A Strategic Model of Magical Thinking: Axioms and Analysis*

Brendan Daley† Philipp Sadowski‡

February 13, 2015

Abstract

This paper suggests a novel way to import the approach of axiomatic theories of individual choice into strategic settings, focusing on the class of symmetric 2×2 games, 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 across 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.

Key words: Prisoners’ dilemma, coordination games, magical thinking, cooperation

JEL Codes: C7, D8

*An earlier working version circulated under the title, “Foundations for Cooperation in the Prisoners’ Dilemma.” We 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, Larry Samuelson, Todd Sarver, Eldar Shafir, Ran Spiegler, Tomasz Strzalecki, and Curtis Taylor for useful discussions. We also thank seminar audiences at Berkeley, Bocconi, Duke, Harvard/MIT, Princeton, NES Moscow, the Yale Conference on Social Choice, and BRIC 2014 for their comments.

†The Fuqua School of Business, Duke University. e-mail: bd28@duke.edu
‡Department of Economics, Duke University. e-mail: p.sadowski@duke.edu
1 Introduction

This paper suggests a novel way to import the approach of axiomatic theories of individual choice into game-theoretic settings, focusing on the class of symmetric $2 \times 2$ games. The novelty lies in the behavioral data to which our axioms apply. First, the axioms concern players’ preferences over actions in each game played as a one-shot game with anonymity, rather than preferences over outcomes. Second, 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.

We provide a representation result that establishes the equivalence between the axioms and the equilibrium play of a behavioral model. Specifically, the behavioral model features magical thinking: players behave as if their expectations about their opponents’ behavior vary with their own choices. Further, the model’s parameters are (essentially) uniquely identified from behavior.

The contribution of the paper is therefore threefold. One, we provide a tractable and empirically plausible theory of magical thinking, a phenomenon that has received attention in psychology and philosophy (see Section 2), 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 Prisoners’ Dilemma (PD), the Battle of the Sexes, Hawk-Dove (also known as Chicken), and the Stag Hunt.

Two, distinct from typical work in applied or behavioral game theory, we present a representation result that delineates the precise behavioral content of the model (namely, the axioms) and show that the model’s parameters can be identified from behavior. The representation result allows for the evaluation and empirical testing of the model, and facilitates its comparison to alternative theories. Identification 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 a subset of games can be used to identity the parameters of the model, which can be used in turn to generate predictions for (not-yet-observed) behavior in a different subset of games. All of this is important for applied work.

Three, a key component of our approach is that the axioms apply to players’ preferences over actions (rather than outcomes). Axiomatizing this type of data confers the following benefits.

1. 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.
2. This type of data is straightforward and common to collect in experiments.\(^1\)

3. 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.

**Outline**

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.\(^2\) We begin our analysis on PD games in order to better build intuition, to demonstrate that our model makes behavioral predictions distinct from other explanations of cooperative behavior in PD games,\(^3\) and to show that the model’s parameters can be identified from behavior in PD games alone. We then extend our analysis to all symmetric 2×2 games, where the model’s predictions continue to align with experimental evidence. 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.

Within the as-if model, players believe their action choices stochastically influence the choices of their anonymous opponents. More specifically, each player \(i\) in the collection of players, \(I\), is endowed a type, \(\alpha_i\), and there is a CDF over types, \(F\), from which players perceive types to be i.i.d. draws. Given any PD, player \(i\) forms the following nonstandard beliefs. He assigns probability \(\alpha_i\) 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 \(\alpha_i\) as \(i\)’s degree of magical thinking. A player with \(\alpha_i = 0\) corresponds

---

\(^1\)To facilitate comparison with experimental findings, we interpret game-payoffs in monetary terms throughout. However, we also emphasize that our theory does not rely on this interpretation, and applies equally well when game-payoffs are denoted in utils, as is customary in game theory. In particular, our model explains cooperation in games that remain PDs even when game-payoffs are denoted in utils that already account for any preferences over the opponent’s physical payout (in contrast to explanations such as altruism or inequity aversion).

\(^2\)Of course, cooperation is easier to explain in the repeated PD, provided players are sufficiently patient. 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 that while observed cooperation levels are higher when cooperative equilibria exist, substantial levels of cooperation (over 30% across all rounds) remain even when the unique equilibrium strategy is to always defect.

\(^3\)In the Supplemental Online Appendix, we demonstrate that models assuming the other-regarding preferences of altruism of Ledyard (1995); Levine (1998), inequity-aversion of Fehr and Schmidt (1999), and reciprocity of Rabin (1993) are all logically distinct from our model on the domain of PD games by showing that they violate our axioms.
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 Inter-Player 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 also consistent with a player’s willingness to cooperate being responsive to the true cost of doing so. We compare our axioms to the relevant findings in the experimental literature and find them to be broadly consistent (Section 4).

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 PDs. 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. Finally, note that in the representation, $F$ is the common belief among players regarding the distribution that types are drawn from. In Appendix B, 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 (Section 5.1). 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 (Section 5.1). 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 (Section 6). However, the prediction is more nuanced than risk dominance in that whether the payoff-
dominant Nash equilibrium is too risky or not depends on the (perceived) distribution of types, $F$.

It should be noted that the model’s ability to capture all of these findings does not owe to any flexibility across different games. Our representation result shows 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 non-standard 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 (and closely related variants) as well as other prominent games in our domain, there remains significant evidence of non-standard behavior that is not explained by these theories, but is consistent with our model of magical thinking, as we discuss in Sections 4, 5, and the Supplemental Appendix.

The remainder of the paper is organized as follows. Section 2 reviews how magical thinking relates to topics in psychology, philosophy, and economics. Focusing first on PD games, Section 3 presents our model, axioms, representation theorem, and comparative statics, and Section 4 compares our theory to experimental evidence and alternative theories of play. Sections 5 and 6 extend our analysis to all symmetric $2 \times 2$ games. Section 7 compares our axiomatic methodology to alternative approaches and offers an outlook for future research. Proofs are found in Appendix A. Appendix B extends our representation result to collections with infinitely many players and provides a characterization of $F$ being empirically valid.

2 Magical Thinking and Related Notions

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.\(^5\)

In strategic settings, two results are particularly relevant. First, 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.\(^5\)

---

5The 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)).
Unsurprisingly, in the sequential game almost no second-movers cooperated after observing defection by their opponent. 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. While this finding is inconsistent with reciprocity being the sole motivation for cooperation, it 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. Second, our model is in line with a well-documented belief asymmetry: subjects who defect are more likely to predict that their opponents will also defect. (Further discussion of experimental evidence is found in Section 4.)

Throughout, we refer to magical thinking as the belief that one’s action choice influences one’s opponent’s action choice. A related, and observationally equivalent, notion is found in a normative debate in philosophy that concerns Newcomb’s Paradox (Nozick, 1969; Gardner, 1974) and extends to the PD if one presumes a notion of self similarity as described by Rubinstein and Salant (2014) (and citations therein). 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 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.

In contrast, causal decision theorists, who presumably include nearly all economists, 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 (see Joyce (1999) for a thorough treatment).

We mention this debate here not because we will participate in it—the nature of the behavioral data we consider rules out that magical thinking be anything other than a cognitive error—but to highlight that a number of intelligent, serious individuals have reasoned in such a manner. Finally, a similar idea is apparent in common casual reasoning, such as, “I contribute/recycle/volunteer because if I didn’t, then how could I believe that others are...”

---

6 While this may seem consistent with a reciprocity motive—players cooperate if and only if they expect cooperation—it is inconsistent with an equilibrium model of reciprocity (see Section 4 for more).

7 Here self-similarity, or projection bias, refers to the belief that others are likely to make similar judgements and choices as one’s self, though these concepts can be more general—for example, Madarász (2012, 2014) studies models in which agents project that others will have access to the same information as they do.

8 Nozick (1969) observes about Newcomb’s Paradox that, “To almost everyone, it is perfectly clear what should be done. The difficulty is that these people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly.” Experimental evidence suggestive of evidentiary reasoning is found in Quattrone and Tversky (1984), and a particular operationalization of such reasoning is studied in Masel (2007).
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. Also, 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 illustrated 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. In addition, the interpretation of magical thinking is more in line with the evidence discussed above and in Section 4.

In an evolutionary setting with cooperating and defecting types, Orbell and Dawes (1991) explain the survival of cooperating types when agents both decide whether to play PDs or opt out, and assign greater likelihood to others being of their same type. Finally, Bernheim and Thomadsen (2005) consider an alternative explanation for behavior that appears to result from magical thinking, but is instead based upon limited memory, anticipatory utility, and coordinating signals.

Of course, as we stress throughout, this paper is also distinguished by providing a tight axiomatic characterization of our behavioral model. It is worth pointing out that, while we focus on symmetric 2×2 games for tractable axiomatic analysis, our conception of magical thinking in games can be extended to other environments. At a very general level, the two key components are that players believe their actions choices have stochastic influence over the decisions of others and that equilibrium beliefs are biased as a result (evidence for each is discussed in Sections 2 and 4). In principle, players could have arbitrary (magical) beliefs about how their choices affect others; however, we believe a natural starting point is for them 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, 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). In addition, in our model players believe that their action choice, no matter how it is reached (e.g., through private randomization or based on a privately known type), may influence the action choice of others. This is one reason we reach very different predictions about play in the PD compared to Roemer (2010), wherein players evaluate mixed strategy profiles compared to others that feature proportionately higher or lower probabilities of defection by both players.
symmetric games provide a setting in which “the same action” is meaningful strategically. Next, in games with more players, there would be the added modelling choice of which other players $i$ believes he is influencing and whether or not there is correlation in his perceived influencing. 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. These modeling choices will likely need to be tailored to the setting at hand, but the two key features would remain.

3 Prisoners’ Dilemma Games

3.1 The Domain

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). 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 $p - y$ 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 $g \in PD^0$, or, if it is useful to be more explicit about its payoffs, as $(r, p, x, y)$.

![Figure 1: An Arbitrary Prisoners’ Dilema in $PD^0$](image)

We will consider a finite collection of players, indexed by $I := \{1, ..., n\}$, 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.

3.2 The Behavioral Model

We present the behavioral model, or as-if representation, first and then the axioms in Section 3.3. Compared to axiomatic theories of individual choice, the most notable procedural difference is the necessity to conduct equilibrium analysis (Section 3.2.2) in order to apply our representation.\footnote{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. However, each choice problem can be interpreted as a single-player game, with the notions of optimization and...}
3.2.1 Setup

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 \(g \in PD^0\), player \(i\) evaluates the expected payoff of action \(a_i \in \{c, d\}\) as:

\[
V_i(c) = \alpha_i \cdot r + (1 - \alpha_i) \left[ P_i \cdot (p - y) + (1 - P_i) \cdot r \right]
\]

\[
V_i(d) = \alpha_i \cdot p + (1 - \alpha_i) \left[ P_i \cdot p + (1 - P_i) (r + x) \right]
\]

(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 \(\alpha_i\) that his opponent will match whatever 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 \(\alpha_i\).

Given a game \(g \in PD^0\), player \(i\)’s strategy is a mapping \(\sigma_i : [0, 1] \to [0, 1]\), where \(\sigma_i(\alpha_i)\) is the probability that player \(i\) chooses \(d\). Throughout, we consider only symmetric equilibria, defined as follows.

Definition 3.1 Fix any CDF \(F\) and \(g \in PD^0\). An equilibrium is a pair \((\sigma, P)\), where \(\sigma\) is a measurable function from \([0, 1]\) to \([0, 1]\) and \(P \in [0, 1]\), such that, for all \(i \in I\),

1. \(\sigma_i = \sigma\)

2. The support of \(\sigma(\alpha_i)\) is a subset of \(\arg\max\ \{V_i(c), V_i(d)\}\), as given by (1).

3. \(P = \int_0^1 \sigma(\alpha) dF(\alpha)\)

The first two requirements are standard: the first is the symmetry condition, and the second states that each player must choose an action that yields him the highest expected payoff, given his 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 play identically to \(i\). \[\text{equilibrium coinciding. Therefore, the standard result is identical to showing that the axioms are equivalent to the decision-maker playing an equilibrium in every (single-player) game where payoffs are defined by the utility representation.} \]

\[\text{Because players in the model seek to maximize their expected payoff (albeit, with non-standard beliefs),} \]
3.2.2 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 \((\sigma, P)\) is an equilibrium of the game \((r, p, x, y) \in PD^0\), then it is also an equilibrium of the game \(\kappa(r + \xi, p + \xi, x, y) \in PD^0\) for all \(\kappa > 0\) and \(\xi \in \mathbb{R}\).

All proofs are located in Appendix A. From the lemma, the set of equilibria is identical in games \((r, p, x, y)\) and \((1, 0, \frac{r}{r-p}, \frac{y}{r-p})\); the latter being the positive affine transformation of the former with \(\kappa = \frac{1}{r-p} > 0\) and \(\xi = -p\). Let \(PD \subset PD^0\) denote the subset of games in which \(r\) and \(p\) are normalized to 1 and 0, respectively, with \((x, y) \in PD\) 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 3.2.

**Definition 3.2** An equilibrium is a **cutoff equilibrium** if its strategy profile is of the form: \(\sigma(\alpha) = 1\) if \(\alpha < \alpha^*\), and \(\sigma(\alpha) = 0\) if \(\alpha > \alpha^*\), for some \(\alpha^* \in [0, 1]\).

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

Fixing any \((x, y) \in PD\), 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 \(\alpha_i\). Then, in equilibrium, \(P_i = P = F(\alpha^*)\), and the cutoff type, \(\alpha^*\), is indifferent between \(c\) and \(d\). So, the set of equilibria is identical to the set of solutions to the equation\(^{12}\)

\[
V_i(c|\alpha_i = \alpha^*) = \alpha^* \cdot 1 + (1 - \alpha^*) [F(\alpha^*) \cdot (-y) + (1 - F(\alpha^*)) \cdot 1]
= \alpha^* \cdot 0 + (1 - \alpha^*) [F(\alpha^*) \cdot 0 + (1 - F(\alpha^*)) \cdot (1 + x)] = V_i(d|\alpha_i = \alpha^*) \quad (2)
\]

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**).

---

\(^{12}\)Definition 3.2 does not specify the behavior of the cutoff type, who is indifferent between \(c\) and \(d\). We will not always distinguish equilibria that have the same cutoff, but in which the cutoff type plays different strategies since this type has measure zero and the distinction has no effect on payoffs.
Figure 2: A distribution $F \in \mathcal{F}$ that does not satisfy Condition $S$. An equilibrium exists for all $(x, y) \in PD$. The equilibrium is unique for the games in panels (a) and (b), but the game in panel (c) has three equilibria.

Condition $S$ restricts how steep $F$ can be, by limiting its reverse hazard rate, in a manner that depends on $\alpha$. For example, the CDF $F(\alpha) = \alpha^{1/k}, k \geq 1$, satisfies the condition, even though $F'(\alpha) F(\alpha) \to \infty$ 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 better understand Proposition 2, a bit of algebra reveals that for $x \neq y$, (2) is equivalent to

$$F(\alpha^*) = T(\alpha^*|x, y) := \frac{\alpha^* - (1 - \alpha^*) x}{(1 - \alpha^*)(y - x)}. \quad (3)$$

When $x > y$, $T$ is strictly decreasing (and continuous) in $\alpha$. It must intersect $F$, a continuous CDF, exactly once. So, Condition $S$ is not required for uniqueness on this subset of games.

However, when $x < y$, $T$ is strictly increasing in $\alpha$. Now there is the potential for multiple solutions and, in fact, there will be for some games if $F$ does not satisfy Condition $S$. To gain intuition for the potential multiplicity of equilibria, first note that for type $\alpha_i$, defection carries the cost of $r - p = 1$ with perceived probability $\alpha_i$, 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 increasing in $F(\alpha^*)$ (i.e., the probability that one’s opponent defects if his choice is made independently). 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 to maintain uniqueness are those with the smallest $x$ values, which is used to derive the tightness of Condition $S$ for uniqueness (see the proof in Appendix A).

\[\text{For } x = y, (2) \text{ has a unique solution: } \alpha^* = \frac{x}{1+x} = \frac{y}{1+y}.\]
3.3 The Axioms

We now present the axioms, doing so without reliance on the model. The data we consider is 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 $M_i^0 = PD^0 \setminus (D_i^0 \cup C_i^0)$ for which $i$ is indifferent in his choice of $d$ or $c$. We denote by $\overline{D}_i^0 = PD^0 \setminus C_i^0$ and $\overline{C}_i^0 = PD^0 \setminus D_i^0$ the sets of games for which $i$ weakly prefers $d$ or $c$, respectively. The primitive of our analysis is the collection of pairs $(D_i^0, C_i^0)_{i \in I}$, which fully summarizes the behavior of all players in $I$.

Our first four axioms consider individual behavior. It can be noted that a player who adheres to the standard prediction of always defecting, $D_i^0 = PD^0$, 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 (for example, Jones et al. (1968)) and also in many other games (Camerer and Hogarth, 1999; Kocher et al., 2008). 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. Of course, the axiom is also consistent with the interpretation of game-payoffs as vNM utilities as is customary in game theory.

---

14 As in contexts of individual choice on a continuous domain, one might employ a finite grid on $PD^0$ to elicit preferences in the lab. Players can then be randomly matched, and one game on the grid selected randomly to pay players based on. If players expressed strict preference, their choice is implemented. If a player expressed indifference in the game, his action will be implemented randomly with weights of the player’s own choice. This is necessary to accommodate that the randomization may matter to other players (e.g., the collection may be playing a mixed strategy equilibrium), an issue that does not arise in the single-agent context.

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 $g \in PD^0$, $i$ ranks the actions in $\{d, c\}$. Each ranking is a complete binary relation $\succ_i^g$. 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 These vNM utilities can be elicited from choices between lotteries over physical allocations, and then used to translate the nominal payoffs of any physical game into utility payoffs.
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 \(PD \subset PD^0\). 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 \(D_i = D_i^0 \cap PD\), and analogously for \(C_i, M_i, D_i^c\), and \(C_i^c\).

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 \((x, y) \in D_i\), \((x', y') \geq (x, y)\), 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.

**Axiom 4 (Convexity)**
For all \(i \in I\), if \(\{(x, y), (x', y')\} \subset D_i\) and \(\gamma \in (0, 1)\), then \((\gamma x + (1 - \gamma) x', \gamma y + (1 - \gamma) y') \in D_i;\) and analogously for \(C_i\).

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. By supposition, this change does not alter \(i\)'s strict preference for defection. Axiom 4 states that starting from \((x, y)\), a smaller change in the payoffs—one at the same rate, but on a smaller scale (that of \((1 - \gamma) < 1\)—should, again, not alter player \(i\)'s preference. Of course, this means \(D_i\) and \(C_i\) are convex sets.

---

\[17\] We note two things here. First, 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. Second, similar to Axiom 1, Axiom 4 hints that uncertainty enters a player’s objective linearly. Because subjective payoff uncertainty in our domain would depend on one’s beliefs about the actions of others, it is difficult to motivate either axiom in this way without reference to our behavioral model. However, it may be insightful to see how Axiom 4 is implied by the representation (Theorem 1). Within the behavioral model, players are expected-utility maximizers who view the choice between \(c\) and \(d\) in a given game as a choice between lotteries over payoffs with subjective probabilities given by their (non-standard) beliefs. Letting \(q_{ij}^c, q_{ij}^d\) be the subjective probabilities player \(i\) assigns to \(a_j = c\) in
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. It seems natural that the interpretation of such an axiom would be contingent on at least some basic properties of individual behavior; in our case this will be monotonicity (Axiom 3).

The inter-player 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.

**Definition 3.3** For $H, H' \subset PD$ we write $H < H'$ if, for all $(x, y) \in H$ and $(x', y') \in H'$, $x < x'$ and $y < y'$.

**Axiom 5 (Inter-player 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) $(x' + \varepsilon, y' - \delta) \in C_j$.

![Figure 3: Depiction of Axiom 5](image)

Notice that (i) holds, so (ii)-(iv) imply (v).

To see that the axiom 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.\(^{18}\) Second, (ii) and (iii) imply that $i$ is willing game $g$ given $a_i = c$ and $a_i = d$ respectively, if $g \in M_i$, then $q_d^g(1) + (1 - q_d^g)(-y) = q_c^g(1 + x) + (1 - q_c^g)(0)$. Further, in any two games $g, g' \in M_i$, $q_d^g = q_d^{g'}$ and $q_c^g = q_c^{g'}$. Writing out the analogous indifference equation for game $g'$, one finds that $y - y' = \frac{-q_d^g}{1 - q_d^g}(x - x')$, so $M_i$ forms a line in $PD$ and $C_i$ and $D_i$ are convex. Notice that this linearity in probabilities does not determine the player’s attitude toward risk under the interpretation of game-payoffs as vNM utilities.

\(^{18}\) 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, $\overline{D}_j \cap H \subseteq D_i \cap H$ and $\overline{C}_i \cap H \subseteq C_j \cap H$. 

13
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, $\varepsilon$, in the exploitative motive and the same reduction, $\delta$, 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 now provide two complementary views of the axiom: the first comparing the attitudes of players $i$ and $j$, the second comparing the objective incentives for defection in the games where $i$ and $j$ are, respectively, on the cusp of changing their behavior. In terms of the first, a possible intuition for the pattern that player $i$, who defects more often than $j$, may also be more responsive to the exploitative motive $x$ is that both comparisons seem to capture features of a more aggressive disposition on the part of player $i$.

The second view of the axiom takes into account the behavior of other players. 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.

### 3.4 The Representation Theorem

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

**Definition 3.4** For $I' \subseteq I$, 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 $g \in PD^0$ there exists an equilibrium such that for all $i \in I'$, with $V_i$ as defined by (I):

1. $g \in C_i^0$ if and only if $\{c\} = \text{argmax}_{\{c, d\}}\{V_i(c), V_i(d)\}$

2. $g \in D_i^0$ if and only if $\{d\} = \text{argmax}_{\{c, d\}}\{V_i(c), V_i(d)\}$
Theorem 1 The primitive \((D_0^i, C_0^i)_{i \in I}\) 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 \(i \in I\), \(\alpha_i\) 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 \(\alpha_i\) (the degree to which he exhibits magical thinking) is uniquely determined by the primitive and will, in fact, only depend on \((D_0^i, C_0^i)\) as we sketch below. Further, the quantiles of \(F\) at all \(\alpha_i\) 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 believe the \(\alpha\)-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 Appendix B 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. A reader might therefore object to the terminology that a model can explain behavioral data if the data is always consistent with one of the model’s equilibria (Definition 3.4) 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 \(g \in PD^0\) (and all equilibria are cutoff equilibria—Proposition 1). It is immediate, therefore, that the representation satisfies the following, more stringent requirement.

Definition 3.5 For \(I' \subseteq I\), the behavior of the players in \(I'\), \((D_0^i, C_0^i)_{i \in I'}\), can be strongly explained by the behavioral model \([F, (\alpha_i)_{i \in I}]\) if the requirements of Definition 3.4 must hold in all equilibria.

Corollary 1 Theorem 1 is also valid if “explained” is replaced with “strongly explained.”

\[\text{15}\] 

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.
Figure 4: Panel (a) depicts a player $i$’s behavior in $PD$, for whom $C_i, D_i \neq \emptyset$. Panel (b) depicts the $M$-lines for four distinct players—note how they “fan out.”

### 3.4.1 Sketch of Proof of Theorem 1

A couple of notational definitions will simplify exposition for the remainder of the paper.

**Definition 3.6** Let $F_S$ denote the set of CDFs in $F$ that satisfy Condition $S$. Let $\alpha_{-i}$ denote an arbitrary assignment of types to players in $I \setminus \{i\}$ (i.e., $(\alpha_j)_{j \in I \setminus \{i\}}$).

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 $(x, y) \in PD$ is of cutoff form, where the cutoff, $\alpha^*$, is characterized by (2). To find the set of games in $PD$ for which $i$ is indifferent between $c$ and $d$, we fix $\alpha_i \in (0, 1)$ and solve (2) for $y$ as a function of $x$ to get

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

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

In addition, see 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 $\alpha_i$; the former by Condition $S$, the latter by $F \in F$. 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 4(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 $[F_i, (\alpha_i, \alpha_{-i})]$ satisfying $F \in \mathcal{F}$ and $F(\alpha_i) = F_i$. Further, $\alpha_i$ and $F(\alpha_i)$ 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) $D_i = PD$, ii) $C_i = PD$, or iii) $M_i \neq \emptyset$. 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 $\gamma$ such that $M_i = \{(x, y) \in PD | y = \gamma(x)\}$, $C_i = \{(x, y) \in PD | y < \gamma(x)\}$, and $D_i = \{(x, y) \in PD | y > \gamma(x)\}$. Finally, Convexity of $C_i$ and $D_i$ mean $\gamma$ is linear, so can be summarized by two scalars that we denote $int_i$ and $slp_i$: $M_i = \{(x, y) \in PD | y = int_i - slp_i \cdot x\}$.

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

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

Inverting the bijection $(int_i, slp_i) = \left( \frac{\alpha_i}{1 - \alpha_i} F_i, \frac{1 - F_i}{F_i} \right)$ establishes the first part of the proof.

For the second part, consider two players $i$ and $j$, such that $M_i, M_j \neq \emptyset$ and who satisfy Axiom 5. This means $int_i < int_j$ implies $slp_i > slp_j$. The translation of this condition under the bijection yields that $0 < \alpha_i < \alpha_j < 1$ implies $F_i < F_j \leq \frac{F_j}{\alpha_j(1 - \alpha_j)}$. 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,

\footnotemark[20]

\footnotetext[20]{To see this, note that $int_i < int_j$ implies that there are games $(x, y), (x + \varepsilon, y - \delta), (x', y')$, and $(x' + \varepsilon, y' - \delta)$ that satisfy (i) $(x, y), (x + \varepsilon, y - \delta) \in M_i$, (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 $slp_i > slp_j$.}
but merely allow for these properties. This is because the data of a finite number of players generates values for $F$ at only a finite number of points (Appendix B 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 3.2). 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.

### 3.5 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 3.7** Let $|A|$ be the size of any finite set of players $A$. Consider two arbitrary sets of players $A$ and $\tilde{A}$ such that $|A| = |\tilde{A}|$.

- We say that, in $H \subset PD$, the set of players $A$ defects (weakly) more than $\tilde{A}$ if $|\{i \in A | (x, y) \in D_i\}| \geq |\{j \in \tilde{A} | (x, y) \in D_j\}|$ 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 $\tilde{A}$ if $A$ defects more than $\tilde{A}$ in $\{(x, y) | x \geq y\}$ and $\tilde{A}$ defects more than $A$ in $\{(x, y) | x \leq y\}$.

The notion of defects more is straightforward. For singletons $A = \{i\}$ and $\tilde{A} = \{j\}$, it is simply: in $H \subset PD$, player $i$ defects (weakly) more than player $j$ if $D_j \cap H \subseteq D_i \cap H$. 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 $\tilde{A}$ 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 $F \in F_S$, a player of type $\alpha$ cooperates more in $PD$ than does a player of type $\tilde{\alpha}$ if and only if $\alpha \geq \tilde{\alpha}$. 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 representation.

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

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

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

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

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

e) For any $(x, y) \in PD$, $\alpha_{x,y}^* \leq \tilde{\alpha}_{x,y}^*$ 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 $\tilde{F}$ (see Appendix B), or if she is interested in understanding how much cooperation the players themselves predict as their common belief about the distribution of $\alpha$-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 $\alpha$ 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. $\tilde{F}$ means more cooperation from the $F$ population than from the $\tilde{F}$ 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 $\alpha$-type takes advantage of increased cooperation in the populace by defecting in more games when facing $F$ than when facing $\tilde{F}$.

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 analysts 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 3.2.2), 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 \(\tilde{I}\) such that \(|I| = |\tilde{I}| = n\), and whose behavior is described by \([F, (\alpha_j)_{j \in I}]\) and \([\tilde{F}, (\tilde{\alpha}_j)_{j \in \tilde{I}}]\), respectively, with \(F, \tilde{F} \in \mathcal{F}_S\) and each collection ordered by increasing \(\alpha\)-values.

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

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

### 4 Play in the PD: Evidence and Alternative Theories
Starting with [Rapoport and Chammah (1965)](http://dx.doi.org/10.1086/265823) (henceforth RC65), experimentalists have attempted to understand how the payoffs in the Prisoners’ Dilemma affect observed levels of cooperation (both in the aggregate and within-subject). The main findings are that a substantial proportion of subjects choose to cooperate (see [Dawes and Thaler (1988)](http://dx.doi.org/10.1086/265823) for a survey), but that cooperation monotonically decreases with the motives to defect: \(x\) and \(y\).

For example, [Charness et al. (2014)](http://dx.doi.org/10.1086/265823) 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](http://dx.doi.org/10.1086/265823); [Engel and Zhurakhovska, 2012](http://dx.doi.org/10.1086/265823)), consistent with Axiom 3. In addition, for the stakes used in experiments, evidence indicates that a positive affine transformation of all payoffs has little effect on the level of cooperation in the Prisoners’ Dilemma (for example, [Jones et al., 1968](http://dx.doi.org/10.1086/265823)), or on play in games more generally ([Camerer and Hogarth, 1999](http://dx.doi.org/10.1086/265823); [Kocher et al., 2008](http://dx.doi.org/10.1086/265823)), consistent with Axiom 1.
Again starting with [RC65], 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 is increasing in $\frac{r-p}{r-p+x+y}$ (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 linear, as they are in our model, owing to Axioms 2-4. The empirical support for these indices then provides evidence in support of Axioms 1-4. However, the indices do not predict the differing slopes of these level curves as implied by Axiom 5 (illustrated in Figure 4(b)), meaning our axioms/model provide a more nuanced prediction. We are unaware of studies that provide detailed enough data to test the predictions of our model against the predictions based on these indices.

We can distinguish the predictions of our model from alternative models based on other-regarding preferences in two ways. First, as pointed out in the Introduction and verified in the Supplemental Appendix, canonical models with altruism, inequity-aversion, and reciprocity violate our axioms, and hence make different predictions on our domain.21 Second, we now discuss how our predictions differ in settings that are natural extensions of our domain (which for brevity we do not formalize here):

1. The most obvious difference is that the 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., 2011).

2. Shafir and Tversky’s (1992) observation that the level of cooperation by second-movers is lower in the sequential Prisoners’ Dilemma than in the standard, simultaneous-move version of the game—even if the first-mover cooperates (see Section 2)—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.

21 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 one, 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.
cooperate following cooperation than in the simultaneous-move game.  

3. In a similar vein, Morris et al. (1998) find that the temporal order of moves affects cooperation levels 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 provide no rationale for this discrepancy, as play should be invariant to this strategically irrelevant difference in the games.

4. 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. This feature is implied by the interpretation of our model (see Figure 5), but absent from models with other-regarding preferences. While, there may seem to be a sense in which this is consistent with reciprocity—players are more likely to cooperate when they expect cooperation from others—it is clearly inconsistent with standard notions of equilibrium, even if players care about reciprocity. Either the cooperators are too optimistic or the defectors are too pessimistic about their opponents’ behavior.

Figure 5: Predicted probability of opponent defection as it varies with \( \alpha \) for fixed \( g \in PD^0 \).

---

22 In Shafir and Tversky (1992) 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.

23 See Dawes et al. (1977); Orbell and Dawes (1991); Engel and Zhurakhovskaya (2012); Rubinstein and Salant (2014). Rubinstein and Salant (2014) suggest that players’ ex-post reported beliefs will more accurately reflect the beliefs their choices were based upon in Prisoners’ Dilemma games, which feature a dominant strategy, than in games such as Hawk-Dove, where either action is a best response to some belief. In the latter case, they argue, a player might want to “strategically justify” his play ex post, thereby biasing his reported beliefs in favor of the opponent’s action that makes his own play the best response.
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 goods 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 PD⁰, 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 (1)-(4) above. Third, in the subsequent sections 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 in order 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.

Remark 1 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 non-zero level. For a sample of studies Table 1 reports each of their featured one-shot games $(x, y) \in PD$ (modulo positive affine transformations), the approximate number of periods after which stabilization was reached, and the approximate average levels of cooperation thereafter. 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 own $\alpha$-types based on play. Because they are not in fact magical, the updating will be systematically biased downward, leading subjects to cooperate less.

5 More Games

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 now 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, $(D^0_i, C^0_i)_{i \in I}$, as well as $(M^0_i, \overline{D}^0_i, \overline{C}^0_i)_{i \in I}$, be extended to this larger class of games in the obvious way. Note that $c$ (respectively, $d$) still corresponds to the action leading to the better (worse) symmetric outcome, but outside the Prisoners’ Dilemma we no longer refer to it as “cooperate” (“defect”). Finally, define $G = \{(r, p, x, y)|r = 1, p = 0, x > 0\} \subset G^0$, with arbitrary element $(x, y)$, and, as before, $D_i = D^0_i \cap G$ and analogously for $C_i, M_i, \overline{D}_i$, and $\overline{C}_i$.

In addition to Prisoners’ Dilemma games, $G^0$ contains other well-known games. Hawk-Dove (also known as Chicken) games are the subset of $G^0$ in which $y \in (- (r - p), 0)$. Action $c$ corresponds to $dove$ and $d$ to $hawk$. Though not always labeled as a symmetric game, any Battle of the Sexes game can be approximated within $G^0$: for appropriately chosen $x > 0 > y, x \neq -y$, fix $p = 0$ and take the limit as $r \to 0$. That is, $c$ and $d$ do not correspond

---

26This could perhaps be because the $\alpha_i$ 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 subjects pools or other environmental features. That play stabilizes at non-zero levels of cooperation suggests that the lower bound on $\alpha_i$ is believed to be positive by some $i$.  

Table 1: Studies of One-Shot Prisoners’ Dilemma with Random, Anonymous Rematching
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).

It will be useful to compare the predictions of our model to Nash equilibrium in the standard model (hereafter, simply Nash equilibrium). It is easy to derive that every game \( g \in G^0 \) has a unique symmetric Nash equilibrium.

5.1 The Behavioral Model

In terms of the behavioral model, nothing in its setup needs to be altered to accommodate all games in \( G^0 \). Further, the equilibrium characterization results in Lemma 1 and Propositions 1, 2, and 3 remain valid by simply replacing each instance of \( PD^0 \) (respectively, \( PD \)) with \( G^0 (G) \).

In addition, we establish the following result. For any game \( g \in G^0 \), let \( \pi_g \) be the probability with which a player selects \( d \) in the symmetric Nash equilibrium of \( g \). The corresponding probability in our behavioral model is \( F(\alpha^*_g) \).

**Proposition 5** For any \( g \in G^0 \), \( \pi_g > F(\alpha^*_g) \). 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 producing 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 \( \alpha \)-types still differ in their expectations over opponent play). Additionally, these limits constitute equilibria of the behavioral model when \( r = p \), as in Battle of the Sexes games.

Notably, the common prediction of our model and the standard model is substantiated by the experimental studies of Battle of the Sexes games. Hence, we provide a unified model that accommodates the experimental evidence in the Prisoners’ Dilemma (see Section 4), the Hawk-Dove game, and the Battle of the Sexes. This serves as another key distinction between our model and models of other-regarding preferences discussed in Section 4, each of

\[27\] Moreover, if \( r = p \) in \( g \), then in any equilibrium of the behavioral model the (\( F \)-measured) likelihood of \( d \) is \( \pi_g \). However, if there is no payoff difference between the symmetric outcomes, the cutoff property is no longer a requirement for equilibrium—if \( r = p \) there is no reason that players with higher \( \alpha \)-types are more drawn to \( c \). Hence, any strategy \( \sigma \) such that \( \int_\alpha \sigma(\alpha)dF(\alpha) = \pi_g \) constitutes an equilibrium.

\[28\] Camerer (2003, Chap 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.”
which predict patterns of play in the Battle of the Sexes that differ from the prediction of the standard model.\footnote{This result is not difficult to demonstrate. We omit the analysis for the sake of brevity.}

In terms of the usual interpretation given to a Battle of the Sexes game, for a magical thinker, both $c$ and $d$ are self-defeating: by being “selfish” and choosing your preferred event, you believe it more likely that your opponent will also choose his preferred event, but also likewise if you try to be “accommodating” and choose your opponent’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.\footnote{As in the standard model, there exist asymmetric equilibria as well. However, such equilibria seem implausible in a truly symmetric, anonymous, one-shot setting, as neither player would have any way of knowing if he were “player 1” or “player 2.”}

One could imagine an alternative conception of magically 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. While, \textit{a priori}, one may have little preference for one conception over the other, this is precisely why models and axioms are useful. Our theory, applied to a large class of games, leads us to our conception of magical thinking in the Battle of the Sexes. In addition, this alternative model would predict that players select their preferred event more frequently than is found in the experimental evidence.

5.2 The Axioms and Representation

Notice that each of the Axioms \ref{ax:1}--\ref{ax: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 $c$.

\textbf{Axiom 6 (Sensitivity to Benefits from Action $c$)}

\begin{quote}
For all $i \in I$, if $(r, p, x, y) \in C^0_i$ and $r' > r$, then $(r', p, x, y) \in C^0_i$.
\end{quote}

It is not difficult to show that the representation in Theorem 1 satisfies Axiom 6 on $PD^0$, meaning Axiom 6 is implied by Axioms \ref{ax:1}--\ref{ax:5} on this domain. On $G^0$, this is no longer the case (a fact we verify in Appendix \ref{app:1}, Fact \ref{fact:1}). Notice that the axiom is consistent with
the experimental evidence discussed in the preceding subsection. Further, in line with the axiom, RC65 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 $r$.32

By adding Axiom 6, the representation result of Theorem 1 extends to $G^0$.

**Theorem 2** The primitive $(D_i^0, C_i^0)_{i \in I}$, on $G^0$, satisfies Axioms 1-6 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 $i \in I$, $\alpha_i$ and $F(\alpha_i)$ are unique.

The extended representation also satisfies the stronger criterion of Definition 3.5.

**Corollary 2** Theorem 2 is also valid if “explained” is replaced with “strongly explained.”

### 6 Yet More Games

We now extend the domain to all symmetric $2 \times 2$ games, $\{(r, p, x, y) | r \geq p\}$. First, for the (nongeneric) subset in which $r = p$, reasoning analogous to that given in Section 5.1 shows that our model of magical thinking is not behaviorally distinct from the standard model.33

Consider then the (generic) subset $\{(r, p, x, y) | r > p\}$. On this subset, Lemma 1 remains valid, and Axiom 1 seems equally appealing. Therefore, once again, we normalize $r = 1, p = 0$, and refer to a game as $(x, y)$. This allows us to represent the set of games as $\mathbb{R}^2$ in Figure 6. Section 3 analyzed Prisoners’ Dilemma games (quadrant I), and Section 5 analyzed the games in $G$ (the union of quadrant I and IV). Focusing only on $G$, each of the four solid lines in the figure corresponds to a sample $M_i$ for some player $i$, with $C_i$ and $D_i$ being the strict-lower and strict-upper contour sets of $M_i$, respectively.

Let us next consider games in which $x \leq 0$ and $y \leq 0$ (quadrant III of Figure 6). 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. It is straightforward to verify that this behavior corresponds to the unique equilibrium of our behavioral model. In

---

31 It is easy to derive that $\pi_g$ is independent of $r$, meaning 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, their play is consistent with Axiom 6.

32 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 RC65. This evidence is also summarized in Table 1 of Steele and Tedeschi (1967).

33 The only exception is the (measure-zero) set of $\alpha = 1$ players, who are indifferent between $c$ and $d$ for any game in which $r = p$. 

---

27
addition, for each player $i$ such that $M_i \cap G \neq \emptyset$, this behavior is a consequence of Axioms 1-6 when the primitive is likewise extended. Under the (seemingly mild) additional requirement that in the extended domain $\overline{C}_i \neq \emptyset$ for all $i \in I$, the representation result extends with only minor alteration. \footnote{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 6 and maintain the original class of behavioral models.}

The final quadrant of the figure, II ($x \leq 0$ and $y > 0$), 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)).

Our behavioral model is consistent with all players selecting $c$ in all games with $x \leq 0$ and $y > 0$, 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 the lower envelope of the extensions of the $M_i$-lines from quadrant I, as shown in Figure 6 (for the case of $F(\alpha) = \sqrt{\alpha}$). 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

Figure 6: Depiction of the set of symmetric 2×2 games in which $r = 1$ and $p = 0$. 

34If 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 6 and maintain the original class of behavioral models.
inclusion) the more magical thinking there is in the population (in the sense of a first-order
stochastically dominant shift of \( F \)). For games on the other side of the boundary, where
there is a tradeoff between the overall payoffs (higher under \((c,c)\)) and riskiness, our model
does not make a unique prediction and can accommodate a significant proportion of players
selecting \( d \). As with the Prisoners’ Dilemma, Battle of the Sexes, and Hawk-Dove then, our
model comports with the experimental findings.

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 \(|x| < y\). Similarly, our model uniquely predicts the choice of
\( c \) by all players when it is not “too risky,” where “too risky” is characterized by a monotonic
boundary. However, that boundary is more nuanced than the fixed linear one implied by risk
dominance, as it depends on the (perceived) distribution of \( \alpha \)-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.

How can our axioms be extended to the games with \( x \leq 0 \) and \( y > 0 \)? 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 in the figure 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 1-6 on quadrants I and IV.

7 Methodological Discussion and Conclusion

Our approach connects behavioral axioms on the observed play of a collection of players to an *as-if* representation that suggests a procedural interpretation of individual behavior and an equilibrium solution 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 one 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; and 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 (2012)).

Relative to the benefits of our axiomatic approach that were listed as items (1)-(3) in Section 1, the standard approach has the following differences. In contrast to (1), 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 (3), as it is not possible to derive that behavior corresponds to any particular equilibrium notion, or that prior beliefs are common, by axiomatizing the objective function of each player separately. Finally, we have interpreted game-payoffs as private physical payouts (e.g., dollars) to facilitate comparison with experimental findings. Typically, in game theory, game-payoffs are interpreted as vNM utilities over the allocation vector of physical payouts. This alternative view poses no difficulty for our theory. In contrast, the standard approach cannot explain cooperative behavior in PD games if game-payoffs represent vNM utilities over physical allocations, as
then each player should like higher payoffs for himself and be indifferent toward the 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.

Segal and Sobel (2007, 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,” 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. In contrast to the standard approach their model can, for instance, explain cooperation in the Prisoners’ Dilemma even under the interpretation of game-payoffs as vNM utilities over physical allocations. At the same time, their model differs from (1) and (3) 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 (2), this is not data that is commonly collected, and faces the potential difficulty 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 a subject how they would like to play in a particular game under anonymity.

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. 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 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

---

35See 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 non-cooperative 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 expected 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, §12.3), and Blonski et al. (2011) for an application to equilibrium selection in repeated games. Outside the axiomatic literature, Bergemann et al. (2011) 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.
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, both about the opponent’s type and action choice) are derived from behavior in a strategic setting. In future research, we plan to explore the question of behavioral foundations for beliefs and equilibrium play in Bayesian games.

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 (Section 6). While our representation features an equilibrium concept, this does not have to 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 which may, or may not, have the strategic flavor of equilibrium.

36 This is related to a literature exploring econometric procedures for testing and estimating equilibrium models of bidding behavior in auctions, most closely Guerre et al. (2000). However, these papers do not describe the testable/estimable properties of the data as intuitive/empirically plausible axioms independent of the model. In addition, the data is taken to be the entire joint distribution of bids, which is interpreted as an approximation to observing a large number of independent auctions drawn from the same value-generating process.

37 In addition, our behavioral model introduces the possibility of equilibrium multiplicity even in the Prisoners’ Dilemma, 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.
A Proofs

Proof of Lemma 1. Fix any \((r, p, x, y) \in PD^0\), and suppose that \((\sigma, P)\) is an equilibrium according to Definition 3.1. Then, for any player \(i\),

\[
V_i(c) - V_i(d) = \alpha_i[r - p + (1 - P)x + Py] - [(1 - P)x + Py],
\]

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 \(\alpha_i\) when the payoffs are transformed to \(\kappa(r + \xi, p + \xi, x, y)\), where \(\kappa > 0\). Then,

\[
V_i(c) - V_i(d) = \alpha_i[\kappa r + \kappa \xi - \kappa p - \kappa \xi + (1 - P)\kappa x + P\kappa y] - [(1 - P)\kappa x + P\kappa y] = \kappa(\alpha_i[r - p + (1 - P)x + Py] - [(1 - P)x + Py])
\]

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

Proof of Proposition 1. Claim (i): Fix any \((x, y) \in PD\), and suppose that \((\sigma, P)\) is an equilibrium according to Definition 3.1. Then,

\[
V_i(c) - V_i(d) = \alpha_i[1 + (1 - P)x + Py] - [(1 - P)x + Py]
\]

Player \(i\) strictly prefers \(c\), strictly prefers \(d\), or is indifferent if (6) is positive, negative, or zero respectively. For any \(P \in [0, 1]\), if \(\alpha_i = 1\), then (6) is positive, and ii) (6) is linear in \(\alpha_i\). It follows that the equilibrium must be a cutoff equilibrium and that \(\alpha^* < 1\). Suppose now that \(\alpha^* = 0\). Then, by Definition 3.1 \(P = 0\). 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). The argument given in Section 3.2.2 demonstrates the existence of a solution.

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 Section 3.2.2 demonstrate uniqueness of the solution for any \(F \in F\). 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 \(\alpha^* \in (0, 1)\).

\[
F'(\alpha^*) \leq \frac{F(\alpha^*)}{\alpha^* - (\alpha^*)^2} = \frac{T(\alpha^*|x, y)}{(\alpha^* - (\alpha^*)^2) = \frac{\alpha^*(1 + x) - x}{\alpha^*(1 - \alpha^*)^2(y - x)}
\]

Further,

\[
T'(\alpha^*|x, y) = \frac{1}{(1 - \alpha^*)^2(y - x)}
\]

It is a matter of simple algebra to see that the rightmost term in (7) is strictly less than (8) for any \(x, y, \alpha^*\) such that \(0 < x < y\) and \(\alpha^* \in (0, 1)\). Hence, at any solution to \(T\), \(T\) 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 \((x, y) \in PD\) 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 \( \alpha \) 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 \( x \), \((x, Y(x|\alpha, F(\alpha))) \in PD \). Finally, it is straightforward that

\[
\lim_{x \to 0} \left( T'(\alpha|x, Y(x|\alpha_0, F(\alpha_0))) \right)_{|\alpha=\alpha_0} = \frac{F(\alpha_0)}{\alpha_0 - \alpha_0^2}
\]

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 \( \alpha_0 \) for the game \((x, Y(x|\alpha_0, F(\alpha_0))) \). ■

**Proof of Theorem 1.**

Representation \( \implies \) 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 \( A-2 \) is satisfied. To verify that the primitive satisfies the remaining axioms it is sufficient to focus only on \( PD \) and \((D_i, C_i)_{i \in I}\).

Proposition 1 and 2 immediately imply the following. First, if \( \alpha_i = 0 \), then \( D_i = PD \), and if \( \alpha_i = 1 \), then \( C_i = PD \). Second, if \( \alpha_i \in (0,1) \), then \( M_i = \{(x,y) \in PD | \alpha_{x,y}^* = \alpha_i \} \). 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

\[
M_i = \{(x,y) \in PD | \alpha_{x,y}^* = \alpha_i \} = \left\{ (x,y) \in PD | y = \frac{\alpha_i}{1-\alpha_i} F(\alpha_i) - x \left( \frac{1 - F(\alpha_i)}{F(\alpha_i)} \right) \right\}
\]

That is, \( M_i \) forms a line in \( PD \). Define, \( int_i = \frac{\alpha_i}{(1-\alpha_i)F(\alpha_i)} \), and \( slp_i = \frac{1-F(\alpha_i)}{F(\alpha_i)} \). It follows that if \( 0 < \alpha_j < \alpha_i < 1 \), then \( int_i \geq int_j \), and \( slp_i < slp_j \). 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,

\[
\frac{d}{d\alpha} \left( \frac{\alpha}{(1-\alpha)F(\alpha)} \right) = \frac{F(\alpha) - (\alpha - \alpha^2)F'(\alpha)}{(1-\alpha)F(\alpha)^2} \geq 0 \forall \alpha \in (0,1) \iff \text{Condition } S
\]

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 \((x,y) \in MU_i \). From Proposition 2 there exits unique \( \alpha_{x,y}^* \), and it is distinct from \( \alpha_i \) by \((x,y) \notin M_i \). From the argument above, whenever \( \alpha_j \leq \alpha_i \) then \( ML_j \subseteq ML_i \). Therefore, \( \alpha_{x,y}^* > \alpha_i \). By the cutoff form of the equilibrium, \((x,y) \in D_i \). Therefore, \( D_i = MU_i \). An analogous argument establishes \( C_i = ML_i \).

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 \( C_i = \{(x,y) \in PD | y < int_i - x \cdot slp_i \} \) and \( D_i = \{(x,y) \in PD | y > int_i - x \cdot slp_i \} \) 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
\((x', y') \in MU_i = D_i\). To verify Axiom 4, suppose both \((x, y)\) and \((x', y')\) are elements of \(D_i\). Then,

\[
\begin{align*}
y > int_i - x \cdot slp_i \implies \gamma y > \gamma (int_i - x \cdot slp_i) \\
y' > int_i - x' \cdot slp_i \implies (1 - \gamma)y' > (1 - \gamma)(int_i - x' \cdot slp_i)
\end{align*}
\]

\[
\implies \gamma y + (1 - \gamma)y' > \gamma (int_i - x \cdot slp_i) + (1 - \gamma)(int_i - x' \cdot slp_i)
\]

\[
\implies \gamma y + (1 - \gamma)y' > int_i - (\gamma x + (1 - \gamma)x') slp_i
\]

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 \(C_j = PD\) and the axiom is trivial. If \(\alpha_j < 1\), then \(0 < \alpha_i < \alpha_j\) implies \(slpj < slpi\) (above). Further, by hypotheses (ii) and (iii), \(slpi \leq \frac{\delta}{\varepsilon}\). Finally, \(slpj < \frac{\delta}{\varepsilon}\) and \((x', y') \in C_j\) 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–A.2 demonstrate that if \((D_i, C_i)\) satisfies Axioms 2–4 then there is a unique value for \(\alpha_i\) and a unique scalar \(F_i\) such that any behavioral model \([F, (\alpha_i, \alpha_{-i})]\) with \(F \in F\) and \(F(\alpha_i) = F_i\) can explain the behavior of player \(i\). Lemmas A.3–A.5 then show that there exists \(F \in F_S\) that simultaneously satisfies the required values for all \(i \in I\). Therefore, by Proposition 2 for all \((x, y) \in PD\), 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 \(\alpha_i\)'s 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 4 and \(C_i \neq \emptyset, D_i \neq \emptyset\), then for any \((x, y) \in C_i\), \((x', y') \in D_i\), and continuous path \(p : [0, 1] \to PD\) 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 \(\bar{t} = \sup\{t | p(t') \in C_i \land t' \in [0, t]\}\). Because \((x, y)\) is an arbitrary element of \(C_i\), it is sufficient to show that \(p(\bar{t}) \in M_i\). Suppose that \(p(\bar{t}) \not\in C_i\). Then, by definition of \(\bar{t}\), for any \(\varepsilon > 0\) there exists \(t \in (\bar{t}, \bar{t} + \varepsilon)\) such that \(p(t) \not\in C_i\). 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 \(p(t) \not\in D_i\). 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 3–4. If \(D_i \neq \emptyset\) and \(C_i \neq \emptyset\), then there is a unique pair \((int_i, slp_i) \in (0, \infty)^2\) such that \(D_i = \{(x, y) \in PD | y > int_i - slp_i \cdot x\}\) and \(C_i = \{(x, y) \in PD | y < int_i - slp_i \cdot x\}\). If \(D_i = \emptyset\), then \(C_i = PD\); and if \(C_i = \emptyset\), then \(D_i = PD\).

**Proof.** Consider the three possible cases.

Case 1: \(D_i \neq \emptyset\) and \(C_i \neq \emptyset\): Axiom 2 implies that not only is \(M_i\) nonempty, but non-singleton (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 see that \(M_i \cap \{(x, y) \in PD | x \in [x_1, x_2]\}\) must consist of a strictly decreasing function \(y\), where \(\delta y(x_1) = y_1\) and \(\delta y(x_2) = y_2\). For any \(x \in [x_1, x_2]\), if \(y > \delta y(x)\) then \((x, y) \in D_i\), and if \(y \in (0, \delta y(x))\) then \((x, y) \in C_i\). Hence, Axiom 1 implies that \(\delta y\) is linear. Let \(slp_i := \frac{y_1 - y_2}{x_2 - x_1} \in (0, \infty)\) and \(int_i := (y_1 + slp_i \cdot x_1) \in (0, \infty)\).
Since the above applies to any pair of games in $M_i$, all games in $M_i$ must fall on the same line: $M_i \subseteq \{(x, y) \in PD | y = \text{int}_i - \text{slp}_i \cdot x\}$. But, if the inclusion were strict, Axiom $2^2$ would be violated (again, see Fact $A.1$). Hence, $M_i = \{(x, y) \in PD | y = \text{int}_i - \text{slp}_i \cdot x\}$. The claimed structures of $D_i$ and $C_i$ follow from Axiom $3^3$.

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

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

Lemma A.2 Fix any player $i$. If $(D_i, C_i)$ satisfies Axioms $2^2, 4^4$, 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 $F \in F$ and $F(\alpha_i) = F_i$. Further, $(\alpha_i, F_i)$ is given by the following:

$$
(\alpha_i, F_i) = \begin{cases} 
\left(\frac{\text{int}_i}{1 + \text{int}_i + \text{slp}_i}, \frac{1}{1 + \text{slp}_i}\right) & \text{if } D_i, C_i \neq \emptyset \\
(1,1) & \text{if } D_i = \emptyset \\
(0,0) & \text{if } C_i = \emptyset
\end{cases}
$$

(9)

Proof. First, $\text{int}_i, \text{slp}_i > 0$ (Lemma $A.1$) implies that $(\alpha_i, F_i)$ from (9) is always in $[0,1]^2$. Consider, again, the three possible cases.

Case 1: $D_i \neq \emptyset$ and $C_i \neq \emptyset$: Recall from Proposition $1^1$ that in the behavioral model, for any $(x, y) \in PD$, each equilibrium is of cutoff form, with $\alpha^*_{x,y}$ being any solution to (2). So, it is sufficient to show that for arbitrary $(x, y) \in PD$ and $F \in F$ such that $F(\alpha_i) = F_i$: i) $(x, y) \in M_i$ if and only if $\alpha_i$ solves (2); ii) $(x, y) \in C_i$ implies that there exists $\alpha \in (0, \alpha_i)$ such that $\alpha$ solves (2); iii) $(x, y) \in D_i$ implies that there exists $\alpha \in (\alpha_i, 1)$ such that $\alpha$ 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 $y = \frac{F(\alpha_i) - \alpha_i}{F(\alpha_i) - F(\alpha)} x$. The pair of equations $\text{int}_i = \frac{\alpha_i}{F(\alpha_i) - \alpha_i F(\alpha)}$ and $\text{slp}_i = \frac{1 - F(\alpha_i)}{F(\alpha)}$ 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 (9), it follows that $d(\alpha_i) > 0$. Notice that $d(0) = x (F(0) - 1) - y F(0) = -x < 0$. $F \in F$ 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 Proposition $1^1$).

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 (9), it follows that $d(\alpha_i) < 0$. Notice that $d(1) = 1$. $F \in F$ 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 Proposition $1^1$).

The following is relevant for the next two cases. In the behavioral model, for any $F \in F$, Proposition $1^1$ 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 $F \in F$
and any \( \alpha \in (0, 1) \), the game \((x, y) = \left( \frac{\alpha}{1-\alpha}, \frac{\alpha}{1-\alpha} \right) \) is in PD and has a unique equilibrium cutoff \( \alpha^*_x = \alpha \).

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

Case 3: \( C_i = \emptyset \): From above, in the behavioral model, for any \( F \in \mathcal{F} \), a player \( i \) strictly prefers \( d \) in every \((x, y) \in PD \) 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 \( D_i, C_i, D_j, C_j \) 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, \( M_i = \{(x, y) \in PD | y = \text{int}_i - \text{slp}_i \cdot x \} \) and \( M_j = \{(x, y) \in PD | y = \text{int}_j - \text{slp}_j \cdot x \} \). Fix any \((x, y) \in M_i\) and for \( \varepsilon \in (0, \frac{\text{int}_i}{\text{slp}_i}) \), 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 \( \varepsilon \) there exists \((x', y')\) such that \((x', y') \in M_j\) and \( \{ (x, y), (x + \varepsilon, y - \delta) \} \subset \{ (x', y'), (x' + \varepsilon, y' - \delta) \} \). By Axiom 5, \((x' + \varepsilon, y' - \delta) \in C_j\). \((x, y) \in PD \) if \( \text{int}_j - \text{slp}_j \cdot x \). 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 2-5 and \( D_i \neq \emptyset \), \( C_i \neq \emptyset \). Assign \((\alpha_i, F_i)\) as done by (9) in Lemma A.2. Define the function \( H_i : [0, 1] \rightarrow \mathbb{R} \cup \{ \infty \} \) as \( H_i(a) = F_i \frac{\alpha_i(1-a_i)}{\alpha_i(1-a)} \) 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) \( \lim_{a \rightarrow 1} H_i(a) = \infty \), and (v) \( H_i'(\alpha_i) = \frac{F_i}{\alpha_i - \alpha_i^*} \).

**Proof.** Direct calculations. ■

**Lemma A.4** Fix any two players \( i \) and \( j \) such that \((D_i, C_i)\) and \((D_j, C_j)\) satisfy Axioms 2-5. Assign \((\alpha_i, F_i), (\alpha_j, F_j)\) as done by (9) in Lemma A.2. The following 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 \( \alpha_j \in \{0, 1\} \) 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

\[
(\text{int}_j, \text{slp}_j) \in \{(\text{int}, \text{slp})|\text{int} \leq \text{int}_i \text{ and } \text{slp} \geq \text{slp}_i\} \cup \{(\text{int}, \text{slp})|\text{int} \geq \text{int}_i \text{ and } \text{slp} < \text{slp}_i\}
\]

\[
\cup \{(\text{int}, \text{slp})|\text{int} = \text{int}_i \text{ and } \text{slp} = \text{slp}_i\}
\]
Inverting the bijection from (9) in Lemma A.2

\[(\alpha_j, F_j) \in \left\{ (\alpha, \phi) \mid \frac{\alpha}{(1-\alpha)\phi} \leq \frac{\alpha_i}{(1-\alpha_i)F_i} \text{ and } \frac{1-\phi}{\phi} > \frac{1-F_i}{F_i} \right\} \]

\[\cup \left\{ (\alpha, \phi) \mid \frac{\alpha}{(1-\alpha)\phi} \geq \frac{\alpha_i}{(1-\alpha_i)F_i} \text{ and } \frac{1-\phi}{\phi} < \frac{1-F_i}{F_i} \right\} \]

\[\cup \left\{ (\alpha, \phi) \mid \frac{\alpha}{(1-\alpha)\phi} = \frac{\alpha_i}{(1-\alpha_i)F_i} \text{ and } \frac{1-\phi}{\phi} = \frac{1-F_i}{F_i} \right\} \]

Rearranging and using Definition A.1 gives,

\[(\alpha_j, F_j) \in \left\{ (\alpha, \phi) \mid H_i(\alpha) \leq \phi \text{ and } \phi < F_i \right\} \cup \left\{ (\alpha, \phi) \mid H_i(\alpha) \geq \phi \text{ and } \phi > F_i \right\} \]

\[\cup \left\{ (\alpha, \phi) \mid H_i(\alpha) = \phi \text{ and } \phi = F_i \right\} \]

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

\[(\alpha_j, F_j) \in \left\{ (\alpha, \phi) \mid \alpha \in [0, \alpha_i) \text{ and } \phi \in [H_i(\alpha), F_i) \right\} \cup \left\{ (\alpha, \phi) \mid \alpha \in (\alpha_i, 1] \text{ and } \phi \in (F_i, H_i(\alpha)) \right\} \]

\[\cup \left\{ (\alpha, \phi) \mid \alpha = \alpha_i \text{ and } \phi = F_i \right\} \]

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 (\alpha_i, F_i), (\alpha_j, F_j) as done by (9), 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.

\[H_i'(a) - H_j'(a) = \frac{F_i \cdot \alpha_j(1-\alpha_i) - F_j \cdot \alpha_i(1-\alpha_j)}{\alpha_i \cdot \alpha_j(1-a)^2} \quad (10)\]

By Lemma A.4, either (10) 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 (10) 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 (9) in Lemma A.2. There exists F \in 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 \ldots \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_i}. Set F(0) = 0, F(1) = 1, F'(1) = 0. Next, we fill in F in the intervals between the pairs to produce a strictly increasing, differentiable CDF that satisfies Conditions 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 (\bar{z}, \tilde{z}) if i) f_1(\bar{z}) = f_2(\bar{z}), ii) f_1'(\bar{z}) = f_2'(\bar{z}), iii) f_1(\tilde{z}) = f_3(\tilde{z}), and iv) f_1'(\tilde{z}) = f_3'(\tilde{z}).

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

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 (\alpha_k, \alpha_{k+1}), which satisfies all of the necessary properties (see Fact A.2).
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 $\alpha_k$. 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 $\varepsilon$, 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 $H_{\alpha,F}$ be the $H$-function of a hypothetical player with $(\alpha_i, F_i) = (\alpha, F)$. By (v) of Fact A.2, it is sufficient to demonstrate that $F'(\alpha) \leq H_{\alpha,F}(\alpha)$ 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 $\alpha$. Hence, the inequality must be satisfied. Finally, if $\varepsilon$ is small enough, then since $E > H_{k+1}$ on $(\alpha^0 - \varepsilon, \alpha^0 + \varepsilon)$, in order to smoothly paste with $H_{k+1}$, it must be that $E' < H_{k+1}'$ on this interval. From Corollary A.1, $H_{\alpha,F}(\alpha) > H_{k+1}'$, which establishes the inequality.

b) In this case, let $\tilde{L}$ be the line that passes through $(\alpha_k, F_k)$ and $(\alpha_{k+1}, F_{k+1})$, so $\tilde{L} \leq L'$ by hypothesis. Next, let $\tilde{L}_\delta$ be the line that passes through the midpoint between $(\alpha_k, F_k)$ and $(\alpha_{k+1}, F_{k+1})$ with slope $\tilde{L}' - \delta$. For any $\delta > 0$ small enough, there exists $\varepsilon > 0$ small enough such that $(H_k, \tilde{L}_\delta)$ can be smoothly pasted by elliptical arc $E_1$ on $(\alpha_k, \alpha_k + \varepsilon)$, and $(\tilde{L}_\delta, H_{k+1})$ can be smoothly pasted by elliptical arc $E_2$ on $(\alpha_k - \varepsilon, \alpha_{k+1})$. By construction, for sufficiently small $\delta$ and $\varepsilon$, setting $F = E_1$ on $(\alpha_k, \alpha_k + \varepsilon)$, $F = \tilde{L}_\delta$ on $[\alpha_k + \varepsilon, \alpha_{k+1}]$, 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 (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 $\varepsilon$ is sufficiently small.

4. For $\alpha \in (\alpha_m, 1)$, there exists an elliptical arc $E$ that smoothly pastes $(H_m, 1)$ on this interval. Set $F = E$ here to satisfy differentiability, strict monotonicity, and Condition S (since it is concave, the same argument from (3.a) applies).

\begin{lemma}
Lemma A.6 If $(D_i, C_i)_{i \in I}$ satisfies Axioms 3-5, then it can be explained by a behavioral model $[F, (\alpha_i)_{i \in I}]$, where $F \in F$ satisfies Condition S. Furthermore, $(\alpha_i, F(\alpha_i))_{i \in I}$ is unique.
\end{lemma}

\textbf{Proof.} Immediate corollary of Lemmas A.1 and A.5.

\begin{lemma}
Lemma A.7 If $(D_i^0, C_i^0)_{i \in I}$ satisfies Axioms 7-12, then it can be explained by a behavioral model $[F, (\alpha_i)_{i \in I}]$, where $F \in F$ satisfies Condition S. Furthermore, $(\alpha_i, F(\alpha_i))_{i \in I}$ is unique.
\end{lemma}

\textbf{Proof.} Immediate corollary of Lemmas 1 and A.6.

\textbf{Proof of Proposition 3.} The proof is ordered as follows: (a) $\iff$ (c), (b) $\iff$ (c), (a) $\iff$ (e), (d) $\iff$ (e).

(a) $\iff$ (c): Suppose $F$ f.o.s.d. $\tilde{F}$. Recall that if $x = y$, then $\alpha^*_{x,y} = \tilde{\alpha}^*_{x,y}$ (Section 3.2.2), implying $F(\alpha^*_{x,y}) \leq \tilde{F}(\tilde{\alpha}^*_{x,y})$ 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 $\tilde{F}(\tilde{\alpha}^*_{x,y}) = T(\tilde{\alpha}^*_{x,y}|x,y)$. Further, if $x > y$, then $T(0|x,y) > 0 = F(0) = \tilde{F}(0)$,
and $T(|x, y)$ is strictly decreasing. Hence, by f.o.s.d., both $\alpha^*_{x,y} \leq \alpha^*_{x,y}$ and $\tilde{F}(\alpha^*_{x,y}) \geq F(\alpha^*_{x,y})$. If $x < y$, then $T(0, x, y) < 0 = F(0) = \tilde{F}(0)$, and $T$ is strictly increasing (but intersecting $F$ and $\tilde{F}$ each exactly once since both satisfy Condition S). Hence, by f.o.s.d., both $\alpha^*_{x,y} \geq \alpha^*_{x,y}$ and $\tilde{F}(\alpha^*_{x,y}) \geq F(\alpha^*_{x,y})$. So (a) implies (c).

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

(b) $\iff$ (c): Notice that $k_{x,y}$ and $\tilde{k}_{x,y}$ 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 $(1 - F(\alpha^*_{x,y}))$ respectively. Because $n$ is common between the two random variables, a simple “coupling” argument (Lindvall 2002, Ch. 1) establishes that $k_{x,y}$ f.o.s.d. $\tilde{k}_{x,y}$ if and only if $1 - F(\alpha^*_{x,y}) \geq 1 - F(\alpha^*_{x,y})$.

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

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

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

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

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

For part (b), first suppose that $I$ is more influenced by $x$ relative to $y$ than is $\tilde{I}$. Then the behavior of the two collections agree on the subset of games $\{(x, y) \in PD | x = y\}$. By part (a) of the proposition, $\alpha_i |_{x = \tilde{I}} = (\tilde{\alpha}_i) |_{x = \tilde{I}}$. Immediately, if $\alpha_i = 0, 1$, then $F(\alpha_i) = \tilde{F}(\alpha_i)$. In addition, $I$ defecting more in $\{(x, y) \in PD | x = y\}$ than does $\tilde{I}$ 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 $\tilde{F}$, i.e., $slp_i \geq \tilde{slp}_i$ (Lemma A.1). Next, by Lemma A.2 $F(\alpha_i) = \frac{1}{1+slp_i} \leq \frac{1}{1+\tilde{slp}_i} = \tilde{F}(\alpha_i)$.

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

**Proofs for the Extended Domain, $G^0$**

We first establish the representation result (Theorem 2), then Proposition 5. The representation proof uses the following preliminary lemma.
Axiom 6' For all $i \in I$, if $(x, y) \in C_i$ and $\kappa \in (0, 1)$, then $\kappa (x, y) \in C_i$.

Lemma A.8 Under Axiom 1, Axioms 6 and 6 are equivalent.

Proof. Suppose that Axioms 1 and 6 hold and that $\kappa \in (0, 1)$. Then,

$$(x, y) \in C_i \implies (1, 0, x, y) \in C_i^0 \implies \left(\frac{1}{\kappa}, 0, x, y\right) \in C_i^0 \implies \kappa \left(\frac{1}{\kappa}, 0, x, y\right) \in C_i^0 \implies (1, 0, \kappa x, \kappa y) \in C_i^0 \implies (\kappa x, \kappa y) \in C_i \implies \kappa (x, y) \in C_i.$$

Hence, Axiom 6 is implied. Now, suppose that Axioms 1 and 6 hold and that $r' > r$. Then,

$$(r, p, x, y) \in C_i^0 \implies (1, 0, \frac{x}{r - p}, \frac{y}{r - p}) \in C_i^0 \implies \left(\frac{x}{r - p}, \frac{y}{r - p}\right) \in C_i \implies (1, 0, \frac{x}{r - p}, \frac{y}{r - p}) \in C_i^0 \implies (r', p, x, y) \in C_i^0.$$

Hence, Axiom 6 is implied. ■

Fact A.3 Axioms 15 $\implies$ Axiom 6 on $PD^0$. Axioms 13 $\not\implies$ Axiom 6 on $G^0$.

Proof. Relying on Lemma A.8 we consider whether or not Axioms 15 imply Axiom 6 on $PD$ and $G$ for the first and second claims respectively. For the first claim, fix player $i$, for whom $(D_i \cap PD, C_i \cap PD)$ satisfies Axioms 24 with $C_i \cap PD \neq \emptyset$. Then, from the proof of Theorem 1 we have that either $C_i \cap PD = PD$ or $M_i \cap PD = \{(x, y) \in PD | y < int_i - sli \cdot x\}$ and $C_i \cap PD = \{(x, y) \in PD | y < int_i - sli \cdot x\}$, where $int_i, sli$ are positive constants. In either case, Axiom 6 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 24. However, $(D_i, C_i)$ fails Axiom 6: 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 2 First, note that Lemma 1 and Propositions 12 (and their proofs) remain valid when each $PD^0$ and $PD$ are replaced by $G^0$ and $G$ respectively.

Representation $\implies$ Axioms: Given that Lemma 1 and Propositions 12 extend to the larger domain, the proof that the representation satisfies Axioms 15 is completely analogous to that provided for Theorem 1. Using Lemma A.8 we are left to verify that Axiom 6 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 $(x, y) \in C_i$, then $y \leq int_i - \alpha_i \cdot sli$, where $int_i, sli \geq 0$. It follows that, for any $\kappa \in (0, 1)$, $\kappa y \leq \kappa (int_i - \alpha_i \cdot sli) < int_i - (\kappa \alpha) sli$. 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 $D_i \neq \emptyset$ and $C_i \neq \emptyset$. Such a player can be characterized by a pair $(int_i, sli)$, where $sli > 0$. When the domain was $PD$, $int_i > 0$ immediately. This is no longer immediate when the domain is $G$. However, it is ensured by Axiom 6. Suppose to the contrary that $int_i \leq 0$. Take now $(x, y) \in M_i \subset C_i$, which must then satisfy $y = int_i - x \cdot sli < 0$. But then, for any $\kappa \in (0, 1)$,
\( \kappa y = \kappa (int_i - x \cdot slp_i) \geq int_i - (kx)slp_i. \) Hence, \( \kappa (x, y) \not\in ML_i = C_i, \) violating Axiom \( 6 \) (and therefore also Axiom \( 6 \) by Lemma \( A.8 \)). With this established, the remainder of the proofs follows identical steps to those in the proof of Theorem \( 1. \)

**Proof of Proposition 5.** First consider \( g \in PD^0, \) so \( \pi_g = 1. \) By Propositions \( 1.2 \) (and Lemma \( 1. \)), \( \alpha_y \in (0, 1), \) implying \( F(\alpha^*_g) < 1 = \pi_g. \) Second, consider \( g \in G^0 \setminus PD^0, \) so \( y \leq 0 \) and \( \pi_g = \frac{x}{x-y}. \)

The analog to \( (3) \) 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,

\[
\pi_g - F(\alpha^*_g) = \frac{\alpha^*_g(r-p)}{(1-\alpha^*_g)(x-y)} > 0. \tag{11}
\]

The inequality is due to \( \alpha^*_g \in (0, 1), \) \( r > p, \) and \( x > 0 \geq y. \)

For the general limit result, observe that Lemma \( 1 \) implies that it is sufficient to establish the following: 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 following terms: \( \alpha^\varepsilon = F^{-1}(1-\varepsilon), \) \( K_1 = \frac{\alpha^\varepsilon}{1-\alpha^\varepsilon}, \) and, letting \( (int_{\alpha^\varepsilon}, slp_{\alpha^\varepsilon}) \) be the \( (int_i, slp_i) \) generated by the equilibrium behavior of a player \( i \) with \( \alpha_i = \alpha^\varepsilon, K_2 = \max\{int_{\alpha^\varepsilon}, \frac{int_{\alpha^\varepsilon}}{slp_{\alpha^\varepsilon}}\}. \)

Setting \( K = \max\{K_1, K_2\}, \) establishes the claim. To see this, suppose that \( y > 0 \) and \( x + |y| > K. \) Then \( y > K - x \geq K_2 - x \geq int_{\alpha^\varepsilon} - x \cdot slp_{\alpha^\varepsilon}. \) 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 \( (11), \) 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. \( \blacksquare \)

**B Large Collections and Empirically Valid F**

We refer to the commonly perceived distribution of types \( F \) as being *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. \(^{38}\) It is easy to see that this precludes the observation of almost all \( \alpha \)-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 \( 3.4, \) it is difficult to give behavioral meaning to the empirical validity of \( F \) when the analyst’s data is generated by a finite collection.

\(^{38}\)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.
To address both of these issues, in this appendix 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 \( (D_i, C_i)_{i \in I} \). 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 \( B.1 \) immediately apply to the larger set of behavioral data, but they are more restrictive in the following sense.

**Definition B.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 B.1** The primitive \( (D_i, C_i)_{i \in I} \) satisfies Axioms \( B.1 \) if and only if it can be explained by a behavioral model \( [F, (\alpha_i)_{i \in I}] \in \mathcal{M} \). Furthermore, for all \( i \in I \), \( \alpha_i \) 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.\(^{39} \) We show that it is both necessary and sufficient for uniqueness of the equilibrium cutoff in all 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 B.2** Given any \( (\alpha_i)_{i \in I} \), let \( \hat{F} \) 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 \( \hat{F}(\alpha_i) \), 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 \textit{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., \( F = \hat{F} \)), 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 that there be mass points clustered on any given \( (D, C) \) pair. Formally, for arbitrary \((D, C)\), let \( \mathcal{I}(D, C) \) be the Lebesgue measure of the set \( \{ i \in I | (D_i, C_i) = (D, C) \} \).

\(^{39} \)Extending results to \( PD^0 \) is trivial via Axiom \( B.1 \) and an analogous exercise can be similarly carried out for \( G \) (and \( G^0 \)).

\(^{40} \)If \( F \) is differentiable, then Conditions \( S \) \( \iff \) (ii).
Axiom 7 (Smooth Data)
For all \((D, C)\) such that \(D \neq PD\) and \(C \neq PD\), \(\mathcal{L}(D, C) = 0\).

Next, in our behavioral model, player \(i\) compares the perceived benefit of cooperation, \(\alpha_i\), with the perceived cost of cooperation, \((1 - \alpha_i) \left( x(1 - P(x, y)) + yP(x, y) \right)\), 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 B.3 Given \((D_i, C_i)_{i \in I}\), for each \((x, y) \in PD\), define \(\widehat{P}(x, y)\) as the Lebesgue measure of the set \(\{i \in I \mid (x, y) \in D_i\}\), and let \(Q(x, y) := x(1 - \widehat{P}(x, y)) + y\widehat{P}(x, y)\). \(^{41}\)

Because \(i\) cannot, in fact, directly influence his opponent’s action choice, for each \((x, y) \in PD\), \(Q(x, y)\) 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 \(M_i\).

Axiom 8 (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. \(^{42}\)

Proposition B.2 The primitive \((D_i, C_i)_{i \in I}\) satisfies Axioms 7-8 if and only if there exists \((\alpha_i)_{i \in I}\) such that (i) \(\left[ \hat{F}, (\alpha_i)_{i \in I} \right]\) explains \((D_i, C_i)_{i \in I}\) and (ii) \(\left[ F, (\alpha_i)_{i \in I} \right]\) \(\in\mathcal{M}\). Furthermore, for all \(i \in I\), \(\alpha_i\) is unique.

Given Proposition B.1, this shows that Axioms 7 and 8 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: \(F(\alpha_i) = F(\alpha_i)\) for all \(i \in I\). It is slightly more subtle to see that Axiom 8 is needed to ensure they are also correctly calibrated in their

\(^{41}\) 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 14).

\(^{42}\) 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.
beliefs about those types not in the collection (i.e., they do not assign them positive probability). This is because the original axioms require continuity of on [0,1]. If Axiom fails, then will not be continuous on [0,1)—there still exists behavioral models in that can explain the data, but none with .

B.1 Proofs

Proof of Proposition B.1

Representation \(\Rightarrow\) 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 \(\text{"}F \in F\)\) with \(\text{"}(i)\) of Definition B.1\) and \(\text{"}Condition S\)\) with \(\text{"}(ii)\) of Definition B.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 \((x,y) \in PD\) 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 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:

\[
x = \frac{\alpha_1^*(1-\alpha_2^*)F(\alpha_2^*) - \alpha_2^*(1-\alpha_1^*)F(\alpha_1^*)}{(1-\alpha_1^*)(1-\alpha_2^*)(F(\alpha_2^*) - F(\alpha_1^*))}, \quad y = x + \frac{\alpha_2^* - \alpha_1^*}{(1-\alpha_1^*)(1-\alpha_2^*)(F(\alpha_2^*) - F(\alpha_1^*))}
\]

(12)

The denominator of \(x\) is positive. However, the numerator of \(x\) is weakly negative by condition (ii) of Definition B.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 \(F(\alpha') > F(\alpha)\), which implies \(F(\alpha') > F(\alpha)\). Then, setting \(\alpha_1^* = \alpha\) and \(\alpha_2^* = \alpha'\), \((x,y)\) as given by (12) 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. Lemma A.1 remains valid. Lemma A.2 must be modified as follows:

**Lemma B.2** 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 \([F_i, (\alpha_i, \alpha_{-i})]\) such that \(F\) is continuous on \([0,1]\) and \(F(\alpha_i) = F_i\). Further, \(\alpha_i\) is unique, and \(F_i\) is unique if and only if \(C_i \neq \emptyset\), as follows:

\[
(\alpha_i, F_i) = \begin{cases} 
\left(\frac{\int_{i}^{\text{int}t_i+s\beta_i}}{1+s\beta_i}, \frac{1}{1+s\beta_i}\right) & \text{if } D_i, C_i \neq \emptyset \\
(1, 1) & \text{if } D_i = \emptyset \\
(0, K_i), K_i \in [0,1] & \text{if } C_i = \emptyset
\end{cases}
\]

(13)

**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 \(y < \text{int}_i - s\beta_i \cdot x\). 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 (13), 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).

\( \text{iii) Suppose that } (x, y) \in D_i. \) 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 \([13]\), 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 \( d(\alpha) = 0 \) and is therefore an equilibrium cutoff in the game \((x, y) \in D_i \) (by the analog of Proposition 1).

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

Lemmas [A.3] and [A.4] with references to Lemma [A.2] now made to Lemma [B.2] also remain valid. Hence, for any collection whose data satisfy Axioms [2-5] using \([13]\), each player \( i \) can be assigned a unique \( \alpha_i \) and corresponding quantile \( F_i \), that is also unique if \( \alpha_i > 0 \), and \( i \text{'s behavior can be explained by any model } [F_i, (\alpha_i, \alpha_i)] \text{ such that } F_i \text{ satisfies (i) of Definition B.1 and } F(\alpha_i) = F_i. \)

We now show that there exists a model in \( \mathcal{M} \) that simultaneously explains the behavior of all \( i \in I \). Let \( \Delta^+ := \{\alpha_i > 0 | i \in I\} \). The four lemmas \([A.1] B.2 A.3 A.4\) imply that any behavioral model that satisfies \( F(\alpha_i) = F_i \) for all \( i \in I \) also satisfies (iii) and (iv) of Definition B.1 as well as (i) and (ii) \textbf{restricted to the domain } \( \Delta^+ \), \text{ where continuous on } \( \Delta^+ \) \text{ means: for every } \alpha^0 \in \Delta^+, \text{ every sequence } \( (\alpha^m)_{m \in \mathbb{N}} \), \( \alpha^m \in \Delta^+ \) \text{ for all } m, \text{ such that } \lim_{m \to \infty} \alpha^m = \alpha^0 \text{ also satisfies } \lim_{m \to \infty} F(\alpha^m) = F(\alpha^0). \text{ All that remains is to establish existence by extending } F \text{ from } \Delta^+ \text{ to } [0,1) \text{ preserving continuity, (weak) monotonicity, and condition (ii) of Definition B.1.} \text{ 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 PD. \text{ 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 \text{ and } (\alpha_i)_{i \in I} \text{ can explain } (D_i, C_i)_{i \in I}. \)

To extend \( F \) from \( \Delta^+ \) to \([0,1)\), consider arbitrary \( \alpha^0 \in [0,1) \setminus \Delta^+. \) There are three exhaustive cases. First, if there exists a sequence \( (\alpha^m)_{m \in \mathbb{N}} \), \( \alpha^m \in \Delta^+ \) \text{ for all } m, \text{ such that } \lim_{m \to \infty} \alpha^m = \alpha^0 \text{, simply assign } F(\alpha^0) = \lim_{m \to \infty} F(\alpha^m). \) Second, let \( \alpha := \inf(A^+) \) and \( \bar{\alpha} := \sup(A^+) \). If \( \alpha^0 < \alpha \), assign \( F(\alpha^0) = F(\alpha) \), and if \( \alpha^0 > \bar{\alpha} \), assign \( F(\alpha^0) = F(\bar{\alpha})\) notice that even if \( \alpha, \bar{\alpha} \notin \Delta^+ \), \( F(\alpha), F(\bar{\alpha}) \) are assigned in the previous case. Third, and finally, if \( \alpha^0 \) does not contain a sequence converging to \( \alpha^0 \in [\alpha, \bar{\alpha}] \), then \( \alpha := \sup\{\alpha \in A^+ | \alpha < \alpha^0\} < \alpha^0 < \bar{\alpha} := \inf\{\alpha \in A^+ | \alpha > \alpha^0\}. \) Notice that even if \( \alpha^0, \alpha \notin \Delta^+ \), \( F(\alpha^0), F(\bar{\alpha}) \) are assigned in the first case. Let \( L^0 \) be the line that passes through both \( (\alpha^0, F(\alpha^0)) \) and \( (\bar{\alpha}, F(\bar{\alpha})) \). For all \( \alpha \in (\alpha^0, \bar{\alpha}) \), 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 \([\alpha^0, \bar{\alpha}]\) preserves condition (ii) on this interval, and then in general since condition (ii) is transitive. \]

\textbf{Proof of Proposition B.2.} \textit{Representation \( \Rightarrow \) Axioms: Consider a collection } \( I \text{ that satisfies the representation.} \) The fact that \( F = \tilde{F} \) is irrelevant for the proof that \((D_i, C_i)_{i \in I} \) satisfies Axioms [2-5], so this is established by Proposition B.1. Next, \( F = \tilde{F} \), which is continuous on \([0,1)\), immediately implies Axiom \( 7. \) Finally, verifying Axiom \( 8 \) is a straightforward calculation: fix any player \( i \) such that \( M_i \neq \emptyset \) and recall that \( M_i = \{ (x, y) \in PD | y = \left(1 - \frac{\alpha_i}{1 - \alpha_i}F(\alpha_i) \right) - x \left(1 - F(\alpha_i) \right) \}. \) Therefore, for any \( (x, y) \in M_i \), we
can substitute the expression for \( y \) into \( Q(x, y) \) to get,

\[
Q(x, y) = x(1 - \hat{P}(x, y)) + \left( \frac{\alpha_i}{(1 - \alpha_i)F(\alpha_i)} - x \left( \frac{1 - F(\alpha_i)}{F(\alpha_i)} \right) \right) \hat{P}(x, y)
\]  

(14)

Given that \( F = \hat{F} \) and that \( \alpha_i \) is the cutoff type for \((x, y) \in M_i\), \( \hat{P}(x, y) = F(\alpha_i) \); so (14) simplifies to \( Q(x, y) = \frac{\alpha_i}{1 - \alpha} \), which does not vary with \((x, y)\).

**Axioms \( \Rightarrow \) Representation:** The proof of Proposition B.1 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 \( \alpha_i \) and \( F(\alpha_i) = F_i \) are given by (13) (and therefore \( \alpha_i \) is unique and, if \( \alpha_i > 0 \), so is \( F(\alpha_i) \)). Therefore, let \((\alpha_i)_{i \in I} \) be as given by (13), and \( \hat{F} \) be the resultant CDF. It is sufficient to show that 1) for all \( i \) such that \( \alpha_i > 0 \), \( \hat{F}(\alpha_i) = F_i \), and 2) \([\hat{F}, (\alpha_i)_{i \in I}] \in \mathcal{M}\).

To see the first, notice that the structure of the \((D_i, C_i)_{i \in I} \) characterized by Lemmas A.1, B.2, A.3, and A.4 implies that for any \( i \) such that \( M_i \neq \emptyset \), \( \hat{P}(x, y) \) is constant and equal to \( \lim_{\alpha \uparrow \alpha_i} F(\alpha) \) along \( M_i \). By Axiom 7, \( \lim_{\alpha \uparrow \alpha_i} \hat{F}(\alpha) = \hat{F}(\alpha_i) \). Consider \( i \) such that \( \alpha_i \in (0, 1) \), so \( M_i \neq \emptyset \). For \((x, y) \in M_i\),

\[
Q(x, y) = x(1 - \hat{P}(x, y)) + y\hat{P}(x, y)
= x(1 - \hat{F}(\alpha_i)) + y\hat{F}(\alpha_i)
= x(1 - \hat{F}(\alpha_i)) + (\text{int}_i - x \cdot \text{slp}_i)\hat{F}(\alpha_i)
\]

By Axiom 8, \( Q \) is constant along \( M_i \), so \( \hat{F}(\alpha_i) = \frac{1}{1 + \text{slp}_i} = F_i \). If instead, \( \alpha_i = 1 \), then because \( \hat{F} \) is a CDF on \([0, 1]\), \( \hat{F}(\alpha_i) = 1 = F_i \).

To see the second, we need to show that \([\hat{F}, (\alpha_i)_{i \in I}] \) satisfies the four requirements of Definition B.1. Axiom 7 implies (i), and Lemmas B.2 and A.4 imply (iii) and (iv). For (ii), notice that if \( \alpha \) and \( \alpha' \) 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 \( \frac{\hat{F}(\alpha')}{\hat{F}(\alpha)} > \frac{\alpha'(1 - \alpha)}{\alpha(1 - \alpha')} \). Since \( \hat{F} \) 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 \( \alpha \leq \alpha_i < \alpha_j \leq \alpha' \), \( \hat{F}(\alpha_i) - \hat{F}(\alpha) < \varepsilon \), and \( \hat{F}(\alpha') - \hat{F}(\alpha_j) < \varepsilon \). Hence, by our supposition that \( \frac{\hat{F}(\alpha')}{\hat{F}(\alpha)} > \frac{\alpha'(1 - \alpha)}{\alpha(1 - \alpha')} \), for \( \varepsilon \) small enough,

\[
\frac{\hat{F}(\alpha_j)}{\hat{F}(\alpha_i)} > \frac{\alpha'(1 - \alpha)}{\alpha (1 - \alpha')} \geq \frac{\alpha_j (1 - \alpha_i)}{\alpha_i (1 - \alpha_j)}
\]

As we just discussed, Lemma A.4 implies that \( \frac{\hat{F}(\alpha_j)}{\hat{F}(\alpha_i)} \leq \frac{\alpha_j (1 - \alpha_i)}{\alpha_i (1 - \alpha_j)} \), producing a contradiction. ■
References


and Social Psychology, 35(1), 1–11.


