Strategic Choice of Preferences: The Persona Model

David H. Wolpert
Julian Jamison
David Newth
Michael Harre

SFI WORKING PAPER: 2011-08-030

SANTA FE INSTITUTE

Strategic Choice of Preferences: the Persona Model

David H. Wolpert,[1*], Julian Jamison,[2]
David Newth,[3] Michael Harre[4]

[1]Santa Fe Institute, 1399 Hyde Park Road Santa Fe, NM 87501
Center for Nonlinear Studies, MS B258, Los Alamos National Laboratory, Los Alamos, NM 87545
MS 269-1, NASA Ames Research Center, Moffett Field, CA, 94035, USA
[2]Dept of Economics, Yale University, PO Box 208268, New Haven, CT 06520
[3]CSIRO Centre for Complex Systems Science, CSIRO Marine and Atmospheric Research,
PYE Laboratory, Clunies Ross Street, Canberra ACT, Australia
[4]The Centre for the Mind, University of Sydney, Australia

[*]To whom correspondence should be addressed; E-mail: david.h.wolpert@nasa.gov

August 19, 2011

Abstract

Recent work in several fields has established that humans can adopt binding "behavioral" preferences and convincingly signal those preferences to other humans, either via their behavior or via their body language / tone of voice. In this paper we model the strategic implications of this ability. Our thesis is that through a person's lifetime they (perhaps subconsciously) learn what such signaled, binding behavioral preferences result in the highest value of their actual preferences, given the resultant behavior of other players. We argue that this "persona" model may explain why many interpersonal preferences have the particular form they do. As an illustration, we use the persona model to explain cooperation in non-repeated versions of the Prisoner's Dilemma (PD). We also provide quantitative predictions

to distinguish this explanation of cooperation from simply assuming people have actual preferences biased towards cooperation. In particular, we show that the persona model predicts a "crowding out" phenomenon in the PD, in which introducing incentives to cooperate causes players to stop cooperating instead. We also use the persona model to predict a tradeoff between the robustness of cooperation in the PD and the benefit of that cooperation.

Keywords: Non-rationality, Single shot games, Prisoner's Dilemma, Traveler's Dilemma, Schelling, Emotions

JEL Codes: C70, C72, D03

1 Introduction

1.1 Behavioral preference models

Suppose that a human could adopt binding "behavioral" preferences and convincingly signal those preferences to other humans, either via their behavior or via their body language / tone of voice. In general, which such preferences they adopt will affect what strategies are chosen by their opponents in any given game. Accordingly, the behavioral preferences someone adopts will affect the value of their *actual*, underlying preferences, in any game the play. Accordingly, if a human had this ability to adopt binding preferences that are convincingly signaled, they should exploit it to increase the value of their underlying preferences.

In this paper we investigate this hypothesis. Our thesis is that through a person's lifetime they (perhaps subconsciously) learn what such signaled, binding behavioral preferences result in the highest value of their actual preferences, by observing the resulting behavior of other players. We argue that this "persona" model of human behavior may explain many types of (what appear to be) non-rational behavior.

To be more precise, consider a noncooperative game where there are exogenously provided utility functions for all the players that are publicly known, e.g., monetary payoffs to the participants in a game theory experiment. Say that for each such exogenous utility function the differences between the possible values of the function over all game outcomes are very small, so that to first order, they are linear functions of the player's true utility function. Very often a real human player in such a game will not behave in a way that maximizes their exogenous utility function, given the behavior of the other players. This is true even in single-shot games among strangers, where every player knows they will never interact with their opponent(s) again, and knows that their play in the game will not be observed by third parties. (See (Camerer (2003), Gachter and Herrmann (2009)), and references therein.)

A lot of research has modeled such non-rational behavior by hypothesizing that humans have **behavioral** preferences, which differ from their exogenous preferences, and that they maximize behavioral preferences rather than exogenous preferences. In particular, this is the approach of the work on other-regarding / interdependent social preferences (Sobel (2005), Bergstrom (1999), Kockesen, Ok, and Sethi (2000)).

Here we will refer to such models as **behavioral preference models**, with the non-rational behavior given by simultaneous maximization of every player's

behavioral preferences called a **behavioral preference equilibrium** of an associated **behavioral game**. The original game defined by the exogenously provided preferences will be called the **exogenous game**.

Two important questions are unaddressed in this set of results on behavioral preference models:

1. How do the players acquire common knowledge of one another's behavioral preferences before start of play? After all, in the types of behavioral preference equilibria discussed above, the preferences of each player i differs from the expected value of their publicly visible exogenously provided utility. What process leads players to conclude that the preferences of their opponents are not those exogeneous utilities?

2. Why do humans use a particular set of behavioral preferences for any given noncooperative game, when those behavioral preferences differ from their exogenous preferences? In particular, the literature on interdependent preferences does not consider why a human should try to optimize *any* set of interdependent preferences that differ from their exogenously provided preferences, never mind explain why they should try to optimize the particular preferences they appear to. What process leads people to bindingly adopt behavioral preferences instead of exogenous preferences?

In this paper we present a simple argument that we claim answers these two questions in many circumstances. Our answer is based on the fact, established by recent work in several fields, that in many contexts humans do have the ability to adopt binding preferences that are convincingly signaled to others (Bechara and Damasio (2004), Tamir (2010), Ekman (2007), Frith (2008), Brinol, K., and K. (2010)). It appears that this ability is facilitated by neurological feedback mechanisms that force agreement between signaled preferences and those controlling behavior (Tom, Pettersen, Lau, Burton, and Cook (1991), Brinol and Petty (2003), Bechara and Damasio (2004), Tamir (2010)). In fact, as a result of this forced agreement, even if the signaled preferences are externally imposed on a person, they can cause that person to modify their behavior to agree with the signaled preferences, e.g., when the signal is an imposed article of clothing (Gino, Norton, and Ariely (2010)).

There are numerous examples of such necessarily honest preferences signals. One derives from the fact that in humans (unlike other species), the whites of the eyes are visible. This means that where we look telegraphs our attention (which in turn provides much information about our preferences).

In other examples, the signal of preferences is not fully under conscious control. In particular they may be via unconscious body language or tone of voice (Frank (1987), Frith (2008)). Some such signals exploit the anomalously large number of facial muscles humans have compared to other species: the "purpose" of those muscles seems to be to provide highly informative signals concerning emotional states, i.e., concerning preferences. (See for example the copious experimental data establishing unconscious control of facial expressions (Ekman (2007)).)

The signal of an adopted set of preferences does not have to be via gaze, body language or tone of voice however. Indeed, so long as a particular set of preferences adopted by a person cannot be easily discarded (i.e., it is indeed binding), then those preferences will necessarily be signaled via many kinds of publicly observable actions by that person.

Clearly this ability provides a process that answers the first of the two questions raised above, at least in certain circumstances. To see how it can also answer the second question, note that changing a player's behavioral preferences changes the equilibrium strategy profile, assuming those changed preferences are known to the other players. In particular, for a fixed set of behavioral preferences for all players other than i, by varying i's behavioral preferences, 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 exogenous preferences. In this way the possible behavioral preferences of i can be ranked according to i's exogenous preferences.

In a nutshell, our hypothesis is that over the course of their lifetime a human learns what behavioral preferences out of a candidate set of such preferences have the highest such rank in terms of their exogenous preferences, for any type of game (given that those behavioral preferences will be made known to other players via the signaling mechanisms discussed above). In this way, their behavioral preferences are determined endogenously, in a purely rational way, as the ones that optimize their exogenous preferences. We refer to such endogenously determined preferences as "personas".

We make no assumption about whether people consciously choose what persona to adopt just before start of play, or whether their lifetime learning of what persona is best and their internalization of that persona is completely subconscious. Both processes are modeled with our framework. Nor do we claim that the persona phenomenon is always the sole cause of people adopting the behavioral preferences they do, or that the associated signaling of preferences is the sole way players can come to know one another's behavioral preferences. In particular, repeated games and associated social norms can often be involved in both

processes. It may even be that natural selection of the human genome plays a role, as is implicitly assumed in much recent work. (See sections 5.3 and 6 below for comments on this work.) Rather our claim is that the persona phenomenon often contributes to determining the behavioral preferences that people adopt. Moreover, it does so in a way that is far simpler to model and analyze than are repeated games or evolutionary processes, which can often only be analyzed via computer simulation.

The kind of learning of personas across a lifetime that we posit can be seen as an instance of the Baldwin effect (Dennett (2003)). In the Baldwin effect, natural selection does not provide individuals of a species with a particular behavior, but rather with the ability to learn over their lifetimes what behavior to adopt. Our hypothesis is that natural selection has provided people with the ability to adopt binding behavioral preferences that are convincingly signaled to other people. Given this ability, a person would then learn, based on their interactions with their (social) environment, which behavioral preferences are best to adopt for different kinds of games. Assuming that one of the possible behavioral preferences an individual can learn are just the exogenous preferences, the expected value of those exogenous preferences after learning their best behavioral preferences would always be at least as large as the value they would have if the person were forced to always use their exogenous preferences as their behavioral preferences. (This inequality is why natural selection would provide them the ability to learn their behavioral preferences in the first place.) This use of the Baldwin effect provides natural selection a way to optimize behavioral preferences that is more flexible than directly setting those preferences, as for example is assumed in the Evolution of Preferences (EOP) framework, discussed in Sec. 5.3 below.

From now on, as shorthand, we use the term **persona game** to refer an extension of an exogenously provided game in which before start of play, all players choose a persona from their associated **persona set**. They then use the persona they chose to play a behavioral game. This two-stage process is an equilibrium, in that players choose their personas in the first stage so that the outcome of the second stage's behavioral game maximizes their exogenous payoffs. (More generally, in the first stage players choose a distribution over their persona set, which is then sampled in the exogenous game.) A formal definition of persona games is provided below.

1.2 Paper overview

This paper is an analysis of the strategic implications of the ability of a person to adopt a binding persona that is signaled to their opponents before start of play, and in particular of the implications for the question of why humans have interdependent preferences. We don't analyze the evolutionarily stability of the ability to adopt a binding persona, but take it as an empirical fact. Nor do we analyze the physical process whereby the players gain common knowledge of one another's binding personas before start of play, given that at start of play they are told one another's exogenous preferences instead. While these issues are quite important, they are secondary to this paper's focus. They are also not analyzed in the literature summarized in Sec. 5.3, e.g., the EOP literature. (See (Wolpert and Jamison (2011), Wolpert, Jamison, Newth, and M. (2008)) for a high-level discussion of the physical processes by which personas are signaled.)

We begin in Sec. 2 by providing a semi-formal definition of persona games, to ground intuition. We then present our notation, followed by a fully formal definition of persona games, for the simplest case of perfect signaling. We also establish some of the basic mathematical properties of persona games in this section.

In Sec. 3 we focus on personas 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".

Next in this section we present a novel prediction concerning real-world play in 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.

We end this section with some concrete predictions that can be tested in future experiments.

Next, in Sec. 4, we extend the persona framework 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 their persona, and therefore modify the behavior of the other players in later stages.

Then in Sec. 5, we discuss extensions of the persona framework, and how that framework can also describe other kinds of scenarios, e.g., 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 exogenous game to what persona the player chooses for that game.

Finally, the persona model is related to other work in Sec. 5.3. (We delay this section to the end since much of it involves details of the main discussion of the persona model.)

Concluding comments are in Sec. 6.

In (Wolpert and Jamison (2011), Wolpert et al. (2008)) we show that empirical data on the Traveler's Dilemma (Basu (1994), Becker, Carter, and Naeve (2005), Capra, Goeree, Gomez, and Holt (1999)) is explained by persona games, and briefly describe how the persona games framework explains empirical data on the Ultimatum Game. Those papers also show how a variant of the persona game framework can provide explanations for the values of the exponent in the logit QRE.

Throughout this paper, just like most of the work on interdependent preferences and on EOP, for simplicity we restrict attention to finite games. Also like most of the work on interdependent preferences and on EOP, we do not consider the issue of why some persona sets arise rather than others. (Discussion of why persona sets must be finite in the real world arise below in Prop. 2 and Sec. 5.1.) Our focus in this paper is on what the implications are of the ability of people to adopt personas, and in particular on how this may explain why people adopt some parameter values rather than others in other-regarding preferences.

2 Persona games

2.1 Semi-formal motivation of persona games

To ground the intuition, we start in this subsection with a semi-formal definition of the simplest type of persona games.

Like in EOP, in a persona game each player i has an associated exogenous utility function $u_i : X \to \mathbb{R}$, where X is the finite joint move space. Also like in

EOP, in persona games each player i has a set of possible behavioral preferences, i.e., personas, which we 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 are expected utilities, the usual choice in EOP, in persona games we also allow personas other than expected utilities.[1]

Just like in EOP, in persona games the exogenous 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 like in EOP, each player i honestly signals that a_i to the other players.[2] Then in the second stage the players play a NE of the behavioral game, i.e., a NE of the strategic form game with joint move space X and player utility functions given by those signaled personas.[3] The mixed strategy profile of that NE determines the expected values of the players' exogenous utility functions.

We assume that before signaling their personas, each player knows their own set A_i, the sets $\{A_j\}$ of the other players, and knows the exogenous 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. They do so to maximize the associated expected value of their exogenous utility, as evaluated under the NE of the behavioral games.

Note that the players in the first stage of a persona game can be viewed as playing a game. 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 exogenous 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".

It is important to emphasize that no player in the first stage bindingly chooses

[1] 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 their moves. See also (von Widekind (2008)).

[2] A process essentially equivalent to such signaling is implicit in the conventional analysis of non-repeated 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 all of the work on other-regarding preferences.

[3] Note that just like in EOP, in persona games every player i is assumed to play the behavioral game using the persona that they signaled, i.e., it is assumed that the signals were binding. The same kind of assumption arises in games of signaled binding commitments, e.g.,(Renou (2008)). See Sec. 4.2 below and the discussion of the relation between EOP and persona games in (Wolpert and Jamison (2011), Wolpert et al. (2008)).

what move they will play in the behavioral game of the second stage. A player i's choice of persona, by itself, does not determine i's move in the behavioral game. It is the profile of *all* the players' persona choices that jointly determines i's move in the behavioral game (as well as determining the moves of all of i's opponents). A choice of a persona by i is more analogous to a signaled commitment by i of what move i will make in the behavioral game in response to any of the possible similar signals of i's opponents. (Personas are both more flexible and analytically simpler than such signaled commitments of signal-contingent moves however; see the discussion in Sec. 5.2.)

Note that in persona games we assume that the exogenous 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 the work on interdependent preferences, EOP, etc.) Physically, the determination of these exogenous factors may occur through an evolutionary process, either biological and/or cultural. However — like EOP — we do not model such processes here.

We refer to each A_i as a **persona set**. The persona sets humans use in field experiments appear to be quite elaborate. It seems that they are often at least partially determined by social conventions and norms. In addition, it seems that people sometimes use a different persona set for different joint exogenous utility functions u.[4]

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

2.2 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 $\mathcal{N}$, e.g., $\{X_i\}$ is shorthand for $\{X_i : i \in \mathcal{N}\}$. For

[4]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. Ex. **??** in Sec. 5 illustrates a multi-game situation that is sometimes called "interdependent preferences" in this other sense, and shows how to formalize it in terms of multiple persona sets.

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

Given any set of N finite (pure) strategy spaces, $\{X_i\}$, we follow the terminology of (Wolpert (2009)) and refer to any function that maps $\Delta_X \to \mathbb{R}$ as an **objective 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 preferences. As another example, a **free utility** objective function for player i is one of the form $\mathbb{E}_q(u) + TS(q_i)$ where $S(.)$ is the Shannon entropy and $T \geq 0$. [5] Use of objective functions means

[5]The free utility is central to the analysis in (Wolpert (2009)). It is also central to the variant

the framework encompasses non-expected utility theory by construction.

Any pair of a set of finite strategy spaces and an associated set of one preferences 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 $i \in \mathscr{N}$.

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

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

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}.

2.3 Persona worlds

Define a **persona world** as any triple

$$(\{X_i : i \in \mathscr{N}\}, \{U_i : i \in \mathscr{N}\}, \{A_i : i \in \mathscr{N}\}) \tag{1}$$

where $\{X_i\}$ is a set of N finite strategy spaces, $\{U_i\}$ is an associated set of N preferences for $X = \bigtimes_i X_i$, and each A_i is a set of preferences 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 i'th player, with A_i the associated **persona set**. Note that any such persona is a mapping from $\Delta_{\mathcal{X}}$ into $\mathbb{R}$.[6]

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, i.e., that are of the form $P^A(a) = \prod_{i \in \mathscr{N}} P_i^A(a_i)$. 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, i.e., that are of the form $P(x \mid a) = \prod_{i \in \mathscr{N}} P(x_i \mid a_i)$ for all $a \in A$. We make the analogous definition for $\Delta_{\mathcal{X}_{-i}|A_{-i}}$. We refer to (X, U) as an **exogenous game**, and any (X, a) for some $a \in A$ as a **behavioral game**.

---

of persona games that explains the exponents of the logit QRE (Wolpert et al. (2008), Wolpert and Jamison (2011)).

[6] At the expense of more notation, we could extend the domains of each preferences 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. In particular, this would allow us to model scenarios in which different personas impose different computational costs on the player adopting them.

2.4 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 i'th 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:

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

$$\int da_i da_{-i}\, \hat{P}(a_i) P(a_{-i}) U_i \left[\prod_{j \in \mathcal{N}} q(X_j^a) \right] \; > \; \int da_i da_{-i}\, P(a_i) P(a_{-i}) U_i \left[\prod_{j \in \mathcal{N}} q(X_j^a) \right] \quad (2)$$

and

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

$$a_i \left[\hat{q}(X_i) \prod_{j \neq i} q(X_j^a) \right] \; > \; a_i \left[q(X_i^a) \prod_{j \neq i} q(X_j^a) \right]. \quad (3)$$

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 Eq. (3), 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 Eq. (3) for that a, but is not part of a solution to the encompassing set of coupled equations given by Eq. (2) together with Eq. (3).

The following result is proven in App. B:

Theorem 1 *Let (X, U, A) be a persona world where all the exogenous utilities are expected utilities 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 App. B for details.

In addition, persona games become degenerate for any persona set that includes the indifferent preferences, $\mathscr{I}$ which has the same value for all arguments q. (This preferences 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 exogenous utilities are expected utilities and all the persona sets include the indifferent preferences. 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 is all-indifferents. 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 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

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

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 Eq.'s 2 and 3.

First note that Eq. 3 is obeyed for the given persona profile $\vec{\mathscr{I}}$, since no behavioral game player can improve the value of their indifferent persona by changing their strategy. Next, hypothesize that some player i changes their 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 their behavioral game strategy, player j has not improved the value of their preferences. ■

So in particular, persona games become degenerate in this sense for infinite persona sets, since such persona sets in particular contain the indifferent persona. This degeneracy is not much of a concern however, since there are already many physical reasons why infinite persona sets cannot occur in the real world, independent of these formal issues. (See Sec. 5.)

3 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, e.g., 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.

3.1 The Prisoner's Dilemma

Let $\{u_j : j = 1, \ldots N\}$ be the utility functions of the original N-player exogenous game. Have the persona set of player i be specified by a set of distributions $\{\rho_i\}$, 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, e.g., a completely unselfish inequity aversion could be modeled as the persona $[(N-1)u_i - \sum_{j \neq i} u_j]^2$.)

In the case of the PD exogenous game, such 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 their own utility and for their opponent's utility). Consider the PD where the utility function bimatrix (u^R, u^C) is

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

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

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

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

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

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)

(7)

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 exogenous game, with utility functions

$$\begin{bmatrix} (\beta,\beta) & (0,\alpha) \\ (\alpha,0) & (\gamma,\gamma) \end{bmatrix} \tag{8}$$

where $(\mathbf{R},\mathbf{D})$ is (defect, defect), so $\alpha > \beta > \gamma > 0$. We are interested in what happens if the persona sets of both players is 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

$$R_1 \equiv \beta - s\alpha, \tag{9}$$

$$R_2 \equiv \gamma - (1-s)\alpha, \tag{10}$$

$$B \equiv \beta - \gamma. \tag{11}$$

Working through the algebra (see App. A), we first see that neither the non-rational nor the antirational persona will ever be chosen. We also see that for joint cooperation in the behavioral game (i.e., $(\mathbf{L},\mathbf{T})$) to be a NE under the $(\mathscr{C},\mathscr{C})$ joint persona choice, we need $R_1 > 0$ (see App. 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, then 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 Eq.* (8)*, and each player i has the persona set $A_i = \{U_i, sU_i + (1-s)U_i\}$ for some $0 \le s \le 1$. In the associated (unique, pure move) extended persona equilibrium, the joint persona move is $(\mathscr{C},\mathscr{C})$ followed by $(\boldsymbol{L},\boldsymbol{T})$ whenever $s \in (1-\frac{\gamma}{\alpha}, \frac{\beta}{\alpha}]$.*

For our range on allowed s's to be non-empty requires that $\gamma > \alpha - \beta$. Intuitively, this means that player i's defecting in the exogenous 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, etc., can result in joint cooperation being an equilibrium of the infinitely repeated PD.

Now say that one changes the exogenous game of Eq. 8 by adding a penalty $-c < 0$ to the payoff of every player i if they defect, giving the bimatrix

$$\begin{bmatrix} (\beta,\beta) & (0,\alpha-c) \\ (\alpha-c,0) & (\gamma-c,\gamma-c) \end{bmatrix} \tag{12}$$

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 exogenous 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 Eq.* (12)*, and each player i has the persona set $A_i = \{U_i, sU_i + (1-s)U_i\}$ for some $0 \le s \le 1$. 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 $(\boldsymbol{L},\boldsymbol{T})$. 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)). This is the only formal explanation of crowding out in single-shot games as rational behavior that we are aware of.[7]

Return now to the $c = 0$ PD in Eq. 8, so that if both players defect, each player's utility is γ. For this exogenous game, when the extended persona game equilibrium is $(\mathscr{C},\mathscr{C})$ followed by $(\mathbf{L},\mathbf{T})$, the benefit to each player of playing the persona game rather than playing the exogenous game directly is B. Comparing this to 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 Eq.* (8)*, and each player i has the persona set $A_i = \{U_i, sU_i + (1-s)U_i\}$ for some $0 \le s \le 1$. Then $R_1 + R_2 + B \le 1$.*

[7] In (Bar-Gill and Fershtman (2005)) there is a complicated EOP model that shows how crowding out effects could arise in the asymptotic limit after a change to the payoff matrices. In the field, those effects instead arise almost instantaneously — the situation considered by our analysis.

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 exogenous 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 large means that both γ and s are (relatively) large. These conditions guarantee something concerning your opponent: they are not so inclined to cooperate that it benefits you to take advantage of them and be selfish. On the other hand, having R_1 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 it would interesting extension would be to expand 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 deferred to future work.

3.2 Experimental Predictions

There are many predictions our model of the PD makes that could be tested experimentally. One could test the predictions for what parameters of PD games do (not) result in cooperation. One could also test the predictions for what parameters do (not) result in crowding out.

Note in both of these predictions the importance of non-anonymity, to allow signaling of personas. Concretely, this would require at a minimum that the players are in close proximity to each other. It may even prove important to allow them to talk and "get to know each other" before and during their play of the game.

This importance of non-anonymity provides another, informal prediction, that in repeated game versions of the PD with anonymity, due to no persona signaling, NE play is more likely to arise quickly, and defection to arise.

None of these predictions of the persona framework have yet been compared to experimental data, since we have not found data in the literature that directly matches the scenarios of our predictions. However some preliminary work has been done elsewhere comparing predictions of the persona framework to experimental data on a different scenario. In particular, experimental data concerning "bounded rational" behavior in the traveler's dilemma (Basu (2007), Capra et al. (1999), Basu (1994), Rubinstein (2004), Becker et al. (2005), Goeree and Holt

(1999)) has been shown in (Wolpert and Jamison (2011), Wolpert et al. (2008)) to be well explained by persona games. See also the discussion section below.

4 Persona games with imprecise persona signaling

4.1 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 their own persona, but only has 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. It also provides a robustness check on our model.

Such noisy situations are inherently more complicated than noise-free persona games. So in the interests of space, here we only show that a formal definition of persona game equilibria 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 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 that 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 its 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 (i.e.,

conditioned on the information known to player i), can be written as

$$\begin{aligned} Pr(x^{a_i,z_i}) &= \int da_{-i}dz_{-i}\, Pr(a_{-i},z_{-i} \mid a_i,z_i)Pr(x \mid a,z) \\ &= \int da_{-i}dz_{-i}\, Pr(a_{-i},z_{-i} \mid a_i,z_i) \prod_{j\in\mathscr{N}} q(x_j^{a_j,z_j}). \end{aligned} \tag{13}$$

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

$$\begin{aligned} Pr(a,z) &= Pr(z \mid a)P(a) \\ &= \prod_{i\in\mathscr{N}} Q_i^{a_{-i}}(z_i) \prod_{k\in\mathscr{N}} P(a_k) \\ &= \prod_{i\in\mathscr{N}} \left[Q_i^{a_{-i}}(z_i)P(a_i)\right]. \end{aligned} \tag{14}$$

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 Eq. (14),

$$\begin{aligned} Pr(a_{-i},z_{-i} \mid a_i,z_i) &= \frac{Pr(a,z)}{\int dz'_{-i}da'_{-i}\, Pr(a_i,z_i,a'_{-i},z'_{-i})} \\ &= \frac{Q_i^{a_{-i}}(z_i) \prod_{k\neq i}\left[Q_k^{a_{-k}}(z_k)P(a_k)\right]}{\int dz'_{-i}da'_{-i}\, Q_i^{a'_{-i}}(z_i) \prod_{n\neq i}\left[Q_n^{a_i,a'_{-n,-i}}(z'_n)P(a'_n)\right]} \\ &= \frac{Q_i^{a_{-i}}(z_i) \prod_{k\neq i}\left[Q_k^{a_{-k}}(z_k)P(a_k)\right]}{\int da'_{-i}\, Q_i^{a'_{-i}}(z_i) \prod_{n\neq i} P(a'_n)}. \end{aligned} \tag{15}$$

So combining Eq.'s (13) and (15) gives

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

The proof of part (i) of the proposition proceeds analogously. Combining that part (i) with Eq. (15) 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 i'th component of each a_{-k} is implicitly set to a_i. Comparing equations (13) and part (i) of the proposition establishes part (ii) of the proposition. ■

Prop. 6(ii) addresses a potential concern that the framework presented above may be a non-sensical way to model noisy persona games. Intuitively, it says that in the behavioral game, player i chooses their 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:

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

$$\begin{aligned} &\int dadz\ \hat{P}(a_i)P(a_{-i}) \prod_{k\in\mathcal{N}} Q_k^{a_{-k}}(z_k)\ U_i\left[\prod_{j\in\mathcal{N}} q(X_j^{a_j,z_j})\right] > \\ &\qquad \int dadz\ P(a_i)P(a_{-i}) \prod_{k\in\mathcal{N}} Q_k^{a_{-k}}(z_k)\ U_i\left[\prod_{j\in\mathcal{N}} q(X_j^{a_j,z_j})\right] \end{aligned} \tag{16}$$

and

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

$$\int da_{-i} dz_{-i} \; Pr(a_{-i}, z_{-i} \mid a_i, z_i) a_i \left[\hat{q}(X_i^{a_i, z_i}) q(X_{-i}^{a_{-i}, z_{-i}}) \right] \; > \int da_{-i} dz_{-i} \; Pr(a_{-i}, z_{-i} \mid a_i, z_i) a_i \left[q(X_i^{a_i, z_i}) q(X_{-i}^{a_{-i}, z_{-i}}) \right]. \tag{17}$$

where $Pr(a_{-i}, z_{-i} \mid a_i, z_i)$ is given by Eq. (15).

As an alternative, we could replace Eq. (17) with the notationally simpler condition that $\nexists \hat{q} \in \Delta_{X_i}$:

$$a_i \left[\hat{q}(X_i^{a_i, z_i}) Pr(X_{-i}^{a_i, z_i}) \right] \; > \; a_i \left[q(X_i^{a_i, z_i}) Pr(X_{-i}^{a_i, z_i}) \right]. \tag{18}$$

where $Pr(X_{-i}^{a_i, z_i})$ is given by Eq. (16). 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 Prop. 6(i), when all persona 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 App. B, and are omitted here in the interests of space.

4.2 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 straight-forward to prove that in this case 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 exogenous 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 non-rational persona $a_i \neq U_i$.

Empirically, sticky personas seem to be common in real 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 real human will refuse to consider any alternative behavior "once they make up their minds" about how they will behave. Similarly, many real humans "refuse to change their ways", whether or not new data seems to indicate that their behavior is rational.[8]

We can 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 behavioral game stage 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.[9]

For reasons of space, we do not discuss sticky personas any further in this paper. The interested reader is referred to (Wolpert et al. (2008), Wolpert and Jamison (2011)).

5 General discussion of persona games

5.1 Computational issues

Learning a persona equilibrium typically involves far more computational work than calculating the equilibria of the associated exogenous game (since behavioral equilibria have to be calculated for all persona profiles). This has many

[8]There are several possible reasons for such stickiness. First, note that it takes a lot of computational effort to calculate optimal persona distributions. (Crudely speaking, for every possible joint persona, one has to 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 only recompute their persona distributions infrequently compared to 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.)

[9]In more nuanced versions of the analysis, the players can also change their persona mixed strategies between behavioral games, but only by a small amount. Alternatively, they can change their personas 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.

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, who have a lot of interactions with other organisms that can also play persona games. Colloquially speaking, we might characterize such a species who plays persona games well as having "high social intelligence". (Candidate species would be higher primates, corvids, and cetaceans.)

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

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 exogenous game they encounter. (Just think about how many games you play during a typical day, and imagine calculating the precisely optimal persona for every such game.) Rather they might use a simple rule to map any pair {an exogenous game, a specification of which player they are 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 exogenous game a person might play to a persona for them 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 exogenous games rather than just one. This makes the definition intrinsically more complicated. See (Wolpert et al. (2008), Wolpert and Jamison (2011)) for details.

5.2 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, Kalai, Lehrer, and Samet (2008), Jackson and Wilkie (2005), Myerson (1991)). Games of contingent commitments consist of two steps, just like in the persona framework. In the simplest version of those games, in the

first step every player i chooses a contingent commitment from an associated set of possible commitments. Every one of those possible commitments is of the form "if each of the other players send the following signals to me, then I commit to play the following strategy". After all players choose their commitments, the players honestly signal those commitments to one another. Then in the second stage, the joint strategy determined by the joint commitment signaled by the players is implemented 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 changes, 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 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 non-rationality in a single game. However that 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 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 if those predictions are accurate predictions of the outcomes of repeated games. (This is the subject of future work.)

5.3 Relation to previous work

Viewed quite broadly, variants of the persona-based explanation of non-rationality go back at least to the 1950's (Raub and Voss (1990), Kissinger (1957), Schelling (1960)), and arguably back to antiquity (Schelling (1960)). In particular, they played a prominent role in formulation of cold war policies like mutual assured discussion.

However this early work was semi-formal, while there are many subtleties that a fully formal persona game framework must address. For example, for some exogenous 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 be somehow summarized or one of them must be somehow selected. These subtleties are not addressed in the early, semi-formal work.

Moreover, much of that early, semi-formal work allows infinite persona sets. However there are problems with assuming real humans have infinite persona sets, For example, in many strategic situations it takes a very large computational effort for a player to calculate / learn their optimal persona. (Crudely speaking, for every possible joint persona, the player has to ascertain the associated behavioral game equilibria, and only *then* can they determine their persona game equilibrium play.) 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 your possible personas before choosing their own.

Note also that the processes typically used for signaling personas (body language, repeated behavior, etc.) 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 not possible for those processes to quickly and accurately convey a persona choice by a player 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, semi-formal work. Indeed, the fact that not all possible personas are contained in a real person's persona set is taken for granted in the literature on fitting finite-dimensional parameters of interdependent preferences to experimental data; in none of that literature is there an attempt to fit a completely arbitrary, infinite-dimensional set of preferences to the experimental data. (See Sec. 5.1 for further discussion of the computational issues of persona games and their real-world implications.)

A body of work more closely related to the persona model is reputational models of bargaining (Myerson (1991)). In those models, to quote (Wolitzky (2010)), bargaining players may "commit to a range of possible bargaining strategies before the start of bargaining.... {where} the probabilities with which the players are committed to various bargaining postures (either ex ante or after a stage where players strategically announce bargaining postures) are common knowledge, and ... play constitutes an equilibrium.... {The commitments arise because} the bargainer displays emotions that suggest an unwillingness to modify her posture". (See also (Myerson (1991), Abreu and Gul (2000), O. and Jehiel (2002)).) Persona games can be seen as an extension of such models beyond bargaining scenarios, where the player commitments are to personas in general rather than just to bargaining strategies.

Following the appearance of an early version of this paper (Wolpert et al. (2008)), a preprint of another body of work appeared that is related to the persona concept (Winter, Garcia-Jurado, Mendez-Naya, and Mendez-Nay (2009)). While fully formal, the model in (Winter et al. (2009)) allows infinite persona sets, just like the early, semi-formal work, with the attendant problems. Indeed, in (Winter et al. (2009)), those problems mean that the analysis for more than three players is trivial, so the focus is on two-player games. In addition, the model in (Winter et al. (2009)) does not consider extensions like noisy persona games, where the players do not perfectly observe one another's personas. Nor do they consider "sticky" personas, which are multi-stage games where a player is not allowed to change their persona between stages, so their behavior in the early stages can signal their persona, without any need for body language or the like. Both such extensions of persona games are considered below.

There are other, fully formal studies that are loosely related to the persona concept (Frank (1987), Israeli (1996), Becker (1976), De Long, Schleifer, Summers, and Wadmann (1990)). These studies each apply a model of limited scope to a restricted (and often complicated) scenario. For example, many of these studies focus so much 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 full gen-

erality. Also recently, some work has appeared (Renou (2008)) that is related to the "binding commitments" interpretation of personas discussed in Sec. 5 below. More recently still, some work appeared that can be viewed as an investigation of personas in an experimental context (Andrade and Ho (2009)).

Persona games are also related to games of contingent commitments, and to games of binding commitments. This relation is discussed at length below in Sec. 5.2.

If we interpret persona games as a variant of the Baldwin effect, they can be viewed as a two-timescale game where the longer timescale transpires across an individual's lifetime. This perspective connects persona games to yet other bodies of work. For example, the analysis in (Fudenberg and Levine (2006)) posits 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 or she cares about.

Finally, persona games are similar to the work on Evolution Of Preferences (EOP) (Huck and Oechssler (1999), Heifetz, Shannon, and Spiegel (2007), Samuelson (2001), Dekel, Ely, and O. (2007), Guth and Yaari (1995), Bester and Guth (2000)). EOP hypothesizes that humans repeatedly play the exact same game over evolutionary timescales, and that there are genes on the human chromosome controlling behavioral preferences for that particular game. It then further hypothesizes that those genes controlling preferences undergo selective pressure to maximize reproductive fitness. Typically the behavioral preferences that maximize reproductive fitness are not the same as reproductive fitness. (This is due to how other humans change their behavior in response to changes in a person's behavioral preferences, just like in persona games.) Accordingly, if we could establish that the exogenous preferences in laboratory experiments are affine transformations of reproductive fitness, this EOP hypothesis might form a foundation for explaining interdependent preferences.

Loosely speaking, the long timescale in persona games is an individual's lifetime, whereas it is multiple generations in the EOP framework. As a result, the (technically challenging) Evolutionary Stable Strategy equilibrium concept of the EOP framework is replaced by the (technically straight-forward) Nash equilibrium concept in the persona game framework. An extensive discussion of the relation between the persona game framework and the EOP framework can be found in (Wolpert and Jamison (2011), Wolpert et al. (2008)). In particular, it is shown in detail there how persona games circumvent some of the technical difficulties with

the EOP framework, how they circumvent some of the difficulties in relating the EOP framework to real-world natural selection, and how persona games are far simpler to analyze.

Finally, (Wolpert and Jamison (2011)) is complementary to the current paper, providing material absent from this paper. That paper motivates the persona framework in terms of two-timescale games and EOP. It then contains a detailed discussion of the physical processes by which personas are signaled. It also applies the persona framework to make precise predictions for the Traveler's Dilemma, showing that they agree with stylized facts from the literature. In addition that work shows how the persona framework provides an explanation for the particular value of the logit QRE exponent found experimentally. Finally, that paper also contains a substantial analysis of sticky personas. and of personality games.

6 Final Comments

Both humans and some animals sometimes exhibit what appears to be non-rational 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 non-rational, and leave it at that. Under this response, essentially the best we can do is catalog the various types of non-rationality that arise in experiments (loss aversion, framing effects, the endowment effect, sunken cost fallacy, confirmation bias, reflection points, other-regarding preferences, uncertainty aversion, etc.) 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 non-rationality 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 non-rationality of humans in experiments does not reflect "inadequacies" of the human subjects. Rather it reflects inadequacies in us 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.

An example of a body of work that adopts this point of view is evolutionary game theory. The idea in evolutionary game theory is that humans (or other an-

imals) 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. In fact, it has recently become popular to extend such reasoning to explain some types of apparently non-rational behavior like altruism by invoking kin selection or even group selection. In such explanations the behavior that evolution selects for actually harms the individual displaying it, but benefits a group of which the individual is a member.[10]

The persona framework is also based on the view that the apparent game and the real game differ. In the persona game framework, the apparent game is the exogenous 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 they believe 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 they are playing against such a program, they should always play perfectly rationally, in contrast to their behavior when playing against humans. This distinction between behavior when playing computers and playing humans agrees with much experimental data, e.g., concerning the Ultimatum Game (Camerer and Fehr (2006), Camerer (2003), Nowak, Page, and Sigmund (2000)).

What happens if the players in a persona game are unfamiliar with the meaning of each others' signals, say due to coming 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 that happens, since in the behavioral game they each do what would be optimal if their opponents were using that misconstrued persona — but their opponents aren't doing that. This frustration can be viewed as a rough model of what is colloquially called a "culture gap" (Chuah, Hoffman, Jones, and Williams (2007)).

While here we have only considered personas involving degrees of altruism, there is no reason not to expect other kinds of persona sets in the real world. For example, as mentioned briefly above, in (Wolpert et al. (2008), Wolpert and

[10]Despite its popularity, it is worth bearing in mind that direct determination of individual behavior via natural selection is an extreme hypothesis. It implicitly supposes that each of us has genes that completely fix certain aspects of our behavior, in that it is very difficult for us to overcome such genetically encoded behavior, despite the negative effects it presumably has on us individually. In short, as far as such behavior is concerned, we are slaves to our nature. Arguably such an hypothesis has almost eugenic connotations.

Jamison (2011)), the persona game framework was investigated for the case where personas are parametrized by the values of the logit exponents in a QRE. It was found that the resultant predictions for behavioral game play agree closely with experimental data for the Ultimatum Game. This suggests that the persona framework might explain why the logit exponents have the values found in experimental data, just as the persona framework might explain why interdependent preferences parameters have the values found in experimental data. More generally, 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 non-rational behavior. Rather they are complementary to 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, etc., and adopts persona distribution $P(a_i)$ for those games. By playing all those games i would grow accustomed to adopting $P(a_i)$. Accordingly, if i plays new instances of γ where signaling is prevented, they might at first continue to adopt distribution $P(a_i)$. However as they keep playing signal-free versions of γ, they might realize that $P(a_i)$ makes no sense. This would lead them to adopt the fully rational persona instead. If after doing that they were to play a version of γ where signaling was no longer prevented, they could be expected to return to $P(a_i)$ fairly quickly. This behavior agrees with experimental data (Cooper, DeJong, Forsythe, and Ross (1996), Dawes and Thaler (1988)).

ACKNOWLEDGMENTS: We would like to thank Raymond Chan, Nils Bertschinger, Nihat Ay, and Eckehard Olbrich for helpful discussion.

APPENDIX A: PRISONER'S DILEMMA

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

$$\begin{bmatrix} (\beta,\beta) & (0,\alpha) \\ (\alpha,0) & (\gamma,\gamma) \end{bmatrix} \tag{19}$$

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 them. Furthermore, we restrict attention (whenever possible) to NE of the behavioral game that are in pure strategies for both players.

First, it is the 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 their opponent's optimal response would be to choose the rational persona himself (leading to defection on his part in the behavioral game), since defection is dominant and hence a best response to any *fixed* behavioral-game strategy. But in this case they would prefer to also be rational, yielding γ rather than $\gamma/2$ as their exogenous 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 exogenous payoff matrix above, we can describe the behavioral game – if both players choose $\mathscr{C}$ – as follows:

$$\begin{bmatrix} (\beta,\beta) & ((1-s)\alpha, s\alpha) \\ (s\alpha,(1-s)\alpha) & (\gamma,\gamma) \end{bmatrix} \tag{20}$$

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:

$$\begin{bmatrix} (\beta,\beta) & (0, s\alpha) \\ (\alpha,(1-s)\alpha) & (\gamma,\gamma) \end{bmatrix} \tag{21}$$

In order for Row not to prefer to deviate in this way (and then play their 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 non-empty. 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 they 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

$$\begin{bmatrix} (\beta,\beta) & ((1-s_1)\alpha, s_2\alpha) \\ (s_1\alpha,(1-s_2)\alpha) & (\gamma,\gamma) \end{bmatrix} \tag{22}$$

Analogously to before, 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 Eq.'s ((2), (3)) 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)).[11]

To begin, let U_i be the expectation of a utility function $u_i : X \to \mathbb{R}$. Then the left-hand side in Eq. (2) can be written as an integral over $N + N|A|$ variables,

$$\int da' \int \prod_{j \in \mathcal{N}} \Big(\prod_{a'' \in A} dx_j^{a''} \Big) \, u_i(x_1^{a'}, x_2^{a'}, \ldots) \hat{P}(a'_i) \prod_{j \neq i} P(a'_j) \prod_{j \in \mathcal{N}} \Big(\prod_{a'' \in A} q(x_j^{a''}) \Big) \qquad (23)$$

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 $a \in A$, i.e., 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:

$$t_i(a', x_1^{A^1}, x_1^{A^2}, \ldots x_1^{A^{|A|}}, x_2^{A^1}, x_2^{A^2}, \ldots) \;\triangleq\; u_i(x_1^{a'}, x_2^{a'}, \ldots). \qquad (24)$$

This allows us to rewrite the integral in Eq. (23) as

$$\int da' \int \prod_{j \in \mathcal{N}} \Big(\prod_{m=1}^{|A|} dx_j^{A^m} \Big) \, t_i(a', x_1^{A^1}, \ldots, x_2^{A^1}, \ldots) \hat{P}(a'_i) \prod_{j \neq i} P(a'_j) \prod_{j \in \mathcal{N}} \Big(\prod_{m=1}^{|A|} q(x_j^{A^m}) \Big) \qquad (25)$$

[11] 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.

This integral is the expected value of the function t_i over the joint moves $(a', x_1^{A^1}, \ldots, x_2^{A^1}, \ldots)$, evaluated under a product distribution over those moves, $\hat{P}(a'_i)P(a'_{-i}) \prod_{j \in \mathscr{N}} (\prod_{a'' \in A} q(x_j^{a''}))$. 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 Eq. (3) as

$$\left(\int da' \; P(a') \int \prod_{j \in \mathscr{N}} (\prod_{a'' \in A} dx_j^{a''}) \; u_i^a(x_i^a, x_{-i}^a) \hat{q}(x_i^a) \prod_{a'' \neq a} q(x_i^{a''}) \prod_{j \neq i} (\prod_{a'' \in A} q(x_j^{a''})) \right) + (\beta_{a_i})^{-1} \mathscr{S} \left[\hat{q}(X_i^a) \right] \tag{26}$$

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 alla', a'' such that $a'' \neq a$.

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

$$v_i^a(a', x_1^{A^1}, x_1^{A^2}, \ldots x_1^{A^{|A|}}, x_2^{A^1}, x_2^{A^2}, \ldots) \;\; \triangleq \;\; u_i^a(x_1^{a'}, x_2^{a'}, \ldots). \tag{27}$$

Having done this, the integral in Eq. (26) 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, $\hat{P}(a'_i)P(a'_{-i}) \prod_{j \in \mathscr{N}} (\prod_{a'' \in A} q(x_j^{a''}))$. 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 Eq. (26) to be a free utility for behavioral layer (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 Coroll. 1 and Prop. 1 of (Wolpert (2009)), we see that the objective game Γ has a type I NE, as claimed.

References

Abreu, D. and F. Gul (2000): "Bargaining and reputation," *Econometrica*, 85–117.

Andrade, E. and T. Ho (2009): "Gaming emotions in social interactions," *Journal of Consumer Research*, 36.

Bar-Gill, O. and C. Fershtman (2005): "Public policy with endogenous preferences," *Journal of Public Economic Theory*, 7, 841–857.

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*.

Bechara, A. and A. Damasio (2004): "The somatic marker hypothesis: A neural theory of economic decision," *Games and Economic Behavior*, 52, 336–372.

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.

Brinol, P., D. K., and S. K. (2010): "The role of embodied change in perceiving and processing facial expressions of others," *Behavioral and Brain Sciences*, 437–438.

Brinol, P. and R. Petty (2003): "Overt head movements and persuasion: a self-validation analysis," *Journal of Personality and Social Psychology*, 84, 1223–1239.

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

Camerer, C. 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., 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. DeJong, R. Forsythe, and T. 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. and R. Thaler (1988): "Anomalies: Cooperation," *Journal of Economic Perspectives*, 2, 187–197.

De Long, J., A. Schleifer, L. Summers, and R. Wadmann (1990): "Noise trader risk in financial markets," *Journal of Political Economy*, 98, 703–738.

Dekel, E., J. Ely, and Y. O. (2007): "Evolution of preferences," *Review of Economic Studies*, 74, 685–704.

Dennett, D. (2003): "The baldwin effect, a crane, not a skyhook," in *Evolution and learning: The Baldwin effect reconsidered*, MIT Press, 69–106.

Ekman, P. (2007): *Emotions Revealed*, Holt Paperbacks.

Frank, R. (1987): "If homo economicus could choose his own utility function, would he want one with a conscience?" *The American Economic Review*, 77, 593–604.

Frith, C. (2008): "Social cognition," *Philosophical Transactions of the Royal Society B*, doi:10,1098/rstb.2008.0005.

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.

Gino, F., M. Norton, and D. Ariely (2010): "The counterfeit self: the deceptive costs of faking it," *Psychological Science*, 21.

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.

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.

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 payements 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.

Mackay, D. (2003): *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press.

Myerson, R. 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.

O., C. and P. Jehiel (2002): "On the role of outside options in bargaining with obstinate parties," *Econometrica*, 1477–1517.

Raub, W. and T. Voss (1990): "Individual interests and moral institutions," in M. Hechter, K.-D. Opp, and R. Wippler, eds., *Social institutions, their emergence, maintenance and effects*, Walter de Gruyter Inc.

Renou, L. (2008): "Committment games," *Games and Economic Behavior*.

Rubinstein, A. (2004): "Instinctive and cognitive reasoning: a study of response times," Http://arielrubinstein.tau.ac.il/papers/Response.pdf.

Samuelson, L. E. (2001): *Journal of Economic Theory*, 97, special issue on "Evolution of Preferences".

Schelling, T. (1960): *The strategy of conflict*, Harvard university press.

Sobel, J. (2005): "Interdependent preferences and reciprocity," *Journal of Economic Literature*, 43.

Tamir, M. (2010): "The maturing field of emotion regulation," *Emotion Review*.

Tom, G., P. Pettersen, T. Lau, T. Burton, and J. Cook (1991): "The role of overt head movement in the formation of affect," *Basic and Applied Social Psychology*, 12, 281–289.

Tversky, A. (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.

Wolitzky, A. (2010): "Reputational bargaining under knowledge of rationality," .

Wolpert, D. and J. Jamison (2011): "Schelling formalized: Strategic choices of non-rational personas," in *Evolution and Rationality: Decisions, Cooperation and Strategic Behaviour*, Cambridge University Press.

Wolpert, D., J. Jamison, D. Newth, and H. M. (2008): "Schelling formalized: Strategic choices of non-rational personas," Papers.ssrn.com/1172602.

Wolpert, D. H. (2009): "Trembling hand perfection for mixed quantal response / nash equilibria," *International Journal of Game Theory*, in press.