothing$. It must be that $M_i = \varnothing$. Suppose to the contrary that some $(x, y) \in M_i$. Then, by Axiom 3, for $x' > x$, $(x', y) \in D_i$: a contradiction. Hence, $C_i = \text{PD}$.

Case 3: $C_i = \varnothing$. It must be that $M_i = \varnothing$. Suppose to the contrary that some $(x, y) \in M_i$. Consider then a game (x', y) where $x' \in (0, x)$. By $C_i = \varnothing$, $(x', y) \in \overline{D}_i$. Axiom 3 implies that $(x, y) \in D_i$: a contradiction. Hence, $D_i = \text{PD}$. □

Lemma A.2. *Fix any player i. If (D_i, C_i) satisfies Axioms 2–4, then there exists a unique pair $(\alpha_i, F_i) \in [0, 1]^2$ such that (D_i, C_i) can be explained by any behavioral model $[F, (\alpha_i, \alpha_{-i})]$ such that $F \in \mathcal{F}$ and $F(\alpha_i) = F_i$. Further, (α_i, F_i) is given by*

$$(\alpha_i, F_i) = \begin{cases} \left(\dfrac{\text{int}_i}{1 + \text{int}_i + \text{slp}_i}, \dfrac{1}{1 + \text{slp}_i} \right) & \text{if } D_i, C_i \neq \varnothing, \\ (1, 1) & \text{if } D_i = \varnothing, \\ (0, 0) & \text{if } C_i = \varnothing. \end{cases} \tag{10}$$

Proof. First $\text{int}_i, \text{slp}_i > 0$ (Lemma A.1) implies that (α_i, F_i) from (10) is always in $[0,1]^2$. Consider, again, the three possible cases.

Case 1: $D_i \neq \varnothing$ and $C_i \neq \varnothing$. Recall from Proposition 1, that in the behavioral model, for any $(x,y) \in \text{PD}$, each equilibrium is of cutoff form, with $\alpha^*_{x,y}$ being any solution to (2). So, it is sufficient to show that for arbitrary $(x,y) \in \text{PD}$ and $F \in \mathcal{F}$ such that $F(\alpha_i) = F_i$, (i) $(x,y) \in M_i$ if and only if α_i solves (2), (ii) $(x,y) \in C_i$ implies that there exists $\alpha \in (0, \alpha_i)$ such that α solves (2), and (iii) $(x,y) \in D_i$ implies that there exists $\alpha \in (\alpha_i, 1)$ such that α solves (2). We take them in turn.

(i) By Lemma A.1, $(x,y) \in M_i \iff y = \text{int}_i - \text{slp}_i \cdot x > 0$. Solving (2) for y gives $y = \frac{\alpha_i}{F(\alpha_i)-\alpha_i F(\alpha_i)} - \frac{1-F(\alpha_i)}{F(\alpha_i)}x$. The pair of equations $\text{int}_i = \frac{\alpha_i}{F(\alpha_i)-\alpha_i F(\alpha_i)}$ and $\text{slp}_i = \frac{1-F(\alpha_i)}{F(\alpha_i)}$ has a unique solution: $\alpha_i = \frac{\text{int}_i}{1+\text{int}_i+\text{slp}_i}$ and $F(\alpha_i) = \frac{1}{1+\text{slp}_i}$. This establishes the claim.

(ii) Suppose that $(x,y) \in C_i$. By Lemma A.1, this implies that $y < \text{int}_i - \text{slp}_i \cdot x$. Let $d(\alpha) := V(c|\alpha = \alpha^*) - V(d|\alpha = \alpha^*) = \alpha[1 + (1 - F(\alpha))x + F(\alpha)y] - [(1-F(\alpha))x + F(\alpha)y]$. Using the assignments of $(\alpha_i, F(\alpha_i) = F_i)$ from (10), it follows that $d(\alpha_i) > 0$. Notice that $d(0) = x(F(0) - 1) - yF(0) = -x < 0$. Because $F \in \mathcal{F}$, d must be continuous on $[0, \alpha_i]$. Hence, there exists $\alpha \in (0, \alpha_i)$ that achieves $d(\alpha) = 0$ and is therefore an equilibrium cutoff in the game $(x,y) \in C_i$ (by Proposition 1).

(iii) Suppose that $(x,y) \in D_i$. By Lemma A.1, this implies that $y > \text{int}_i - \text{slp}_i \cdot x$. Using the assignments of $(\alpha_i, F(\alpha_i) = F_i)$ from (10), it follows that $d(\alpha_i) < 0$. Notice that $d(1) = 1$. Because $F \in \mathcal{F}$, d must be continuous on $[\alpha_i, 1]$. Hence, there exists $\alpha \in (\alpha_i, 1)$ that achieves $d(\alpha_i) = 0$ and is therefore an equilibrium cutoff in the game $(x,y) \in D_i$ (by Proposition 1).

The following text is relevant for the next two cases. In the behavioral model, for any $F \in \mathcal{F}$, Proposition 1 establishes that a player with type $\alpha_i = 1$ strictly prefers c in all equilibria of all games, and that a player with type $\alpha_i = 0$ strictly prefers d in all equilibria of all games. Further, for any $F \in \mathcal{F}$ and any $\alpha \in (0,1)$, the game $(x,y) = (\frac{\alpha}{1-\alpha}, \frac{\alpha}{1-\alpha})$ is in PD and has a unique equilibrium cutoff $\alpha^*_{x,y} = \alpha$.

Case 2: $D_i = \varnothing$. From above, in the behavioral model, for any $F \in \mathcal{F}$, a player i strictly prefers c in every $(x,y) \in \text{PD}$ if and only if his type is $\alpha_i = 1$. Further, for all $F \in \mathcal{F}$, $F(1) = 1$.

Case 3: $C_i = \varnothing$: From above, in the behavioral model, for any $F \in \mathcal{F}$, a player i strictly prefers d in every $(x,y) \in \text{PD}$ if and only if his type is $\alpha_i = 0$. Further, for all $F \in \mathcal{F}$, $F(0) = 0$. □

Lemma A.3. *Fix any two players i and j such that (D_i, C_i) and (D_j, C_j) satisfy Axioms 2–5 and D_i, C_i, D_j, and C_j are all nonempty. Using* $(\text{int}_i, \text{slp}_i)$, $(\text{int}_j, \text{slp}_j)$ *from Lemma A.1, if* $\text{int}_i < \text{int}_j$, *then* $\text{slp}_i > \text{slp}_j$.

Proof. By Lemma A.1, $M_i = \{(x,y) \in \text{PD} | y = \text{int}_i - \text{slp}_i \cdot x\}$ and $M_j = \{(x,y) \in \text{PD} | y = \text{int}_j - \text{slp}_j \cdot x\}$. Fix any $(x,y) \in M_i$, and for $\varepsilon \in (0, \frac{y}{\text{slp}_i})$, let $\delta = \varepsilon \cdot \text{slp}_i$. It follows that $(x+\varepsilon, y-\delta) \in M_i$. Next, $\text{int}_i < \text{int}_j$ implies that for sufficiently small choices of x and ε there exists (x', y') such that $(x', y') \in M_j$ and $\{(x, y), (x+\varepsilon, y-\delta)\} < \{(x', y'), (x'+\varepsilon, y'-\delta)\}$. By Axiom 5, $(x'+\varepsilon, y'-\delta) \in C_j = \{(x, y) \in \text{PD} | y < \text{int}_j - \text{slp}_j \cdot x\}$. Thus, $\text{slp}_j < \frac{\delta}{\varepsilon} = \text{slp}_i$. □

Definition A.1. Fix any player i such that (D_i, C_i) satisfies Axioms 2–4, $D_i \neq \varnothing$, and $C_i \neq \varnothing$. Assign (α_i, F_i) as done by (10) in Lemma A.2. Define the function $H_i : [0, 1] \to \mathbb{R} \cup \infty$ as $H_i(a) = F_i \frac{a(1-\alpha_i)}{\alpha_i(1-a)}$ for $a \in [0, 1)$ and $H_i(1) = \infty$.

Fact A.2. *For all i such that H_i is defined, (i) $H_i(0) = 0$, (ii) H_i is strictly increasing, differentiable, and strictly convex on $[0, 1)$, (iii) $H_i(\alpha_i) = F_i$, (iv) $\lim_{a\to 1} H_i(a) = \infty$, and (v) $H_i'(\alpha_i) = \frac{F_i}{\alpha_i - \alpha_i^2}$.*

The proof is by direct calculations.

Lemma A.4. *Fix any two players i and j such (D_i, C_i) and (D_j, C_j) satisfy Axioms 2–5. Assign (α_i, F_i) and (α_j, F_j) as done by (10) in Lemma A.2. The following statements are valid:*

(i) If $\alpha_j < \alpha_i < 1$, then $F_j \in [H_i(\alpha_j), F_i)$.

(ii) If $0 < \alpha_i < \alpha_j$, then $F_j \in (F_i, H_i(\alpha_j)]$.

(iii) If $\alpha_i = \alpha_j$, then $F_j = F_i$.

Proof. First, if $\alpha_i \in \{0, 1\}$, then (i) and (ii) have no implications, and (iii) is immediate from Lemma A.2. Now fix $\alpha_i \in (0, 1)$. If $\alpha_j \in \{0, 1\}$, then the claims follow from Fact A.2. If $\alpha_j \in (0, 1)$, then from Lemma A.3,

$$\begin{aligned}(\text{int}_j, \text{slp}_j) \in {} & \{(\text{int}, \text{slp}) | \text{int} \leq \text{int}_i \text{ and slp} > \text{slp}_i\} \cup \{(\text{int}, \text{slp}) | \text{int} \geq \text{int}_i \text{ and slp} < \text{slp}_i\} \\ & \cup \{(\text{int}, \text{slp}) | \text{int} = \text{int}_i \text{ and slp} = \text{slp}_i\}.\end{aligned}$$

Inverting the bijection from (10) in Lemma A.2,

$$\begin{aligned}(\alpha_j, F_j) \in {} & \left\{(\alpha, \phi) \middle| \frac{\alpha}{(1-\alpha)\phi} \leq \frac{\alpha_i}{(1-\alpha_i)F_i} \text{ and } \frac{1-\phi}{\phi} > \frac{1-F_i}{F_i}\right\} \\ & \cup \left\{(\alpha, \phi) \middle| \frac{\alpha}{(1-\alpha)\phi} \geq \frac{\alpha_i}{(1-\alpha_i)F_i} \text{ and } \frac{1-\phi}{\phi} < \frac{1-F_i}{F_i}\right\} \\ & \cup \left\{(\alpha, \phi) \middle| \frac{\alpha}{(1-\alpha)\phi} = \frac{\alpha_i}{(1-\alpha_i)F_i} \text{ and } \frac{1-\phi}{\phi} = \frac{1-F_i}{F_i}\right\}.\end{aligned}$$

Rearranging and using Definition A.1 gives

$$\begin{aligned}(\alpha_j, F_j) \in {} & \{(\alpha, \phi) | H_i(\alpha) \leq \phi \text{ and } \phi < F_i\} \cup \{(\alpha, \phi) | H_i(\alpha) \geq \phi \text{ and } \phi > F_i\} \\ & \cup \{(\alpha, \phi) | H_i(\alpha) = \phi \text{ and } \phi = F_i\},\end{aligned}$$

which, by (ii) and (iii) of Fact A.2, is equivalent to

$$\begin{aligned}(\alpha_j, F_j) \in &\{(\alpha, \phi) | \alpha \in [0, \alpha_i) \text{ and } \phi \in [H_i(\alpha), F_i)\}\\ &\cup \{(\alpha, \phi) | \alpha \in (\alpha_i, 1] \text{ and } \phi \in (F_i, H_i(\alpha)]\}\\ &\cup \{(\alpha, \phi) | \alpha = \alpha_i \text{ and } \phi = F_i\}.\end{aligned}$$

This establishes the result. □

Corollary A.1. *Fix any two players i and j such (D_i, C_i) and (D_j, C_j) satisfy Axioms 2–5. Assigning (α_i, F_i), (α_j, F_j) as done by (10), if $0 < \alpha_i < \alpha_j < 1$, then either $H_i = H_j$ or $H_i'(a) > H_j'(a)$ for all $a \in [0, 1)$.*

Proof. We have

$$H_i'(a) - H_j'(a) = \frac{F_i \cdot \alpha_j(1 - \alpha_i) - F_j \cdot \alpha_i(1 - \alpha_j)}{\alpha_i \cdot \alpha_j(1 - a)^2}. \tag{11}$$

By Lemma A.4, either (11) is zero for all $a \in [0, 1)$, in which case $H_i = H_j$ since $H_i(0) = H_j(0)$ from Fact A.2, or (11) is positive for all $a \in [0, 1)$. □

Lemma A.5. *Fix a primitive $(D_i, C_i)_{i\in I}$ that satisfies Axioms 2–5. Assign $(\alpha_i, F_i)_{i\in I}$ as done by (10) in Lemma A.2. There exists $F \in \mathcal{F}_S$ with $F(\alpha_i) = F_i$ for all $i \in I$.*

Proof. Order and (re-)index the distinct pairs featuring $\alpha_i \in (0, 1)$ such that $(0, 0) \ll (\alpha_1, F_1) \ll (\alpha_2, F_2) \ll \cdots \ll (\alpha_m, F_m) \ll (1, 1)$, where $m \leq n$ and the ordering is strict by Lemma A.4. For each $k \in \{1, 2, \ldots, m\}$, set $F(\alpha_k) = F_k$ and $F'(\alpha_k) = \frac{F_k}{\alpha_k - \alpha_k^2}$. Set $F(0) = 0$, $F(1) = 1$, and $F'(1) = 0$. Next, in F we fill in the intervals between the pairs to produce a strictly increasing, differentiable CDF that satisfies Condition S. In doing so, we say that a differentiable function f_1 *smoothly pastes* the ordered pair of differentiable, increasing functions (f_2, f_3) on an interval $(\underline{z}, \overline{z})$ if (i) $f_1(\underline{z}) = f_2(\underline{z})$, (ii) $f_1'(\underline{z}) = f_2'(\underline{z})$, (iii) $f_1(\overline{z}) = f_3(\overline{z})$, and (iv) $f_1'(\overline{z}) = f_3'(\overline{z})$.

Step 1. On $(0, \alpha_1)$, set $F = H_1$, which satisfies all of the necessary properties (see Fact A.2).

Step 2. Identify all $k \in \{1, 2, \ldots, m-1\}$, such that $H_k = H_{k+1}$. For all such k, set $F = H_k$ on (α_k, α_{k+1}), which satisfies all of the necessary properties (see Fact A.2).

Step 3. Fix arbitrary $k < m$ such that $H_k \neq H_{k+1}$, and let L be the linear function tangent to H_k at α_k. There are two cases: (a) $L(\alpha_{k+1}) < F_{k+1}$ or (b) $L(\alpha_{k+1}) \geq F_{k+1}$.

(a) In this case, L intersects H_{k+1} at some $\alpha^0 \in (\alpha_k, \alpha_{k+1})$, where $L' < H_{k+1}'(\alpha^0)$. Now for any $\varepsilon > 0$ small enough, there exists an elliptical arc E that smoothly pastes (L, H_{k+1}) on $(\alpha^0 - \varepsilon, \alpha^0 + \varepsilon)$. By construction, for sufficiently small ε, setting $F = L$ on $(\alpha_k, \alpha^0 - \varepsilon]$, $F = E$ on $(\alpha^0 - \varepsilon, \alpha^0 + \varepsilon)$, and $F = H_{k+1}$ on $[\alpha^0 + \varepsilon, \alpha_{k+1})$, satisfies differentiability and strict monotonicity. To see that it satisfies Condition S, let $\tilde{H}_{\alpha,F}$ be the H_i function of a hypothetical player with $(\alpha_i, F_i) = (\alpha, F)$. By (v) of Fact A.2, it is sufficient to demonstrate that $F'(\alpha) \leq \tilde{H}'_{\alpha,F(\alpha)}(\alpha)$ for all $\alpha \in (\alpha_k, \alpha_{k+1})$. Notice that Corollary A.1 implies that this holds with equality on $[\alpha^0 + \varepsilon, \alpha_{k+1})$. For $\alpha \in (\alpha_k, \alpha^0 - \varepsilon]$, F is (weakly) concave and crosses the strictly convex function $H_{\alpha,F(\alpha)}$ from above at α. Hence, the inequality must be satisfied. Finally, if ε is small enough, then since $E > H_{k+1}$ on $(\alpha^0 - \varepsilon, \alpha^0 + \varepsilon)$, so as to smoothly paste with H_{k+1}, it must be that $E' < H'_{k+1}$ on this interval. From Corollary A.1, $\tilde{H}'_{\alpha,F(\alpha)} > H'_{k+1}$, which establishes the inequality.

(b) In this case, let $\hat{L}$ be the line that passes through (α_k, F_k) and (α_{k+1}, F_{k+1}), so $\hat{L}' \leq L'$ by hypothesis. Next, let $\hat{L}_\delta$ be the line that passes through the midpoint between (α_k, F_k) and (α_{k+1}, F_{k+1}) with slope $\hat{L}' - \delta$. For any $\delta > 0$ small enough, there exists $\varepsilon > 0$ small enough such that $(H_k, \hat{L}_\delta)$ can be smoothly pasted by elliptical arc E_1 on $(\alpha_k, \alpha_k + \varepsilon)$, and $(\hat{L}_\delta, H_{k+1})$ can be smoothly pasted by elliptical arc E_2 on $(\alpha_k - \varepsilon, \alpha_{k+1})$. By construction, for sufficiently small δ and ε, setting $F = E_1$ on $(\alpha_k, \alpha_k + \varepsilon)$, $F = \hat{L}_\delta$ on $[\alpha_k + \varepsilon, \alpha_{k+1} - \varepsilon]$, and $F = E_2$ on $(\alpha_{k+1} - \varepsilon, \alpha_{k+1})$ satisfies differentiability and strict monotonicity. The arguments that it satisfies Condition S are analogous to those made in Step 3(a) above, since F is weakly concave on $(\alpha_k, \alpha_{k+1} - \varepsilon)$, and $E_2' < H'_{k+1}$ on $(\alpha_{k+1} - \varepsilon, \alpha_{k+1})$ when ε is sufficiently small.

Step 4. For $\alpha \in (\alpha_m, 1)$, given the properties of H_m from Fact A.2, there exist $\alpha^0 \in [\alpha_m, 1)$ and an elliptical arc E that smoothly pastes $(H_m, 1)$ on $(\alpha^0, 1)$ and is concave. Set $F = H_m$ on $(\alpha_m, \alpha^0]$ and $F = E$ on $(\alpha_0, 1)$ to satisfy differentiability, strict monotonicity, and Condition S (by the same arguments from Step 3). □

Lemma A.6. *If $(D_i, C_i)_{i\in I}$ satisfies Axioms 2–5, then it can be explained by a behavioral model $[F, (\alpha_i)_{i\in I}]$, where $F \in \mathcal{F}$ satisfies Condition S. Furthermore, $(\alpha_i, F(\alpha_i))_{i\in I}$ is unique.*

The proof is an immediate corollary of Lemmas A.1–A.5.

Lemma A.7. *If $(D_i^0, C_i^0)_{i\in I}$ satisfies Axioms 1–5, then it can be explained by a behavioral model $[F, (\alpha_i)_{i\in I}]$, where $F \in \mathcal{F}$ satisfies Condition S. Furthermore, $(\alpha_i, F(\alpha_i))_{i\in I}$ is unique.*

The proof is an immediate corollary of Lemmas 1 and A.6.

Proof of Proposition 3. The proof is ordered: (a) $\Longleftrightarrow$ (c), (b) $\Longleftrightarrow$ (c), (a) $\Longleftrightarrow$ (e), (d) $\Longleftrightarrow$ (e).

(a) $\Longleftrightarrow$ *(c)*. Suppose F f.o.s.d. $\widetilde{F}$. Recall that if $x = y$, then $\alpha^*_{x,y} = \widetilde{\alpha}^*_{x,y}$ (Section 2.1.1), implying $F(\alpha^*_{x,y}) \leq \widetilde{F}(\widetilde{\alpha}^*_{x,y})$ by f.o.s.d. If $x \neq y$, then the cutoffs are implicitly characterized by $F(\alpha^*_{x,y}) = T(\alpha^*_{x,y}|x,y)$ and $\widetilde{F}(\widetilde{\alpha}^*_{x,y}) = T(\widetilde{\alpha}^*_{x,y}|x,y)$. Further, if $x > y$, then $T(0|x,y) > 0 = F(0) = \widetilde{F}(0)$, and $T(\cdot|x,y)$ is strictly decreasing. Hence, by f.o.s.d., both $\widetilde{\alpha}^*_{x,y} \leq \alpha^*_{x,y}$ and $\widetilde{F}(\widetilde{\alpha}^*_{x,y}) \geq F(\alpha^*_{x,y})$. If $x < y$, then $T(0|x,y) < 0 = F(0) = \widetilde{F}(0)$, and T is strictly increasing (but intersecting F and $\widetilde{F}$ each exactly once since both satisfy Condition S). Hence, by f.o.s.d., both $\widetilde{\alpha}^*_{x,y} \geq \alpha^*_{x,y}$ and $\widetilde{F}(\widetilde{\alpha}^*_{x,y}) \geq F(\alpha^*_{x,y})$. So (a) implies (c).

Now suppose that (a) does not hold; there exists $\alpha^0 \in (0,1)$ such that $F(\alpha^0) > \widetilde{F}(\alpha^0)$. In the game $x = y = \frac{\alpha^0}{1-\alpha^0}$, we have that $\alpha^*_{x,y} = \widetilde{\alpha}^*_{x,y} = \alpha^0$, which then violates (c).

(b) $\Longleftrightarrow$ *(c)*. Notice that $k_{x,y}$ and $\widetilde{k}_{x,y}$ are binomial random variables with n "trials" (i.e., each players' action) and probabilities of "success" (i.e., cooperation) of $(1 - F(\alpha^*_{x,y}))$ and $(1 - \widetilde{F}(\widetilde{\alpha}^*_{x,y}))$, respectively. Because n is common between the two random variables, a simple "coupling" argument Lindvall (2002, Chapter 1) establishes that $k_{x,y}$ f.o.s.d. $\widetilde{k}_{x,y}$ if and only if $1 - F(\alpha^*_{x,y}) \geq 1 - \widetilde{F}(\widetilde{\alpha}^*_{x,y})$.

(a) $\Longleftrightarrow$ *(e)*. That (a) implies (e) is shown in the proof of (a) $\Longleftrightarrow$ (c) above. Now suppose that (a) does not hold; there exists $\alpha^0 \in (0,1)$ such that $F(\alpha^0) > \widetilde{F}(\alpha^0)$. In the game $x = y = \frac{\alpha^0}{1-\alpha^0}$, we have that $\alpha^*_{x,y} = \widetilde{\alpha}^*_{x,y} = \alpha^0$. Because F, $\widetilde{F}$, and T are all continuous in y, and T is decreasing in α when $x > y$, there exits an $\varepsilon > 0$ small enough such that in the game $(x,y) = (\frac{\alpha^0}{1-\alpha^0}, \frac{\alpha^0}{1-\alpha^0} - \varepsilon) \in \mathrm{PD}$, $\alpha^*_{x,y} < \widetilde{\alpha}^*_{x,y}$, which violates (e).

(d) $\Longleftrightarrow$ *(e)*. That (e) implies (d) follows from the cutoff nature of equilibrium behavior (Propositions 1 and 2). Now suppose that (e) does not hold in that there exists $x_0 \leq y_0$ such that $\alpha^*_{x_0,y_0} > \widetilde{\alpha}^*_{x_0,y_0}$. Let $\alpha^0 \in (\widetilde{\alpha}^*_{x_0,y_0}, \alpha^*_{x_0,y_0})$. Then

$$\{(x_0, y_0)\} \subset C_{\alpha^0,\widetilde{F}} \cap D_{\alpha^0,F} \cap \{(x,y)|x \leq y\} \neq \varnothing,$$

violating (d). A symmetric argument applies if there exists $x_0 \geq y_0$ such that $\alpha^*_{x_0,y_0} < \widetilde{\alpha}^*_{x_0,y_0}$. □

Proof of Proposition 4. When $x = y$, the unique solution to (2) is $\alpha^* = \frac{x}{1+x}$. Hence, Proposition 1 implies that, for any $(x,x) \in \mathrm{PD}$ and any player i (of any collection), $(x,x) \in D_i$ if and only if $\alpha_i < \frac{x}{1+x}$. Part (a) of the proposition follows.

For part (b), first suppose that I is more influenced by x relative to y than is $\widetilde{I}$. Then the behaviors of the two collections agree on the subset of games $\{(x,y) \in \mathrm{PD}|x = y\}$. By part (a) of the proposition, $(\alpha_i)_{i \in I} = (\widetilde{\alpha}_i)_{i \in \widetilde{I}}$. Immediately, if $\alpha_i = 0, 1$, then $F(\alpha_i) = \widetilde{F}(\widetilde{\alpha}_i)$. In addition, I defecting more in $\{(x,y) \in \mathrm{PD}|x \geq y\}$ than does $\widetilde{I}$ implies that, for each i with $\alpha_i \in (0,1)$, it must be that the M_i line is weakly steeper under F than under $\widetilde{F}$, i.e., $\mathrm{slp}_i \geq \widetilde{\mathrm{slp}}_i$ (Lemma A.1). Next, by Lemma A.2, $F(\alpha_i) = \frac{1}{1+\mathrm{slp}_i} \leq \frac{1}{1+\widetilde{\mathrm{slp}}_i} = \widetilde{F}(\widetilde{\alpha}_i)$.

Second, suppose that for all $i \leq n$, $\alpha_i = \widetilde{\alpha}_i$ and $F(\alpha_i) \leq \widetilde{F}(\widetilde{\alpha}_i)$. Immediately, if $\alpha_i = 0, 1$, then i's behavior is the same in I and $\widetilde{I}$. For each i with $\alpha_i \in (0,1)$, $M_i \cap \widetilde{M}_i = \{(\frac{\alpha_i}{1-\alpha_i}, \frac{\alpha_i}{1-\alpha_i})\}$. Also, $\mathrm{slp}_i = \frac{1-F(\alpha_i)}{F(\alpha_i)} \geq \frac{1-\widetilde{F}(\widetilde{\alpha}_i)}{\widetilde{F}(\widetilde{\alpha}_i)} = \widetilde{\mathrm{slp}}_i$. Hence, for any $(x,y) \in \mathrm{PD}$, if $x \geq y$, then $(x,y) \in \widetilde{D}_i \implies (x,y) \in D_i$, and if $x \leq y$, then $(x,y) \in D_i \implies (x,y) \in \widetilde{D}_i$, establishing that I is more influenced by x relative to y than is $\widetilde{I}$. □

Proof of Proposition 5. First, the proof of Proposition 1, Claim (i) remains valid and implies that all equilibria are cutoff with $\alpha^* < 1$ and that $\alpha^* = 0$ is an equilibrium cutoff if and only if $x \leq 0$. Second, if $x > 0$, the existence, uniqueness, and characterization of the equilibrium cutoff follow identically from the proofs of Propositions 1 and 2. Third, any $\alpha \in (0, 1)$ is an equilibrium cutoff if and only if $g \in \tilde{M}_\alpha$ (the argument given in the proof of Theorem 1, “Representation $\Longrightarrow$ Axioms,” immediately extends to establish this). Fix now $x \leq 0$. By definition of B, (i) if $y < B(x)$, then there does not exist an interior equilibrium cutoff in game (x, y), and (ii) if $y = B(x)$, then there does exist an interior equilibrium cutoff in game (x, y). The final case is $y > B(x)$. For $\alpha \in (0, 1)$, define $y^*(\alpha|x) := \frac{\alpha}{(1-\alpha)F(\alpha)} - x(\frac{1-F(\alpha)}{F(\alpha)})$ or, equivalently, $(x, y^*(\alpha|x)) \in \tilde{M}_\alpha$. For any $y > B(x)$, there exists $\alpha_1 < \alpha_2$ such that $y^*(\alpha_1|x) = B(x) < y < y^*(\alpha_2|x)$, where the final inequality follows from $\lim_{\alpha\to 1} \frac{\alpha}{(1-\alpha)F(\alpha)} = \infty$ and $\lim_{\alpha\to 1} \frac{1-F(\alpha)}{F(\alpha)} = 0$. Since $y^*(\cdot|x)$ is continuous, the intermediate value theorem implies that there exists $\alpha_3 \in (\alpha_1, \alpha_2)$ such that $y^*(\alpha_3|x) = y$, meaning α_3 in an interior equilibrium cutoff in the game (x, y). □

Proof of Proposition 6. First consider $g \in \mathrm{PD}^0$, so $\pi_g = 1$. By Propositions 1 and 2 (and Lemma 1), $\alpha_g^* \in (0, 1)$, implying $F(\alpha_g^*) < 1 = \pi_g$. Second, consider g such that $x > 0 \geq y$. It is straightforward to obtain $\pi_g = \frac{x}{x-y}$. The analog to (7) in which r, p have not been normalized is $F(\alpha_g^*) = \frac{\alpha_g^*(r-p)-(1-\alpha_g^*)x}{(1-\alpha_g^*)(y-x)}$. Therefore,

$$\pi_g - F(\alpha_g^*) = \frac{\alpha_g^*(r-p)}{(1-\alpha_g^*)(x-y)} > 0. \tag{12}$$

The inequality is due to $\alpha_g^* \in (0, 1)$ (by Proposition 5), $r > p$, and $x > 0 \geq y$.

For the general limit result, observe that Lemma 1 implies that it is sufficient to establish that if $r = 1$ and $p = 0$, then for any $\varepsilon > 0$, there exists $K \in \mathbb{R}_+$ such that, if $x + |y| > K$, then $\pi_{x,y} - F(\alpha_{x,y}^*) < \varepsilon$. Fix $\varepsilon > 0$, and define the terms $\alpha^\varepsilon = F^{-1}(1 - \varepsilon)$, $K_1 = \frac{\alpha^\varepsilon}{(1-\alpha^\varepsilon)\varepsilon}$, and, letting $(\mathrm{int}_{\alpha^\varepsilon}, \mathrm{slp}_{\alpha^\varepsilon})$ be the $(\mathrm{int}_i, \mathrm{slp}_i)$ generated by the equilibrium behavior of a player i with $\alpha_i = \alpha^\varepsilon$, $K_2 = \max\{\mathrm{int}_{\alpha^\varepsilon}, \frac{\mathrm{int}_{\alpha^\varepsilon}}{\mathrm{slp}_{\alpha^\varepsilon}}\}$.

Setting $K = \max\{K_1, K_2\}$ establishes the claim. To see this, suppose that $y > 0$ and $x + |y| > K$. Then $y > K - x \geq K_2 - x \geq \mathrm{int}_{\alpha^\varepsilon} - x \cdot \mathrm{slp}_{\alpha^\varepsilon}$. Hence, $(x, y) \in D_{\alpha^\varepsilon}$ and $\alpha_{x,y}^* > \alpha^\varepsilon$. Therefore, $F(\alpha_{x,y}^*) > F(\alpha^\varepsilon) = 1 - \varepsilon$ and $\pi_{x,y} - F(\alpha_{x,y}^*) = 1 - F(\alpha_{x,y}^*) < \varepsilon$. Suppose instead that $y \leq 0$ and $x + |y| > K$. First, if $\alpha_{x,y}^* > \alpha^\varepsilon$, then $1 \geq \pi_{x,y} > F(\alpha_{x,y}^*) > F(\alpha^\varepsilon) = 1 - \varepsilon$, and the result holds. Second, if $\alpha_{x,y}^* \leq \alpha^\varepsilon$, then by (12), we have

$$\pi_{x,y} - F(\alpha_{x,y}) = \frac{\alpha_{x,y}^*}{(1-\alpha_{x,y}^*)(x-y)} \leq \frac{\alpha^\varepsilon}{(1-\alpha^\varepsilon)(x-y)} < \frac{\alpha^\varepsilon}{(1-\alpha^\varepsilon)K_1} = \varepsilon,$$

establishing the claim. □

Proof of Proposition 7. Let $r = p$. If $\alpha_i = 1$, $V_i(c) = V_i(d)$ and player i is indifferent between c and d. However, such players are measure zero, and their behavior has no effect on the claims in the proposition. For the remainder, focus then on players with $\alpha_i \in [0, 1)$, and therefore $\mathrm{sign}(V_i(c) - V_i(d)) = \mathrm{sign}((1 - \alpha_i)[P_i \cdot (-y) + (1 - P_i) \cdot x]) =$

$\operatorname{sign}(P_i \cdot (-y) + (1-P_i) \cdot x)$, which is independent of α_i. Suppose now that (σ, P) is an equilibrium. If $\sigma(d|\alpha)$ is constant in α on $[0,1)$, then by condition (iii) of Definition 1, $\sigma(d|\alpha) = P$ for $\alpha \in [0,1)$. Therefore, assigning probability P to d is a best response to P, regardless of α_i, implying $\pi_g = P$ characterizes a symmetric Nash equilibrium. If $\sigma(d|\alpha)$ is not constant in α on $[0,1)$, then, since preferences are independent of α_i, for any $\alpha_i \in [0,1)$, *any* mixture over d and c is a best response to P. Again, then, $\pi_g = P$ characterizes a symmetric Nash equilibrium. Now suppose that π_g characterizes a symmetric Nash equilibrium. Let $F(\tilde{\alpha}) = \pi_g$, and let $\tilde{\sigma}(d|\alpha) = 1$ if $\alpha < \tilde{\alpha}$ and $= 0$ otherwise. It is trivial to verify that $(\tilde{\sigma}, \pi_g)$ is an equilibrium according to Definition 1. Finally, existence of an equilibrium follows from the existence of a symmetric Nash equilibrium Osborne and Rubinstein (1994, Section 20.4). □

Proof of Proposition 8. Fix $g \in \Gamma$ and a (candidate) equilibrium (σ, P). Let $\alpha < \alpha'$ and $\overline{k}(\alpha) := \max\{k : \sigma(k|\alpha) > 0\}$. Hence, for all $k < \overline{k}(\alpha)$,

$$V_i\big(\overline{k}(\alpha)|\alpha_i = \alpha\big) \geq V_i(k|\alpha_i = \alpha),$$

$$\alpha s\big(\overline{k}(\alpha)\big) + (1-\alpha) \sum_{k' \in A} P\big(k'\big) v\big(\overline{k}(\alpha), k'\big) \geq \alpha s(k) + (1-\alpha) \sum_{k' \in A} P\big(k'\big) v\big(k, k'\big),$$

$$\frac{\alpha}{1-\alpha} \underbrace{\big(s\big(\overline{k}(\alpha)\big) - s(k)\big)}_{>0} \geq \sum_{k' \in A} P\big(k'\big)\big(v\big(k, k'\big) - v\big(\overline{k}(\alpha), k'\big)\big),$$

$$\frac{\alpha'}{1-\alpha'} \big(s\big(\overline{k}(\alpha)\big) - s(k)\big) > \sum_{k' \in A} P\big(k'\big)\big(v\big(k, k'\big) - v\big(\overline{k}(\alpha), k'\big)\big),$$

$$\alpha' s\big(\overline{k}(\alpha)\big) + \big(1-\alpha'\big) \sum_{k' \in A} P\big(k'\big) v\big(\overline{k}(\alpha), k'\big) > \alpha' s(k) + \big(1-\alpha'\big) \sum_{k' \in A} P\big(k'\big) v\big(k, k'\big),$$

$$V_i\big(\overline{k}(\alpha)|\alpha_i = \alpha'\big) > V_i\big(k|\alpha_i = \alpha'\big).$$

Because A is finite, but the set of α-types is continuous, σ must therefore be increasing as described in Definition 6. Equilibrium existence is then an immediate application of the argument in Athey (2001) (Theorem 1, as the above establishes that its single crossing condition holds in our model). The only difference is that, since we are looking for a symmetric fixed point, we apply Kakutani's fixed point theorem to the single-player best-response correspondence (in that paper's notation, $\Gamma_i : \Sigma_i \to \Sigma_i$), rather than to the two-player best-response correspondence (i.e., $(\Gamma_1, \Gamma_2) : \Sigma_1 \times \Sigma_2 \to \Sigma_1 \times \Sigma_2$). □

Proof of Proposition 9. Fix a supermodular pure dilemma game $g \in \Gamma$ and let $W(k, l, \alpha) := \alpha s(k) + (1-\alpha) v(k, l)$ (i.e., $V_i(k)$ given $\alpha_i = \alpha$ and $P(l) = 1$). Then W has increasing differences in (k, α), and has increasing differences in (k, l). To see the first, let $k' \geq k$ and $\alpha' \geq \alpha$:

$$\begin{aligned}&\big(W\big(k', l, \alpha'\big) - W\big(k, l, \alpha'\big)\big) - \big(W\big(k', l, \alpha\big) - W(k, l, \alpha)\big)\\ &\quad = \big(\alpha' - \alpha\big)\big(\underbrace{s\big(k'\big) - s(k)}_{\geq 0} + \underbrace{v(k, l) - v\big(k', l\big)}_{\geq 0}\big) \geq 0.\end{aligned}$$

To see the second, let $k' \geq k$ and $l' \geq l$:

$$\begin{aligned}&\big(W(k',l',\alpha) - W(k,l',\alpha)\big) - \big(W(k',l,\alpha) - W(k,l,\alpha)\big)\\&\quad = (1-\alpha)\underbrace{\Big(\big(v(k',l') - v(k,l')\big) - \big(v(k',l) - v(k,l)\big)\Big)}_{\geq 0} \geq 0.\end{aligned}$$

Both statements in the proposition are then implications of Van Zandt and Vives (2007) (VZV). Part (i) follows from VZV Theorem 14 (they note that for symmetric games/models such as ours, the greatest and least equilibria are symmetric (VZV p. 346)). Part (ii) follows from VZV Proposition 16. □

Proof of Proposition 10. Fix $g \in \Gamma$ that has increasing returns to joint cooperation and a (candidate) equilibrium (σ, P). From Proposition 8, σ is increasing. Suppose now that $\alpha_1^* < \alpha_K^*$. Then, by Definition 6, there exists $k \notin \{0, K\}$ and $\alpha > \alpha'$ such that $\sigma(k|\alpha) = \sigma(k|\alpha') = 1$. So $V_i(k|\alpha_i = \alpha') \geq V_i(k-1|\alpha_i = \alpha')$, which implies $V_i(k|\alpha_i = \alpha) > V_i(k-1|\alpha_i = \alpha)$ (see the proof of Proposition 8). It follows that

$$\alpha s(k) + (1-\alpha)\sum_{k'\in A} P(k')v(k,k') > \alpha s(k-1) + (1-\alpha)\sum_{k'\in A} P(k')v(k-1,k'),$$

$$1 > \frac{(1-\alpha)}{\alpha}\sum_{k'\in A} P(k')\frac{v(k-1,k') - v(k,k')}{s(k) - s(k-1)}.$$

Increasing returns to joint cooperation imply

$$\sum_{k'\in A} P(k')\frac{v(k-1,k') - v(k,k')}{s(k) - s(k-1)} \geq \sum_{k'\in A} P(k')\frac{v(k,k') - v(k+1,k')}{s(k+1) - s(k)}.$$

Hence,

$$1 > \frac{(1-\alpha)}{\alpha}\sum_{k'\in A} P(k')\frac{v(k,k') - v(k+1,k')}{s(k+1) - s(k)},$$

$$\alpha s(k+1) + (1-\alpha)\sum_{k'\in A} P(k')v(k+1,k') > \alpha s(k) + (1-\alpha)\sum_{k'\in A} P(k')v(k,k'),$$

$$V_i(k+1|\alpha_i = \alpha) > V_i(k|\alpha_i = \alpha),$$

which contradicts that $\sigma(k|\alpha) = 1$ in equilibrium. Therefore, $\alpha_1^* = \alpha_K^*$. So, at most a measure-zero set of α-types assigns positive probability to any action other than 0 or K. This has no effect on the best responses of other types, so equilibrium analysis is identical to that done in the PD game $(r, p, x, y) = (s(K), s(0), v(0,K) - s(K), s(0) - v(K,0)) \in \text{PD}^0$. □

References

Ahn, T. K., Elinor Ostrom, David Schmidt, Robert Shupp, and James Walker (2001), "Cooperation in PD games: Fear, greed, and history of play." *Public Choice*, 106, 137–155. [926]

al-Nowaihi, Ali and Sanjit Dhami (2015), "Evidential equilibria: Heuristics and biases in static games of complete information." *Games*, 6, 637–676. [939]

Ali, S. Nageeb and Charles Lin (2013), "Why people vote: Ethical motives and social incentives." *American Economic Journal: Microeconomics*, 5, 73–98. [939]

Andreoni, James (1989), "Giving with impure altruism: Applications to charity and Ricardian equivalence." *Journal of Political Economy*, 97, 1447–1458. [929]

Andreoni, James and John H. Miller (1993), "Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence." *Economic Journal*, 103, 570–585. [938]

Aoyagi, Masaki and Guillaume Fréchette (2009), "Collusion as public monitoring becomes noisy: Experimental evidence." *Journal of Economic Theory*, 144, 1135–1165. [938]

Athey, Susan (2001), "Single crossing properties and the existence of pure strategy equilibria in games of incomplete information." *Econometrica*, 69, 861–890. [950]

Aumann, Robert J. (1964), "Mixed and behavior strategies in infinite extensive games." In *Advances in Game Theory* (Melvin Dresher, Lloyd S. Shapley, and Albert W. Tucker, eds.), 627–650, Princeton University Press, Princeton. [915]

Bereby-Meyer, Yoella and Alvin E. Roth (2006), "The speed of learning in noisy games: Partial reinforcement and the sustainability of cooperation." *American Economic Review*, 96, 1029–1042. [938]

Bergemann, Dirk, Stephen Morris, and Satoru Takahashi (2017), "Interdependent preferences and strategic distinguishability." *Journal of Economic Theory*, 168, 329–371. [937]

Bernheim, B. Douglas and Raphael Thomadsen (2005), "Memory and anticipation." *Economic Journal*, 115, 271–304. [939]

Blonski, Matthias, Peter Ockenfels, and Giancarlo Spagnolo (2011), "Equilibrium selection in the repeated prisoner's dilemma: Axiomatic approach and experimental evidence." *American Economic Journal: Microeconomics*, 3, 164–192. [937]

Bó, Perdro Dal, Andrew Foster, and Louis Putterman (2010), "Institutions and behavior: Experimental evidence on the effects of democracy." *American Economic Review*, 100, 2205–2229. [938]

Bowley, A. L. (1924), *The Mathematical Groundwork of Economics: An Introductory Treatise*. Clarendon Press. [939]

Camerer, Colin F. (2003), *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press. [933]

Camerer, Colin F. and Ernst Fehr (2003), "Measuring social norms and preferences using experimental games: A guide for social scientists." In *Foundations of Human Sociality: Experimental and Ethnographic Evidence From* 15 *Small-Scale Societies* (Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis, eds.), 55–95, Oxford University Press. [938]

Camerer, Colin F. and Robin M. Hogarth (1999), "The effects of financial incentives in experiments: A review and capital-labor-production framework." *Journal of Risk and Uncertainty*, 19, 7–42. [926]

Capraro, Valerio and Joseph Y. Halpern (2015), "Translucent players: Explaining cooperative behavior in social dilemmas." [939]

Capraro, Valerio, Jillian J. Jordan, and David G. Rand (2014), "Heuristics guide the implementation of social preferences in one-shot prisoner's dilemma experiments." *Scientific Reports*, 4, 6790. [936]

Charness, Gary, Luca Rigotti, and Aldo Rustichini (2016), "Payoff parameters and cooperation in the prisoner's dilemma." [926, 927]

Clark, Kenneth and Martin Sefton (2001), "The sequential prisoner's dilemma: Evidence on reciprocation." *Economic Journal*, 111, 51–68. [928]

Coate, Stephen and Micheal Conlin (2004), "A group rule-utilitarian approach to voter turnout: Theory and evidence." *American Economic Review*, 94, 1476–1504. [939]

Cooper, Russel, Douglas V. DeJong, Robert Forsythe, and Thomas W. Ross (1996), "Cooperation without reputation: Experimental evidence from prisoner's dilemma games." *Games and Economic Behavior*, 12, 187–218. [938]

Dawes, Robyn M., Jeanne McTavish, and Harriet Shaklee (1977), "Behavior, communication, and assumptions about other peoples' behavior in a commons dilemma situation." *Journal of Personality and Social Psychology*, 35, 1–11. [928]

Dawes, Robyn M. and Richard H. Thaler (1988), "Anomalies: Cooperation." *Journal of Economic Perspectives*, 2, 187–197. [926, 928]

Dillenberger, David and Philipp Sadowski (2012), "Ashamed to be selfish." *Theoretical Economics*, 7, 99–124. [936]

Ellingsen, Tore, Magnus Johannesson, Johanna Mollerstrom, and Sara Munkhammar (2012), "Social framing effects: Preferences or beliefs?" *Games and Economic Behavior*, 76, 117–130. [928]

Engel, Christoph and Lilia Zhurakhovska (2016), "When is the risk of cooperation worth taking? The prisoner's dilemma as a game of multiple motives." *Applied Economics Letters*, 23, 1157–1161. [926, 928]

Feddersen, Timothy and Alvaro Sandroni (2006), "A theory of participation in elections." *American Economic Review*, 96, 1271–1282. [939]

Fehr, Ernst and Klaus M. Schmidt (1999), "A theory of fairness, competition, and cooperation." *Quarterly Journal of Economics*, 114, 817–868. [927, 936]

Fudenberg, Drew, David G. Rand, and Anna Dreber (2012), "Slow to anger and fast to forgive: Cooperation in an uncertain world." *American Economic Review*, 102, 720–749. [910]

Gilboa, Itzhak and David Schmeidler (1989), "Maxmin expected utility with a non-unique prior." *Journal of Mathematical Economics*, 18, 141–153. [913]

Harsanyi, John C. and Reinhard Selten (1988), *A General Theory of Equilibrium Selection in Games*. MIT Press, Cambridge, MA. [912, 932]

Jeffrey, Richard C. (1983), *The Logic of Decision*, second edition. University of Chicago Press. [913]

Jones, Brooks, Matthew Steele, and James Gahagan (1968), "Matrix values and cooperative behavior in the prisoner's dilemma game." *Journal of Personality and Social Psychology*, 8, 148–153. [926]

Joyce, James M. (1999), *The Foundations of Causal Decision Theory*. Cambridge University Press. [914]

Kocher, Martin G., Peter Martinsson, and Martine Visser (2008), "Does stake size matter for cooperation and punishment?" *Economics Letters*, 99, 508–511. [926]

Kreps, David M., Paul Milgrom, John Roberts, and Robert Wilson (1982), "Rational cooperation in the finitely repeated prisoners' dilemma." *Journal of Economic theory*, 27, 245–252. [910]

Langer, Ellen J. (1975), "The illusion of control." *Journal of Personality and Social Psychology*, 32, 311–328. [913]

Ledyard, John (1995), "Public goods: A survey of experimental research." In *Handbook of Experimental Economics* (John H. Kagel and Alvin E. Roth, eds.), 111–194, Princeton University Press. [927]

Levine, David K. (1998), "Modeling altruism and spitefulness in experiments." *Review of Economic Dynamics*, 1, 593–622. [927]

Lewis, David (1979), "Prisoners' dilemma is a Newcomb problem." *Philosophy & Public Affairs*, 8, 235–240. [913]

Lindvall, Torgny (2002), *Lectures on the Coupling Method*, second edition. Dover Publications. [948]

List, John A. (2006), "Friend or foe? A natural experiment of the prisoner's dilemma." *Review of Economics and Statistics*, 88, 463–471. [926]

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green (1995), *Microeconomic Theory*. Oxford University Press. [919]

Masel, Joanna (2007), "A Bayesian model of quasi-magical thinking can explain observed cooperation in the public good game." *Journal of Economic Behavior & Organization*, 64, 216–231. [939]

McKelvey, Richard D. and Thomas R. Palfrey (1995), "Quantal response equilibrium for normal form games." *Games and Economic Behavior*, 10, 6–38. [927]

Milgrom, Paul R. and Robert J. Weber (1985), "Distributional strategies for games with incomplete information." *Mathematics of Operations Research*, 10, 619–632. [915]

Morris, Michael W., Damien L. H. Sim, and Vittorio Girotto (1998), "Distinguishing sources of cooperation in the one-round prisoner's dilemma: Evidence for cooperative decisions based on the illusion of control." *Journal of Experimental Social Psychology*, 34, 494–512. [928]

Nozick, Robert (1969), "Newcomb's problem and two principles of choice." In *Essays in Honor of Carl G. Hempel* (Nicholas Rescher, ed.), volume 24 of Synthese Library, 114–146, Springer, Netherlands. [913]

Orbell, John and Robyn M. Dawes (1991), "A 'cognitive miser' theory of cooperators' advantage." *American Political Science Review*, 85, 515–528. [928, 939]

Osborne, Martin J. and Ariel Rubinstein (1994), *A Course in Game Theory*. The MIT Press. [950]

Pakes, Ariel (2008), *Theory and Empirical Work on Imperfectly Competitive Markets*. [939]

Pigou, Arthur Cecil (1924), *The Economics of Welfare*, second edition. Macmillan. [939]

Quattrone, George A. and Amos Tversky (1984), "Causal versus diagnostic contingencies: On self-deception and on the voter's illusion." *Journal of Personality and Social Psychology*, 46, 237–248. [914]

Rabin, Matthew (1993), "Incorporating fairness into game theory and economics." *American Economic Review*, 83, 1281–1302. [927, 937]

Rapoport, Anatol and Albert M. Chammah (1965), *Prisoner's Dilemma: A Study in Conflict and Cooperation*. University of Michigan Press. [926]

Roemer, John E. (2010), "Kantian equilibrium." *Scandinavian Journal of Economics*, 112, 1–24. [939]

Roemer, John E. (2013), "Kantian optimization: An approach to cooperative behavior." Unpublished, Cowles Foundation Discussion Paper No. 1854R. [939]

Rohde, Kirsten I. M. (2010), "A preference foundation for Fehr and Schmidt's model of inequality aversion." *Social Choice and Welfare*, 34, 537–547. [936]

Rubinstein, Ariel and Yuval Salant (2014), "They do what I do: Positive correlation in ex-post beliefs." Unpublished, Tel Aviv University. [928, 931]

Rubinstein, Ariel and Yuval Salant (2016), "Isn't everyone like me?: On the presence of self-similarity in strategic interactions." *Judgment and Decision Making*, 11, 168–173. [913]

Saito, Kota (2013), "Social preferences under uncertainty: Equality of opportunity versus equality of outcome." *American Economic Review*, 103, 3084–3101. [936]

Savage, Leonard J. (1954), *The Foundations for Statistics*, Dover Books on Mathematics Series. John Wiley and Sons. [938]

Segal, Uzi and Joel Sobel (2007), "Tit for tat: Foundations of preferences for reciprocity in strategic settings." *Journal of Economic Theory*, 136, 197–216. [937]

Segal, Uzi and Joel Sobel (2008), "A characterization of intrinsic reciprocity." *International Journal of Game Theory*, 36, 571–585. [937]

Shafir, Eldar and Amos Tversky (1992), "Thinking through uncertainty: Nonconsequential reasoning and choice." *Cognitive Psychology*, 24, 449–474. [913, 928]

Steele, Matthew W. and James T. Tedeschi (1967), "Matrix indices and strategy choices in mixed-motive games." *Journal of Conflict Resolution*, 11, 198–205. [926, 927]

Straub, Paul G. (1995), "Risk dominance and coordination failures in static games." *Quarterly Review of Economics and Finance*, 35, 339–363. [931]

Thomson, William (2001), "On the axiomatic method and its recent applications to game theory and resource allocation." *Social Choice and Welfare*, 18, 327–386. [937]

Van de Assen, Martijn J., Dennie van Dolder, and Richard H. Thaler (2012), "Split or steal? Cooperative behavior when the stakes are large." *Management Science*, 58, 2–20. [926]

Van Zandt, Timothy and Xavier Vives (2007), "Monotone equilibria in Bayesian games of strategic complementarities." *Journal of Economic Theory*, 134, 339–360. [951]

Von Neumann, John and Oscar Morgenstern (1944), *Theory of Games and Economic Behavior*. Princeton University Press, Princeton. [936]

Co-editor Johannes Hörner handled this manuscript.

Manuscript received 17 February, 2015; final version accepted 29 April, 2016; available online 4 May, 2016.

Supplement to "Magical thinking: A representation result"

(*Theoretical Economics*, Vol. 12, No. 2, May 2017, 909–956)

Brendan Daley
The Fuqua School of Business, Duke University

Philipp Sadowski
Department of Economics, Duke University

This supplement contains extended formal results for Daley and Sadowski (2017) (henceforth DS16). Specifically, Section S.1 establishes that in prisoners' dilemma (PD) games, the model of DS16 is logically distinct from three models that employ well known forms of other-regarding preferences: altruism (Ledyard 1995, Levine 1998), inequity aversion (Fehr and Schmidt 1999), and reciprocity (Rabin 1993). Section S.2 provides an axiomatic characterization of F—the perceived distribution of types in the model—being empirically valid when there are infinitely many players. Section S.3 extends the axiomatic analysis to symmetric 2×2 games beyond PD games. All references to numbered sections/axioms/results/etc. are from DS16, unless otherwise indicated.

S.1. Models with other-regarding preferences

Consider the class of games denoted PD from Section 2, in which each game is parameterized by a pair $(x, y) \in R^2_{++}$.[1] The representation result (Theorem 1) establishes that under a condition on the slope of F, the data generated by the unique equilibrium behavior in PD of any such model satisfies four axioms, and for any data set that satisfies the axioms there exists a model, satisfying the same slope condition on F, that can explain it.[2]

Of course, there may exist other equivalent representations. As well-known models employing what are referred to as "other-regarding preferences" can sometimes accommodate cooperation by some players in some games in PD, they may seem candidates for this equivalence. In this section, we demonstrate that the models endowed with three of the most popular forms of other-regarding preferences are logically distinct from our model on PD.

Let u_i, u_j be the payoffs to players i and j as specified by the outcome of a two-player game. In each of the three models, player i seeks to maximize a different objective, which we denote v_i.

Brendan Daley: bd28@duke.edu
Philipp Sadowski: p.sadowski@duke.edu

[1] Because the purpose of this supplement is to demonstrate that the alternative models are behaviorally distinct from that of DS16, it suffices to establish the result on the subclass of games $\text{PD} \subset \text{PD}^0$.

[2] Recall that the four axioms are Axioms 2–5, as Axiom 1 is needed only for the larger set of games PD^0.

DOI: 10.3982/TE2099

1. *Altruism.* As proposed by Ledyard (1995) and further studied by Levine (1998): $v_i = u_i + \alpha_i u_j$, where $0 \leq \alpha_i < 1$; player i may care about his opponent's payoff, but not more than his own.

2. *Inequity Aversion.* As proposed by Fehr and Schmidt (1999): $v_i = u_i - \alpha_i \max\{u_j - u_i, 0\} - \delta\alpha_i \max\{u_i - u_j, 0\}$, where $0 \leq \alpha_i$ and $0 < \delta < \min\{1, \frac{1}{\alpha_i}\}$; player i may dislike inequity, but dislikes it more if his is the smaller payoff, and is not willing to "burn" his own payoff to create equity.[3]

3. *Reciprocity.* As proposed by Rabin (1993), player i cares about how "fair" he and his opponent are being to one another. Fixing the action of player i, a_i, how fair player j is being to player i is captured by the "kindness" function $K_j(a_j|a_i)$. In the prisoners' dilemma, once a_i is fixed all outcomes are Pareto optimal. In this case,

$$K_j(a_j|a_i) = \frac{u_i(a_i, a_j) - \frac{1}{2}\big(u_i^h(a_i) + u_i^l(a_i)\big)}{u_i^h(a_i) - u_i^l(a_i)},$$

where $u_i^h(a_i)$ and $u_i^l(a_i)$ are, respectively, the highest and lowest possible payoffs to i given a_i. Finally, $v_i = u_i + \alpha_i K_j(1 + K_i)$, where $\alpha_i \geq 0$.

The original specifications of these models did not include heterogeneity in the degree to which players are other-regarding. To incorporate heterogeneity into these models, in each we assume there is a common prior that α-types are drawn i.i.d. from a continuous distribution with support $[\underline{\alpha}, \overline{\alpha}]$, where $\underline{\alpha} \geq 0$, and CDF F. Complete homogeneity can be thought of as a limiting case as $(\overline{\alpha} - \underline{\alpha}) \to 0$. The equilibrium notion remains as in Definition 1, with $V_i(\cdot)$ suitably adapted to each model.

It is not our goal here to provide a comprehensive analysis of these models (which, while doable, would require a considerably longer treatment), but to establish the following.

Proposition S.1. *Fix any model of those described above and an equilibrium, (σ_g, P_g), for each game $g \in$ PD and consider the resultant data of all collections of arbitrary size n. Either, for all collections I, $D_i =$ PD for all $i \in I$, or there is a positive measure of collections (according to the common prior, F) each of whose data violates Axioms 2–5.*[4]

The result is proved in the subsequent analysis.

S.1.1 *Altruism*

Fixing any $(x, y) \in$ PD and an equilibrium (σ, P),

$$V_i(c|x, y, P) = (1 - P)(1 + \alpha_i \cdot 1) + P\big(-y + \alpha_i(1 + x)\big),$$

$$V_i(d|x, y, P) = (1 - P)\big(1 + x + \alpha_i(-y)\big) + P(0 + \alpha_i \cdot 0).$$

[3] One could consider a more general version in which the $\delta\alpha_i$ term is replaced by β_i. That is, players can have two-dimensional types. This would not alter our result.

[4] Further, the proposition remains valid if collections are formed via i.i.d. draws from any distribution with support $[\underline{\alpha}, \overline{\alpha}]$, even if its CDF differs from the one perceived by the players, F.

Therefore, $V_i(c|x, y, P) - V_i(d|x, y, P) = \alpha_i(1 + Px + (1-P)y) - (1-P)x - Py$. This expression is strictly increasing in α_i for all (x, y), P. Hence, all equilibria are cutoff equilibria.

For any given $(x, y) \in \text{PD}$, there exists an equilibrium with cutoff type α if and only if given $\alpha_i = \alpha$, $V_i(c|x, y, F(\alpha)) = V_i(d|x, y, F(\alpha))$. For any α, let $\tilde{M}_\alpha$ be the set of games in which there exists an equilibrium in which α is the cutoff type. Algebraically,

$$\tilde{M}_\alpha = \left\{(x, y) \in \text{PD} | y = \frac{\alpha}{(1+\alpha)F(\alpha) - \alpha} - \frac{1 - (1+\alpha)F(\alpha)}{(1+\alpha)F(\alpha) - \alpha} \cdot x\right\}.$$

Clearly, for all $i \in I$, $M_i \subset \tilde{M}_{\alpha_i}$.

We now argue that for any F, there exists a (generic) collection drawn from its support whose equilibrium play violates the axioms. First, let $\alpha^0 < \overline{\alpha}$ be the unique solution to $F(\alpha^0) = \frac{1}{1+\alpha^0}$. For all $\alpha \in [\alpha^0, \overline{\alpha}]$, $\tilde{M}_\alpha$ forms a line in PD that is weakly *upward* sloping. So, for any player i with $\alpha_i \in [\alpha^0, \overline{\alpha}]$ to be consistent with *Continuity* (Axiom 2) and *Monotonicity* (Axiom 3), it must be that $M_i = \varnothing$.[5] Second, fix arbitrary $\alpha \in [\alpha^0, \overline{\alpha}]$. Simple algebra shows that in the game $(\frac{\alpha}{1-\alpha}, \frac{\alpha}{1-\alpha}) \in \text{PD}$, α is the unique equilibrium cutoff, so must be in M_i for any i such that $\alpha_i = \alpha$. Hence, any player drawn from a high enough quantile of the distribution will have a violation.

The intuition for this is easy to see. Suppose that $\alpha_i = \overline{\alpha}$, so $F(\alpha_i) = 1$. Then, if in game (x, y), i is indifferent between c and d, all other players are choosing d. Therefore, i's indifference condition is $V_i(c|x, y, 1) = -y + \alpha_i(1 + x) = V_i(d|x, y, 1) = 0$. An increase in x *increases* $V_i(c)$ because it increases i's opponent's payoff, which i values altruistically. This makes player i strictly prefer c to d, and violates *Monotonicity*.

S.1.2 *Inequity aversion*

Fixing any $(x, y) \in \text{PD}$ and an equilibrium (σ, P),

$$V_i(c|x, y, P) = (1-P)(1 - \alpha_i \cdot 0) + P\big(-y - \alpha_i(1 + x + y)\big),$$
$$V_i(d|x, y, P) = (1-P)\big(1 + x - \delta\alpha_i(1 + x + y)\big) + P(0 - \alpha_i \cdot 0).$$

Therefore, $V_i(c|x, y, P) - V_i(d|x, y, P) = \alpha_i(1 + x + y)(\delta - P(1+\delta)) - (1-P)x - Py$. This expression is negative for $\alpha_i = 0$, monotonic in α_i, and increasing in α_i if and only if $P < \frac{\delta}{1+\delta} \leq \frac{1}{2}$. This immediately implies that all players defecting regardless of type (i.e., $P = 1$) is an equilibrium for any $(x, y) \in \text{PD}$. It also implies that if, for a given game, there exists an equilibrium in which a type cooperates, then it is a cutoff equilibrium where the cutoff type α^* must satisfy $F(\alpha^*) < \frac{\delta}{1+\delta} \leq \frac{1}{2}$.

Fix now any player i with α_i such that $F(\alpha_i) > \frac{1}{2}$. From above, $M_i = \varnothing$. Notice, though, that in any game, in any equilibrium where any type cooperates, player i cooperates. Therefore, we have the following two cases:

Case 1: Suppose $C_i = \varnothing$. Then, by the previous paragraph, in every game players are coordinating on the "all defect" equilibrium. Therefore, $D_j = \text{PD}$ for all $j \in I$, consistent with Proposition S.1.

[5] Suppose not, and that $(x, y) \in M_i$. Then to satisfy Axiom 3, (i) all other $(x', y') \in \tilde{M}_{\alpha_i}$ cannot be in M_i (so $M_i = \{(x, y)\}$), and (ii) $C_i \neq \varnothing$ and $D_i \neq \varnothing$. But then Axiom 2 is clearly violated.

Case 2: Suppose $C_i \neq \varnothing$. Then, given $M_i = \varnothing$, for player i to satisfy *Continuity* (Axiom 2), it must be that $D_i = \varnothing$. We now show that this cannot hold. To see this notice that (i) $V_i(c|x,y,P) - V_i(d|x,y,P)$ is monotonic (in fact, linear) in P, and (ii) $V_i(c|x,y,1) - V_i(d|x,y,1) = -y - \alpha_i(1+x+y) < 0$ for all α_i and $(x,y) \in \mathrm{PD}$. Therefore, if $V_i(c|x,y,0) - V_i(d|x,y,0) < 0$, then there is no equilibrium for game (x,y) in which i cooperates.

$$V_i(c|x,y,0) - V_i(d|x,y,0) = \delta\alpha_i(1+y) + x(-1+\delta\alpha_i).$$

Since $\delta\alpha_i < 1$, this is negative if $x > \frac{\delta\alpha_i(1+y)}{1-\delta\alpha_i}$. For any fixed y, there exist large enough x-values to satisfy this inequality for all α_i. Hence, $D_i \neq \varnothing$, violating Axiom 2.

S.1.3 *Reciprocity*

It is easy to calculate that for any pair of players i, j and $(x,y) \in \mathrm{PD}$, regardless of a_i, $K_j(a_j = d|a_i) = -\frac{1}{2}$ and $K_j(a_j = c|a_i) = \frac{1}{2}$. So, fixing any $(x,y) \in \mathrm{PD}$ and an equilibrium (σ, P),

$$V_i(c|x,y,P) = (1-P)\left(1 + \frac{3}{4}\alpha_i\right) + P\left(-y - \frac{3}{4}\alpha_i\right),$$

$$V_i(d|x,y,P) = (1-P)\left(1 + x + \frac{1}{4}\alpha_i\right) + P\left(0 - \frac{1}{4}\alpha_i\right).$$

From here, the analysis is analogous to that performed for inequity-averse players. $V_i(c|x,y,P) - V_i(d|x,y,P) = \alpha_i(\frac{1}{2} - P) - (1-P)x - Py$. This expression is negative for $\alpha_i = 0$, monotonic in α_i, and increasing in α_i if and only if $P < \frac{1}{2}$. This immediately implies that all players defecting regardless of type (i.e., $P = 1$) is an equilibrium for any $(x,y) \in \mathrm{PD}$. It also implies that if, for a given game, there exists an equilibrium in which a type cooperates, then it is a cutoff equilibrium where the cutoff type α^* must satisfy $F(\alpha^*) < \frac{1}{2}$.

Fix now any player i with α_i such that $F(\alpha_i) > \frac{1}{2}$. From above, $M_i = \varnothing$. Notice, though, that in any game, in any equilibrium where any type cooperates, player i cooperates. Therefore, we have the following two cases:

Case 1: Suppose $C_i = \varnothing$. Then, by the previous paragraph, in every game players are coordinating on the "all defect" equilibrium. Therefore, $D_j = \mathrm{PD}$ for all $j \in I$, consistent with Proposition S.1.

Case 2: Suppose $C_i \neq \varnothing$. Then, given $M_i = \varnothing$, for player i to satisfy *Continuity* (Axiom 2), it must be that $D_i = \varnothing$. We now show that this cannot hold. To see this notice that (i) $V_i(c|x,y,P) - V_i(d|x,y,P)$ is monotonic (in fact, linear) in P, and (ii) $V_i(c|x,y,1) - V_i(d|x,y,1) = -(\frac{\alpha_i}{2} + y) < 0$ for all α_i and $(x,y) \in \mathrm{PD}$. Therefore, if $V_i(c|x,y,0) - V_i(d|x,y,0) < 0$, then there is no equilibrium for game (x,y) in which i cooperates:

$$V_i(c|x,y,0) - V_i(d|x,y,0) = \frac{\alpha_i}{2} - x.$$

This is negative if $x > \frac{\alpha_i}{2}$. Hence, $D_i \neq \varnothing$, violating Axiom 2.

S.2. Large collections and empirically valid F

We say that F, the commonly perceived distribution of types in the model, is *empirically valid* if it agrees with empirical distribution of types in the collection. If so, magical thinking is the sole source of error in players' beliefs, and we refer to them as being *calibrated*. One issue that arises in our context, but not in axiomatic theories of individual choice, is the lack of data in the primitive itself. There, the primitive is typically assumed to be the agent's preference relation over all possible acts/choices. While our primitive includes each player's preferences over actions in all games in the domain, the collection of players is assumed to be finite.[6] It is easy to see that this precludes the observation of almost all α-types in $[0,1]$ and therefore the recovery of a unique F from the primitive. In addition, even if adhering to the population/sample interpretation discussed in Section 2.3, it is difficult to give behavioral meaning to the empirical validity of F when the analyst's data is generated by a finite collection.

To address both of these issues, in this supplement we let the collection of players be the interval $I = [0,1]$, endowed with the Lebesgue measure. This can be thought of as an approximation of an arbitrarily large collection or of drawing an arbitrarily large (and therefore completely representative) random sample in the population/sample interpretation, or as simply satisfying a theoretical curiosity. For simplicity, we consider the domain to be PD, and primitive $(D_i, C_i)_{i\in I}$.[7] In order for analysis to be tractable, we assume that the following are Lebesgue measurable: for all $(x,y) \in$ PD, the sets $\{i \in I | (x,y) \in D_i\}$ and $\{i \in I | (x,y) \in C_i\}$, and for any arbitrary individual behavior (D, C), the set $\{i \in I | (D_i, C_i) = (D, C)\}$.

Axioms 2–5 immediately apply to the larger set of behavioral data, but they are more restrictive in the following sense.

Definition S.1. Let $\mathcal{M}$ be the set of behavioral models, $[F, (\alpha_i)_{i\in I}]$, for which (i) F is continuous on $[0,1)$, (ii) if $\alpha < \alpha'$, then $F(\alpha') \leq F(\alpha)\frac{\alpha'(1-\alpha)}{\alpha(1-\alpha')}$, (iii) if, for $i \in I$, $\alpha_i \in (0,1)$, then $F(\alpha_i) \in (0,1)$, and (iv) if, for $\{i,j\} \subset I$, $\alpha_i < \alpha_j$, then $F(\alpha_i) < F(\alpha_j)$.

Proposition S.2. *The primitive $(D_i, C_i)_{i\in I}$ satisfies Axioms 2-5 if and only if it can be explained by a behavioral model $[F, (\alpha_i)_{i\in I}] \in \mathcal{M}$. Furthermore, for all $i \in I$, α_i is unique, and if $\alpha_i > 0$, then $F(\alpha_i)$ is unique.*

First, the convenient assumption that F is differentiable has no behavioral content in the case of finite I, but is no longer without loss of generality when I is a continuum. Consequently, the class of behavioral models $\mathcal{M}$ does not require differentiability. Further, (ii) is the meaningful content of Condition S without differentiability.[8] We show that it is both necessary and sufficient for uniqueness of the equilibrium cutoff in all

[6] There are common experimental techniques to circumvent the requirement of collecting infinite data on individual choice. In particular, infinite data can be approximated by finite data, indifference points can be elicited directly, or the individual can be asked to specify a decision rule. In contrast, the concern about the number of players in the sample is novel.

[7] Extending results to PD^0 is trivial via Axiom 1.

[8] If F is differentiable, then Conditions $S \iff$ (ii).

games. Second, while full support is not implied by the axioms when I is finite, it does encompass (iii) and (iv) (which are now joint restrictions on F and $(\alpha_i)_{i\in I}$). Finally, notice that atoms at $\alpha = 0, 1$ are permitted.

Definition S.2. Given any $(\alpha_i)_{i\in I}$, let $\widehat{F}$ be the CDF of types in I.

If the analyst views I as a perfectly representative sample of a grand population, then it is easy to evaluate whether or not $F(\alpha_i)$ is empirically valid for any $\alpha_i > 0$: simply compare the uniquely recovered value $F(\alpha_i)$ to $\widehat{F}(\alpha_i)$, which is identical to the population CDF by hypotheses. Any disagreement between the two represents miscalibration of the players.

There are two concerns with this evaluation method. First, it is *ad hoc* in that the analyst compares objects derived from the representation, instead of testing properties of the primitive directly. Second, the analyst cannot be sure that players are correctly calibrated regarding $F(\alpha)$ for $\alpha \notin (\alpha_i)_{i\in I}$. We now establish the behavioral content of the empirical validity of F (i.e., $F = \widehat{F}$), thereby eliminating both concerns.

Our first additional axiom rules out atoms of players with identical, nonextreme behavioral data. That is, there may be positive masses of players who strictly prefer to defect in all games, or strictly prefer to cooperate in all games. But, of all the players who exhibit both weak preference for defection and weak preference for cooperation somewhere within PD, it would seem nongeneric for a mass of them to cluster on any given (D, C) pair. Formally, for arbitrary (D, C), let $\mathcal{L}(D, C)$ be the Lebesgue measure of the set $\{i \in I | (D_i, C_i) = (D, C)\}$.

Axiom 6 (Smooth Data). *For all (D, C) such that $D \neq \text{PD}$ and $C \neq \text{PD}$, $\mathcal{L}(D, C) = 0$.*

Next, in our behavioral model, player i compares the perceived benefit of cooperation, α_i, with the perceived cost of cooperation, $(1-\alpha_i)(x(1-P(x, y))+yP(x, y))$, where $P(x, y)$ is the perceived probability that a random opponent in I will defect contingent on not being influenced by i. If $(x, y) \in M_i$, then i is indifferent between c and d, so $x(1 - P(x, y)) + yP(x, y) = \frac{\alpha_i}{1-\alpha_i}$. That is, $x(1 - P(x, y)) + yP(x, y)$ is constant on M_i. If player i is correctly calibrated, then the perceived probability $P(x, y)$ should coincide with the empirical frequency of defection in the population.

Definition S.3. Given $(D_i, C_i)_{i\in I}$, for each $(x, y) \in \text{PD}$, define $\widehat{P}(x, y)$ as the Lebesgue measure of the set $\{i \in I | (x, y) \in D_i\}$, and let $Q(x, y) := x(1 - \widehat{P}(x, y)) + y\widehat{P}(x, y)$.[9]

Because i cannot, in fact, directly influence his opponent's action choice, for each $(x, y) \in \text{PD}$, $Q(x, y)$ represents the true expected (opportunity) cost of cooperating in (x, y) against a random opponent in I. Our final axiom captures correct calibration by requiring this true cost of cooperation to be constant on M_i.

[9] Notice that we are interpreting $\widehat{P}(x, y)$ as the empirical analog of $P(x, y)$. Within the context of our axioms this is valid. However, in general, the empirical frequency of defection in game (x, y) may depend on the implementation of actions by players for which $(x, y) \in M_i$. This can be accommodated in a straightforward manner (see footnote 15).

Axiom 7 (Willingness to Pay for Own Cooperation). *For all $i \in I$, if $\{(x, y), (x', y')\} \subset M_i$, then $Q(x, y) = Q(x', y')$.*

To motivate the axiom without invoking the representation, imagine that there is a grand population, and that over time i plays various games in PD against random opponents from the population. In addition, I is a perfectly representative sample from this population. If player i cooperates in a given game, he does so at a cost to his own game payoff due to some nonstandard feature affecting his choice behavior, commonly referred to as a *bias* (not necessarily magical thinking). The axiom states that there is a single level for this true cost such that i is equally drawn to playing optimally (defecting) or being overcome by his bias to play suboptimally (cooperating). That is, $Q(x, y)$ for arbitrary $(x, y) \in M_i$, is the maximum cost associated with cooperation that i can endure.[10]

Proposition S.3. *The primitive $(D_i, C_i)_{i\in I}$ satisfies Axioms 2–7 if and only if there exists $(\alpha_i)_{i\in I}$ such that (i) $[\widehat{F}, (\alpha_i)_{i\in I}]$ explains $(D_i, C_i)_{i\in I}$ and (ii) $[\widehat{F}, (\alpha_i)_{i\in I}] \in \mathcal{M}$. Furthermore, for all $i \in I$, α_i is unique.*

Given Proposition S.2, this shows that Axioms 6 and 7 are the behavioral content of empirical validity. In fact, the role of each of the two can be isolated. Axiom 7 is the content of players being correctly calibrated about the types in the collection: $F(\alpha_i) = \widehat{F}(\alpha_i)$ for all $i \in I$. It is slightly more subtle to see that Axiom 6 is needed to ensure they are also correctly calibrated in their beliefs about those types *not* in the collection (i.e., they do not assign them positive probability). This is because the original axioms (2–5) require continuity of F on $[0, 1)$. If Axiom 6 fails, then $\widehat{F}$ will not be continuous on $[0, 1)$—there still exist behavioral models in $\mathcal{M}$ that can explain the data, but none with $F = \widehat{F}$.

S.2.1 *Proofs*

Proof of Proposition S.2. *Representation $\implies$ Axioms*: Consider a collection I that satisfies the representation. Our first step is to establish the analogs of Propositions 1–2 in this setting generated by replacing every appearance of "$F \in \mathcal{F}$" with "(i) of Definition S.1," and "Condition S" with "(ii) of Definition S.1." The proof of the modified version of Proposition 1 follows easily. To prove the modified version of Proposition 2, let F satisfy (i), and first suppose that condition (ii) is also satisfied. For the purpose of contradiction, suppose there exists $(x, y) \in \text{PD}$ that has two equilibrium cutoffs $\alpha_1^* < \alpha_2^*$, each of which satisfy (2). Writing out these two linear equations we can attempt to solve

[10] Recall that PD normalizes the payoff from (c, c) and (d, d). If considering all of PD^0 and applying Axiom 1, this maximum cost would be interpreted on a relative scale: if the stakes are higher all-around, then this maximum cost is likewise higher. This is consistent with an interpretation that i perceives gaining something from cooperating that scales with the game's payoffs. It is inconsistent with a bias such as *inattention* or *cognitive costs*, where i chooses cooperation only when the stakes are too small to bother figuring out the correct choice.

for x and y. Notice that when $F(\alpha_1^*) = F(\alpha_2^*)$, the two equations are inconsistent, and there is no solution, contradicting the hypotheses. If $F(\alpha_1^*) < F(\alpha_2^*)$, then solving for x and y yields unique values:

$$\begin{aligned} x &= \frac{\alpha_1^*(1-\alpha_2^*)F(\alpha_2^*) - \alpha_2^*(1-\alpha_1^*)F(\alpha_1^*)}{(1-\alpha_1^*)(1-\alpha_2^*)(F(\alpha_2^*) - F(\alpha_1^*))}, \\ y &= x + \frac{\alpha_2^* - \alpha_1^*}{(1-\alpha_1^*)(1-\alpha_2^*)(F(\alpha_2^*) - F(\alpha_1^*))}. \end{aligned} \tag{S.1}$$

The denominator of x is positive. However, the numerator of x is weakly negative by condition (ii) of Definition S.1. Therefore, $(x, y) \notin$ PD, contradicting the hypothesis. Hence, condition (ii) is sufficient for uniqueness of the cutoff. Second, to see that it is necessary, suppose that it is not satisfied, so there exists $\alpha < \alpha'$ such that $F(\alpha') > F(\alpha)\frac{\alpha'(1-\alpha)}{\alpha(1-\alpha')}$, which implies $F(\alpha') > F(\alpha)$. Then, setting $\alpha_1^* = \alpha$ and $\alpha_2^* = \alpha'$, (x, y) as given by (S.1) is in PD and simultaneously satisfies (2) for both types. Hence, there exists a game in which the equilibrium cutoff is not unique.

With this established, the remainder of the proof is analogous to the one used for Theorem 1.

Axioms $\Longrightarrow$ *Representation*: The proof follows the same steps as for Theorem 1. Lemma A.1 remains valid. Lemma A.2 must be modified as follows:

Lemma S.1. *Fix any player i. If (D_i, C_i) satisfies Axioms 2–4, then there exists a pair $(\alpha_i, F_i) \in [0,1]^2$ such that (D_i, C_i) can be explained by any behavioral model $[F, (\alpha_i, \alpha_{-i})]$ such that F is continuous on $[0, 1)$ and $F(\alpha_i) = F_i$. Further, α_i is unique, and F_i is unique if and only if $C_i \neq \varnothing$, as follows:*

$$(\alpha_i, F_i) = \begin{cases} \left(\dfrac{\text{int}_i}{1 + \text{int}_i + \text{slp}_i}, \dfrac{1}{1 + \text{slp}_i}\right) & \text{if } D_i, C_i \neq \varnothing, \\ (1, 1) & \text{if } D_i = \varnothing, \\ (0, K_i), K_i \in [0, 1] & \textit{if } C_i = \varnothing. \end{cases} \tag{S.2}$$

Proof. The proof is completely analogous to the proof of Lemma A.2 except in the following places: Case 1, parts (ii) and (iii); Case 3.

Case 1:

(ii) Suppose that $(x, y) \in C_i$. By Lemma A.1, this implies that $y < \text{int}_i - \text{slp}_i \cdot x$. Let $d(\alpha) := V(c|\alpha = \alpha^*) - V(d|\alpha = \alpha^*) = \alpha[1 + (1 - F(\alpha))x + F(\alpha)y] - [(1 - F(\alpha))x + F(\alpha)y]$. Using the assignments of $(\alpha_i, F(\alpha_i) = F_i)$ from (S.2), it follows that $d(\alpha_i) > 0$. Notice that $d(0) = x(F(0) - 1) - yF(0) < 0$ for any $F(0) \in [0, 1)$. F continuous on $[0, 1)$ implies that d is continuous $[0, \alpha_i]$. Hence, there exists $\alpha \in (0, \alpha_i)$ that achieves $d(\alpha) = 0$ and is therefore an equilibrium cutoff in the game $(x, y) \in C_i$ (by the analog of Proposition 1).

(iii) Suppose that $(x, y) \in D_i$. By Lemma A.1, this implies that $y > \text{int}_i - \text{slp}_i \cdot x$. Using the assignments of $(\alpha_i, F(\alpha_i) = F_i)$ from (S.2), it follows that $d(\alpha_i) < 0$. Further,

$\lim_{\alpha \uparrow 1} d(\alpha) = 1$. F continuous on $[0, 1)$ implies that d is continuous on $[\alpha_i, 1)$. Hence, there exists $\alpha \in (\alpha_i, 1)$ that achieves $d(\alpha_i) = 0$ and is therefore an equilibrium cutoff in the game $(x, y) \in D_i$ (by the analog of Proposition 1).

Case 3: In the behavioral model, for any F continuous on $[0, 1)$, a player i strictly prefers d in every $(x, y) \in \text{PD}$ if and only if his type is $\alpha_i = 0$. Given $\alpha_i = 0$, the value of $F(0)$ is irrelevant for i's behavior, so it cannot be determined. □

Lemmas A.3 and A.4, with references to Lemma A.2 now made to Lemma S.1, also remain valid. Hence, for any collection whose data satisfy Axioms 2–5, using (S.2), each player i can be assigned a unique α_i and corresponding quantile F_i, that is also unique if $\alpha_i > 0$, and i's behavior can be explained by any model $[F, (\alpha_i, \alpha_{-i})]$ such that F satisfies (i) of Definition S.1 and $F(\alpha_i) = F_i$.

We now show that there exists a model in $\mathcal{M}$ that simultaneously explains the behavior of all $i \in I$. Let $A^+ := \{\alpha_i > 0 | i \in I\}$. The four lemmas (A.1, S.1, A.3, A.4) imply that any behavioral model that satisfies $F(\alpha_i) = F_i$ for all $i \in I$ also satisfies (iii) and (iv) of Definition S.1 as well as (i) and (ii) *restricted to the domain* A^+, where continuous on A^+ means: for every $\alpha^0 \in A^+$, every sequence $(\alpha^m)_{m \in \mathbb{N}}$, $\alpha^m \in A^+$ for all m, such that $\lim_{m \to \infty} \alpha^m = \alpha^0$ also satisfies $\lim_{m \to \infty} F(\alpha^m) = F(\alpha^0)$. All that remains is to establish existence by extending F from A^+ to $[0, 1)$ preserving continuity, (weak) monotonicity, and condition (ii) of Definition S.1. From the proof of the opposite direction above, these properties imply that the behavioral model emits a unique equilibrium cutoff in all games $(x, y) \in \text{PD}$. This ensures that in each game there is an equilibrium consistent with the behavior of all players; hence, the behavioral model using this assignment of F and $(\alpha_i)_{i \in I}$ can explain $(D_i, C_i)_{i \in I}$.

To extend F from A^+ to $[0, 1)$, consider arbitrary $\alpha^0 \in [0, 1) \setminus A^+$. There are three exhaustive cases. First, if there exists a sequence $(\alpha^m)_{m \in \mathbb{N}}$, $\alpha^m \in A^+$ for all m, such that $\lim_{m \to \infty} \alpha^m = \alpha^0$, simply assign $F(\alpha^0) = \lim_{m \to \infty} F(\alpha^m)$. Second, let $\underline{\alpha} := \inf(A^+)$ and $\overline{\alpha} := \sup(A^+)$. If $\alpha^0 < \underline{\alpha}$, assign $F(\alpha^0) = F(\underline{\alpha})$, and if $\alpha^0 > \overline{\alpha}$, assign $F(\alpha^0) = F(\overline{\alpha})$—notice that even if $\underline{\alpha}, \overline{\alpha} \notin A^+$, $F(\underline{\alpha})$, $F(\overline{\alpha})$ are assigned in the previous case. Third, and finally, if A^+ does not contain a sequence converging to $\alpha^0 \in [\underline{\alpha}, \overline{\alpha}]$, then $\underline{\alpha}^0 := \sup\{\alpha \in A^+ | \alpha < \alpha^0\} < \alpha^0 < \overline{\alpha}^0 := \inf\{\alpha \in A^+ | \alpha > \alpha^0\}$. Notice that even if $\underline{\alpha}^0, \overline{\alpha}^0 \notin A^+$, $F(\underline{\alpha}^0)$, $F(\overline{\alpha}^0)$ are assigned in the first case. Let L^0 be the line that passes through both $(\underline{\alpha}^0, F(\underline{\alpha}^0))$ and $(\overline{\alpha}^0, F(\overline{\alpha}^0))$. For all $\alpha \in (\underline{\alpha}^0, \overline{\alpha}^0)$, assign $F(\alpha) = L^0(\alpha)$. It is immediate that these assignments preserve continuity and monotonicity in each case and also condition (ii) in the first and second cases. For the assignments in the third case, it is trivial to verify that the linearity of F between $[\underline{\alpha}^0, \overline{\alpha}^0]$ preserves condition (ii) on this interval, and then in general since condition (ii) is transitive. □

Proof of Proposition S.3. *Representation* $\Longrightarrow$ *Axioms*: Consider a collection I that satisfies the representation. The fact that $F = \widehat{F}$ is irrelevant for the proof that $(D_i, C_i)_{i \in I}$ satisfies Axioms 2–5, so this is established by Proposition S.2. Next, $F = \widehat{F}$, which is continuous on $[0, 1)$, immediately implies Axiom 6. Finally, verifying Axiom 7 is a straightforward calculation: fix any player i such that $M_i \neq \varnothing$ and recall that

$M_i = \{(x,y) \in \mathrm{PD} | y = \frac{\alpha_i}{(1-\alpha_i)F(\alpha_i)} - x(\frac{1-F(\alpha_i)}{F(\alpha_i)})\}$. Therefore, for any $(x,y) \in M_i$, we can substitute the expression for y into $Q(x,y)$ to get,

$$Q(x,y) = x\big(1-\widehat{P}(x,y)\big) + \left(\frac{\alpha_i}{(1-\alpha_i)F(\alpha_i)} - x\left(\frac{1-F(\alpha_i)}{F(\alpha_i)}\right)\right)\widehat{P}(x,y). \tag{S.3}$$

Given that $F = \widehat{F}$ and that α_i is the cutoff type for $(x,y) \in M_i$, $\widehat{P}(x,y) = F(\alpha_i)$; so (S.3) simplifies to $Q(x,y) = \frac{\alpha}{1-\alpha}$, which does not vary with (x,y).

Axioms $\Longrightarrow$ *Representation*: The proof of Proposition S.2, establishes that if $(D_i, C_i)_{i\in I}$ satisfies Axioms 2–5, then it can be explained by any model $[F, (\alpha_i)_{i\in I}] \in \mathcal{M}$, where α_i and $F(\alpha_i) = F_i$ are given by (S.2) (and therefore α_i is unique and, if $\alpha_i > 0$, so is $F(\alpha_i)$). Therefore, let $(\alpha_i)_{i\in I}$ be as given by (S.2), and $\widehat{F}$ be the resultant CDF. It is sufficient to show that 1) for all i such that $\alpha_i > 0$, $\widehat{F}(\alpha_i) = F_i$, and 2) $[\widehat{F}, (\alpha_i)_{i\in I}] \in \mathcal{M}$.

To see the first, notice that the structure of $(D_i, C_i)_{i\in I}$ characterized by Lemmas A.1, S.1, A.3, and A.4 implies that for any i such that $M_i \neq \varnothing$, $\widehat{P}(x,y)$ is constant and equal to $\lim_{\alpha\uparrow\alpha_i}\widehat{F}(\alpha)$ along M_i. By Axiom 6, $\lim_{\alpha\uparrow\alpha_i}\widehat{F}(\alpha) = \widehat{F}(\alpha_i)$. Consider i such that $\alpha_i \in (0,1)$, so $M_i \neq \varnothing$. For $(x,y) \in M_i$,

$$\begin{aligned} Q(x,y) &= x\big(1-\widehat{P}(x,y)\big) + y\widehat{P}(x,y) \\ &= x\big(1-\widehat{F}(\alpha_i)\big) + y\widehat{F}(\alpha_i) \\ &= x\big(1-\widehat{F}(\alpha_i)\big) + (\mathrm{int}_i - x\cdot\mathrm{slp}_i)\widehat{F}(\alpha_i). \end{aligned}$$

By Axiom 7, Q is constant along M_i, so $\widehat{F}(\alpha_i) = \frac{1}{1+\mathrm{slp}_i} = F_i$. If instead, $\alpha_i = 1$, then because $\widehat{F}$ is a CDF on $[0,1]$, $\widehat{F}(\alpha_i) = 1 = F_i$.

To see the second, we need to show that $[\widehat{F}, (\alpha_i)_{i\in I}]$ satisfies the four requirements of Definition S.1. Axiom 6 implies (i), and Lemmas S.1 and A.4 imply (iii) and (iv). For (ii), notice that if α and α' are elements of $(\alpha_i)_{i\in I}$, then the property holds due to Lemma A.4 and if $\alpha = 0$ or $\alpha' = 1$, the property is trivial. Consider now an arbitrary pair $0 < \alpha < \alpha' < 1$, and for the purpose of contradiction suppose that $\frac{\widehat{F}(\alpha')}{\widehat{F}(\alpha)} > \frac{\alpha'(1-\alpha)}{\alpha(1-\alpha')}$. Since $\widehat{F}$ is the CDF of $(\alpha_i)_{i\in I}$, and is continuous on $[\alpha, \alpha']$, for any $\varepsilon > 0$, there must exist $\{i,j\} \subset I$ such that $\alpha \le \alpha_i < \alpha_j \le \alpha'$, $\widehat{F}(\alpha_i) - \widehat{F}(\alpha) < \varepsilon$, and $\widehat{F}(\alpha') - \widehat{F}(\alpha_j) < \varepsilon$. Hence, by our supposition that $\frac{\widehat{F}(\alpha')}{\widehat{F}(\alpha)} > \frac{\alpha'(1-\alpha)}{\alpha(1-\alpha')}$, for ε small enough,

$$\frac{\widehat{F}(\alpha_j)}{\widehat{F}(\alpha_i)} > \frac{\alpha'(1-\alpha)}{\alpha\big(1-\alpha'\big)} \ge \frac{\alpha_j(1-\alpha_i)}{\alpha_i(1-\alpha_j)}.$$

As we just discussed, Lemma A.4 implies that $\frac{\widehat{F}(\alpha_j)}{\widehat{F}(\alpha_i)} \le \frac{\alpha_j(1-\alpha_i)}{\alpha_i(1-\alpha_j)}$, producing a contradiction. □

S.3. Axiomatic analysis beyond the PD

The defining feature of the prisoners' dilemma is that there are strict gains to a player for selecting d whether his opponent is playing c or d (i.e., $x, y > 0$). We first enlarge our domain by relaxing the latter. That is, we consider games in which there are strict gains from unilaterally deviating away from the better symmetric outcome. To do so, let $G^0 = \{(r, p, x, y) | r > p, x > 0\}$, with labels as in Figure 1, and let our primitive, $(D_i^0, C_i^0)_{i \in I}$, as well as $(M_i^0, \overline{D}_i^0, \overline{C}_i^0)_{i \in I}$, be extended to this larger class of games in the obvious way. Finally, define $G = \{(r, p, x, y) | r = 1, p = 0, x > 0\} \subset G^0$, with arbitrary element (x, y), and, as before, $D_i = D_i^0 \cap G$ and analogously for C_i, M_i, $\overline{D}_i$, and $\overline{C}_i$. Notice that G is the union of the games in quadrants I and IV of Figure 4.

Each of the Axioms 1–5 can be applied verbatim on this larger class of games (simply replace each PD^0 and PD with G^0 and G, respectively). In addition, with the caveat of changing all instances of “cooperate” and “defect” to “play c” and “play d,” respectively, the interpretations of each of the axioms are also unchanged.

We introduce an additional axiom. Fixing all other payoff parameters, the societal benefit from (either or both) players selecting c, the action corresponding to the better symmetric outcome, is increasing in r. The following axiom requires that increases in r should increase the propensity to select c.

Axiom 8 (Sensitivity to Benefits from Action c). *For all $i \in I$, if $(r, p, x, y) \in \overline{C}_i^0$ and $r' > r$, then $(r', p, x, y) \in C_i^0$.*

It is not difficult to show that the representation in Theorem 1 satisfies Axiom 8 on PD^0, meaning Axiom 8 is implied by Axioms 1–5 on this domain. On G^0, this is no longer the case.

Fact S.1. *Axioms 1–5 $\Longrightarrow$ Axiom 8 on* PD^0. *Axioms 1-5 $\Longrightarrow$ Axiom 8 on G^0.*

Notice that the axiom is consistent with the experimental evidence discussed in Section 5.1.[11] Further, in line with the axiom, Rapoport and Chammah (1965) and Minas et al. (1960) compare behavior across different Prisoners' Dilemma games and provide evidence that the fraction of players selecting c indeed increases with r.[12]

By adding Axiom 8, the representation result of Theorem 1 extends to G^0.

Theorem S.1. *The primitive $(D_i^0, C_i^0)_{i \in I}$, on G^0, satisfies Axioms 1–5 and 8 if and only if it can be explained by a behavioral model $[F, (\alpha_i)_{i \in I}]$, where $F \in \mathcal{F}$ satisfies Condition S. Furthermore, for all $i \in I$, α_i and $F(\alpha_i)$ are unique.*

The extended representation also satisfies the more stringent definition of *can explain* attained if the requirements of Definition 4 must instead hold in *all* equilibria (see Section 2.3).

[11] It is easy to derive that for any hawk–dove game $(r, 0, x, y)$, $x \neq -y$, the game $(0, 0, x, y)$ is a battle of the sexes game with the same symmetric Nash equilibrium. Hence, insofar as subjects adhere to the symmetric Nash equilibrium in battle of the sexes games, but play c more frequently than in the symmetric Nash equilibrium in hawk–dove games (see Section 5.1), their play is consistent with Axiom 8.

[12] Up to adding constants (as permitted once we assume Axiom 1), see games labeled G4 and G5 in Minas et al. (1960) and games numbered 1 and 4 in Rapoport and Chammah (1965). This evidence is also summarized in Table 1 of Steele and Tedeschi (1967).

Next, one can extend the domain to include games in which $x \leq 0$ and $y \leq 0$ (i.e., quadrant III of Figure 4 when r and p are normalized). In these games c is *both* the action leading to the better symmetric outcome and a dominant strategy (even without magical thinking), with the dominance being strict on the interior of the quadrant. It seems natural that all players should choose c then, as they do in the our behavioral model (Section 5.1.1). In addition, for each player i such that $M_i \cap G \neq \varnothing$, this behavior is a consequence of Axioms 1–5 and 8 when the primitive is likewise extended. Under the (seemingly mild) additional requirement that in the extended domain $\overline{C}_i \neq \varnothing$ for all $i \in I$, the representation result extends with only minor alteration.[13]

How can our axioms be extended to the games with $x \leq 0$ and $y > 0$ (i.e., quadrant II of Figure 4 when r and p are normalized)? We suggest three possible ways. First, and most immediately, one can add an axiom that specifies c as the preferred action for all players whenever $x \leq 0$ and restrict our other axioms to games with $x > 0$. Second, one can extend our theory as discussed in the context of quadrant-III games, but additionally weaken Axiom 5 to allow the extended M_i-lines to intersect when $x \leq 0$. It can then be shown that the resulting representation in terms of our behavioral model would entail that, in each game, each player selects his action in accordance with an equilibrium, implying his choice is rationalizable (but not all players will play in accordance with the same equilibrium when there are multiple).

Third, one could try to really capture if/when there is multiplicity. For instance, suppose players would be willing to participate in different profiles of play (as would be the case if they actually conceived of multiple equilibria). How could this manifest itself in behavior? Since our primitive requires each player to rank d and c for every possible game, one would need to consider a richer primitive. One possibility mirrors the menu-choice approach in theories of individual choice. The analyst could instruct players that they will face an anonymous opponent in a game in period 2. In period 1, the analyst could ask players to specify for each game whether they are willing to commit to d, to c, or whether they have a preference for flexibility in the sense that they do not want to precommit to an action choice for period 2. Such preference for flexibility could be interpreted as the anticipation of coordination on an equilibrium based on some state of the world that is unobserved (or indecipherable) by the analyst and that realizes between periods 1 and 2. One could try to formulate axioms that restrict period-1 preferences over menus of actions across games and players to ensure that multiplicity is consistent with our model. In particular, the axioms should correspond to Axioms 1–5 and 8 on quadrants I and IV.

S.3.1 *Proofs*

The representation proof uses the following preliminary lemma.

Axiom 8$'$. *For all $i \in I$, if $(x, y) \in \overline{C}_i$ and $\kappa \in (0, 1)$, then $\kappa(x, y) \in C_i$.*

[13] If extending the axioms verbatim, the representation will require that $\alpha_i \neq 0$ for all $i \in I$. Since this event already has probability one according to any $F \in \mathcal{F}$, no other change to the corresponding behavioral model is required. Alternatively, one could slightly relax the extensions of Axioms 3 and 8 and maintain the original class of behavioral models.

Lemma S.2. *Under Axiom 1, Axioms 8 and 8′ are equivalent.*

Proof. Suppose that Axioms 1 and 8 hold and that $\kappa \in (0,1)$. Then

$$\begin{aligned}(x,y)\in\overline{C}_i \quad &\Longrightarrow \quad (1,0,x,y)\in\overline{C}_i^0 \quad \underset{\text{Axiom 8}}{\Longrightarrow} \quad \left(\frac{1}{\kappa},0,x,y\right)\in C_i^0\\ &\underset{\text{Axiom 1}}{\Longrightarrow} \quad \kappa\left(\frac{1}{\kappa},0,x,y\right)\in C_i^0 \quad \Longrightarrow \quad (1,0,\kappa x,\kappa y)\in C_i^0\\ &\Longrightarrow \quad (\kappa x,\kappa y)\in C_i \quad \Longrightarrow \quad \kappa(x,y)\in C_i.\end{aligned}$$

Hence, Axiom 8′ is implied. Now, suppose that Axioms 1 and 8′ hold and that $r' > r$. Then

$$\begin{aligned}(r,p,x,y)\in\overline{C}_i^0 \quad &\underset{\text{Axiom 1}}{\Longrightarrow} \quad \left(1,0,\frac{x}{r-p},\frac{y}{r-p}\right)\in\overline{C}_i^0 \quad \Longrightarrow \quad \left(\frac{x}{r-p},\frac{y}{r-p}\right)\in\overline{C}_i\\ &\Longrightarrow \quad \frac{1}{r-p}(x,y)\in\overline{C}_i \quad \underset{\text{Axiom 8}'}{\Longrightarrow} \quad \frac{1}{r'-p}(x,y)\in\overline{C}_i\\ &\Longrightarrow \quad \left(\frac{x}{r'-p},\frac{y}{r'-p}\right)\in C_i \quad \Longrightarrow \quad \left(1,0,\frac{x}{r'-p},\frac{y}{r'-p}\right)\in C_i^0\\ &\underset{\text{Axiom 1}}{\Longrightarrow} \quad \left(r',p,x,y\right)\in C_i^0.\end{aligned}$$

Hence, Axiom 8 is implied. □

Proof of Fact S.1. Relying on Lemma S.2, we consider whether or not Axioms 2–5 imply Axiom 8′ on PD and G for the first and second claims, respectively. For the first claim, fix player i, for whom $(D_i \cap \text{PD}, C_i \cap \text{PD})$ satisfies Axioms 2–4, with $\overline{C_i} \cap \text{PD} \neq \varnothing$. Then, from the proof of Theorem 1, we have that either $C_i \cap \text{PD} = \text{PD}$ or $M_i \cap \text{PD} = \{(x,y) \in \text{PD} | y = \text{int}_i - \text{slp}_i \cdot x\}$ and $C_i \cap \text{PD} = \{(x,y) \in \text{PD} | y < \text{int}_i - \text{slp}_i \cdot x\}$, where int_i, slp_i are positive constants. In either case, Axiom 8′ follows immediately. For the second claim, consider a player i with $M_i = \{(x,y) \in G | y = -1 - x\}$, and C_i and D_i being the strict-lower and strict-upper contour sets of M_i, respectively. It is immediate that (D_i, C_i) satisfies Axioms 2–4. However, (D_i, C_i) fails Axiom 8′: for any $(x,y) \in M_i$, $\frac{1}{2}(x,y) \in D_i$. The fact that $(D_j, C_j)_{j\in I}$ satisfies Axiom 5 does not rule out the existence of such a player, meaning the result is established. □

Proof of Theorem S.1. First, note that Lemma 1 and Propositions 1 and 2 (and their proofs) remain valid when each PD^0 and PD are replaced by G^0 and G, respectively.

Representation $\Longrightarrow$ *Axioms*: Given that Lemma 1 and Propositions 1 and 2 extend to the larger domain, the proof that the representation satisfies Axioms 1–5 is completely analogous to that provided for Theorem 1. Using Lemma S.2, we are left to verify that Axiom 8′ is satisfied. First, if $\alpha_i = 0$, then $D_i = G$ so the axiom is vacuous; and if $\alpha_i = 1$, then $C_i = G$ so the axiom is trivial. Second, if $\alpha_i \in (0,1)$ and $(x,y) \in \overline{C}_i$, then $y \leq \text{int}_i - x \cdot \text{slp}_i$, where $\text{int}_i, \text{slp}_i > 0$. It follows that, for any $\kappa \in (0,1)$, $\kappa y \leq \kappa(\text{int}_i - x \cdot \text{slp}_i) < \text{int}_i - (\kappa x)\text{slp}_i$. Hence, $\kappa(x,y) \in ML_i = C_i$, verifying the axiom.

Axioms $\implies$ *Representation*: The only aspect of the proof that is not completely analogous to that given for Theorem 1 is in extending the following aspect of Lemma A.1. Consider a player i for which $D_i \neq \varnothing$ and $C_i \neq \varnothing$. Such a player can be characterized by a pair $(\text{int}_i, \text{slp}_i)$, where $\text{slp}_i > 0$. When the domain was PD, $\text{int}_i > 0$ immediately. This is no longer immediate when the domain is G. However, it is ensured by Axiom 8. Suppose to the contrary that $\text{int}_i \leq 0$. Take now $(x, y) \in M_i \subset \overline{C}_i$, which must then satisfy $y = \text{int}_i - x \cdot \text{slp}_i < 0$. But then, for any $\kappa \in (0, 1)$, $\kappa y = \kappa(\text{int}_i - x \cdot \text{slp}_i) \geq \text{int}_i - (kx)\text{slp}_i$. Hence, $\kappa(x, y) \notin ML_i = C_i$, violating Axiom $8'$ (and therefore also Axiom 8 by Lemma S.2). With this established, the remainder of the proofs follows identical steps to those in the proof of Theorem 1. □

References

Daley, B. and P. Sadowski (2017), "Magical thinking: A representation result." *Theoretical Economics*, 12, 909–956. [1]

Fehr, Ernst and Klaus M. Schmidt (1999), "A theory of fairness, competition, and cooperation." *Quarterly Journal of Economics*, 114, 817–868. [1, 2]

Ledyard, John (1995), "Public goods: A survey of experimental research." In *Handbook of Experimental Economics* (John H. Kagel and Alvin E. Roth, eds.), 111–194, Princeton University Press. [1, 2]

Levine, David K. (1998), "Modeling altruism and spitefulness in experiments." *Review of Economic Dynamics*, 1, 593–622. [1, 2]

Minas, J. S., A. Scodel, D. Marlowe, and H. Rawson (1960), "Some Descriptive Aspects of Two-Person Non-Zero-Sum Games. II." *Journal of Conflict Resolution*, 4, 193–197. [11]

Rabin, Matthew (1993), "Incorporating fairness into game theory and economics." *American Economic Review*, 83, 1281–1302. [1, 2]

Rapoport, Anatol and Albert M. Chammah (1965), *Prisoner's Dilemma: A Study in Conflict and Cooperation*. University of Michigan Press. [11]

Steele, Matthew W. and James T. Tedeschi (1967), "Matrix indices and strategy choices in mixed-motive games." *Journal of Conflict Resolution*, 11, 198–205. [11]

Co-editor Johannes Hörner handled this manuscript.

Manuscript received 17 February, 2015; final version accepted 29 April, 2016; available online 4 May, 2016.