Wolpert, David H.; Jamison, Julian; Newth, David; Harre, Michael

Working Paper

Strategic choice of preferences: The persona model

Working Papers, No. 10-10

Provided in Cooperation with:
Federal Reserve Bank of Boston

Suggested Citation: Wolpert, David H.; Jamison, Julian; Newth, David; Harre, Michael (2010) : Strategic choice of preferences: The persona model, Working Papers, No. 10-10, Federal Reserve Bank of Boston, Boston, MA

This Version is available at:
https://hdl.handle.net/10419/55569

Working Papers

FEDERAL RESERVE BANK OF BOSTON

No. 10-10

Strategic Choice of Preferences: The Persona Model

David H. Wolpert, Julian Jamison, David Newth, and Michael Harre

Abstract

We introduce a modification to the two-timescale games studied in the evolution of preferences (EOP) literature. In this modification, the strategic process occurring on the long timescale is learning by an individual across his or her lifetime, not natural selection operating on genomes over multiple generations. This change to the longer timescale removes many of the formal difficulties of EOP models and allows us to show how two-timescale games can provide endogenous explanations for why humans sometimes adopt interdependent preferences and sometimes exhibit logit quantal response functions. In particular, we show that our modification to EOP explains experimental data in the Traveler's Dilemma. We also use our modification to show how cooperation can arise in nonrepeated versions of the Prisoner's Dilemma (PD). We then show that our modification to EOP predicts a "crowding out" phenomenon in the PD, in which introducing incentives to cooperate causes players to stop cooperating instead. We also use our modification to predict a tradeoff between the robustness and the benefit of cooperation in the PD.

Keywords: nonrationality, single-shot games, Prisoner's Dilemma, Traveler's Dilemma, Schelling, emotions

JEL codes: C70, C72, D03

David H. Wolpert is a senior computer scientist at the NASA Ames Research Center and a consulting professor in the aeronautics and astronautics department of Stanford University, Julian Jamison is a senior economist at the Federal Reserve Bank of Boston's Center for Behavioral Economics, David Newth is a research scientist with the CSIRO Centre for Complex Systems Science, and Michael Harre is a post-doctoral fellow with the Centre for the Mind at the University of Sydney. Their e-mail addresses are david.h.wolpert@nasa.gov, julian.jamison@bos.frb.org, david.newth@csiro.au, and mike@centreforthemind.com, respectively.

We would like to thank Nihat Ay, Nils Bertschinger, Raymond Chan, Johannes Horner, Stefano Lovo, Eckehard Olbrich, and Tanya Rosenblat for helpful discussion, as well as Kyle Carlson, Lynn Conell-Price, and Suzanne Lorant for comments on the manuscript.

This paper, which may be revised, is available on the web site of the Federal Reserve Bank of Boston at http://www.bos.frb.org/economic/wp/index.htm.

This version: September 24, 2010

RESEARCH CENTER FOR BEHAVIORAL ECONOMICS AND DECISIONMAKING

1 Introduction

1.1 Behavioral preference models

It is well established that even in an anonymous single-shot game where every player knows he will never interact with his opponent(s) again, human players often exhibit "nonrational" behavior. (See Camerer (2003), Gachter and Herrmann (2009), and references therein.) To state this more precisely, often in an anonymous single-shot game where there are exogenously provided (often material) underlying preferences, humans do not maximize those *underlying* preferences. A great deal of research has modeled such nonrational behavior by hypothesizing that humans have *behavioral* preferences that differ from their underlying preferences and that they maximize behavioral preferences rather than maximizing underlying preferences. We will refer to such models as *behavioral preference models*, with the nonrational behavior given by simultaneous maximization of every player's behavioral preferences as a *behavioral preference equilibrium*. Different kinds of behavioral preference models arise for different choices of how to formalize the underlying and behavioral preferences.

Perhaps the most prominent example of a behavioral preference model is the work on interdependent / other-regarding / social preferences (Sobel, 2005; Bergstrom, 1999; Kockesen et al., 2000). In that work, both the underlying and the behavioral preferences are formalized as expectations of von Neumann - Morgenstern utility functions. Accordingly, these behavioral preference models presume that people do not maximize expected underlying utility subject to play of their opponents, but instead maximize expected behavioral utility. Often in this interdependent preferences work the behavioral utility function of player i is a parameterized combination of i's underlying utility function and the underlying utility functions of i's opponents. A typical analysis in this work seeks to find parameters of such behavioral utility functions that provide a good fit for some experimental data.

Other work has explored behavioral preference models when the behavioral preferences are not expected utilities. An example is the (logit) quantal response equilibrium (QRE). Suppose we have a finite strategic-form game where the mixed strategy of player i over moves $x_i \in X_i$ is $q_i(x_i)$. Then, under the QRE,

$$q_i(x_i) \propto e^{\lambda_i \mathbb{E}(u_i|x_i)} \quad \forall \text{ players } i. \tag{1}$$

The QRE was originally motivated purely theoretically, using a model of a player's uncertainty

in his own utility (McKelvey and Palfrey, 1995, 1998). However the majority of work on the QRE since its introduction has treated it as a parametric model of nonrational behavior. In this interpretation, $\lambda_i \in \mathbb{R}$ is a parameter describing "how rational" or "how smart" player i is, and this parameter is fit to experimental data (Camerer, 2003).[1]

The QRE cannot be expressed as a behavioral preference model where behavioral preferences are expected utilities; in general, the solution to equation (1) is not a Nash equilibrium for any set of utilities. However, define

$$S(q_i) \equiv -\sum_{x_i} q_i(x_i)\ln(q_i(x_i)), \tag{2}$$

the Shannon entropy of player i's mixed strategy, q_i (Cover and Thomas, 1991; Mackay, 2003). Then define the *free utility* of player i by

$$\mathscr{F}(q_i) \equiv \lambda_i \mathbb{E}(u_i) + S(q_i). \tag{3}$$

When all players i simultaneously maximize their free utilities, we have a QRE (Wolpert, 2009; Meginniss, 1976). Accordingly, when players adopt a QRE with a particular set of rationalities $\{\lambda_i\}$, their mixed strategy profile is a behavioral preferences equilibrium where all players simultaneously maximize their associated (behavioral) free utilities, rather than maximize the expected values of some behavioral utility functions. (Wolpert, 2007; Meginniss, 1976; Fudenberg and Kreps, 1993; Fudenberg and Levine, 1993; Shamma and Arslan, 2004; Hopkins, 2002; Anderson et al., 2001).

From this perspective, a typical analysis in the QRE literature that fits (the rationality parameters specifying) a logit QRE to experimental data is fitting experimental data with a behavioral preference model. In this sense, it is equivalent to the work in the interdependent preferences literature that fits experimental data with a behavioral model of interdependent utility preferences. In both cases, the basic premise is that in the real world, players work to maximize their behavioral preferences rather than their underlying ones. Also, in both cases, much of the literature is concerned with using experimental data to estimate what precise behavioral preferences the players work to maximize.

Define an *objective function* as a map from the space of possible mixed strategy profiles to $\mathbb{R}$ (Wolpert, 2009). As an example, given a utility function u, the expectation of u under a mixed

[1] Often in this work λ_i is taken to be independent of i.

strategy profile is an objective function. Similarly, a free utility is an objective function. From now on, we restrict attention to player preferences (both underlying and behavioral) that can be expressed as objective functions. So for us, a behavioral preference model explains human behavior by hypothesizing that players maximize behavioral objective functions (subject to play of others) rather than maximize underlying objective functions.

In this interdependent preferences and QRE experimental work, the task of the researcher is simply to ascertain the parameters of real-world behavioral objective functions from data. Two important issues are unaddressed in that work.

The first such issue is how the players acquire common knowledge of one another's behavioral objective functions before the start of play. (After all, in the types of behavioral preference equilibria discussed above, the objective function of each player i differs from his expected underlying utility, $\mathbb{E}(u_i)$.) This issue is particularly pronounced in nonrepeated games, and even more so when the games are played anonymously.

The second issue is how to explain why the parameters of the behavioral objective functions have the values they do. The interdependent preferences and QRE experimental work does not consider the issue of *why* a human should try to optimize a particular behavioral objective function rather than his underlying objective function.

In this paper, we address this second issue. We start by noting that by definition, the strategy profile that the players adopt in any strategic scenario is an equilibrium of the game specified by the players' behavioral objective functions, not the game specified by their underlying objective functions. Therefore, changing the values of the parameters in the behavioral objective functions changes the equilibrium strategy profile. In particular, for a fixed set of behavioral objective function parameters for all players other than i, by varying the parameters of i's behavioral objective function, we create a set of equilibrium profiles of the associated behavioral games. The profiles in that set can be ranked in terms of player i's underlying objective function. In this way, the possible values of the parameters in i's behavioral objective function can be ranked according to i's underlying objective function.

In a nutshell, our thesis is simply that over the course of a lifetime a human learns what values of the parameters of his behavioral objective function have the highest such rank in terms of his underlying objective function. In this way, the parameters of an individual's behavioral objective function are determined endogenously, in a purely rational way, as the values that optimize his underlying objective functions.

1.2 Evolution of Preferences

To introduce our approach, we first discuss earlier work that also addresses the issue of what determines players' behavioral objective functions. This earlier work is evolution of preferences (EOP), also known as "indirect evolution" or "two-timescale games" (Huck and Oechssler, 1999; Heifetz et al., 2007; Samuelson, 2001.; Dekel et al., 2007; Guth and Yaari, 1995; Bester and Guth, 2000).[2]

Loosely speaking, EOP models human behavior using two coupled strategic scenarios involving two different timescales. First, there is an exogenously provided set of pure strategy spaces of the players. In addition to these strategy spaces, there is an exogenously supplied underlying utility function for each of the players. Each player will choose a distribution across his space of pure strategies to play a "short timescale" game. Critically though, the players do not choose their mixed strategies to maximize their underlying utility function. (In this, they appear to be irrational.) Rather each player i chooses a mixed strategy to maximize his behavioral utility function.

In turn, the behavioral utility functions of the players are determined in a "long timescale" process. The idea is that the behavioral utility function of each player i is set in this long timescale process so that the resultant play by all players in the short timescale game, according to their behavioral utility functions, maximizes each i's underlying utility function. EOP concentrates on situations where the process occurring on the longer timescale is biological natural selection, operating on genomes that encode behavioral preferences and change across multiple generations (hence the name, "evolution of preferences").

More precisely, EOP models all involve an infinite sequence of two-player games γ_t. Each game has the same joint move space X. There is also a set of underlying player utilities over X, $\{u_i\}$. Finally, there is also an infinite population of players.

At iteration t, two players are chosen randomly from the infinite population to play the game γ_t. However, in EOP it is not the strategy $x_i \in X_i$ of each chosen player i that is fixed at the beginning of iteration t, as in much of evolutionary game theory. Rather it is the preferences of player i that are fixed at the beginning of each iteration t. In other words, a behavioral utility function $b_i(t)$ from a set B_i is fixed at the beginning of each iteration t for both players i engaged in γ_t chosen to play the game at that iteration.

[2]Most of the EOP literature has concentrated on behavioral utility functions and has not considered the broader issue of behavioral objective functions. See von Widekind (2008) for some preliminary EOP work concerning behavioral objective functions.

Given the joint behavioral utility $b(t) \equiv \{b_i(t)\}$, the game γ_t proceeds in two stages. In the first stage, each player i honestly signals his current behavioral utility $b_i(t)$ to the other player. In the second stage, the players play γ, using the signaled joint behavioral utility $b(t)$. This means that in the second stage, the players adopt a NE of the behavioral game at t, that is, of the game specified by the value of b at t. So, in EOP, the strategy of player i in game iteration t is not pre-fixed, but depends on signals from the other players at t. Next, in analogy to much of evolutionary game theory, in EOP the mixed strategy profile over X at iteration t (which, in EOP, is the NE of the behavioral game at t) is evaluated under the actual, joint underlying utility u. This gives values of the underlying expected utilities of the players for game iteration t, and iteration t ends.

In EOP, natural selection uses the expected underlying utilities from the final stage of the two-stage game at iteration t as "fitness" values, to guide the evolution of player behavioral utilities to the next iteration $t + 1$. This is exactly analogous to how, in much of evolutionary game theory, the expected values of underlying utilities guide the evolution of player strategies. Accordingly, in EOP the appropriate equilibrium concept for the long timescale process in which the joint behavior utility b is determined is evolutionarily stable strategies (ESS).

Note that in EOP it is implicitly assumed that player i cannot signal one behavioral utility $b_i \neq u_i$ and then play the behavioral game using utility u_i; if he could engage in dishonest signaling, then play would simply collapse to the NE of the underlying utilities. So, although it is not mentioned in the EOP literature, there must be some implicit mechanism that forces honest signaling. In other words, like the interdependent preferences and QRE experimental work, the EOP framework takes it as given that players can acquire common knowledge of one another's binding objective functions before the start of play.

1.3 Difficulties with EOP

The central insight of EOP — that games occur on more than one timescale — is a very powerful one. Unfortunately though, calculations concerning EOP are quite complex. Because of this, much of the formal work in EOP restricts attention to a game (or set of coupled games) with a single symmetric utility function shared by all the players. For similar reasons of tractability, much of EOP concerns only two-player games, often with only binary move spaces. Also, for mathematical tractability, EOP typically requires that the population be infinite. However, the equilibria of natural selection when the population is infinite can differ substantially from the equilibria when the population is finite, even if the finite population is very large (Fogel et al.,

1997; Ficici et al., 2005). All of these simplifications restrict the ability of EOP to make predictions concerning either laboratory or field experiments.

Other difficulties with EOP arise in the simple scenarios where one can do the calculations. For example, in some EOP games, the evolutionary dynamic process has no equilibrium at all; EOP cannot make predictions for such games. A similar difficulty is that the results in EOP typically vary with the initial characteristics of the population being evolved.

These and other difficulties have prevented EOP from being used to analyze real-world behavior. In particular, they have prevented EOP from being used to analyze either the results of fitting interdependent preference parameters to experimental data or the results of fitting logit QRE parameters to experimental data.

1.4 Changing the long timescale in EOP

In this paper, we introduce a modification of this earlier version of EOP. In this modification we allow for broader sets of behavioral objective functions than just behavioral utilities. More importantly, we also change the long timescale process to involve learning by a single individual across his lifetime, rather than natural selection occurring across multiple individuals' lifetimes. This modification allows us to replace the ESS solution concept with the NE solution concept.

By making this modification, we avoid all the mathematical difficulties of the earlier work on EOP; calculations become far more tractable. This enables us to extend the analysis of two-timescale games to situations far beyond those that the earlier EOP frameworks can analyze, for example, to games with more than two players, more than binary move spaces, and so on. Indeed, changing the long timescale enables us to make many theoretical predictions for the outcomes of game theory experiments, something EOP cannot do.

Note that persona games are not the first model in economics to consider two-timescale games where both the long and the short timescales occur within a single individual's lifetime. For instance, Fudenberg and Levine (2006) posit a dual-self model in which a long-run self can (at a cost) alter the baseline preferences of the current period's myopic short-run self. This differs from the persona framework in that it models a decision problem rather than a strategic setting and in that the agent choosing preferences at any given time is doing so for only a subset of the outcomes that he cares about.

To avoid confusion with the earlier EOP work where the long timescale process is natural selection, we use the term *persona* to indicate the behavioral objective function learned in the

long timescale process we consider here. We refer to the associated two-timescale game as a *persona game*.

1.5 Paper overview

We begin Section 2 by discussing how persona games appear to arise in the real world, as "emotions" or "moods" that people learn to adopt for real-world games. This discussion addresses an issue implicitly assumed in both EOP and the experimental work on interdependent preferences and the QRE: what is the physical process whereby the players gain common knowledge of one another's objective functions before the start of play?

Next in that section we provide a semiformal definition of persona games, along with illustrations of them. This presentation discusses some of the differences between the persona game framework and the EOP framework. We end this section by relating persona games to work other than EOP.

As mentioned above, EOP assumes perfect signaling of behavioral objective functions. In Section 3 we present a fully formal definition of persona games for this simplest case of perfect signaling. We also establish some of the basic mathematical properties of persona games.

In Section 4 we consider persona behavioral objective functions that are interdependent preference utility functions. We show how persona games involving such personas can explain altruistic behavior. We concentrate our analysis on the Prisoner's Dilemma (PD). In particular, we demonstrate how the Pareto efficient (cooperate, cooperate) outcome of the PD can arise as jointly rational behavior of a persona game — but only for certain ranges of PD bimatrices.

In this section we also derive a crowding out effect for the PD (Bowles, 2008). This shows that crowding out can be explained without relying on ad hoc notions of "different kinds of moral sentiments."

We end this section with a novel prediction concerning real-world play of the PD: an unavoidable tradeoff between the utility gain to the players for jointly cooperating rather than jointly defecting and the robustness of a persona equilibrium that induces such joint cooperation. The sobering implication is that the greater the benefit of an equilibrium where people are altruistic, the more fragile that equilibrium.

In Section 5 we show that empirical data on the Traveler's Dilemma (Basu, 1994; Becker et al., 2005; Capra et al., 1999) are explained by persona games, and briefly describe how the persona games framework explains empirical data on the Ultimatum Game.

Next, in Section 6, we extend the formalism to the case of noisy signaling of personas. We also extend it to the case where signaling is done via behavior. More precisely, we extend the formalism to a multi-stage game scenario, where a persona is chosen by player i in the first stage and cannot be modified afterward. In this scenario, the behavior of player i in early stages can signal his persona, and therefore modify the behavior of the other players in later stages. (This second extension takes our analysis beyond the kind of signaling of behavioral objective functions assumed in EOP.)

Then, in Section 7, we discuss extensions of the persona game framework and how that framework can also describe other kinds of scenarios, for example, those involving binding contracts. In that section we also discuss general issues about how persona games can transpire in the real world. In particular, it is in this section that we discuss "personalities," which are mappings from an underlying game to the persona the player chooses for that game.

Final comments are in Section 8.

Throughout this paper, just as in most of the work in EOP, for simplicity we restrict attention to finite games. Also, as in most of the work in EOP, we do not consider the issue of how personas are signaled among players before start of the behavioral game.[3] Furthermore, we do not consider the issue of why some persona sets arise rather than others. In this, we are just like the interdependent preferences literature, which simply takes the parametric form of the interdependent preferences as exogenous. (Discussion of why persona sets must be finite in the real world arise below in Proposition 2 and Section 7.1.) Our focus in this paper is on the implications of people's ability to adopt personas and, in particular, on how this may explain why people's behavior exhibits some parameter values rather than others in other-regarding preferences and QREs.

2 Persona games

2.1 Adopting and signaling personas

As illustrated by the QRE and interdependent preferences work, often people adopt and signal a behavioral objective function before the play of a game. Colloquially, often before they play

[3] It is not just EOP and persona games that do not consider this issue. Almost all papers that involve conventional normal form games also take it for granted that somehow the utility functions the players will use in a game are communicated before the start of play (that is, common knowledge holds), without analyzing how that communication arises.

a game, human beings have adopted a binding "persona" that fixes how they will interact with the other players in the game and have signaled that persona to the other players. Interestingly, it seems that this is often done so that the outcome of the resultant game is better for them than would otherwise be the case. This phenomenon of varying the persona we adopt to improve the outcome of the behavioral game is so widespread it has even entered common discourse, for example, in the statement "the workplace is full of chameleons who adopt a different persona each day" (Stern, 2008).

In some circumstances a person may choose his persona, signal it to others, and commit to using it, all in a conscious manner. An example of this is whenever we interact with a very young child by acting dumber than we are. Although adult - child interactions are quite elaborate games, it seems that this "acting dumb" behavior corresponds roughly to shrinking the exponent in one's logit quantal response function for the duration of the game with the child and that it results in a more efficient QRE equilibrium of that game with the child than would arise if the adult did not "act dumb."

Such conscious signaling of personas seems to occur most often in interactions with young children. This may be because young children are too immature in their social skills to realize that they are interacting with a consciously contrived artificial persona. With adults, such conscious adoption and signaling of personas typically does not work, because the signaled persona is usually not believable to someone with fully developed social skills.

This does not mean that adults do not choose a persona, signal it to others, and commit to using it during play of a game. However, with adults it seems that all this is rarely done consciously, but rather in a semi-conscious or even unconscious manner. Such semi-consciously adopted personas may be "moods" or "emotions" that are induced by the person's knowledge of the game he is about to play. Such personas, once adopted, can only be discarded with substantial time and effort.

In most circumstances, learning which such persona is best suited to a given game is something that an individual gains over his entire lifetime via personal experience, communication with others, and so on, and becomes manifested as a "habit." All such learning processes occur on a long timescale. In contrast, the effects of the learned persona — the effects of the habit — occur during the much shorter timescale on which the game is played.

In fact, it appears that there is a neural feedback mechanism forcing signaled and actual personas to match. Loosely speaking, if you *act* a certain way, then you start to *feel* that way (for example, during communication; see Stephens et al. (2010)). Our very biochemistry helps to

ensure that the persona governing our behavior matches the one we signal. This feedback helps to stabilize personas, making them difficult to discard. It underlies the common observations that people cannot "change their mood on arbitrarily short timescales," and therefore their actions often "betray the way they feel," that is, signal their persona. (See the discussion in Frank (1987); Frith (2008), and of costly signaling in general in Spence (1973, 1977); Lachmann et al. (2001).) In short, it would appear that humans have some ability, at least at a subconscious level, to adopt a persona / emotion that is binding on them and that they are forced to semi-honestly signal.[4]

To give a graphic example, it appears that in the Ultimatum Game, the rate of rejection of low offers by responders when they have little time to decide whether to reject is far higher than when they have more time to decide (Grimm and Mengel, 2010). In other words, on timescales so short that an adopted persona cannot be easily discarded (in this case the persona of refusing to allow oneself to be treated unfairly), such a persona governs an individual's publicly observable actions. However, on longer timescales, a person can, with effort, discard the adopted persona, and instead maximize his utility.

The semi-conscious signal of such a mood / persona can occur in the person's tone of voice, body language, and so on.[5] Empirically, it is typically quite difficult for someone to consciously fake such a bodily signal of a mood / persona. For example, it typically takes a lot of study —- perhaps years of study — for an aspiring stage actor to learn how to be completely convincing. Indeed, in the technique of "method acting," a thespian does not so much learn how to falsify the signal of his persona to match that of his character, as learn how to adopt the persona of that character — in which case the signal he sends to the audience is, in fact, honest.

There are numerous examples of adaptations that facilitate such signaling of personas. One is the fact that in humans (but not in other species), the whites of the eyes are visible, so that where we look telegraphs our attention (which, in turn, provides much information about our emotion / mood). Another is the anomalously large number of human facial muscles (compared with those of other species) whose "purpose" seems to be to signal emotions. (See, for example, the copious experimental data establishing unconscious control of facial expressions (Ekman, 2007).)

Note though that the signal of a semi-consciously adopted persona does not necessarily have

[4]Of course, a person's mood / persona can change drastically and suddenly in response to shocking new information. For example, if you were in an altruistic mood, and were then told that the person with whom you were interacting tortures animals, your mood would change. The point here is that absent such a shock, personas cannot change quickly.

[5]Such persona signals are not only faked but overly exaggerated, when one consciously adopts a persona for an interaction with a young child.

to be via body language. Indeed, so long as a particular persona adopted by someone cannot be easily discarded (that is, it is indeed binding), then that persona will necessarily be signaled via many kinds of publicly observable actions by that person. From this perspective, body language signals of personas are simply adjuncts to persona signals in the form of publicly observable behavior, helping reduce noise in those persona signals. The crucial thing is the ability to adopt a binding persona in the first place.

The question addressed in this paper is: what are the strategic implications of this ability to adopt a binding persona, assuming the person can find some way to signal that persona. Loosely speaking, how can a person (or at least his subconscious habits) exploit the fact that he can make binding commitments to adopt a persona? Is it ultimately advantageous to someone to adopt (and honestly signal) a finite rationality? Is it ultimately advantageous to someone to adopt (and honestly signal) an interdependent utility function? What kinds of "nonrational" behavior can be explained endogenously this way?

Given this focus, in this paper we do not analyze the evolutionary stability of the empirical fact that people have the ability to adopt a binding persona. Nor do we analyze why some people are better than others at obscuring such persona signals, or why some people are better than others at choosing their personas consciously rather than subconsciously. Such issues are important, but secondary to this paper's focus. (They also are not analyzed in the EOP literature.)

2.2 Semiformal definition of persona games

Stated informally as above, variants of the persona-based explanation of nonrationality predate EOP, going back at least to the 1950s (Raub and Voss, 1990; Kissinger, 1957; Schelling, 1960), and arguably back to antiquity (Schelling, 1960). In particular, they played a prominent role in the formulation of cold war policies such as mutual assured destruction. In this paper we introduce a formal framework for this persona-based explanation of nonrationality, by modifying the long timescale of EOP. To ground the intuition, we begin in this subsection with a semiformal definition of persona games.

As in EOP, in a persona game each player i has an associated underlying utility function $u_i : X \to \mathbb{R}$, where X is the finite joint move space. Also as in EOP, in persona games each player i has a set of possible behavioral objective functions, that is, personas, which we here write as A_i. Each such persona maps the set of possible mixed strategy profiles over X into $\mathbb{R}$. In addition to personas that take the form of expected utilities, the usual choice in EOP, in persona games we

also allow personas other than expected utilities.[6]

Just as in EOP, in persona games the underlying utility functions and sets of possible personas are used to define a two-stage game. In the first stage, every player i samples an associated distribution $P(a_i \in A_i)$ to get a persona a_i. Next, just as in EOP, each player i honestly signals that a_i to the other players.[7] Then, in the second stage, the players play a NE of the behavioral game, that is, a NE of the strategic-form game with joint move space X and player objective functions given by those signaled personas.[8] The mixed strategy profile of that NE determines the expected values of the players' underlying utility functions, just as in EOP.

EOP considers an infinite sequence of such two-stage games, where the personas are modified at the end of each game by natural selection. This is where persona games differ from EOP. In persona games we replace EOP's repeated games with a single game. In addition, we introduce the common knowledge assumption. In the scenarios we consider here, this means that before signaling his persona, each player knows his own set A_i, the sets $\{A_j\}$ of the other players, and the underlying utility functions of all players. Each player i then uses this information to choose the distribution $P(a_i)$ to be sampled to generate the signaled a_i (rather than having evolution set $P(a_i)$, as in EOP). Players do so to maximize the associated expected value of their underlying utility, as evaluated under the NE of the behavioral games.

This modification means that the distributions $P(a_i)$ themselves are NE, of the full, multi-stage game. In contrast, in EOP the distributions $P(a_i)$ change from one game to the next, stabilizing only if they settle on an ESS.

Note that the players in the first stage of the persona model are simply playing a standard game, properly interpreted. Their joint move is the joint persona they adopt a. The utility function of player i in this game is the mapping from all possible a's to the expected underlying utility of the (NE of the) behavioral game specified by a. It is this full game that we have in mind when we use the term "persona game."

[6]This extra breadth is needed, for example, to model the irrational player, who prefers a uniform mixed strategy to all other possible mixed strategies, rather than just being indifferent over his moves. See also von Widekind (2008).

[7]A process essentially equivalent to such signaling is implicit in the conventional analysis of nonrepeated complete information normal form games. "Common knowledge" of utility functions in such a game implicitly requires some sort of communication of all the utility functions among all the players before the game begins, communication that must be honest, in that it leads the opponent(s) of each player i to the correct conclusion about i's utility function. In particular, this requirement holds in most of the work on other-regarding preferences.

[8]Note that just as in EOP, in persona games every player i is assumed to play the behavioral game using the persona that he signals, that is, it is assumed that the signals are binding. The same kind of assumption arises in games of signaled binding commitments, for example, (Renou, 2008). See Sections 6.2 and 2.4 below.

Note that in persona games we assume that the underlying utility function u_i of every player i is provided exogenously, as is the set of i's possible adopted personas A_i. (This assumption is also made in EOP.) Physically, the determination of these exogenous factors may occur through an evolutionary process, either biological and/or cultural. However — as in EOP — we do not model such processes here.

We refer to each A_i as a *persona set*. The persona sets that humans use in field experiments appear to be quite elaborate. It seems that they are often at least partly determined by social conventions and norms. In addition, it seems that people sometimes use different persona sets for different joint underlying utility functions u.[9]

Here, for simplicity, we do not address such complexities. Nor do we address the issue of why some persona sets, but not others, arise in human society. (This limitation is shared with the work on interdependent preferences) Rather, we concentrate on some simple, "natural" personas suggested by the interdependent preferences and QRE literatures, to investigate the explanatory power of persona games.

2.3 Illustrations of persona games

Adopting a persona that disagrees with one's true utility would seem to be nonrational. To illustrate how it can actually be rational, first consider a game involving two players, Row and Col, each of whom can choose one of two moves, (Up, Down) ($\{\mathbf{U}, \mathbf{D}\}$) for Row, and (Left, Right) ($\{\mathbf{L}, \mathbf{R}\}$) for Col. Let the utility function bimatrix (u^R, u^C) be

$$\begin{bmatrix} (0,0) & (6,1) \\ (5,5) & (4,6) \end{bmatrix}, \tag{4}$$

where, as usual, in each cell Row's utility comes first and Col's second. The joint move $(\mathbf{U}, \mathbf{R})$ is the only Nash equilibrium of this game. At that NE, $\mathbb{E}(u^C) = 1.0$. Now, say that rather than being rational, Col adopts a persona where he is perfectly irrational. That is, he adopts a free utility with a λ of 0. This means that he commits to choosing uniformly randomly between his two moves, with no evident concern for the resultant value of his utility function, and therefore no concern for what strategy Row adopts. Then, Row would have expected utility of 3 for playing

[9]The general term "interdependent preferences" is sometimes used in the literature to refer to behavior in multiple games, rather than behavior in just a single game, which is how we use the term in this paper. Example 1 in Section 7 illustrates a multi-game situation that is sometimes called "interdependent preferences" in the other sense, and it shows how to formalize such behavior in terms of multiple persona sets.

U, and of 4.5 for playing **D**. So, if Row were rational, given that Col is irrational, Row would play **D** with probability 1.0. Given that Col plays both columns with equal probability, this in turn would mean that $\mathbb{E}(u^C) = 5.5$.

So, by being irrational rather than rational, Col has improved his expected utility from 1.0 to 5.5. Such irrationality by Col allows Row to play a move that Row otherwise would not be able to play, and that ends up helping Col. This is true even though Col would increase his expected utility by acting rationally rather than irrationally if Row's strategy were fixed (at **D**). The important point is that if Col were to act rationally rather than irrationally while Row's rationality were fixed (at full rationality), then Col would decrease his expected utility. This phenomenon can be seen as a model of the common, real-world scenario in which someone "acts dumber than he is" (by not being fully rational) and benefits by doing so.

Now, consider a game where each player has four possible moves rather than two, and the utility functions (u_R, u_C) are:

$$\begin{bmatrix} (0,6) & (4,7) & (-1,5) & (4,4) \\ (-1,6) & (5,5) & (2,3) & (7,4) \\ (-2,1) & (3,2) & (0,0) & (5,-1) \\ (1,1) & (6,0) & (1,-2) & (6,-1) \end{bmatrix}. \tag{5}$$

A NE of this game is the joint pure strategy where Row plays the bottom-most move, and Col plays the left-most move. In other words, the "joint rationality move" of (rational, rational) results in an ultimate payoff of $(1, 1)$.

In contrast, say that both players are anti-rational in that they both want to minimize their utility functions. Now, an equilibrium occurs if Row plays the top-most row and Col plays the right-most column. The resultant ultimate payoff is $(4, 4)$ rather than $(1, 1)$.

There are also two intermediate cases, where one player is rational and one is anti-rational, (rational, anti-rational) and (anti-rational, rational). Equilibria for those two cases are given by the remaining two entries on the diagonal line through the bimatrix in Table (5) extending from bottom-left to top-right. Plugging in those values results in the following bimatrix of the ultimate

payoffs for the four possible joint personas:

	Col rationality	
	$-\infty$	$+\infty$
Row rationality		
$-\infty$	(4, 4)	(3, 2)
$+\infty$	(2, 3)	(1, 1)

, (6)

where full rationality is indicated by the value $+\infty$, and anti-rationality by $-\infty$.

This bimatrix defines a "persona game," which controls what personas the players should adopt. $(-\infty, -\infty)$ is a dominant NE of this persona game. At this joint persona, neither player would benefit from changing to rational behavior, no matter which persona his opponent adopted. Note, in particular, that $(+\infty, +\infty)$ is not a NE of the persona game. So, if the players are sophisticated enough to play the persona game with each other rather than the underlying game, they will both act anti-rationally. Stated more colloquially, in this game both players have an incentive to be "self-destructive," regardless of whether their opponent is.

In this game, when both players are anti-rational, the resultant utility is better for both of them than when both are rational. So (anti-rational, anti-rational) is not only a jointly dominant choice of rationalities, it is also Pareto efficient. This is not always the case; there are games where (anti-rational, anti-rational) is the jointly dominant choice, but is not efficient. The implication is that in some, but not all, games, good public policy would induce the players to try to be self-destructive.[10]

2.4 Relation between persona games and EOP

Since persona game equilibria are NE, all the powerful techniques for analyzing NE can be used to predict what $P(a_i)$'s arise in the real world. In contrast, in EOP the distributions analogous to $P(a_i)$ are equilibria of a dynamic evolutionary process, and often many techniques for analyzing NE cannot be applied. This is why it is easier to generate theoretical results in the persona framework than in EOP. In particular, in the persona framework, one does not have to restrict

[10]It is important to realize that the benefit to a player of being anti-rational is not the same as the benefit that noncredible threats can provide in certain extensive-form games. A player making a noncredible threat says to his opponent, "If you do α, I'll do something that hurts me — but also hurts you. So you must not do α, and I will exploit that." In contrast, an anti-rational player says "*No matter what you do*, I will do something that will hurt me."

any of the analysis to symmetric games involving few (for example, two) players with small (for example, binary) move spaces, as one often must in the EOP.

The use of NE techniques also means that every game in the persona framework has an equilibrium, as formally established below. So the persona framework can always make a prediction of how humans will behave, in contrast to EOP. Furthermore, no initial characteristics of a population are relevant in the persona framework, and we make no physically impossible assumptions about such a population. Again, this contrasts with EOP.

As mentioned above, since behavioral utilities vary across the population in EOP, there must be some mechanism that before the start of the game at any instant t forces each player i to send an honest signal to all players $-i$ telling them the behavioral utility that player i will use in that time-t game. More precisely, recall that in the EOP u_i is the underlying utility of player i at time t, while $b_i(t)$ is the behavioral utility that i will actually use to decide his move in the game played at t (see Section 1.2). Define $s_i(t)$ as a signal that player i sends (inadvertently or otherwise) to all players $-i$ before the start of the time-t game. Then, common knowledge for that time-t game requires the existence of some mechanism that forces $s_i(t)$ to honestly reflect $b_i(t)$. This mechanism must be unavoidable; if it could be subverted, then there would not be common knowledge in the game at t, and the EOP framework would not apply.[11]

Precisely what this mechanism is that forces honest signaling of $b_i(t)$ is not specified in the EOP. Presumably it involves i's body language, vocal tone, and so on. Nor is it specified how this mechanism comes to exist in the first place. (One possibility is that it arises via some sort of natural selection process.) However the mechanism operates, given that it enforces honest signaling of $b_i(t)$, there is no reason that player i would not learn to take advantage of it. He would do this by strategically choosing what behavioral utility $b_i(t)$ to use — knowing that $b_i(t)$ will be honestly signaled to the other players — so as to optimize the resultant value of $b_i(t)$ in the time-t game. In essence, persona games are analyses of such strategic exploitation of the unavoidably honest signaling mechanism.

In this sense, if EOP is a valid description of real behavior, then persona games should be also. However, the persona game framework goes further. It can also be used to analyze scenarios where the signaling of personas is noisy (Section 6.1), or even non-existent (Section 6.2).

Finally, and perhaps most importantly, the persona framework offers explanations for appar-

[11] Arguably, there must be such signaling even in standard, full-rationality, strategic-form games. After all, for common knowledge to apply, there must be some way that *before* the start of play, each player i can ascertain the utility function that player $-i$ will actually use in the game. While the theory of epistemics, knowledge operators, and so on, is well-developed, it mostly takes such a signaling process as implicit and given.

ent nonrationality even in anonymous single-shot games where the players know they will never again interact. This corroborates the conclusion in Henrich et al. (2006) that "cooperation can (be explained), even among non-kin, in situations devoid of repeat interaction." However, the persona framework shows that this conclusion holds even without punishment and without genes for non-kin altruism (which have not been found on the human chromosome), whereas Henrich et al. (2006) assumes both of those. Cooperation can exist for purely self-interested reasons.

2.5 Previous work related to persona games

There is some early, semiformal work related to the persona concept (Raub and Voss, 1990; Kissinger, 1957; Schelling, 1960). However, there are many subtleties that a fully formal persona game framework must address. For example, for some underlying games and persona sets, some joint personas result in a behavioral game that has more than one NE. To define the associated utility functions in the persona game, either those multiple NE must somehow be summarized or one of them must be selected somehow. These subtleties are not addressed in the early, semiformal work.

Moreover, much of that early, semiformal work allows infinite persona sets. However, in many strategic situations it takes a very large computational effort for a player to calculate his optimal persona. (Crudely speaking, for every possible joint persona, the player has to calculate the associated behavioral game equilibria, and only then can he calculate the persona game equilibria.) Accordingly, having an infinite persona set would often place a very large computational burden on a real-world human player. For a player to have a large persona set would also often place a large burden on the other players in the persona game, who must consider all possible opposing personas before choosing their own.

Note also that the processes typically used for signaling personas (body language, repeated behavior, and so on) have low information channel capacity (Mackay, 2003; Cover and Thomas, 1991). Loosely speaking, those processes cannot convey very much information in a reasonable amount of time. This means that it is physically impossible for those processes to quickly and accurately convey a persona choice if there are too many possible persona choices.

In light of these difficulties, it is not surprising that in the real world persona sets seem to be finite, in contrast to the infinite persona sets considered in the early, semiformal work. Indeed, the finiteness of persona sets is taken for granted in all the literature on fitting parameters of interdependent preferences and/or QRE to experimental data; in none of that literature is there an

attempt to fit a completely arbitrary objective function to the experimental data.

Following the appearance of an early version of this paper (Wolpert et al., 2008), a preprint appeared that is both fully formal and related to the persona concept (Winter et al., 2009). This other work allows infinite persona sets, just as in the early, semiformal work, with the attendant problems. Indeed, in this work, these problems mean that the analysis for more than three players is trivial, so the focus is on two-player games. Furthermore, the focus is restricted to equilibria in expected utilities rather than objective functions. Accordingly, the analysis cannot consider equilibrium concepts like the QRE. In addition, the model in Winter et al. (2009) does not consider extensions like the noisy persona game, where the players do not perfectly observe one another's personas. Nor does it consider "sticky" personas, which are multi-stage games where a player is not allowed to change his persona between stages, and so his behavior in the early stages can signal his persona without any need for body language or the like.

There are other, fully formal studies that are loosely related to the persona concept (Frank, 1987; Israeli, 1996; Becker, 1976; De Long et al., 1990). These studies each apply a model of very limited scope to a restricted (and often rather complicated) scenario (for example, the studies in Israeli (1996); Frank (1987)). Many of these studies focus so narrowly on the PD that their results do not apply to other types of irrationality. As a result, these studies do not concern the persona concept in its full generality.

Also recently, some work has appeared (Renou, 2008) that is related to the "binding commitments" interpretation of personas discussed in Section 7 below. More recently still, some work has appeared that can be viewed as an investigation of personas in an experimental context (Andrade and Ho, 2009).

3 Formal definition of persona games

3.1 Notation

Define $\mathbb{N} \equiv \{1, 2, \ldots\}$, fix a positive integer N, and define $\mathcal{N}$ as the integers $\{1, \ldots, N\}$. we will occasionally use curly brackets to indicate a set of indexed elements where the index set is implicit, being all $\mathcal{N}$; for example, $\{X_i\}$ is shorthand for $\{X_i : i \in \mathcal{N}\}$. For any set Z, $|Z|$ indicates the cardinality of Z. Unless explicitly stated otherwise, we always assume the standard topology for any Euclidean space.

We will use the integral symbol with the measure implicit. So, for example, for finite X,

"$\int_X dx$" implicitly uses a point-mass measure and therefore means a sum. Similarly, we will be loose in distinguishing between probability distributions and probability density functions, using the term "probability distribution" to mean both concepts, with the context making the precise meaning clear if only one of the concepts is meant. We will write $\delta(a, b)$ to mean the Dirac or Kronecker delta function, as is appropriate for the space containing a and b.

We use "$P(.)$" to indicate a probability distribution (or density function as the case might be). An upper-case argument of $P(.)$ indicates the entire distribution (that is, the associated random variable), and a lower-case argument indicates the distribution evaluated at a particular value. When defining a function, the symbol "$\triangleq$" indicates that the definition holds for all values of the listed arguments. So, for example, "$f(a, b) \triangleq \int dc \; r(a)s(b, c)$" means that the definition holds for all values of a and b. (In contrast, "$f(a, b) = \int dc \; r(a)s(b, c)$" is an equation specifying some value(s) of the pair (a, b).)

The unit simplex of possible distributions over a space Z is written Δ_Z. Given two spaces A, B, we write $\Delta_{A \times B}$ to mean the unit simplex over the Cartesian product $A \times B$. Similarly, $\Delta_{A|B}$ indicates the set of all functions from B into Δ_A, that is, the set of all conditional distributions $P(A \mid B)$. Given a set of N finite spaces $\{X_i\}$ we write $X \equiv \bigtimes_{i \in \mathcal{N}} X_i$ and for any $x \in X$, use x_i to indicate the ith component of x, and we use a minus sign before a set of subscripts of a vector to indicate all components of the vector other than the indicated one(s). For example, we write $X_{-i} \equiv \bigtimes_{j \in \mathcal{N} : j \neq i} X_j$ and use x_{-i} to indicate the ordered list of all components of x except for x_i.

We define Δ_X as the set of distributions in $q \in \Delta_X$ that are product distributions, that is, that are of the form $q(x) \triangleq \prod_{i \in \mathcal{N}} q_i(x_i)$. Similarly, we define $\Delta_{X_{-i}}$ as the set of product distributions in $\Delta_{X_{-i}}$. In the usual way, for any $q \in \Delta_X$ and $i \in \mathcal{N}$, we define the distribution $q_{-i} \in \Delta_{X_{-i}}$ as $\prod_{j \in \mathcal{N} : j \neq i} q_j$.

Given any set of N finite (pure) strategy spaces, $\{X_i\}$, we refer to any function that maps $\Delta_X \to \mathbb{R}$ as an *o*bjective function for X. As an example, for a fixed utility function $u : X \to \mathbb{R}$, the expected value of u under $q \in \Delta_X$, $\mathbb{E}_q(u)$ is an objective function. Any pair of a set of finite strategy spaces and an associated set of one objective function U_i for each strategy space i is called a (strategic-form) objective game (Wolpert, 2009). Often, we leave the indices on the elements of an objective game implicit, for example, referring to (X, U) rather than $(\{X_i\}, \{U_i\})$. Unless explicitly stated otherwise, we assume N is finite as is X_i for all $i \in \mathcal{N}$.

Best responses in objective games, extensive form objective games, NE equilibria of objective games, and Bayesian objective games are defined in the obvious way. In particular, we write the

NE of an objective game (X, U) as

$$\mathcal{E}(X, U) \triangleq \{q \in \Delta_{\mathcal{X}} : \forall i \in \mathcal{N}, \forall \hat{q}_i \in \Delta_{X_i}, U_i(\hat{q}_i, q_{-i}) \leq U_i(q_i, q_{-i})\}.$$

We will sometimes be loose with the terminology and refer to player i as "making move $q_i \in \Delta_{X_i}$", even though her pure strategy space is X_i, not Δ_{X_i}.

As an example of an objective game, say each objective function U_i is a free utility, $\mathbb{E}_q(u_i) + \beta_i^{-1}\mathcal{S}(q_i)$ for some associated utility function u_i, so the objective game NE is a QRE. Then, a rational player i can be modeled by setting $\beta_i = \infty$. Similarly, an "irrational" player i, who always plays a uniform mixed strategy, is one with rationality $\beta_i = 0$. An anti-rational player i, who always acts to *hurt* himself, can then be modeled by setting $\beta_i = -\infty$.

3.2 Persona worlds

Define a *persona world* as any triple

$$(\{X_i : i \in \mathcal{N}\}, \{U_i : i \in \mathcal{N}\}, \{A_i : i \in \mathcal{N}\}), \tag{7}$$

where $\{X_i\}$ is a set of N finite strategy spaces, $\{U_i\}$ is an associated set of N objective functions for $X = \bigtimes_i X_i$, and each A_i is a set of objective functions for X. As shorthand, we typically write such a persona world as (X, U, A). For simplicity, in this paper we will always take any A_i to be finite. We refer to an $a_i \in A_i$ as a *persona* of the ith player, with A_i the associated *persona set*. Note that any such persona is a mapping from $\Delta_{\mathcal{X}}$ into $\mathbb{R}$.[12]

We refer to any N-tuple $a = (a_1, \ldots, a_N) \in A \equiv A_1 \times \ldots \times A_N$ of every player's persona as a "joint persona" of the players. We write $\Delta_{\mathcal{A}}$ to mean the members of Δ_A that are product distributions, that is, that are of the form $P^A(a) = \prod_{i \in \mathcal{N}} P_i^A(a_i)$. We define $\Delta_{\mathcal{A}_{-i}}$ similarly. We also define $\Delta_{\mathcal{X}|A}$ to mean the members of $\Delta_{X|A}$ that are product distributions, that is, that are of the form $P(x \mid a) = \prod_{i \in \mathcal{N}} P(x_i \mid a_i)$ for all $a \in A$. We make the analogous definition for $\Delta_{\mathcal{X}_{-i}|A_{-i}}$. We refer to (X, U) as an *underlying game*, and any (X, a) for some $a \in A$ as a *behavioral game*.

[12] At the expense of more notation, we could extend the domains of each objective function U_i to be $\Delta_{\mathcal{X}} \times \Delta_{A_i}$. This would allow us to model scenarios in which a player of the persona game has *a priori* preferences over her possible personas.

3.3 Extended persona games

Consider an N-player persona world (X, U, A) where all spaces are finite, and suppose we have a set of $N + N|A|$ distributions $\{P(A_i) \in \Delta_{A_i}, q_i^a(X_i^a) \in \Delta_{X_i} : i \in \mathcal{N}, a \in A\}$ where each space X_i^a is a copy of X_i. Intuitively, we can view each distribution $P(A_i)$ as the mixed strategy of the ith persona player, and each distribution $q_i^a(X_i^a)$ as the mixed strategy of the behavioral player who corresponds to persona player i when the persona players adopt joint persona a. From now on, to simplify the presentation we will sometimes write $q(X_i^a)$ rather than $q_i^a(X_i^a)$. Say that the following two conditions are met:

$\forall i, \ \nexists \hat{P} \in \Delta_{A_i} \ :$

$$\int da_i da_{-i} \ \hat{P}(a_i) P(a_{-i}) U_i \left[\prod_{j \in \mathcal{N}} q(X_j^a) \right] \ > \ \int da_i da_{-i} \ P(a_i) P(a_{-i}) U_i \left[\prod_{j \in \mathcal{N}} q(X_j^a) \right] \tag{8}$$

and

$\forall i, a \in A, \ \nexists \hat{q} \in \Delta_{X_i} \ :$

$$a_i \left[\hat{q}(X_i) \prod_{j \neq i} q(X_j^a) \right] \ > \ a_i \left[q(X_i^a) \prod_{j \neq i} q(X_j^a) \right]. \tag{9}$$

Then we say that the $N + N|A|$ distributions $\{P(A_i) \in \Delta_{A_i}, q_i^a(X_i^a) \in \Delta_{X_i} : i \in \mathcal{N}, a \in A\}$ form an *extended persona equilibrium* of the persona world (X, U, A).

Intuitively, an extended persona equilibrium is an agent-representation equilibrium of a two-stage game that models the process of players choosing personas, signaling them to one another, and then playing the associated behavioral game. Note that for simplicity, we require that at any such equilibrium (q, P), each component q^a is a NE of the objective game (X, a), even if $P(a) = 0$. (This is analogous to requiring subgame perfection.) On the other hand, though, if for some a where $P(a) \neq 0$ there is more than one q^a satisfying equation (9), some of them may not be the q^a component of an extended persona equilibrium. Such a q^a is a solution to the coupled equations, equation (9), for that a, but is not part of a solution to the encompassing set of coupled equations given by equation (8) together with equation (9).

The following result is proven in Appendix B:

Theorem 1 *Let (X, U, A) be a persona world where all the underlying utilities are expected utili-*

ties and all the personas are either expected utilities or free utilities. Then there exists an extended persona equilibrium of (X, U, A).

In fact, there exists such an equilibrium that is "trembling hand perfect," using a definition appropriate for games involving free utilities. See Appendix B for details.

In addition, persona games become degenerate for any persona set that includes the indifferent objective function $\mathscr{I}$, which has the same value for all arguments q. (This objective function is given by the expectation of a constant-valued utility function.) More precisely, we have the following:

Proposition 2 *Let* (X, U, A) *be a persona world where all the underlying utilities are expected utilities and all the persona sets include the indifferent objective function. Let* x^* *be any pure strategy profile such that for all players i,*

$$U_i(x^*) \geq \min_{x_{-i}}[\max_{x_i}[U_i(x_i, x_{-i})]].$$

Then there exists an extended persona equilibrium of (X, U, A) *in whose behavioral game the players all adopt* x^* *with probability 1.0.*

Proof. Consider the following combination of a persona game profile and a behavioral game profile:

1. Every persona game player i adopts the indifferent persona. Formally, $P(a_i) = \delta(a_i, \mathscr{I})$ for all i;

2. Every behavioral game player i adopts x_i^* whenever the joint persona involves all players choosing to be indifferent. Formally, $q_i^{\vec{\mathscr{I}}}(x_i) = \delta(x_i, x_i^*)$ for all i, where $\vec{\mathscr{I}}$ is the persona profile of all indifferent;

3. If persona player i does not adopt the indifferent persona, then all behavioral players $j \neq i$ conspire to give player i the worst possible outcome for i. Formally, let $\vec{\mathscr{I}}_{-i}$ be the vector indicating that all personas other than a_i are the indifferent persona. Then we require that for any player i, the joint behavioral game strategy for the players other than i obeys

$$q_{-i}^{a_i, \vec{\mathscr{I}}_{-i}}(x_{-i}) = \delta(x_{-i}, x_{-i}^\dagger)$$

for all $a_i \neq \mathscr{I}$, where

$$x^{\dagger}_{-i} \equiv \text{argmin}_{x_{-i}}[\max_{x_i}[U_i(x_i, x_{-i})]].$$

4. The behavioral game strategies q_i^a for all other joint personas a are arbitrary.

By requirements 1 and 2, under the given combination of persona and behavioral game profiles, the behavioral game players do indeed play x^*. We must now confirm that this combination of profiles obeys equations 8 and 9.

First note that equation 9 is obeyed for the given persona profile $\vec{\mathscr{I}}$, since no behavioral game player can improve the value of his indifferent persona by changing his strategy. Next, hypothesize that some player i changes his persona from $\mathscr{I}$. By requirement 3, this means that the all behavioral game players $j \neq i$ will adopt the joint strategy $x^{\dagger}_{-i}$. By hypothesis, however, $U_i(x_i, x^{\dagger}_{-i}) \leq U_i(x^*)$, no matter what move x_i the behavioral player i adopts. So, by changing her behavioral game strategy, player j has not improved the value of her objective function. ■

So, persona games become degenerate in this sense for infinite persona sets, since such persona sets, in particular, contain the indifferent persona. However this degeneracy is not a great concern since there are already many physical reasons why infinite persona sets cannot occur in the real world, independent of these formal issues. (See Section 7.)

4 Persona games with other-regarding personas

To illustrate the breadth of persona games, we now consider personas for a player that involve the utilities of that player's opponents. Such personas allow us to model other-regarding preferences, for example, altruism and equity biases. If a player benefits by adopting a persona with such an other-regarding preference in a particular game, then that other-regarding preference is actually optimal for purely self-regarding reasons.

4.1 Personas and altruism

Let $\{u_j : j = 1, \ldots N\}$ be the utility functions of the original N-player underlying game. Have the persona set of player i be specified by a set of distributions $\{\rho_i\}$, with each distribution ρ_i being an N-dimensional vector written as $(\rho_i^1, \rho_i^2, \ldots, \rho_i^N)$. By adopting persona ρ_i, player i commits to

playing the behavioral game with a utility function $\sum_j \rho_i^j u_j$ rather than u_i. So, pure selfishness for player i is the persona $\rho_i^j = \delta(i, j)$, which equals 1 if $i = j$, 0 otherwise. "Altruism" then is a ρ_i^j that places probability mass on more than one j. ("Inequity aversion" is a slightly more elaborate persona than these linear combinations of utilities; for example, a completely unselfish inequity aversion could be modeled as the persona $[(N-1)u_i - \sum_{j \neq i} u_j]^2$.)

To illustrate this, consider the two-player, two-move underlying game with the following utility functions:

$$\begin{bmatrix} (2,0) & (1,1) \\ (3,2) & (0,3) \end{bmatrix}. \tag{10}$$

There is one joint pure strategy NE of this game, at $(\mathbf{U}, \mathbf{R})$. Say that both players i in the associated persona game have only two possible pure strategies, $\rho_i^j \triangleq \delta(i, j)$ and $\rho_i^j \triangleq 1 - \delta(i, j)$, which we refer to as *selfish* ($\mathscr{E}$) and *saint* ($\mathscr{A}$), respectively. Under the $\mathscr{E}$ persona, a player acts purely in their own interests, while under the $\mathscr{A}$ persona, a player acts purely in his opponent's interests.

As an example, if Row chooses $\mathscr{E}$ while Col chooses $\mathscr{A}$, then the behavioral game equilibrium for the underlying game in Table 10 is $(\mathbf{D}, \mathbf{L})$, since Row's payoff there is maximal. Note that this joint move gives both players a higher utility (3 and 2, respectively) than at $(\mathbf{U}, \mathbf{R})$, the behavioral game equilibrium when both players adopt the selfish persona. Continuing this way, we get the following pair of utility functions for the possible joint persona choices:

	Col ρ	
	$\mathscr{E}$	$\mathscr{A}$
Row ρ		
$\mathscr{E}$	(1, 1)	(3, 2)
$\mathscr{A}$	(0, 3)	(3, 2)

. (11)

The pure strategy NE of this persona game is $(\mathscr{E}, \mathscr{A})$, that is, the optimal persona for Row to adopt is to be selfish, and for Col is to be saintly. Note that both players benefit by having Col be saintly. One implication is that Row would be willing to pay up to 2.0 to induce Col to be saintly. Perhaps more surprisingly, Col would be willing to pay up to 1.0 to be a saint, that is, to

be allowed to completely ignore her own utility function, and work purely in Row's interests.

4.2 The Prisoner's Dilemma

In the case of the PD underlying game, other-regarding personas can lead the players in the behavioral game to cooperate. For example, say that each player i can choose either the selfish persona, or a "charitable" persona, $\mathscr{C}$, under which ρ_i is uniform (so that player i has equal concern for his own utility and for his opponent's utility). Consider the PD where the utility function bimatrix (u^R, u^C) is

$$\begin{bmatrix} (6,0) & (4,4) \\ (5,5) & (0,6) \end{bmatrix}, \tag{12}$$

so (defect, defect) is $(\mathbf{U}, \mathbf{R})$. Then, the utility matrix for a charitable persona is

$$\begin{bmatrix} 3 & 4 \\ 5 & 3 \end{bmatrix}. \tag{13}$$

So, for example, if the row player is selfish and the column player is charitable, the behavioral game is

$$\begin{bmatrix} 6,3 & 4,4 \\ 5,5 & 0,3 \end{bmatrix} \tag{14}$$

with an equilibrium at (defect, defect). The complete persona game is

	Player 2 persona	
	$\mathscr{E}$	$\mathscr{C}$
Player 1 persona		
$\mathscr{E}$	(4, 4)	(4, 4)
$\mathscr{C}$	(4, 4)	(5, 5)

. (15)

The efficient equilibrium of this persona game is for both players to be charitable, a choice that leads them to cooperate in the behavioral game. Note that they do this for purely self-centered reasons, in a game they play only once. This result might account for some of the experimental

data showing a substantial probability for real-world humans to cooperate in such single-play games (Tversky, 2004).

To investigate the breadth of this PD result, consider the fully general, symmetric PD underlying game, with utility functions

$$\begin{bmatrix} (\beta,\beta) & (0,\alpha) \\ (\alpha,0) & (\gamma,\gamma) \end{bmatrix}, \tag{16}$$

where $(\mathbf{D},\mathbf{R})$ is (defect, defect), so $\alpha > \beta > \gamma > 0$. We are interested in what happens if the persona sets of both players are augmented beyond the triple {fully rational persona $\mathscr{E}$, the irrational persona, the anti-rational persona} that was investigated above, to also include the $\mathscr{C}$ persona. More precisely, we augment the persona set of both players i to include a fourth persona $\rho_i u_i + (1-\rho_i)u_{-i}$. For simplicity, we set ρ_i to have the same value s for both players.

Define

$$R_1 \equiv \beta - s\alpha, \tag{17}$$

$$R_2 \equiv \gamma - (1-s)\alpha, \tag{18}$$

$$B \equiv \beta - \gamma. \tag{19}$$

Working through the algebra (see Appendix A), we first see that neither the nonrational nor the anti-rational persona will ever be chosen. We also see that for joint cooperation in the behavioral game (that is, $(\mathbf{L},\mathbf{T})$) to be a NE under the $(\mathscr{C},\mathscr{C})$ joint persona choice, we need $R_1 > 0$ (see Appendix A). If instead $R_1 < 0$, then, under the $(\mathscr{C},\mathscr{C})$ joint persona, either player i would prefer to defect, given that $-i$ cooperates.

Note that R_1 can be viewed as the "robustness" of having joint cooperation be the NE when both players are charitable. The larger R_1 is, the larger the noise in utility values, confusion of the players about utility values, or some similar fluctuation would have to be to induce a pair of charitable players not to cooperate. Conversely, the lower R_1 is, the more "fragile" the cooperation is, in the sense that the smaller a fluctuation would need to be for the players not to cooperate.

Given that $R_1 > 0$, we then need $R_2 > 0$ to ensure that each player prefers the charitable persona to the selfish persona whenever the other player is charitable. R_2 can also be viewed as a form of robustness, this time of the players' both wanting to adopt the charitable persona in the

first place.

Combining provides the following result:

Theorem 3 *Consider a two-player persona world* (X, U, A) *where each* X_i *is binary,* U *is given by the generalized PD with payoff matrix in equation* (16)*, and each player* i *has the persona set* $A_i = \{U_i, sU_i + (1-s)U_i\}$ *for some* $0 \leq s \leq 1$*. In the associated (unique, pure move) extended persona equilibrium, the joint persona move is* $(\mathscr{C}, \mathscr{C})$ *followed by* $(\boldsymbol{U}, \boldsymbol{L})$ *whenever* $s \in (1 - \frac{\gamma}{\alpha}, \frac{\beta}{\alpha}]$.

For our range on allowed s's to be nonempty requires that $\gamma > \alpha - \beta$. Intuitively, this means that player i's defecting in the underlying game provides a larger benefit to i if player $-i$ also defects than it does if $-i$ cooperates. It is interesting to compare these bounds on α, β, and γ to analogous bounds, discussed in Nowak (2006), that determine when direct reciprocity, group selection, and so on can result in joint cooperation being an equilibrium of the infinitely repeated PD.

Now, say that one changes the underlying game of equation 16 by adding a penalty $-c < 0$ to the payoff of every player i if he defects, giving the bimatrix

$$\begin{bmatrix} (\beta, \beta) & (0, \alpha - c) \\ (\alpha - c, 0) & (\gamma - c, \gamma - c) \end{bmatrix}. \tag{20}$$

In other words, one introduces a material incentive c to try to deter defection. Say that $cs > \gamma - (1-s)\alpha$, and that both $\gamma > c$ and $\alpha - \beta > c$. Then, the new underlying game is still a PD, and the new R_1 is still positive, but the new R_2 is negative, where before it had been positive. So, the persona equilibrium will now be $(\mathscr{E}, \mathscr{E})$. This establishes the following result:

Corollary 4 *Consider a two-player persona world* (X, U, A) *where each* X_i *is binary,* U *is given by the generalized PD with payoff matrix in equation* (20)*, and each player* i *has the persona set* $A_i = \{U_i, sU_i + (1-s)U_i\}$ *for some* $0 \leq s \leq 1$*. Then, if* $s \in (1 - \frac{\gamma}{\alpha}, \frac{\beta}{\alpha}]$*, and* $c = 0$*, the extended persona equilibrium is* $(\mathscr{C}, \mathscr{C})$ *followed by* $(\boldsymbol{U}, \boldsymbol{L})$*. If instead* c *is changed so that* $cs > \gamma - (1-s)\alpha, \gamma > c$ *and* $\alpha - \beta > c$*, then, in the extended persona equilibrium both players defect.*

Under the conditions of the corollary, for purely self-interested reasons, adding a material incentive that favors cooperation instead causes defection. This is true even though the players had cooperated before. We have an automatic "crowding out" effect (Bowles, 2008).

Return now to the $c = 0$ PD in equation 16, so that if both players defect, each player's utility is γ. For this underlying game, when the extended persona game equilibrium is $(\mathscr{C}, \mathscr{C})$, followed by $(\mathbf{U}, \mathbf{L})$, the benefit to each player of playing the persona game rather than playing the

underlying game directly is B. Comparing this with the formulas for R_1 and R_2 establishes the following:

Corollary 5 *Consider a two-player persona world* (X, U, A) *where each* X_i *is binary,* U *is given by the generalized PD with payoff matrix in equation* (16)*, and each player* i *has the persona set* $A_i = \{U_i, sU_i + (1-s)U_i\}$ *for some* $0 \leq s \leq 1$*. Then,* $R_1 + R_2 + B \leq 1$.

This sobering result says that there are unavoidable tradeoffs between the robustness of cooperation and the potential benefit of cooperation in the PD, whenever (as here) the underlying game matrix is symmetric and both players can either be selfish or charitable for the same value of s. The more a society benefits from cooperation, the more fragile that cooperation.

To understand this intuitively, note that having R_2 be large means that both γ and s are (relatively) large. These conditions guarantee something concerning your opponent: She is not so inclined to cooperate that it benefits you to take advantage of her and be selfish. On the other hand, having R_1 be large guarantees something concerning you: the benefit to you of defecting when your opponent cooperates is small.

There are many ways the foregoing analysis can be extended. For example, an anonymous referee suggested that expanding each player i's persona set to include multiple ρ_i values. Another extension would be to allow the persona sets to vary among the players. For reasons of space, such extensions are deferred to future work.

5 Comparison with experimental data

To illustrate the persona framework, we provide an explanation for some of the experimental data concerning the Traveler's Dilemma (TD) (Basu, 2007; Capra et al., 1999; Basu, 1994; Rubinstein, 2004; Becker et al., 2005; Goeree and Holt, 1999). The TD models a situation where two travelers fly on the same airline with identical antiques in their baggage, and the airline accidentally destroys both antiques. The airline asks them separately how much the antique was worth, allowing them the answers $\{2, 3, \ldots, 101\}$. To try to induce honesty in their claims, the airline tells the travelers that it will compensate both of them with the lower of their two claims, with a bonus of R for the maker of the lower of the two claims, and a penalty of R for the maker of the higher of the two claims.

To formalize the TD, let $\Theta(z)$ be the Heaviside step function,

$$\Theta(z) = \begin{cases} 0 & z < 0 \\ 1/2 & z = 0 \\ 1 & z > 0. \end{cases} \tag{21}$$

Then, for both players i, the utility function in the TD underlying game is $u_i(x_i, x_{-i}) = (x_i + R)\Theta(x_{-i} - x_i) + (x_{-i} - R)\Theta(x_i - x_{-i})$, where, R is the reward/penalty (for making a low/high claim), x_i is the monetary claim made by player i, and x_{-i} is the monetary claim made by the other player.

The NE of this game is $(2, 2)$, since whatever i's opponent claims, it will benefit i to undercut that claim by 1. However, in experiments (not to mention common sense), this NE never arises. Even when game theoreticians play the TD with one another for real stakes, they tend to make claims that are not much lower than 101, and they almost never make claims of 2. When describing these results, Basu (2007) called for a formalization of "the idea of behavior generated by rationally rejecting rational behavior ... to solve the paradoxes that plague game theory."

Consider a persona game based on the $R = 2$ TD underlying game. It seems that real humans are sometimes irrational (that is, purely random, as discussed above) and sometimes fully rational. So we choose these as the possible personas of the players, indicated by $\rho = 0$ and $\rho = \infty$, respectively.[13]

When both players are fully rational, the expected utility to both is the NE value, 2, that is, $\mathbb{E}(u_i \mid \rho_1 = \infty, \rho_2 = \infty) = 2$ for both players i. Now, say that player i is rational while the other player is irrational. This results in the expected utility $\mathbb{E}(u_i \mid x_i, \beta_i = \infty, \beta_{-i} = 0) = \frac{1}{100}\Big(\Big[\sum_{y=2}^{x_i-1}(y-2)\Big] + x_i + \Big[\sum_{y=x_i+1}^{101}(x_i+2)\Big]\Big)$ for all of i's possible underlying game moves x_i. The (integer) maximum of this is at $x_i \in \{97, 98\}$. The associated expected utility is $\mathbb{E}(u_i \mid \beta_i = \infty, \beta_{-i} = 0) \simeq 49.6$.

Since player i is indifferent between these two moves in the behavioral game, and since player $-i$ plays a uniform mixed strategy regardless of i's move, any behavioral game mixed strategy by i over these two moves is a NE of the behavioral game. However, while the expected underlying utility to player i is the same for either behavioral game move $x_i \in \{97, 98\}$, the expected underlying utility to player $-i$ differs for these two moves. Accordingly, the expected underlying utility to player $-i$ will depend on the (arbitrary) mixing probability s with which player i chooses between $x_i \in \{97, 98\}$ in the behavioral game. Similar calculations hold when it is player 1 who is

[13] In the real world, there are many other personas that also arise in human behavior. Some of these are discussed below. The goal here is to analyze how well even such an extreme persona set can explain observed behavior.

irrational and player 2 who is rational. The mixing probability for player $-i$, which is analogous to s, will be written as q.

Combining, we can express the NE of the behavioral games corresponding to all four possible joint personas with the following set of matrices, where the ordering is (player 1 objective, player 2 objective), the top row is for $\beta_1 = 0$, the left column is for $\beta_2 = 0$, and both q and s range over the interval $[0.0, 1.0]$:

$$\begin{bmatrix} \left(\frac{6967}{200}, \frac{6967}{200}\right) & \left(\frac{s2661+(1-s)2665}{50}, \frac{2479}{50}\right) \\ \left(\frac{2479}{50}, \frac{q2661+(1-q)2665}{50}\right) & (2, 2) \end{bmatrix}. \tag{22}$$

We can now define the full persona game utility functions by uniformly averaging over s and q.[14] Those (rounded) values are:

	Player 2 rationality	
	0	$+\infty$
Player 1 rationality		
0	(34.8, 34.8)	(53.3, 49.6)
$+\infty$	(49.6, 53.3)	(2, 2)

(23)

This persona game has two pure strategy NE, $(\rho_1, \rho_2) = (0, \infty)$ and $(\rho_1, \rho_2) = (\infty, 0)$. The associated distribution $P(x_1)$ for the first of these rationality NE is uniform. The associated $P(x_2)$ instead has half its mass on $x_2 = 97$, and half on $x_2 = 98$. The two distributions for the other pure strategy rationality NE are identical, but with $P(x_1)$ and $P(x_2)$ flipped. (As an aside, note that if one of the players is irrational and the other rational, it is better to be the irrational player rather than the rational one.)

There is also a symmetric mixed strategy NE of this persona game. To calculate the behavioral game distribution associated with that mixed rationality NE, write

$$\begin{aligned} P(x_1, x_2) &= \sum_{\rho_1, \rho_2} P(x_1, x_2 \mid \rho_1, \rho_2) P(\rho_1, \rho_2) \\ &= \sum_{\rho_1, \rho_2} P(x_1, x_2 \mid \rho_1, \rho_2) P(\rho_1) P(\rho_2), \end{aligned} \tag{24}$$

[14] Formally speaking, this is an instance of using a "uniform averaging summarizer." See Appendix C.

where each of the four distributions $P(x_1, x_2 \mid \rho_1, \rho_2)$ (one for each (ρ_1, ρ_2) pair) is evaluated by uniformly averaging the multiple behavioral game NE for that (ρ_1, ρ_2) pair.[15] For both players i the associated marginal distribution is given by $P(x_i) = \sum_{\rho_1,\rho_2} P(x_i \mid \rho_1, \rho_2)P(\rho_1)P(\rho_2)$. Plugging in gives that both rationality players choose $\rho = 0$ with probability 0.78. The associated marginal distributions $P(x_i)$ are identical for both i's: $P(x_i = 2) \simeq 5.8$ percent, $P(x_i = 97) = P(x_i = 98) \simeq 9.5$ percent, and $P(x_i) \simeq 0.8$ percent for all other values of x_i. (Note that because $P(\rho_1, \rho_2)$ is not a delta function, $P(x_1, x_2) \neq P(x_1)P(x_2)$.)

At such a mixed strategy NE of the persona game the persona players choose randomly among some of their possible personas. Formally, the possibility of such a mixed NE is why persona games always have equilibria, in contrast to EOP. Empirically, such a NE can be viewed as a model of "capricious" or "moody" behavior by humans.

In general, to compare experimental data with the predictions of game theory when there are multiple equilibria is an informal exercise. This is just as true here, with the multiplicity of persona game equilibria. We simply note that if we uniformly average over the three NE of the persona game, we get a $P(x)$ that is highly biased to large values of x. This agrees with the experimental data recounted above.

We can perform the same persona game analysis for other values of R besides 2. When R grows, the mixed strategy equilibrium of the persona game places more weight on the persona ∞. This makes $P(x)$ become more weighted towards low values. In fact, the persona game will have a mixed NE only for $0 < R < (25 + 5\sqrt{322})/3$; the associated mixed strategy equilibrium is given by:

$$P(\rho_1 = 0) \quad = \quad P(\rho_2 = 0) \quad = \quad \frac{4(3R^2 - 50R - 2475)}{8R^2 - 13233}. \tag{25}$$

For R outside of this range the two NE at $(\infty, 0)$ and $(0, \infty)$ also vanish, and the only NE is the pure strategy of full rationality (∞, ∞). So, for such values of R, the players are fully rational. These results agree with experimental data (Capra et al., 1999) on what happens as R changes.

[15] Formally speaking, this is another instance of using the "uniform averaging summarizer." See Footnote 14 and Appendix C.

6 Persona games with imprecise persona signaling

6.1 Noisy extended persona games

We now show that one can modify extended persona games to model scenarios where in the behavioral game, each player knows his own persona, but has only noisy information concerning the personas of the other players. (This reflects the real-world fact that when observing others, we are never sure *exactly* what "mood" they are in.) Such noisy situations are inherently more complicated than noise-free persona games. So, in the interests of space, here we show only that a formal definition of persona game equilibria in such situations can be given and establish that such equilibria always exist; we leave examples and more detailed analysis to future work.

While it is possible to express such cases by introducing Nature players (who have endogenously fixed distributions over actions) and information sets in the usual way, it is notationally simpler to do it a bit differently. Consider a tuple (X, U, A, Z) where (X, U, A) is an N-player persona world, all spaces are finite, and $Z = \times_{i \in \mathcal{N}} Z_i$. Fix a set of $\sum_{i \in \mathcal{N}} |A_{-i}|$ distributions $\{Q_i^{a_{-i}} \in \Delta_{Z_i} : i \in \mathcal{N}, a_{-i} \in A_{-i}\}$. Intuitively, for every pair $(a_{-i} \in A_{-i}, z_i \in Z_i)$, $Q_i^{a_{-i}}(z_i)$ is the conditional distribution where behavioral game player i observes datum z_i, given that the joint persona of the other players is a_{-i}. We call the tuple (X, U, A, Z) together with the distributions $\{Q_i^{a_{-i}} \in \Delta_{Z_i} : i \in \mathcal{N}, a_{-i} \in A_{-i}\}$ a *noisy (extended) persona game*.

Suppose that in addition to a noisy extended persona game we have an associated set of distributions $\{P(A_i) \in \Delta_{A_i}, q_i^{a_i,z_i}(X_i^{a_i,z_i}) \in \Delta_{X_i} : i \in \mathcal{N}, a_i \in A_i, z_i \in Z_i\}$, where each space $X_i^{a_i,z_i}$ is a copy of X_i. (We will sometimes write a distribution $q_i^{a_i,z_i}(X_i^{a_i,z_i})$ as $q(X_i^{a_i,z_i})$.) Intuitively, each $P(A_i)$ is the mixed strategy of persona player i, and each $q_i^{a_i,z_i}$ is the mixed strategy that behavioral game player i adopts upon observing his own persona a_i, together with the datum z_i, concerning the personas of the other players.

Assume these distributions together with the $\{Q_i^{a_{-i}} : i \in \mathcal{N}\}$ relate the variables with the conditional independencies implied by the superscripts and arguments. So, in particular, $Pr(x^{a_i,z_i})$, the probability that the behavioral game players make joint move x, conditioned on a_i and z_i, the persona and datum of player i (that is, conditioned on the information known to player i), can be

written as

$$\begin{aligned} Pr(x^{a_i,z_i}) &= \int da_{-i}dz_{-i}\ Pr(a_{-i}, z_{-i} \mid a_i, z_i)Pr(x \mid a, z) \\ &= \int da_{-i}dz_{-i}\ Pr(a_{-i}, z_{-i} \mid a_i, z_i) \prod_{j\in\mathcal{N}} q(x_j^{a_j,z_j}). \end{aligned} \tag{26}$$

Similarly the joint probability of a and z can be written as

$$\begin{aligned} Pr(a, z) &= Pr(z \mid a)P(a) \\ &= \prod_{i\in\mathcal{N}} Q_i^{a_{-i}}(z_i) \prod_{k\in\mathcal{N}} P(a_k) \\ &= \prod_{i\in\mathcal{N}} \left[Q_i^{a_{-i}}(z_i)P(a_i)\right]. \end{aligned} \tag{27}$$

Combining these two expansions establishes the following result:

Proposition 6 *Let (X, U, A, Z) be a noisy persona game. Then for all players i, personas a_i, and data z_i,*

i) $Pr(x_{-i}^{a_i,z_i}) = \int da_{-i}dz_{-i}\ Pr(a_{-i}, z_{-i} \mid a_i, z_i) \prod_{j\neq i} q(x_j^{a_j,z_j})$

ii) $Pr(x^{a_i,z_i}) = q(x_i^{a_i,z_i})Pr(x_{-i}^{a_i,z_i})$

Proof. By equation (27),

$$\begin{aligned} Pr(a_{-i}, z_{-i} \mid a_i, z_i) &= \frac{Pr(a, z)}{\int dz'_{-i}da'_{-i}\ Pr(a_i, z_i, a'_{-i}, z'_{-i})} \\ &= \frac{Q_i^{a_{-i}}(z_i) \prod_{k\neq i} \left[Q_k^{a_{-k}}(z_k)P(a_k)\right]}{\int dz'_{-i}da'_{-i}\ Q_i^{a'_{-i}}(z_i) \prod_{n\neq i} \left[Q_n^{a_i,a'_{-n,-i}}(z'_n)P(a'_n)\right]} \\ &= \frac{Q_i^{a_{-i}}(z_i) \prod_{k\neq i} \left[Q_k^{a_{-k}}(z_k)P(a_k)\right]}{\int da'_{-i}\ Q_i^{a'_{-i}}(z_i) \prod_{n\neq i} P(a'_n)}. \end{aligned} \tag{28}$$

So, combining equations (26) and (28) gives

$$Pr(x^{a_i,z_i}) = \int da_{-i}dz_{-i}\ \prod_{j\in\mathcal{N}} q(x_j^{a_j,z_j}) \frac{Q_i^{a_{-i}}(z_i) \prod_{k\neq i} \left[Q_k^{a_{-k}}(z_k)P(a_k)\right]}{\int da'_{-i}\ Q_i^{a'_{-i}}(z_i) \prod_{n\neq i} P(a'_n)}\ .$$

The proof of part (i) of the proposition proceeds analogously. Combining that part (i) with equation (28) gives

$$\begin{aligned} Pr(x_{-i}^{a_i,z_i}) &= \int da_{-i}dz_{-i} \prod_{j\neq i} q(x_j^{a_j,z_j}) \frac{Q_i^{a_{-i}}(z_i) \prod_{k\neq i} \left[Q_k^{a_{-k}}(z_k)P(a_k)\right]}{\int da'_{-i}\ Q_i^{a'_{-i}}(z_i) \prod_{n\neq i} P(a'_n)} \\ &= \int da_{-i}dz_{-i} \frac{Q_i^{a_{-i}}(z_i) \prod_{k\neq i} \left[q(x_k^{a_k,z_k})Q_k^{a_{-k}}(z_k)P(a_k)\right]}{\int da'_{-i}\ Q_i^{a'_{-i}}(z_i) \prod_{n\neq i} P(a'_n)} \\ &= \frac{\int da_{-i}dz_{-i}\ Q_i^{a_{-i}}(z_i) \prod_{k\neq i} \left[q(x_k^{a_k,z_k})Q_k^{a_{-k}}(z_k)P(a_k)\right]}{\int da'_{-i}\ Q_i^{a'_{-i}}(z_i) \prod_{n\neq i} P(a'_n)}, \end{aligned}$$

where in the integrands, the ith component of each a_{-k} is implicitly set to a_i. Comparing equations (26) and part (i) of the proposition establishes part (ii) of the proposition. ■

Proposition 6(ii) addresses a potential concern that the framework presented above may be a nonsensical way to model noisy persona games. Intuitively, it says that in the behavioral game, player i chooses his distribution "as though" it were independent of the distribution of the other behavioral game players, conditioned on what i knows about the moves of those other players. So, there are no peculiar, unavoidable couplings between i's behavioral move and those of the other players, even when we condition on what i knows.

We say we have a *noisy (extended) persona equilibrium* iff the following conditions hold:

$\forall i,\ \nexists\ \hat{P} \in \Delta_{A_i}$:

$$\int dadz\ \hat{P}(a_i)P(a_{-i}) \prod_{k\in\mathcal{N}} Q_k^{a_{-k}}(z_k)\ U_i\left[\prod_{j\in\mathcal{N}} q(X_j^{a_j,z_j})\right] > \int dadz\ P(a_i)P(a_{-i}) \prod_{k\in\mathcal{N}} Q_k^{a_{-k}}(z_k)\ U_i\left[\prod_{j\in\mathcal{N}} q(X_j^{a_j,z_j})\right] \tag{29}$$

and

$\forall i, a_i \in A_i, z_i \in Z_i,\ \nexists\ \hat{q} \in \Delta_{X_i}$:

$$\int da_{-i}dz_{-i}\ Pr(a_{-i}, z_{-i} \mid a_i, z_i)a_i\left[\hat{q}(X_i^{a_i,z_i})q(X_{-i}^{a_{-i},z_{-i}})\right] > \int da_{-i}dz_{-i}\ Pr(a_{-i}, z_{-i} \mid a_i, z_i)a_i\left[q(X_i^{a_i,z_i})q(X_{-i}^{a_{-i},z_{-i}})\right]. \tag{30}$$

where $Pr(a_{-i}, z_{-i} \mid a_i, z_i)$ is given by equation (28).

As an alternative, we could replace equation (30) with the notationally simpler condition that $\nexists\, \hat{q} \in \Delta_{X_i}$:

$$a_i \left[\hat{q}(X_i^{a_i,z_i}) Pr(X_{-i}^{a_i,z_i})\right] \;>\; a_i \left[q(X_i^{a_i,z_i}) Pr(X_{-i}^{a_i,z_i})\right]. \tag{31}$$

where $Pr(X_{-i}^{a_i,z_i})$ is given by equation (29). This alternative would be appropriate if we were modeling a situation where we expected the mixed strategy of a behavioral game player i having persona a_i to be optimized for the "background" joint distribution over the moves of the other behavioral game players.

By Proposition 6(i), when all personas are expected utilities, these two alternatives are identical. Proofs of the existence of noisy extended persona equilibria for such persona sets proceed analogously to the development in Appendix B, and are omitted here in the interests of space.

6.2 Persona games without signaling

In the extreme case of noisy persona signals, each $Q_i^{a_{-i}}(z_i)$ is independent of a_{-i}. In this case, the signals z_i provide no information concerning a_{-i}. If $U_i \in A_i$, it is straightforward to prove that there is no incentive for any persona player i to choose some $a_i \neq U_i$.

However, in many real-world situations people play multiple successive behavioral games based on the same underlying game, and in addition cannot modify their persona distributions as quickly as those behavioral games are played. Their persona distributions are "sticky," persisting across multiple behavioral games. In such situations, behavior in the early behavioral games can play a similar role to overt persona signals, providing information to the other behavioral players in the later stages about what personas have been adopted. In such a situation, it can again be rational for a persona player i to choose a nonrational persona $a_i \neq U_i$.

Empirically, sticky personas seem to be common in actual human behavior (Frith, 2008). Loosely speaking, "stubborn," "traditional," or "conservative" behavior can be viewed as sticky personas. In an extreme version of sticky personas, sometimes a person will refuse to consider any alternative behavior "once he makes up his mind" about how he will behave. Similarly, many people "refuse to change their ways," whether or not new data seem to indicate that their behavior is rational.[16]

[16]There are several possible reasons for such stickiness. First, note that it takes a great deal of computational effort to calculate optimal persona distributions. (Crudely speaking, for every possible joint persona, one has to

We model sticky personas as an extended persona game that involves multiple, "behavioral game" stages, all played with a persona strategy profile that is set in the first stage. The players in each stage of the behavioral game can observe some aspects of the outcome of all earlier stages. (Such observation can be formalized using information sets, in the usual way.) In general, the persona profile could be sampled only once, before the behavioral games start, or multiple times, as the behavioral games unfold.[17] Here, for simplicity, we consider the former situation and assume that the persona distribution is sampled once, before any of the behavioral games.

Note that that sampled persona will be used in all subsequent behavioral games. So, each persona game player i chooses his persona distribution to optimize a goal that combines the values of his underlying game objective function U_i evaluated at the outcomes of all the behavioral games. For example, the goal of persona player i might be to maximize a discounted sum of the values of U_i evaluated at the mixed strategy profiles $P(x)$ adopted in all of the associated behavioral games. Since in our physical model the persona distribution of a human arises through a long-term learning process, it is reasonable to have such discounting be very weak. In particular, if there are only a finite number of behavioral game stages, it is reasonable to have no discounting at all. In this case, the goal of each persona player i is to maximize a uniform sum of the values U_i under the equilibrium mixed strategy profiles of the associated behavioral games.

In contrast to the long timescale process under which persona strategies are chosen, behavioral game strategies are chosen through a reasoning process that occurs on a short timescale. (Recall that persona games are "two-timescale games.") Accordingly, their discounting factor will typically differ from that used to set the persona distribution. In the extreme case, the behavioral game player i can use infinite discounting. This would mean that at each behavioral game stage, each player i chooses his move to optimize his associated persona a_i, evaluated at that stage, without any concern for the effects on the values of the objective functions in later stages.

calculate the associated behavioral game equilibria, and only then can one calculate the persona game equilibria.) Accordingly, any computational limitations on the players might force them to recompute their persona distributions only infrequently compared with the rate at which the behavioral game at each stage changes. In particular, if each game is played anonymously, so that no signaling can arise via vocal inflection, body language, or the like, then "cognitive inertia" might lead the players to stick with some default persona. (In contrast, if the game is not anonymous, one might expect less stickiness arising from such inertia, since new information about the opponent becomes available between stages.) As another example, the players might be genetically "hard-wired" not to change their persona distributions frequently. (Such a genome might arise via natural selection processes similar to those investigated in EOP.)

[17] In more nuanced versions of the analysis, the players can also change their mixed strategy persona between behavioral games, but only by a small amount. Alternatively, they can change their persona by an arbitrary amount, but the more they change it, the more they pay an internal psychological / computational cost. All such analysis is beyond the scope of this paper.

In the interests of space, in the remainder of this section we do not elaborate the broad theory of signal-free persona games. In particular, no theorems will be presented. Instead, to illustrate some of the core concepts, we investigate a simple pedagogical scenario. In this scenario, each persona player i moves once, before any of the behavioral games, and uses zero discounting, while each of the behavioral game players wants to optimize the value of his persona in his game alone (that is, uses infinite discounting), without concern for the games in the subsequent stages.

To see why a persona player j might want to choose a persona $a_j \neq U_j$, say that a behavioral game player $i \neq j$ in a current stage t is able to observe the move chosen by a behavioral game player j in an earlier stage $t' < t$. In general, such an observation may provide information concerning the persona controlling behavioral game player j in stage t', and thereby provide information concerning the persona controlling behavioral game player j in stage t. In turn, having this information may lead behavioral player i in stage t to choose a different move than he otherwise would. As is usual with personas, this new move by player i may be better for persona player j when the persona choice by j is one that is not fully rational.

To illustrate this, we now explore an example involving two stages of a behavioral game. To help ground the discussion, we first give the fully formal definition of the persona equilibrium for this example.

Consider an N-player persona world (X, U, A) where all spaces are finite and where we have a set of $N + \sum_{i \in \mathscr{N}} |A_i| + |X| \sum_{i \in \mathscr{N}} |A_i|$ distributions $\{P(A_i) \in \Delta_{A_i}, q^{a_i}(X_i^{a_i}) \in \Delta_{X_i} : i \in \mathscr{N}, a_i \in A_i, q^{a_i,x}(X_i^{a_i,x}) \in \Delta_{X_i} : i \in \mathscr{N}, a_i \in A_i, x \in X\}$, where for each i, all spaces of the form $X_i^{a_i}$ or $X_i^{a_i,x}$ are copies of X_i. Intuitively, we can view each distribution $P(A_i)$ as the mixed strategy of the ith persona player. Each distribution $q^{a_i}(X_i^{a_i})$ is the mixed strategy adopted by the player in the first behavioral game stage who corresponds to persona player i, when persona player i adopts persona a_i. Similarly, each distribution $q^{a_i,x}(X_i^{a_i,x})$ is the mixed strategy of the player in the second behavioral game stage who corresponds to persona player i when persona player i adopts persona a_i and the players in the first behavioral game stage choose joint move x.

From now on, to simplify the presentation, we will usually write $q^{a_i}(X_i)$ rather than $q^{a_i}(X_i^{a_i})$ and $q^{a_i,x}(X_i)$ rather than $q^{a_i,x}(X_i^{a_i,x})$. Similarly, we will often write $q^a(X)$ rather than $\prod_{j \in \mathscr{N}} q(X_j^{a_j})$, $q^{a_{-i}}(X_{-i})$ rather than $\prod_{j \neq i} q^{a_j}(X_j^{a_j})$, and $q^{a,x}(X)$ rather than $\prod_{j \in \mathscr{N}} q^{a_j,x_j}(X_j^{a_j,x_j})$.

Say that the following three conditions are met simultaneously:

$\forall i,\ \nexists\, \hat{P} \in \Delta_{A_i}$:

$$\int da\ \hat{P}(a_i)P(a_{-i})\left(U_i[q^a(X)] + \int dx\ q^a(x)U_i[q^{a,x}(X)]\right) \ > \\ \int da\ P(a_i)P(a_{-i})\left(U_i[q^a(X)] + \int dx\ q^a(x)U_i[q^{a,x}(X)]\right), \tag{32}$$

$\forall i, a_i \in A_i,\ \nexists\, \hat{q} \in \Delta_{X_i}$:

$$\int da_{-i}\ P(a_{-i})a_i\left[\hat{q}(X_i)\prod_{j\neq i} q^{a_j}(X_j)\right] \ > \ \int da_{-i}\ P(a_{-i})a_i\left[q^{a_i}(X_i)\prod_{j\neq i} q^{a_j}(X_j)\right], \tag{33}$$

and

$\forall i, a_i \in A_i, x \in X,\ \nexists\, \hat{q} \in \Delta_{X_i}$:

$$\int da_{-i}\ P(a_{-i})q^{a_{-i}}(x_{-i})a_i\left[\hat{q}(X_i)\prod_{j\neq i} q^{a_j,x}(X_j)]\right] \ > \ \int da_{-i}\ P(a_{-i})q^{a_{-i}}(x_{-i})a_i\left[q^{a_i,x}(X_i)\prod_{j\neq i} q^{a_j,x}(X_j)]\right]. \tag{34}$$

These equations specify the equilibrium of an extended persona game involving two successive behavioral games. In both behavioral games the players observe only their own personas, not those of their opponents. In addition, the players in the second behavioral game observe the outcome of the first behavioral game. Finally, the goal of each persona player is to maximize the sum of the values of his objective function evaluated for both behavioral game outcomes. Note that nothing about the persona distribution is directly communicated to the behavioral game players, that is, there is no direct signaling of personas. (As usual, it is natural to consider variants where some of the integrals in equations (32) – (34) are moved inside the objective functions, but we will not do so here.)

Note that the product $P(a_{-i})q^{a_{-i}}(x_{-i})$ occurring in equation (34) is the joint probability that $A_{-i} = a_{-i}$ and $X_{-i} = x_{-i}$. Combining this with Bayes' theorem and the fact that x is fixed (and therefore x_{-i} is also), we see that $P(a_{-i})q^{a_{-i}}(x_{-i})$ is proportional to the conditional probability that $A_{-i} = a_{-i}$, given that $X_{-i} = x_{-i}$. This is how the value of x_{-i} in the first behavioral game provides information concerning a_{-i} to player i in the second behavioral game.

To illustrate equations (32) – (34), consider the following underlying game between a Row player and a Col(umn) player:

$$\begin{bmatrix} (0,0) & (6,1) \\ (5,5) & (0,6) \end{bmatrix} \tag{35}$$

. Let Col's persona set consist of the fully rational and anti-rational personas, indicated by the values $\beta_{Col} \in \{\infty, -\infty\}$, while Row's persona set comprises the singleton of the fully rational persona.

It is straightforward to confirm that the following distributions form an equilibrium of the resultant two-stage, no-signaling persona game, that is, satisfy equations (32) – (34):

1. $P(a_{Col}) \triangleq \delta(\beta_{Col}, -\infty)$.
2. $q^{\infty}(x_{Col}) \triangleq \delta(x_{Col}, \mathbf{R})$.
3. $q^{-\infty}(x_{Col}) \triangleq \delta(x_{Col}, \mathbf{L})$.
4. $q(x_{Row}) \triangleq \delta(x_{Row}, \mathbf{D})$.
5. $q^{\infty,x}(x'_{Col}) \triangleq \delta(x'_{Col}, \mathbf{R})$.
6. $q^{-\infty,x}(x'_{Col}) \triangleq \delta(x'_{Col}, \mathbf{L})$.
7. $q^{x}(x'_{Row}) \triangleq \delta(x'_{Row}, \mathbf{D})$ if $x_{Col} = \mathbf{L}$, $\delta(x'_{Row}, \mathbf{U})$ otherwise.

Under these distributions, with probability 1.0, $a_{Col} = -\infty$ and the two behavioral game joint moves are $x^1 = x^2 = (\mathbf{D}, \mathbf{L})$. The resultant value of the Col persona player's goal, given by the integral on the right-hand side of equation (32), is $5 + 5 = 10$.

Note that if that persona player changed her mixed strategy, so that instead of condition (i) she set $P(a_{Col}) \triangleq \delta(\beta_{Col}, \infty)$, then, with probability 1.0, $a_{Col} = \infty$, $x^1 = (\mathbf{D}, \mathbf{R})$, and $x^2 = (\mathbf{U}, \mathbf{R})$. The resultant value of the Col persona player's goal is $6 + 1 = 7$. This confirms that the Col persona player will elect to be anti-rational rather than rational.

There are also other equilibria of this persona game, that is, other solutions to equations (32) – (34). In particular, there are solutions under which both behavioral game Row players always make move **U**, and thereby force the Col persona player to choose to be rational. However, the choice of the second of those Row players to move **U** even if Col moves **L** is a noncredible threat.

(If the first Col behavioral game player actually did move **L**, then Col must be anti-rational, and therefore the second Col behavioral game player would also make move **L**, to which the best response by the second behavioral game Row player would be **D**, not **U**.)

Since the persona objective functions and the personas in the persona sets are all expected utilities, we can use any of the usual refinements to winnow the set of equilibria.[18] In particular, the distributions of (i)–(vii), under which Col chooses to be anti-rational, form a trembling-hand perfect equilibrium.

As a final comment, note that the sticky persona model explaining (apparently) nonrational play within a particular stage might seem to be related to the folk theorems. There are several important distinctions, however. For example, with sticky personas, there are no infinite sequences. Nor are trigger strategies involved in any sense.

7 General discussion of persona games

7.1 Computational issues

Recall that calculating a persona equilibrium typically involves far more computational work than calculating the equilibria of the associated underlying game. This has many implications for how persona games arise in the real world.

One very broad implication of this fact is that persona games should only arise in a species with advanced cognitive capabilities, whose members have many interactions with other organisms that can also play persona games. Colloquially speaking, we might characterize such a species whose members play persona games well as having "high social intelligence." (Candidate species would be higher primates, corvids, and cetaceans.)

Also for computational reasons, one would expect the persona set of any social animal for any underlying game not to be very large. This is because a large set increases both the computational burden on the player with that set, and the burden on the other players he plays against. This raises the intriguing issue, beyond the scope of this paper, of how persona sets might evolve under natural selection.

[18] Formally, to do this one should recast equations(32)– (34) as a conventional three-stage, extensive-form game, involving six players — two persona players and four behavioral game players — each of whom moves exactly once. The first two players jointly determine a "persona" bit that is observed by the remaining four "behavioral game" players. That bit is also an argument of the utility functions of two of those remaining four players (the ones that correspond to Col).

Now, consider a fixed persona game (X, U, A), involving signaling, that occurs often in a person i's daily life. Assume that there is little variability in the persona equilibrium in the instances of that persona game. As a result, there would be little variability in the associated behavioral game equilibria, and in particular in i's behavioral game mixed strategy. Now, suppose that i participates in an experiment in which the subjects repeatedly engage in a version of this very game, where the players are prevented from signaling personas to one another. Formally, this means that they actually play the underlying game (X, U), while in their previous experience with the game they were repeatedly playing (X, U, A).

In such a scenario, due to "cognitive inertia" the players might at first adopt the mixed strategies that are behavioral game equilibria of (X, U, A), even though they are actually playing (X, U). However, over time that inertia may fade, and the players may realize that they are not actually involved in the kind of persona game to which they are accustomed. This might lead each human in the experiment to eventually play the underlying game (X, U) directly, with no concern for personas and the like. This would mean that each player i has been led to assume that the other players will be fully rational in the behavioral game and also to assume that they expect i to be fully rational. Such an assumption would lead each player i to be fully rational. (This in turn would validate the other players' assumption concerning her rationality, resulting in a sort of common knowledge where there is an equilibrium of assumptions.) Such a gradual increase in rationality is seen in many experiments, for example those involving the PD (Cooper et al., 1996; Dawes and Thaler, 1988).

7.2 Personality games

As a final example, computational issues might prevent a social animal from calculating the optimal persona from some associated persona set, even a limited persona set, for every underlying game he encounters. (Just think about how many games you play during a typical day, and imagine calculating the precisely optimal persona for every such game.) Rather he might use a simple rule to map any pair {an underlying game, a specification of which player he is in that game} to a persona for that game. As an example, a value for the altruism N-vector ρ can be used to map every N-player underlying game a person might play to a persona to adopt for that game.

We can make this more precise by defining "personality games." Formally, these are similar to extended persona games with perfect signaling. The major difference is that in personality games there are multiple underlying games rather than just one. This makes the definition intrinsically

more complicated.

Let G be an indexed set of N-player persona games, with index g. Abusing notation, indicate each persona game in G with its index: $g = \{X^g, U^g, A^g\}$. For each player i, let S_i be an associated set of *personalities*, each of which is a function $s_i : g \in G \rightarrow a_i^g \in A_i^g$. Define $s = (s_1, s_2, \ldots, s_N)$ as a personality profile, and define $s(g)$ as the persona profile adopted by the N players when the persona game is g. Let Γ be a distribution over G.

Say that for each player i we have a distribution P_i over S_i, together with a set of distributions $\{q(X_i^{g,a^g})\}$, one distribution for each possible g and associated profile of the personas of the N players, a^g. Our distributions form a *personality equilibrium* iff $\forall i$, there does not exist a $P_i' \in \Delta(S_i)$ such that

$$\begin{aligned} &\int dg ds_i ds_{-i}\, \Gamma(g) P'(s_i) P(s_{-i}) U_i \left[\prod_{j \in N} q(X_j^{g,s(g)}) \right] \\ &> \int dg ds_i ds_{-i}\, \Gamma(g) P(s_i) P(s_{-i}) U_i \left[\prod_{j \in N} q(X_j^{g,s(g)}) \right] \end{aligned} \tag{36}$$

and $\forall i, g \in G, a^g \in A^g$, there does not exist a q' such that

$$a_i^g \left[q'(X_i) \prod_{j \neq i} q(X_j^{g,a^g}) \right] > a_i^g \left[q(X_i^a) \prod_{j \neq i} q(X_j^{g,a^g}) \right]. \tag{37}$$

Example 1 *As a simple example, let $N = 2$, and let g be a set of 2×2 move games. So, X^g can be taken to be $\mathbb{B} \times \mathbb{B}$ for all games g. Also take U_i^g for all g, i to be the expectation of a utility function u_i^g. Let A_i^g, the persona set for player i in game g, be $\{\alpha_i u_i^g + (1 - \alpha_i) u_{-i}^g : \alpha_i \in \{1, \rho_i\}\}$ for some constant ρ_i that is the same for all games g. So, in each game g, player i can either be selfish or charitable, with the same "charitability" parameter ρ_i in every game. Note that in this example, since X^g is independent of g, and so is A^g (in the sense that the set of possible values of α_i does not vary with g), the only thing that varies between instances of g is the set of underlying game utility functions u^g.*

Finally, assume that both players i have only two possible personalities, one for each possible value of α_i. So, for both players i, there are only two maps in S_i. One of these personalities maps all games to the persona with $\alpha_i = 1$ (so that if player i adopts this personality he always chooses the selfish persona). In the other personality, all games are mapped to the persona with $\alpha_i = \rho_i$ (so that if player i adopts this personality he always chooses the charitable persona).

At the personality equilibrium, each player i will adopt a mixed strategy $P(s_i)$ over whether he adopts his selfish persona or his charitable persona, a distribution that is independent of the underlying game u^g. This equilibrium mixed strategy has the property that player i cannot benefit by changing $P(s_i)$, given the associated equilibrium mixed strategy $P(s_{-i})$ of $-i$ over $-i$'s personas, the associated behavioral game behaviors, and the distribution Γ.

Intuitively, if the distribution of underlying games Γ *that you encounter in your day-to-day life means you should adopt the charitable personality with probability* 1*, then always being charitable is your personality equilibrium. One might speculate that the process by which a human child "becomes socialized" is the process of his settling on a distribution over personalities to use in his life.*

It may be possible to use personalities to make quantitative predictions about many aspects of human societies. In particular, as one goes from one society to another, the distribution Γ over underlying games that the members of society encounter in their daily lives varies. Accordingly, the personality equilibria may vary, even if the personality sets are the same in all societies. This provides a way to predict how human behavior (that is, how adopted personas vary) varies from one society to another. It may be possible to compare such predictions to anthropological data, for example, to an extended version of the data in Henrich et al. (2001) and Henrich et al. (2006).

7.3 Other applications of the persona framework

The persona game framework was motivated above as a way to model scenarios in which players adopt a persona that they then signal (either directly or by their behavior) to one another. The same mathematical framework can also be used to model several other scenarios. This section discusses some of them.

First, the persona framework can be used to model a variant of games of contingent commitments (Kalai et al., 2008; Jackson and Wilkie, 2005; Myerson, 1991). Games of contingent commitments consist of two steps, just as in the persona framework. In the simplest version of these games, in the first step every player *i* chooses a contingent commitment from an associated set of possible commitments. Every one of these possible commitments is of the form "if each of the other players sends the following signals to me, then I commit to play the following strategy." After all players have chosen their commitments, the players honestly signal those commitments to one another. Then, in the second stage, the players implement the joint strategy determined by the joint commitment signaled by the players.

The persona framework can be viewed as a modification of such a contingent commitments game. The behavioral game NE of the persona framework plays the same role as the implemented second-stage strategy profile in a signaled commitments game. The difference between the frameworks is that the persona framework replaces the signals of binding if-then statements with signals of binding personas.

Personas are typically far less mathematically cumbersome than if-then binding commitments (and far less computationally demanding on the players), especially when there are many players. In addition, personas are far more flexible: the possible persona sets of player i in a persona game need not change if the persona sets of the other persona players change, whereas the set of possible binding commitments of a player in a signaled commitments game must always "match" the corresponding sets of the other players.

Games of binding commitments may also involve a first stage in which the players commit not to play some of their possible pure strategies, and honestly signal those commitments to one another (Renou, 2008; Jackson and Wilkie, 2005). These games can be seen as a special case of persona games in which the personas a_i in the set of player i consist of "masked" versions of u_i, where the forsworn pure strategies are given negative infinite utility: $a_i(x) = u_i(x)$ if $x_i \notin X_i' \subset X_i$, and $a_i(x) = -\infty$ otherwise.

More broadly, the persona game framework can be used to model "idealized" principal-agent scenarios, in which each principal i has the power to arbitrarily set the utility function a_i of his agent to any utility in a set A_i, and the agents of the principals then play a NE of the game specified by the agent functions given to them by their associated principals. Mathematically, the principals are persona players, the sets A_i are persona sets, and the agents are behavioral game players.

Finally, here we have presented the persona framework as a way to analyze nonrationality in a single game. However the same mathematics can also be used to model asymptotic behavior in a sequence of repeated games. The advantage of such a model is that it abstracts from many of the complicating details underlying such sequences in the real world and therefore greatly simplifies the mathematical analysis. The legitimacy of such modeling would be established by comparing the results of experiments involving repeated games with the predictions of the persona framework, to see whether those predictions are accurate predictions of the outcomes of repeated games. (This is the subject of future work.)

8 Final comments

Both humans and some animals sometimes exhibit what appears to be nonrational behavior when they play noncooperative games with others (Camerer, 2003; Camerer and Fehr, 2006; Kahneman, 2003). One response to this fact is to simply state that humans are nonrational, and leave it at that. Under this response, essentially the best we can do is catalog the various types of nonrationality that arise in experiments (loss aversion, framing effects, the endowment effect, sunken cost fallacy, confirmation bias, reflection points, other-regarding preferences, uncertainty aversion, and so on) Inherent in this response is the idea that "science stops at the neck," that somehow logic suffices to explain the functioning of the pancreas but not of the brain.

There has been a lot of work that implicitly disputes this and tries to explain apparent nonrationality of humans as actually being rational, if we appropriately reformulate the strategic problem faced by the humans. The implicit notion in this work is that the apparent nonrationality of humans in experiments does not reflect "inadequacies" of the human subjects. Rather it reflects inadequacies in scientists, in our hubristic presumption to know precisely what strategic scenario the human subjects are considering when they act. From this point of view, our work as scientists should be to try to determine just what strategic scenario really faces the human subjects, as opposed to the one that apparently faces them.

One body of work that adopts this point of view is evolutionary game theory. The idea in evolutionary game theory is that humans (or other animals) really choose their actions in any single instance of a game to optimize results over an infinite set of repetitions of that game, not to optimize it in the single instance at hand. The persona framework is based on the same point of view–that the apparent game and the real game differ. In the persona game framework, the apparent game is the underlying game, but the real game the humans play is the persona game.

There are many interesting subtleties concerning when and how persona games arise in the real world. For example, a necessary condition for a real-world player to adopt a persona other than perfect rationality is that he believes that the other players are aware that they can do that. The simple computer programs for maximizing utility that are currently used in game theory experiments do not have such awareness. Accordingly, if a human knows he is playing against such a program, he should always play perfectly rationally, in contrast to his behavior when playing against humans. This distinction between behavior when playing computers and playing humans agrees with much experimental data, for example, concerning the Ultimatum Game (Camerer and Fehr, 2006; Camerer, 2003; Nowak et al., 2000).

What happens if the players in a persona game are unfamiliar with the meaning of one another's signals, say, because they come from different cultures? This might lead them to misconstrue the personas (or more generally persona sets) adopted by one another. Intuitively, one would expect that the players would feel frustrated when this happens, since in the behavioral game each does what would be optimal if his opponents were using that misconstrued persona — but his opponents are not doing that. This frustration can be viewed as a rough model of what is colloquially called a "culture gap" (Chuah et al., 2007).

Persona games provide a very simple justification for irrationality with very broad potential applicability. They also make quantitative predictions that can often be compared with experimental data. (In work currently being written for submission, two of us have found that the predictions of the persona game framework also agree with experimental data for the Ultimatum Game.) While here we have considered only personas involving degrees of rationality and degrees of altruism, there is no reason not to expect other kinds of persona sets in the real world. Risk aversion, uncertainty aversion, reflection points, framing effects, and all the other "irrational" aspects of human behavior can often be formulated as personas.

Even so, persona games should not be viewed as a candidate explanation of all nonrational behavior. Rather they complement other explanations, for example, those involving sequences of games (like EOP). Indeed, many phenomena probably involve sequences of persona games (or more generally, personality games). As an illustration, say an individual i repeatedly plays a face-to-face persona game γ involving signaling, persona sets, and so on, and adopts persona distribution $P(a_i)$ for these games. By playing all these games, i would grow accustomed to adopting $P(a_i)$. Accordingly, if i plays new instances of γ where signaling is prevented, he might at first continue to adopt distribution $P(a_i)$. However, as he keeps playing signal-free versions of γ, he might realize that $P(a_i)$ makes no sense. This would lead him to adopt the fully rational persona instead. If, after doing that he was to play a version of γ where signaling was no longer prevented, he could be expected to return to $P(a_i)$ fairly quickly. This behavior agrees with experimental data (Cooper et al., 1996; Dawes and Thaler, 1988).

APPENDIX A: PRISONER'S DILEMMA

Consider the general Prisoner's Dilemma (PD) underlying game, parameterized as

$$\begin{bmatrix} (\beta,\beta) & (0,\alpha) \\ (\alpha,0) & (\gamma,\gamma) \end{bmatrix} \tag{38}$$

with $\alpha > \beta > \gamma > 0$. Thus, each player's first strategy is cooperation and second strategy is defection. We will explore what outcomes are possible in the corresponding persona game, where we consider persona sets that include charitable personas in addition to rational, irrational, and/or anti-rational ones. For simplicity in the analysis, if there are multiple Nash equilibria of the behavioral game, we presume that each player is individually "optimistic" and considers only the NE outcome that is best for him. Furthermore, we restrict attention (whenever possible) to NE of the behavioral game that involve pure strategies for both players.

First, it is clear that in this game no player would choose an irrational persona (in the formal sense of committing to play both actions with equal probability), assuming the rational persona is always available to both players—as we do throughout. This is because his opponent's optimal response would be to choose the rational persona herself (leading to defection on her part in the behavioral game), since defection is dominant and hence a best response to any *fixed* behavioral-game strategy. But in this case he would prefer to also be rational, yielding γ rather than $\gamma/2$ as his underlying payoff. For exactly analogous reasons, no player would ever choose to be anti-rational (which in the PD is a commitment to cooperate no matter what) and get taken advantage of with a payoff of 0, instead of also choosing to be rational and securing γ.

Thus, from here on we consider only weakly charitable personas, with various parameters ρ_i representing the relative weight on one's own payoff. In general, we study binary persona sets with one element being the rational persona $\mathscr{E}$ ($\rho_i = 1$) and one element being a fixed charitable persona $\mathscr{C}$ ($\rho_i = s_i$), although this too can be relaxed. For now, we take the charitable personas to be symmetric: $s_1 = s_2 = s$, for $s \in [0, 1)$. Given this, and the underlying payoff matrix above, we can describe the behavioral game—if both players choose $\mathscr{C}$—as follows:

$$\begin{bmatrix} (\beta,\beta) & ((1-s)\alpha, s\alpha) \\ (s\alpha,(1-s)\alpha) & (\gamma,\gamma) \end{bmatrix} \tag{39}$$

For mutual cooperation to be a NE here, obviously we need that $R_1 \equiv \beta - s\alpha \geqq 0$. Meanwhile,

if Row chooses $\mathscr{E}$ while Col chooses $\mathscr{C}$, we end up in the following behavioral game:

$$\begin{bmatrix} (\beta,\beta) & (0, s\alpha) \\ (\alpha,(1-s)\alpha) & (\gamma,\gamma) \end{bmatrix}. \tag{40}$$

In order for Row not to prefer to deviate in this way (and then play his dominant strategy of defection), it must be that Col would choose to defect under those circumstances as well (otherwise Row would expect $\alpha > \beta$). That is, we require that $R_2 \equiv \gamma - (1-s)\alpha > 0$. This is a strict inequality because otherwise there would be an equilibrium of the behavioral game in which Col cooperated while Row defected, which would imply (since players are assumed to be optimistic) that Row would strictly prefer to choose $\mathscr{E}$ in the persona stage.

Summing these two inequalities, we see that $R_1 + R_2 = \beta + \gamma - \alpha > 0$, or $\gamma > \alpha - \beta$. This is a necessary and sufficient condition on the parameters of the PD for it to be the case that $(\mathscr{C},\mathscr{C})$ followed by mutual cooperation is an equilibrium of the overall persona game for some value of s. In particular, if the condition holds, then any $s \in (1 - \frac{\gamma}{\alpha}, \frac{\beta}{\alpha}]$ will induce such an outcome; the same condition precisely implies that this interval will be nonempty. Each of these conditions is interpreted more thoroughly in the body of the text.

For instance, we can see immediately at this point that the saintly persona $\mathscr{A}$ ($s = 0$) is never a possibility for producing cooperation in the PD, which makes perfect sense in light of the reasoning above regarding anti-rational personas: it would essentially commit the player to cooperating in the behavioral game, which means he will be taken advantage of—and that will never happen in equilibrium. However, for any fixed $s \in (0,1)$ and any α, we can find parameters β and γ for which cooperation is possible as a result of the persona game with the corresponding charitable personas available. To do so, simply pick $\beta \in (\max(s\alpha,(1-s)\alpha),\alpha)$ and then pick $\gamma \in ((1-s)\alpha,\beta)$.

Finally, we see that all of this analysis is basically the same for asymmetric charity preferences s_1 and s_2, again considered as part of a binary persona set along with $\mathscr{E}$. If each player chooses $\mathscr{C}$, the resulting game is

$$\begin{bmatrix} (\beta,\beta) & ((1-s_1)\alpha, s_2\alpha) \\ (s_1\alpha,(1-s_2)\alpha) & (\gamma,\gamma) \end{bmatrix}. \tag{41}$$

Analogously, we need $\beta \geq s_i\alpha$ and $\gamma > (1-s_i)\alpha$ for $i = 1,2$. If and only if $\gamma > \alpha - \beta$, there will exist some s_1 and s_2 inducing the possibility of cooperation. Likewise, given any $s_1, s_2 \in (0,1)$, we can choose β and then γ as in the previous paragraph (forcing the inequalities to hold for both

s_i). Hence, there is always a nonempty feasible parameter set.

APPENDIX B: EXISTENCE OF EXTENDED PERSONA EQUILIBRIA

We now prove that any persona world (X, U, A) has an extended persona equilibrium if every U_i is an expected utility and every a_i is either an expected utility or a free utility. To do this, it is necessary to re-express the $N + N|A|$ inequalities in equations ((8), (9)) as NE conditions for a (single-stage) objective game involving $N + N|A|$ players. We will see that not only must such an objective game have an NE, it must have a type I perfect NE (Wolpert, 2009).[19]

To begin, let U_i be the expectation of a utility function $u_i : X \to \mathbb{R}$. Then the left-hand side in equation (8) can be written as an integral over $N + N|A|$ variables,

$$\int da' \int \prod_{j \in \mathcal{N}} (\prod_{a'' \in A} dx_j^{a''}) \; u_i(x_1^{a'}, x_2^{a'}, \ldots) \hat{P}(a'_i) \prod_{j \neq i} P(a'_j) \prod_{j \in \mathcal{N}} (\prod_{a'' \in A} q(x_j^{a''})) \tag{42}$$

and similarly for the right-hand side. Note that many distributions in the two integrands marginalize out. For example, for each a' inside the outer integral, $\int dx_j^{a''} q(x_j^{a''}) = 1$ for every $a'' \neq a'$.

Now define a (single-stage) objective game Γ involving $N(1+|A|)$ players, the first N of whom we will call "persona" players, indexed by i, with the remaining $N|A|$ players being "behavioral players," indexed by pairs (i, a). The move space of each persona player i is A_i, and the move space of any player (i, a) is X_i. To simplify the notation, write the set of possible joint moves of the persona players as $A = \{A^1, A^2, \ldots A^{|A|}\}$. (So, each A^k is a joint persona $a \in A$, that is, it is an ordered list, of N separate persona choices, by the N persona players.)

Given this notation, define a utility function over the joint move of all $N(1 + |A|)$ players:

$$t_i(a', x_1^{A^1}, x_1^{A^2}, \ldots x_1^{A^{|A|}}, x_2^{A^1}, x_2^{A^2}, \ldots) \;\; \triangleq \;\; u_i(x_1^{a'}, x_2^{a'}, \ldots). \tag{43}$$

This allows us to rewrite the integral in equation (42) as

$$\int da' \int \prod_{j \in \mathcal{N}} (\prod_{m=1}^{|A|} dx_j^{A^m}) \; t_i(a', x_1^{A^1}, \ldots, x_2^{A^1}, \ldots) \hat{P}(a'_i) \prod_{j \neq i} P(a'_j) \prod_{j \in \mathcal{N}} (\prod_{m=1}^{|A|} q(x_j^{A^m})) \tag{44}$$

This integral is the expected value of the function t_i over the joint moves $(a', x_1^{A^1}, \ldots, x_2^{A^1}, \ldots)$,

[19] Typically, when some objectives are expected utilities and some are free utilities, we have to be careful in how we define "trembling hand perfection," lest some games have trembling hand perfect equilibrium. Class I perfection is such a definition.

evaluated under a product distribution over those moves, $\hat{P}(a'_i)P(a'_{-i}) \prod_{j \in \mathcal{N}} (\prod_{a'' \in A} q(x_j^{a''}))$. Accordingly, we can take this integral to be an expected utility objective for persona player i in objective game Γ.

Similarly, if a_i is a free utility with logit exponent $\beta_i^{a_i}$ and utility function u_i^a, we can recast the left-hand side in equation (9) as

$$\left(\int da' \; P(a') \int \prod_{j \in \mathcal{N}} (\prod_{a'' \in A} dx_j^{a''}) \; u_i^a(x_i^a, x_{-i}^a) \hat{q}(x_i^a) \prod_{a'' \neq a} q(x_i^{a''}) \prod_{j \neq i} (\prod_{a'' \in A} q(x_j^{a''})) \right) \; + \; (\beta_{a_i})^{-1} \mathscr{S} \left[\hat{q}(X_i^a) \right] \tag{45}$$

and similarly for the right-hand side. Just as before, we have an integral over $N + N|A|$ variables. Also similarly to before, $\int dx + j^{a''} q(x_j^{a''}) = 1$ for all a', a'' such that $a'' \neq a$.

Now, define a utility function over the joint move of all $N(1 + |A|)$ players:

$$v_i^a(a', x_1^{A^1}, x_1^{A^2}, \ldots x_1^{A^{|A|}}, x_2^{A^1}, x_2^{A^2}, \ldots) \;\; \triangleq \;\; u_i^a(x_1^{a'}, x_2^{a'}, \ldots). \tag{46}$$

Having done this, the integral in equation (45) becomes the expected value of the function v_i^a over the joint moves $(a', x_1^{A^1}, \ldots, x_2^{A^1}, \ldots)$, evaluated under a product distribution over those moves, $\hat{P}(a'_i)P(a'_{-i}) \prod_{j \in \mathcal{N}} (\prod_{a'' \in A} q(x_j^{a''}))$. Accordingly, we can take this integral to be an expected utility objective for behavioral player (i, a) in objective game Γ. This means we can take the entire expression in equation (45) to be a free utility for behavioral player (i, a) (who sets $q(X_i^a)$ in objective game Γ).

Similar conclusions hold if some of the a_i are expected utilities. Combining and using Corollary 1 and Proposition 1 of Wolpert (2009), we see that the objective game Γ has a type I NE, as claimed.

APPENDIX C: SUMMARIZER PERSONA GAMES

In addition to extended persona equilibria, there are other, simpler ways to formalize persona games. These can prove useful for calculating predictions concerning common experimental scenarios. We introduce one such alternative formalization of persona games in this appendix. It is used in the main text in our analysis of the TD.

Given a persona world and a player $i \in \mathscr{N}$, an associated *summarizer* w_i is a real-valued quantification of i's desire for the objective game specified by any possible joint persona $a \in A$. In this paper, we restrict attention to summarizers w_i that can be expressed as $w_i(a) \triangleq \bar{w}_i[Q(a), U_i]$, where Q is a mapping taking any $a \in A$ to a subset of Δ_X. (As an example, $Q(a)$ could be a set of equilibria of (X, a) for some pre-fixed equilibrium concept.)

Intuitively, such a summarizer says that how much i desires joint persona a is a function of i's objective function and the joint mixed strategies specified by $Q(a)$. An example of such a w_i is the uniform average of the values of U_i evaluated at the NE of the objective game (X, a). We refer to such a summarizer as the *equilibrium-averaging* summarizer. A variation of this summarizer, motivated by the entropic prior and its use in predictive game theory (Wolpert, 2007), is to weight the value of U_i at each NE q in the average by $e^{\alpha \mathscr{S}(q)}/Z$, where α is the entropic prior constant and Z is a normalization constant.

More generally, $\bar{w}_i[Q(a), U_i]$ could reflect other concerns besides the $q \in Q(a)$ and associated values $\{U_i(q) : q \in Q(a)\}$. For example, it could reflect the computational cost to i of calculating $Q(a)$ and then evaluating $U_i(q)$ for all $q \in Q(a)$. As shorthand, we often abbreviate $(\{X_i : i \in \mathscr{N}\}, \{U_i : i \in \mathscr{N}\}, \{A_i : i \in \mathscr{N}\}, \{w_i : i \in \mathscr{N}\})$ as (X, U, A, w). We write $\bar{w}$ to mean the set of associated functions, $\{\bar{w}_i : i \in \mathscr{N}\}$, and refer to w as the "joint summarizer."

From now on, we implicitly assume that for any persona world that we will consider, (X, U, A), for every $a \in A$, $\mathcal{E}(X, a)$ is nonempty. (As an example, this is true when every persona objective function a_i is either an expected utility or a free utility.) Accordingly, given a persona world (X, U, A) and joint summarizer w, every joint persona a specifies a real number for each player i according to the rule

$$W_i(a) \quad = \quad \bar{w}_i[\mathcal{E}(X, a), U_i]. \tag{47}$$

As an example, say that each $a_i \in A_i$ is the free utility for some associated parameter β_{a_i} and some fixed utility u_i. (So A_i is parameterized by a set of nonzero real numbers $\{\beta_{a_i} : a_i \in A_i\}$.)

Then, the joint persona of the players can be viewed as an N-tuple of logit exponents, in the sense that the equilibrium for the behavioral game associated with a joint persona a is a QRE with logit exponents given by $\beta_a \equiv (\beta_{a_i} : i \in \mathscr{N})$. Say we also have each each $U_i(q)$ be the expected value under q of a utility function $u_i : X \to \mathbb{R}$. Also have $\bar{w}_i$ be the equilibrium-averaging summarizer. Then $W_i(a)$ is the uniform average of the expected utilities of player i over the QRE of the game for the vector of logit exponents β_a.

The *summarizer persona game* for persona world (X, U, A) and joint summarizer w is the N-player (noncooperative) utility game where each player i's space of pure strategies is A_i and his utility function is W_i. Since we are assuming that for every $a \in A$ the associated objective game (X, a) has a NE, the summarizer persona game is a well-defined strategic-form noncooperative game, and therefore always has a NE. Such a NE is a product distribution over joint personas, $P^A(a) \triangleq \prod_{i=1}^N P_i^A(a_i)$. Summarizer persona games based on any particular refinement of the NE concept for the behavioral game are defined in the obvious way.

Note that if each w_i is an equilibrium-averaging summarizer, then at any equilibrium $P^A \in \Delta_{\mathcal{A}}$ there are two averages defining the associated expected value of each W_i: the average over a according to $P^A(a)$, and then for each a, the average over all NE $q(x)$ for that a. In addition, if U_i is an expectation of a utility function $u_i(x)$, then for each NE q there is yet another average, of the values of $u_i(x)$ distributed according to $q(x)$.

Summarizer persona games are a convenient way to model the outcome of persona worlds when some of the joint personas have multiple behavioral game equilibria. (We use them this way below, in Section 5.)

Typically, a summarizer equilibrium P^A is invariant under any affine transformation of A_i for any $i \in \mathscr{N}$. To illustrate this, write such a transform as $a_i \to Ca_i + D \; \forall a_i \in A_i$, and consider again the example above where each $a_i \in A_i$ is a free utility. Applying our affine transform to such an A_i is equivalent to multiplying both u_i and $[\beta_{a_i}]^{-1}$ for each $a_i \in A_i$ by C, and then adding D to u_i. Doing this will not affect the value of $e^{\beta_{a_i}[\mathbb{E}_{q_{-i}}(u_i|x_i) - \mathbb{E}_{q_{-i}}(u_i|x_i')}]$ for any x_i, x_i', q_{-i}. Accordingly, such a transform will not affect the set of equilibrium q's specified by any joint set of free utilities, a, that is, it will not affect $\mathcal{E}(X, a)$ for any joint persona a. Therefore, it will not affect any function W_i, and so will not change the summarizer equilibrium P^A. Note that this invariance holds even if only one free utility $a_i \in A_i$ undergoes the affine transform.

Similarly, if a persona a_i is an expected utility, then for every two joint distributions q, q', $a_i(q) > a_i(q')$ iff $Ca_i(q) + D > Ca_i(q') + D$. Accordingly, the NE q_i will be the same for personas a_i and $Ca_i + D$. This in turn means that the NE q will be the same if the second of these personas

replaces the first. So again, the affine transformation does not affect any function W_j, and so will not change the summarizer equilibrium P^A. The other types of persona equilibria discussed below are also typically invariant under an affine transform of any A_i.

The most natural way to implement a summarizer persona game for a persona world (X, U, A) and equilibrium-averaging summarizer w is with a two-step process. First, the N players play the persona game, based on common knowledge in the usual way. This produces a Nash equilibrium, that is, a product distribution over joint personas, $P^A(a)$. Next, that Nash equilibrium is sampled to produce a joint persona a'. At the end of this first step, each player i adopts persona a'_i. The resultant joint persona is then made known to all the players, so that joint persona becomes common knowledge. (See the discussion below on signaling of personas.)

In the second step, in the usual way the N players play (X, a'), the behavioral objective game with their payoff functions set to a'. This produces a set of possible equilibrium joint product distributions $\mathcal{E}(X, a')$. Finally, each player i receives "payoff" $\bar{w}_i[\mathcal{E}(X, a'), U_i]$. Typically, this payoff is actually an expectation value, over all $q \in \mathcal{E}(X, a')$, of the associated value $U_i(q)$. The goal of each player i in such a persona game is to choose $P^A_i(a_i)$ so that, given the choices $P^A_{-i}(a_{-i})$ of the other players, the expected value of the payoff she receives at the end of the second step is as large as possible.

References

Anderson, S. P., J.K. Goeree, and C. A. Holt. 2001. "Minimum-Effect Coordination Games: Stochastic Potential and Logit Equilibrium." *Games and Economic Behavior* 34: 177–199.

Andrade, E., and T. Ho. 2009. "Gaming Emotions in Social Interactions." *Journal of Consumer Research* 36.

Basu, K. 1994. "The Traveler's Dilemma: Paradoxes of Rationality in Game Theory." *American Economic Review* 84: 391–395.

Basu, K. 2007. "The Traveler's Dilemma." *Scientific American*.

Becker, G. 1976. "Altruism, Egoism and Genetic Fitness: Economics and Sociobiology." *Journal of Econonic Literature* 14: 817.

Becker, T., M. Carter, and J. Naeve. 2005. "Experts Playing the Traveler's Dilemma." Universitat Hohenheim Nr. 252/2005.

Bergstrom, T. 1999. "Systems of Benevolent Utility Functions." *Journal of Public Economic Theory* 1: 71–100.

Bester, H., and W. Guth. 2000. "Is Altruism Evolutionarily Stable?" *Journal of Economic Behavior and Organization* 34: 193–209.

Bowles, S. 2008. "Policies Designed for Self-Interested Citizens May Undermine 'The Moral Sentiments'." *Science* 320: 1605.

Camerer, C.F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Camerer, C.F., and E. Fehr. 2006. "When Does Economic Man Dominate Social Behavior?" *Science* 311: 47–52.

Capra, C. M., J. K. Goeree, R. Gomez, and C. H. Holt. 1999. "Anomalous Behavior in a Traveler's Dilemma Game." *American Economic Review* 19: 678–690.

Chuah, S.H., R. Hoffman, M. Jones, and G. Williams. 2007. "Do Cultures Clash? Evidence from Cross-National Ultimatum Game Experiments." *Journal of Economic Behavior and Organization* 64: 35–48.

Cooper, R., D.V. DeJong, R. Forsythe, and T.W. Ross. 1996. "Cooperation Without Reputation: Experimental Evidence from Prisoner's Dilemma Games." *Games and Economic Behavior* 12: 187–218.

Cover, T., and J. Thomas. 1991. *Elements of Information Theory*. New York: Wiley-Interscience.

Dawes, R.M., and R. Thaler. 1988. "Anomalies: Cooperation." *Journal of Economic Perspectives* 2: 187–197.

De Long, J.B., A. Schleifer, L.H. Summers, and R.J. Wadmann. 1990. "Noise Trader Risk in Financial Markets." *Journal of Political Economy* 98: 703–738.

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

Ekman, P. 2007. *Emotions Revealed.* Holt Paperbacks.

Ficici, S., O. Melnik, and J. Pollack. 2005. "A Game-Theoretic and Dynamical-Systems Analysis of Selection Methods in Coevolution." *IEEE Transactions on Evolutionary Computation* 9: 580–602.

Fogel, D., G. Fogel, and P. Andrews. 1997. "On the Instability of Evolutionary Stable Strategies." *BioSystems* 44: 135–152.

Frank, R.H. 1987. "If Homo Economicus Could Choose His Own Utility Function, Would He Want One With a Conscience?" *The American Economic Review* 77(4): 593–604.

Frith, C. 2008. "Social Cognition." *Philosophical Transactions of the Royal Society B* Doi:10,1098/rstb.2008.0005.

Fudenberg, D., and D. Kreps. 1993. "Learning Mixed Equilibria." *Games and Economic Behavior* 5: 320–367.

Fudenberg, D., and D. K. Levine. 1993. "Steady State Learning and Nash Equilibrium." *Econometrica* 61(3): 547–573.

Fudenberg, D., and D. K. Levine. 2006. "A Dual Self Model of Impulse Control." *American Economic Review* 96: 1449–1476.

Gachter, S., and B. Herrmann. 2009. "Reciprocity, Culture and Human Cooperation: Previous Insights and a New Cross-Cultural Experiment." *Philosophical Transactions of the Royal Society B* 354: 791–806.

Goeree, J. K., and C. A. Holt. 1999. "Stochastic Game Theory: For Playing Games, Not Just Doing Theory." *Proceedings National Academy of Sciences* 96: 10564–10567.

Grimm, V., and F. Mengel. 2010. "Let Me Sleep On It; Delay Reduces Rejection Rates in Ultimatum Games." Http://edocs.ub.unimaas.nl/loader/file.asp?id=1491.

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

Heifetz, A., C. Shannon, and Y. Spiegel. 2007. "The Dynamic Evolution of Preferences." *Economic Theory* 32: 251–286.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, and R. McElreath. 2001. "Cooperation, Reciprocity and Punishment in Fifteen Small-scale Societies." *American Economic Review* 91.

Henrich, J., et al. 2006. "Costly Punishment Across Human Societies." *Science* 312: 1767 – 1770.

Hopkins, Edward. 2002. "Two Competing Models of How People Learn in Games." *Econometrica*.

Huck, S., and J. Oechssler. 1999. "The Indirect Evolutionary Approach to Explaining Fair Allocations." *Games and Economic Behavior* 28: 13–24.

Israeli, E. 1996. "Sowing Doubt Optimally in Two-Person Repeated Games." *Games and Economic Behavior* 28: 203–216.

Jackson, M., and S. Wilkie. 2005. "Endogenous Games and Mechanisms: Side Payments Among Players." *Review of Economic Studies* 72: 543–566.

Kahneman, D. 2003. "Maps of Bounded Rationality: Psychology of Behavioral Economics." *American Economic Review* 93: 1449–1475.

Kalai, A., E. Kalai, E. Lehrer, and D. Samet. 2008. "Voluntary Commitments Lead to Efficiency." Northwestern university, Center for mathematical studies in economics and management science, discussion paper 1444.

Kissinger, H. 1957. *Nuclear Weapons and Foreign Policy*. Harper and Brothers.

Kockesen, L., E. Ok, and R. Sethi. 2000. "The Strategic Advantage of Negatively Interdependent Preferences." *Journal of Economic Theory* 92: 274–299.

Lachmann, M., C. Bergstrom, and S. Szamado. 2001. "Cost and Conflict in Animal Signals and Human Language." *Proceedings of the National Academy of Sciences* 98: 13189–13194.

Mackay, D.J.C. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

McKelvey, R. D., and T. R. Palfrey. 1995. "Quantal Response Equilibria for Normal Form Games." *Games and Economic Behavior* 10: 6–38.

McKelvey, R. D., and T. R. Palfrey. 1998. "Quantal Response Equilibria for Extensive Form Games." *Experimental Economics* 1: 9–41.

Meginniss, J. R. 1976. "A New Class of Symmetric Utility Rules for Gambles, Subjective Marginal Probability Functions, and a Generalized Bayes' Rule." *Proceedings of the American Statistical Association, Business and Economics Statistics Section* 471 – 476.

Myerson, Roger B. 1991. *Game Theory: Analysis of Conflict*. Harvard University Press.

Nowak, M., K. Page, and K. Sigmund. 2000. "Fairness Versus Reason in the Ultimatum Game." *Science* 289.5485: 1773.

Nowak, M. A. 2006. "Five Rules for the Evolution of Cooperation." *Science* 314: 1560–1563.

Raub, W., and T. Voss. 1990. "Individual Interests and Moral Institutions." In *Social institutions, their emergence, maintenance and effects*, eds. M. Hechter, K.-D. Opp, and R. Wippler. Walter de Gruyter Inc.

Renou, L. 2008. "Committment Games." *Games and Economic Behavior*.

Rubinstein, Ariel. 2004. "Instinctive and Cognitive Reasoning: a Study of Response Times." Http://arielrubinstein.tau.ac.il/papers/Response.pdf.

Samuelson, L. (Ed.). 2001. "Evolution of Preferences." Special issue, *Journal of Economic Theory* 97(2).

Schelling, T.C. 1960. *The Strategy of Conflict*. Harvard University Press.

Shamma, J.S., and G. Arslan. 2004. "Dynamic Fictitious Play, Dynamic Gradient Play, and Distributed Convergence to Nash Equilibria." *IEEE Trans. on Automatic Control* 50(3): 312–327.

Sobel, J. 2005. "Interdependent Preferences and Reciprocity." *Journal of Economic Literature* 43.

Spence, M. 1973. "Job Market Signaling." *The Quarterly Journal of Economics* 87: 355–374.

Spence, M. 1977. "Consumer Misperceptions, Product Failure and Product Liability." *The Review of Economic Studies* 44: 561–572.

Stephens, G., L. Silbert, and U. Hasson. 2010. "Speaker-Listener Neural Coupling Underlies Successful Communication." *Proceedings of the National Academy of Sciences* 107(32): 14425–14430.

Stern, S. 2008. "Be Yourself But Know Who You Are Meant to Be." *Financial Times* March 17.

Tversky, Amos. 2004. *Preference, Belief, and Similarity: Selected Writings*. MIT Press.

von Widekind, S. 2008. *Evolution of Non-Expected Utility Preferences*. Springer.

Winter, E., O. Garcia-Jurado, J. Mendez-Naya, and L. Mendez-Nay. 2009. "Mental Equilibrium and Rational Emotions." Discussion Paper Series dp521, Center for Rationality and Interactive Decision Theory, Hebrew University, Jerusalem.

Wolpert, D. H. 2007. "Predicting the Outcome of a Game." Submitted. Updated version at papers.ssrn.com/1184325.

Wolpert, D. H. 2009. "Trembling Hand Perfection for Mixed Quantal Response / Nash Equilibria." *International Journal of Game Theory* In press.

Wolpert, D.H., J. Jamison, D. Newth, and Harre M. 2008. "Schelling Formalized: Strategic Choices of Non-Rational Personas." Submitted. papers.ssrn.com/1172602.