It can be smart to be dumb

David H. Wolpert,[1*] Michael Harre[2]

[1]MS 269-1, NASA Ames Research Center,
Moffett Field, CA, 94035, USA
[2]The Centre for the Mind and The School of Information Technologies,
Sydney University, Australia

*To whom correspondence should be addressed; E-mail: dhw@email.arc.nasa.gov.

An important problem in game theory is why bounded rationality occurs, e.g., why altruism to non-kin occurs. Previous explanations involved computational limitations of the players, repeated play among the players etc. As an alternative, we introduce a pre-play step in which each player i chooses a "persona", i.e., a ficticious utility function she commits to use in the game. By adopting different personas, player i induces different moves by the other players in the game. Sometimes by adopting a "bounded rational" persona, player i induces moves by the other players that increase i's true utility function. In such cases, player i's acting in a "bounded rational" way can be optimal for i. This phenomenon can explain some experimental observations concerning the prisoner's dilemma, the ultimatum game, and the traveler's dilemma.

Summary: We show how bounded rational behavior in general, and altruism in particular, can be utility-maximizing in non-repeated, anonymous games.

One of the enduring problems of game theory (*1–3*) is explaining why human and animal players so often exhibit "bounded rationality", in that they seem not to adopt the strategy that optimizes their objective (*4–7*). As a particular example, outside of the biology-specific phenomenon of kin selection, the wide-spread phenomena of altruism and cooperation often appears to be bounded rational.

Many models have been offered to explain bounded rationality in general and non-kin-selection altruism in particular (*1,2,4,5,8,9,9–16*). Some of these rely on *ad hoc* assumptions of bounded computational capabilities of the players. Other involve repeated interactions among the players, where the players have some ability to recognize their opponents from one interaction to the next, at least implicitly. The folk theorems of game theory can be viewed as the earliest examples of these repeated-interaction explanations. More recent examples are models that involve reputation effects, punishment, "loners", negotiating processes, etc.

Here we present a simple and broadly-applicable framework that provides an alternative explanation for bounded rationality. This framework is based on the observation that social organisms often have definite "personas" that they adopt for their interactions with one another. For example, someone might "act dumb" in some social situations but not others. More generally, we "act like a different person" when we interact with (i.e., engage in a game with) our boss, with our spouse, with a child, etc. Often we even consciously decide "how to act" with those people before the interaction.

Say that all players are free to choose such a persona before the start of play. Then for many games, adopting a persona that is bounded rational (e.g., acting dumb) actually results in larger expected utility when the game is played than does adopting a fully rational persona. This is true even for anonymous, non-repeated games, and even when there is no explicit pre-play signaling of personas among the players. In this sense, "bounded rationality" can actually be

utility-maximizing.

We show below how this framework can explain a lot of behavioral game theory experimental data, focusing on the famous Traveler's Dilemma (TD) and Ultimatum Game (UG). We also explore the possible implications of our framework for how to optimally design distributed systems of adaptive agents. Next we show that our framework contains phase transitions, in which infinitesimal changes in a player's choice of persona can completely change the equilibrium of the game. We end with a discussion of the possible relation between our framework and phenomena like "social intelligence" and "culture gaps".

Before presenting our framework, we first review game theory using the example of the famous Prisoner's Dilemma (PD) (*10, 17, 18*). Say we have two players, Row and Col, each of whom can choose one of two moves (also know as "pure strategies"). Write those sets of allowed moves as $\{\mathbf{T}op, \mathbf{D}own\}$ for Row, and $\{\mathbf{L}eft, \mathbf{R}ight\}$ for Col. Both players have a "utility function", which maps any joint move by both players into a real number. For example, in the PD the utility function pairs (u_{Row}, u_{Col}) for the four possible joint moves can be written as the matrix

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

So for example, if Row plays **T** while Col plays **L**, then Row's utility function equals 6 and Col's equals 0.

To play a game each player $i \in \{\text{Row, Col}\}$ independently chooses a "mixed strategy", i.e., a probability distribution $P_i(x_i)$ over her pair of allowed moves. So the expected utility for player i is $\mathbb{E}_P(u_i) = \sum_{x_i, x_{-i}} P_i(x_i) P_{-i}(x_{-i}) u(x_i, x_{-i})$, where $P_{-i}(x_{-i})$ is the mixed strategy of i's opponent. We say that a pair of mixed strategies (P_{Row}, P_{Col}) is a Nash Equilibrium (NE) of the game if for all players i, $\mathbb{E}_P(u_i)$ cannot increase if P_i changes while P_{-i} stays the same. Intuitively, at a NE, neither player could benefit by changing her mixed strategy, in light of her opponent's mixed

strategy. If either player violates this condition, she is said to exhibit "bounded rationality".

As an example, in the PD, there is a unique NE, where Row plays **T** with probability 1.0 and Col plays **R** with probability 1.0. (Given the mixed strategy of Row, Col's expected utility would decrease if she played **L** with non-zero probability, and given the mixed strategy of Col, Row's expected utility would decrease if she played **D** with non-zero probability.) Note though that at the (non-NE) joint move $(\mathbf{D}, \mathbf{L})$, both players have higher expected utility than at the NE. So if they could only both be induced to cooperate with one another and choose that move — and in doing so both be bounded rational— both of the players would benefit. In light of this, often the move **T** by Row and the move **R** by Col are referred to as the "Defect" move, while their other moves are referred to as "Cooperate".

Now say that rather than being rational in the PD, Col were perfectly irrational. This means that she commits to choosing uniformly randomly between the two columns, with no evident concern for the resultant value of her utility function, and in particular with no concern for what strategy Row adopts. Given such irrationality of Col, Row would have expected utility of 5 for playing **T**, and of 2.5 for playing **D**. So if Row were rational, given that Col is irrational, Row would still play **T** with probability 1.0. Given that Col plays both columns with equal probability, this in turn would mean that $\mathbb{E}(u_C) = 2$. Since if Col were rational her expected utility would be 4, being irrational rather than rational would hurt her in the PD.

Now however modify the PD to have the following utility functions (u_R, u_C):

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

Again the pure strategy $(\mathbf{T}, \mathbf{R})$ is the only NE. At that NE, $\mathbb{E}(u_C) = 1.0$. Now though if Col were irrational, Row would have expected utility of 3 for playing **T**, and of 4.5 for playing **D**. So if Row were rational, given that Col is irrational, Row would play **D** with probability 1.0. Given that Col plays both columns with equal probability, this in turn would mean that $\mathbb{E}(u_C) = 5.5$.

So by being irrational rather than rational, Col has improved her expected utility from 1.0 to 5.5. Loosely speaking, such irrationality by Col allows Row to play a move that Row otherwise wouldn't be able to play, and that ends up helping Col. This is true even though Col would increase her expected utility by acting rationally rather than irrationally if Row's *mixed strategy* were fixed (at **D**). The important point is that if Col were to act rationally rather than irrationally while Row's *rationality* were fixed (at full rationality), then Col would decrease her expected utility.

This phenomenon can be seen as a model of the common real-world scenario in which someone "acts dumber than they are" (by not being fully rational), and benefits by doing so. Similar phenomena can model other types of real-world bounded rationality. To illustrate this, further modify our game to have the following utility functions:

$$\begin{bmatrix} (0,0) & (6,1) \\ (5,5) & (0,6) \end{bmatrix} \tag{3}$$

As before, $(\mathbf{T}, \mathbf{R})$ is the sole Nash equilibrium of the game. Again as before, assume that Row is perfectly rational. Then Col doesn't benefit from being irrational, since doing that won't induce Row to flip from **T** to **D**. But now say that Col were *anti*-rational, in that she always chooses the strategy that *minimizes* $\mathbb{E}(u_c)$ (given the strategy of Row), rather than maximizes it. Since **R** is dominant for Col (i.e., gives higher utility regardless of Row's move), this anti-rationality means that Col always chooses move **L**. This in turn causes Row to flip from **T** to **D**, which benefits Col ($\mathbb{E}(u_{Col})$ goes from 1.0 to 5.0). Loosely speaking, in this game, given that Row is rational, Col would benefit from "being her own worst enemy", i.e., from preferentially choosing whatever move is worst for her given Row's move.

It is important to realize that the potential benefit of being anti-rational is not the same as the benefit that non-credible threats can provide in certain extensive form games (*1, 2*). In essence, a player making a non-credible threat says to her opponent, "If you do α, I'll do something

that will hurt me — but will also hurt you. So you must not do α, and I will exploit that". In contrast, a player with $\beta = -\infty$ says "*No matter what you do*, I'll do something that will hurt me", and in certain games, derives benefit from that. In fact, such a benefit can arise even if we restrict attention to "subgame perfect NE", which preclude non-credible threats (see supplemental online materials).

We can model all of these phenomena with a sequence of two games. In the first, "rationality" game, all players decide how rational to be. That joint rationality choice specifies the subsequent "realized" game that the players then play with one another. It is the NE of that subsequent realized game that determines the ultimate joint strategy $P(x)$ of the players. Therefore it is the NE of the realized game that ultimately provides the utility to the players of the rationality game.

As an example, indicate the choice of full rationality as $\beta = \infty$ and anti-rationality as $\beta = -\infty$. Let $\gamma \equiv (X_1, X_2, u_1, u_2)$ be the original, "concrete" game, where X_i is player i's strategy space (e.g., $X_{Row} = \{\mathbf{T}, \mathbf{D}\}$), and u_i is her utility function. The rationality choices made by the two players in the rationality game determine how to modify the concrete game to construct the realized game: the realized game associated with a pair of rationality choices $\{\beta_{Row}, \beta_{Col}\}$ is $\gamma' \equiv (X_1, X_2, u'_1, u'_2)$ where for both players i, $u'_i = u_i$ if $\beta_i = \infty$, and $u'_i = -u_i$ if $\beta_i = -\infty$. Any such realized game γ' has a NE joint strategy $P_{\gamma'}(x)$. (Note that for any i, in general a change to u'_i changes $P_{\gamma'}$.) In turn, that strategy determines $\mathbb{E}_{P_{\gamma'}}(u_i)$ for every player i. This is the expected utility of each player i of the rationality game. So the goal of each player i in the rationality game is to choose a rationality β_i such that, given the rationality choice of her opponent, the NE of the associated realized game γ' maximizes i's expected utility. (See the supplemental online materials for a more detailed formal definition of rationality games.)

To illustrate this, let the concrete γ be the game indicated in Table 3. If $\beta_R = \beta_C = \infty$, then $\gamma' = \gamma$. So the NE of γ' in this case is $(\mathbf{T}, \mathbf{R})$, the NE of γ discussed above. For this joint

rationality, the utilities to the players is $(6, 1)$. We can do an analogous calculation of utilities for the remaining three possible joint rationalities. For this concrete game, doing this generates the following utility bimatrix to the players of the rationality game for their four possible joint rationality choices:

	Col rationality	
	$-\infty$	$+\infty$
Row rationality		
$-\infty$	$(0,\ 0)$	$(0,\ 6)$
$+\infty$	$(5,\ 5)$	$(6,\ 1)$

(4)

(Note that this is just a rearrangement of the entries in the concrete game γ.)

There is a dominant NE of this rationality game at the pure strategy $(+\infty, -\infty)$. Accordingly, if two humans play the rationality game and resultant realized game rather than the single concrete game, then Col will elect to be anti-rational, and Row will elect to be rational; neither would benefit from choosing a different rationality, given the choice of her opponent.

Alternatively, imagine that behavior of two species' playing this game with each other has been determined via repeated plays of the game conducted under natural selection pressures. If those pressures acted on genes whose alleles code for the concrete game strategies, then typically the frequency of those alleles will evolve to the NE of the concrete game (*3, 15*). However if natural selection instead acts on genes whose alleles code for one of the two rationalities $\{-\infty, +\infty\}$, then in general the frequency of those alleles will evolve to the NE of the rationality game. This would manifest itself as an equilibrium strategy over the realized game in which members of the Col species always choose **L**, and members of the Row species always plays **D**; an outcome that differs from the NE of the concrete game.

It is important to note that when a player commits to some rationality, she fixes neither her strategy in the realized game nor the strategies of her opponents in that game. These strategies are only determined by the realized game players after all of the rationality game players have

made their choices.

Now modify our two-player concrete game by allowing each player four possible moves rather than two, and have the utility functions (u_R, u_C) be:

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

A NE of this game is the joint pure strategy where Row plays her bottom-most move, and Col plays her left-most move. An anti-rational equilibrium, where both players try to minimize their utility functions, occurs if Row plays the top-most row and Col plays the right-most column. The remaining two possible joint rationalities of the players correspond to the remaining two entries on the skew-diagonal of the matrix. This results in the following rationality game:

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

(6)

The joint rationality $(-\infty, -\infty)$ of this game is Pareto superior to $(+\infty, +\infty)$, i.e., if *both* players play anti-rationally rather than rationally, then *both* players benefit. Moreover, $(-\infty, -\infty)$ is a (dominant) NE of the rationality game. At that joint rationality, neither player would benefit from changing to rational behavior, given that her opponent were anti-rational. Note in particular that $(+\infty, +\infty)$ is not a NE of the rationality game. So if the players are sophisticated enough to play the rationality game with each other rather than the concrete game, they will both act anti-rationally, and will thereby both benefit.

There are also concrete games where the NE of the associated rationality game is $(-\infty, -\infty)$ but this is not optimal for either player. An example is the following concrete game:

$$\begin{bmatrix} (3,3) & (6,2) & (-1,0) & (4,1) \\ (2,-1) & (5,0) & (-2,-2) & (3,-3) \\ (1,8) & (8,7) & (0,5) & (2,6) \\ (0,3) & (7,4) & (-3,2) & (1,1) \end{bmatrix} \tag{7}$$

The four possible joint rationalities have equilibria lying on the main diagonal of this matrix, so the associated rationality game is:

	Col rationality	
	$-\infty$	$+\infty$
Row rationality		
$-\infty$	(1, 1)	(5, 0)
$+\infty$	(0, 5)	(3, 3)

(8)

This pair of utility matrices is just the PD introduced above (up to irrelevant rescalings, etc.) with "defect-defect" identified as $(-\infty, -\infty)$, and "cooperate-cooperate" as $(+\infty, +\infty)$. So for that concrete game, the players would *not* benefit by being sophisticated enough to play the rationality game. Rather they would be better off simply playing the (NE of the) concrete game.

Many distributed engineered systems can be viewed as games among the distributed subsystems where, loosely speaking, the engineer has the ability to set some aspects of the utility functions of the players and/or of how rational the players are. Examples involving purely artificial players include distributed adaptive control, distributed reinforcement learning (e.g., such systems involving multiple autonomous adaptive rovers on Mars or multiple adaptive telecommunications routers), and more generally multi-agent systems involving adaptive agents (*19–22*). In other instances of such engineered systems some of the players are human beings. Examples here include air-traffic management (*23*), multi-disciplinary optimization (*24, 25*), and in a certain sense, much of mechanism design, and in particular design of auctions (*1, 2, 26*).

The implications of the rationality games based on the concrete game in Table 2 suggests that the performance of some of these engineered systems could be improved if the players

were impeded from playing rationally (e.g., by corrupting their sensor input). The rationality game associated with the concrete games in Table 3 and 5 suggests that some of the players might even improve their performance if they were induced to always act against their own best interests (e.g., by appropriate transformation of their reward signals from their environment).

The focus of the discussion so far has been on pure strategy equilibria of rationality games. At such equilibria, though a player may not be playing rationally in the realized game, her rationality has a single, definite value. In contrast, when the equilibrium of the rationality game is mixed, some players chooses how rational to be in the realized game by randomly sampling their mixed strategies. Intuitively, such players benefit by being "capricious", or "flighty", and keeping their opponents "on their toes" as to whether they will be rational or not.

Such a mixed strategy equilibrium of the rationality game arises when the concrete game is the famous Traveller's Dilemma (TD) (*27–32*). The TD models a situation where two travelers fly on the same airline with an identical antique in their baggage, and the airline accidentally destroys both antiques. The airline asks them separately how much the antique was worth, allowing them the answers $\{2, 3, \ldots, 101\}$. To try to induce honesty in their claims, the airline tells the travelers that it will compensate both of them with the lower of their two claims, with a bonus of R for the maker of the lower of the two claims, and a penalty of R for the maker of the higher of the two claims.

To formalize the TD, let $\Theta(z)$ be the Heaviside step function, $\Theta(z) = \{0, 1/2, 1\}$ for $z < 0, z = 0$ and $z > 0$, respectively. Then for both players i, the utility function in the TD concrete game is $u_i(x_i, x_{-i}) = (x_i + R)\Theta(x_{-i} - x_i) + (x_{-i} - R)\Theta(x_i - x_{-i})$ where $R = 2$ is the reward/penalty (for making a low/high claim), x_i is the monetary claim made by player i, and x_{-i} is the monetary claim made by the other player.

The NE of this game is $(2, 2)$, since whatever i's opponent claims, it will benefit i to undercut that claim by 1. However in experiments (not to mention common sense), this NE never arises.

Even when game theoreticians play the TD with one another for real stakes, they tend to make claims that are not much lower than 101, and almost never make claims of 2.

For simplicity, we consider a rationality game based on the TD concrete game where both players can only choose either to be rational or irrational. When both players are fully rational, the expected utility to both is 2, i.e., $\mathbb{E}(u_i \mid \beta_1 = \infty, \beta_2 = \infty) = 2$ for both players i. Now say that player i is rational while the other player is irrational. This results in the expected utility $\mathbb{E}(u_i \mid x_i, \beta_{-i} = 0) = \frac{1}{100}\left(\left[\sum_{y=2}^{x_i-1}(y-2)\right] + x_i + \left[\sum_{y=x_i+1}^{101}(x_i+2)\right]\right)$ for all of i's possible concrete game moves x_i. The (integer) maximum of this is at $x_i \in \{97, 98\}$. The associated expected utility is $\mathbb{E}(u_i \mid \beta_i = \infty, \beta_{-i} = 0) \simeq 49.6$ (see supplemental online materials). Continuing in this way gives rationality game utility functions with the following (rounded) values:

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

(9)

This rationality game has two pure strategy NE, $(\beta_1, \beta_2) = (0, \infty)$ and $(\beta_1, \beta_2) = (\infty, 0)$. The associated distribution $P(x_1)$ for the first of these rationality NE is uniform. The associated $P(x_2)$ instead has half its mass on $x_2 = 97$, and half on $x_2 = 98$. The two distributions for the other pure strategy rationality NE are identical, just with $P(x_1)$ and $P(x_2)$ flipped. (As an aside, note that if one of the players is irrational and the other rational, it is better to be the *irrational* one of the two players rather than the rational one.)

There is also a symmetric mixed strategy NE of the rationality game, at which both rationality players choose $\beta = 0$ with probability .78. The associated marginal distributions $P(x_i)$ are identical for both i's: $P(x_i = 2) \simeq 5.8\%$, $P(x_i = 97) = P(x_i = 98) \simeq 9.5\%$, and $P(x_i) \simeq 0.8\%$ for all other values of x_i. (Note that because $P(\beta_1, \beta_2)$ is not a delta function, $P(x_1, x_2) \neq P(x_1)P(x_2)$.) Uniformly averaging over these three NE gives a $P(x)$ that is highly

biased to large values of x, in agreement with experimental data.

We can do the same analysis for other values of R besides 2. When R grows, the mixed strategy equilibrium of the rationality game places more weight on the persona ∞. This makes $P(x)$ become more weighted towards low values. In fact, when R gets larger than ~ 38.2, the two pure strategy NE of the rationality game disappear, and the mixed strategy NE reduces to the pure strategy where both players are fully rational. So for such values of R, the players are fully rational. These results agree with experimental data (*28*) on what happens as R changes.

In some real-world scenarios, the rationality that every player adopts is made known to all the other players before play of the realized game. In other situations however, each player in the realized game only knows her own rationality, together with the fact that her opponents are humans. When the rationality game has a mixed strategy NE (as does the TD), this means that no player of the realized game *explicitly* knows the rationalities of any player other than herself.

The framework outlined above has to be extended to deal with such scenarios. Formally, this extension is similar to "Bayesian games" (*1, 2*). The idea is that each player exploits common knowledge (in the game theory sense, of knowing the possible moves of one another, etc.) to infer the rationality mixed strategies of the other players. This extension is presented in the supplemental online materials along with a related one that addresses the possibility of multiple rationality game NE, as occurs for examples in the TD concrete game.

In the scenarios considered so far the set of "personas" that a player can commit to before start of the realized game is some subset of the three rationalities $\{-\infty, 0, +\infty\}$. There is little reason to believe that real human beings are so severely limited in the set of personas they can choose among. One slightly richer space of personas arises by extending the concept of "rationality" to all real numbers: if player i chooses the rationality $\beta_i \in \mathbb{R}$, then she commits to play the mixed strategy

$$P_i(x_i) \quad \propto \quad \exp[\beta_i \mathbb{E}_{P_{-i}}(u_i \mid x_i)] \;\; \forall x_i \tag{10}$$

in the realized game. As an example, the rationality $\beta_i = 0$ means that player i is irrational in the realized game and the limit $\beta_i \to +\infty$ means she is fully rational. (More precisely, one of the NE of the concrete game arises in the realized game in the limit of $\beta_i \to +\infty$ for all players i (*33*).) Similarly $\beta_i \to -\infty$ corresponds to anti-rationality. The set of values β_i allowed to player i can be any (potentially infinite) subset of $\mathbb{R}$, and such sets may vary among the players.[1]

Given any vector $\vec{\beta}$ of such real-values for all of N players in a game, there is always at least one joint mixed strategy $P(x)$ that satisfies Eq. 10 simultaneously for all players i. Such a P is known as a (logit) **Quantal Response Equilibrium** (QRE) (*32, 33, 35, 36*). So each $\vec{\beta} \in \mathbb{R}^N$ specifies a QRE $P(x)$, which in turn specifies expected utilities for all of the players. We can write those expected utilities as $\mathbb{E}(u_i \mid \vec{\beta})$, where i ranges over the players. In this way the QRE concept gives us N mappings from $\mathbb{R}^N$ into $\mathbb{R}$, $f_i : \vec{\beta} \to \mathbb{E}(u_i \mid \vec{\beta})$. Just as with the rationality games discussed above, we can view these N functions f_i as the utilities for N rationality game players, each of whom sets a separate component of the vector $\vec{\beta}$. The rationality games discussed above are just the special case of such a generalized rationality game where each β_i is restricted to some associated subset of $\{-\infty, 0, \infty\}$.

As an example, consider again the two-player game with utility functions given in Table 2. If the set of allowed $\vec{\beta}$ is all of $\mathbb{R}$ for both players, one can show that there is a unique pure strategy equilibrium of the rationality game at the pair $\vec{\beta} = (\ln(5/2) - \epsilon, \infty)$ where ϵ is arbitrarily small (see supplemental online materials). At that $\vec{\beta}$, neither player i can improve her associated expected utility by changing β_i.[2]

[1]See (*34*) for an investigation of the relation between Eq. 10 and a model-independent measure of the "degree of rationality" of a player who adopts a particular mixed strategy when the other players adopt a specified particular (potentially non-equilibrium) mixed strategy. Also see those references for a discussion of the relationship between Eq. 10 and the canonical ensemble of statistical physics. Intuitively speaking, the parameter β_i in Eq. 10 can be viewed as the reciprocal of a "temperature". In this, it is formally true that a rational person is "cold"; they have low temperature. Similarly an irrational person is "hot". In optimization usually such a temperature is viewed as determining a player's exploration / exploitation tradeoff as she searches for what strategy to adopt. (For example this is the case in simulated annealing.) However games like those in Table 3 and 5 indicate temperature plays other roles in optimization as well.

[2]As an aside, such rationality games can have phase transitions, even for concrete games having two players

As another example of this richer set of personas, we now consider the famous Ultimatum Game (UG) (*4, 6, 37–39*). The UG is a two-player game where the first player, the Buyer (B) selects some $x \in [0, 1]$ and offers it to the Seller (S). S can either accept or refuse B's offer. If she accepts, S gets a utility of x, and B gets a utility of $1 - x$. If S refuses the offer, both players get a utility of 0.

The NE for this game is for S to accept whatever offer B makes, and therefore for B to make the minimal possible offer. In experiments though, often S refuses an offer if it is too low, in essence saying "I'd prefer nothing to such an insultingly low offer." In keeping with this, typically B makes an offer substantially higher than the minimum.

For simplicity, we will consider a rationality game based on the UG concrete game where B is always perfectly rational, but S can choose essentially any positive rationality value β. Intuitively, if S is sufficiently irrational (i.e., β is low enough), then for her probability of accepting some particular offer x to be significantly greater than $1/2$, that offer must be substantially greater than 0. So if B knows that S has that "sufficient irrationality", B has to make an offer "substantially greater than 0" to have a probability "significantly greater than $1/2$" of getting any non-zero utility at all. Since S prefers offers that are as large as possible, this means that it is in S's interests to be "sufficiently irrational", and therefore for B to make an offer "substantially greater than 0".

Carrying through the calculation more formally, we find that at the equilibrium of the rationality game S adopts the pure strategy $\beta^* = 7.60$. Any value of β gives a unique associated offer value that is optimal for B, which we write as $x(\beta)$; $x(\beta^*) = 0.21$. The associated probability that S accepts the offer is 0.83. While it is hard to directly compare these results to experimental data (due to idiosyncrasies in the protocols in the experiments in the literature), they are broadly

with two moves each. More precisely, the log of the normalization constant for the QRE distribution (i.e., the log of the QRE analog of a "partition function") can be discontinuous as a function of $\vec{\beta}$ (see supplemental online materials). While in condensed matter physics phase transitions only occur for infinite-particle systems, in game theory they can arise even for only two players.

consistent with that data.

One explanation that has been offered for experimental results concerning the UG is that B has a 'fairness bias' built into her utility function. Other explanations involve *ad hoc* cognitive models that restrict the capabilities of the players. In contrast, the explanation offered by rationality games does not posit modifications to the utility functions of the players, nor limits their cognitive capabilities. In addition, the focus shifts from B to S, saying, in essence, that S would benefit by being spiteful — and therefore will be, which fact B must account for.

Note though that in the rationality game considered above, all of S's allowed personas only concern her own expected utility (via her rationalities). Introspection suggests that in the real world, S sometimes adopts personas that also concern B's expected utility. (This may be why in experiments S sometimes rejects offers as "too low", in conflict with the prediction of the rationality game considered above.) For example, a "fairness bias" can be modeled with such a persona.

This raises the question of whether there are games where a player can benefit by adopting a persona which involves the expected utilities of that player's opponents, just as there are games where she can benefit by adopting a finite rationality persona. Such personas involving the expected utilities of one's opponents might implement many types of "other-regarding preference", including altruism as well as fairness 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.

To investigate this, let $\{u_j : j = 1, \ldots N\}$ be the utility functions of the original N-player concrete game. Then rather than have the possible personas of player i be parameterized by a set of rationality real numbers, $\{\beta_i\}$, have them be parameterized by a set of distributions $\{\rho_i\}$ each of the form $(\rho_i^1, \rho_i^2, \ldots, \rho_i^N)$. By adopting persona ρ_i, player i commits to playing the realized game with a utility function $\sum_j \rho_i^j u_j$ rather than u_i. So pure selfishness for player i is

$\rho_i^j = \delta(i, j)$. Altruism is a ρ_i^j that places probability mass on more than one j. (Fairness is a slightly more elaborate persona than these linear combinations of utilities, e.g., the commitment to play the realized game with a utility function $[(N-1)u_i - \sum_{j \neq i} u_j]^2$.) To distinguish them from persona games in general, we refer to persona games where some personas involve the (concrete game) utilities of multiple playes as "other-regarding games".

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

$$\begin{bmatrix} (2,0) & (1,1) \\ (3,2) & (0,3) \end{bmatrix} \tag{11}$$

There is one pure strategy NE of this game, at $(\mathbf{T}, \mathbf{R})$. Say that both players in the associated other-regarding game only have 2 possible pure strategies, $\rho_i^j \triangleq \delta_{i,j}$ and $\rho_i^j \triangleq 1 - \delta_{i,j}$, which we refer to as s**E**lfish and s**A**intly, respectively. So for example, if Row chooses selfish while Col chooses saintly, then the equilibrium is $(\mathbf{D}, \mathbf{L})$. This gives both players a higher utility than if both were selfish. Continuing this way, we get the following pair of other-regarding player utility functions:

	Col ρ	
	E	**A**
Row ρ		
E	(1, 1)	(3, 2)
A	(0, 3)	(3, 2)

(12)

The pure strategy NE of this other-regarding game is $(\mathbf{E}, \mathbf{A})$, i.e., the optimal persona for Row is to be selfish, and for Col is to be saintly. Note that both players benefit by having Col be saintly. So for example Row would be willing to pay up to 2.0 to induce Col to be saintly. Perhaps more surprisingly, Col would be willing to pay up to 2.0 to be allowed to completely ignore her own utility function, and work purely in Row's interests.

In the case of the PD concrete game, other-regarding personas can lead the players in the realized game to cooperate. For example, say that each player can choose either to be selfish, or

to be "charitable", meaning that ρ is uniform, so that the player has equal concern for her own utility and her opponent's utility. Then for the PD concrete game in Table 1, the equilibrium of the associated other-regarding game is for both players to be charitable. That in turn leads them to cooperate in the realized game. They do this for purely self-centered reasons, in a game they play only once. Such effects might account for some of the experimental data showing a substantial probability for real-world humans to cooperate (*40*). (A detailed exploration of other-regarding games can be found in (*41*).)

Discussion. Persona games highlight the distinction between games against Nature and games with other objective-maximizing players. In particular, when the concrete game is a game against Nature, the rationality equilibrium is always $\beta = \infty$, regardless of the details of the concrete game. This is not the case when the concrete game involves other players.

In addition to having other players in the concrete game, another necessary condition for a player to adopt a persona other than perfect rationality is that she believes that the *other* players are aware that she can do that. The simple computer objective-maximizing algorithms currently used in game theory experiments do not have such awareness. Accordingly, if a human knows she is playing against such an algorithm, she should always play perfectly rationally, in contrast to her behavior when playing against humans. This distinction between behavior when playing computers and playing humans agrees with much experimental data, e.g., concerning the UG game.

Typically that calculating a persona equilibrium involves far more computational work than calculating the equilibria of the associated concrete game. (For every possible joint persona, one has to calculate the associated realized game equilibria, and only *then* can one calculate the persona game equilibria.) Hence, one would expect persona games only in members of a species with advanced cognitive capabilities, who have a lot of interactions with other animals

that can also play persona games. A member of such a species who plays persona games well has "high social intelligence".

For similar reasons, one would expect the persona set of any animal not to be too large. (A large set both increases the burden on the player with that set, and with the other players she plays against.) Computational considerations might also lead an animal to use a similar persona set across those games that she is likely to encounter in her life. Phenomenologically, such a persona is what is similar to what is colloquially called a "personality".

What happens if the players misconstrue the persona (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 realized game they each do what would be optimal if their opponents were using that misconstrued persona — but they aren't. This can be viewed as a rough model of what is colloquially called a "culture gap", or a "generation gap", as the case might be. (See supplemental online materials for futher speculation on anthropological and evolutionary implications of persona games.)

There are many open issues, both experimental and formal, associated with the current persona game framework (see the supplementary online material). Future extensions of the persona game framework will involve integrating it with previous work on bounded rationality involving computational restrictions, repeated games in which there are reputation effects, games on networks, punishment, etc. But even based only on the preliminary results recounted above, it appears that persona games may provide the formalization, recently called for by Basu (*27*), of "the idea of behavior generated by rationally rejecting rational behavior ... (which is necessary) to solve the paradoxes that plague game theory".

Supplemental Online material

In these supplemental online materials we present formal definitions of persona games, and then make the calculations mentioned in the text concerning the UG, TD, and phase transitions. As might be expected given the breadth of scenarios that persona games can model, their formal definition takes some care. We start by introducing our notation.

Notation. Define $\mathbb{N} \equiv \{1, 2, \ldots\}$, fix a positive integer N, and define $\mathcal{N}$ as the integers $\{1, \ldots, N\}$. We will occasionally use curly brackets to indicate a set of indexed elements where the index set is implicit, being all $\mathcal{N}$. For any set Z, $|Z|$ indicates the cardinality of Z.

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.

We will use the integral symbol with the measure implicit. So for example for finite X, "$\int dx_X$" 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.

The unit simplex of possible distributions over a space Z is written Δ_Z. The interior of a set A (with implicit topology) is written A^0, e.g., the interior of Δ_Z is Δ_Z^0. 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 $\prod_{j \in \mathcal{N}: j \neq i} q_j$. Finally, given any finite Z and $p \in \Delta_Z$, we write the Shannon entropy of p as $\mathscr{S}(p) \triangleq -\int_Z p(z) \ln[p(z)]$.

Objective games. Say we are given a set of N (pure) strategy finite spaces, $\{X_i\}$. Then we 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 objective function. Any pair of a set of finite strategy spaces and an associated set of one objective function for each stategy space is called a (strategic form) **objective game**. Often we leave the indices on the elements of an objective game implicit, for example referring to (X, U) rather than $(\{X_i\}, \{U_i\})$.

Objective games provide a clean way to model scenarios in which some of the players use a conventional (expected utility) best response, and the others instead use a quantal response. Note in this regard that one cannot model an irrational player who always plays a uniform mixed strategy as an expected utility best-responder for some appropriate utility function. In particular, if a utility function is independent of x, then *any* distribution over x can arise at equilibrium, depending on the strategies of the other players; optimizing this utility function does not always pick out the uniform distribution associated with irrationality. On the other hand, a player with a free utility objective (i.e., a quantal response player) where $\beta = 0$ always plays a uniform distribution over x, exactly as desired.

Best responses in objective games, extensive form objective games, Nash equilibria of ob-

jective games, and (trembling hand) perfect equilibria of objective games are defined in the obvious way. We write the Nash equilibria of an objective game (X, U) as

$$\mathcal{E}(X, U) \quad \triangleq \quad \{q \in \Delta_{\mathcal{X}} : \forall i \in \mathcal{N}, \forall q'_i \in \Delta_{X_i}, U(q'_i, q_{-i}) \leq U(q_i, q_{-i})\}. \tag{13}$$

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

As an example, let each objective function U_i be an expected utility $\mathbb{E}_q(u_i)$. Then $\mathcal{E}(\{X_i\}, \{U_i\})$ is the set of Nash equilibria of the N-player noncooperative game with pure strategy spaces $\{X_i\}$ and utility functions $\{u_i\}$. We call such an objective game a "utility game". As another example, associate with each player i a non-zero real number β_i. Define each player i's objective function U_i as the **free utility**, $\mathbb{E}_q(u_i) + \beta_i^{-1}\mathcal{S}(q_i)$. Then $\mathcal{E}(X, U)$ is the set of (logit) QRE's of the N-player game with pure strategy spaces $\{X_i\}$ and logit exponents $\{\beta_i\}$ (*20, 33, 42, 43*).[3]

As a final example, an N-player **Bayesian objective game** is any triple (X, A, P^A) where $\{X_i\}$ is a set of N finite strategy spaces, each of the N separate A_i is a set of objective functions with domain Δ_X, and P^A is a product distribution over A, $P^A(a) \triangleq \prod_{i \in \mathcal{N}} P_i^A(a_i)$. For simplicity, in this paper we will always take any A_i to be finite.

An equilibrium of a Bayesian objective game is any set of conditional distributions $\{P_i(X_i \mid A_i) \in \Delta_{X_i|A_i} : i \in \mathcal{N}\}$ where for all players i and associated objectives a_i such that $P_i^A(a_i) \neq 0$, there is no alternative conditional distribution $P'_i(X_i \mid A_i)$ for which

$$\int da_{-i}\ P_{-i}^A(a_{-i}) a_i [P_{-i}(X_{-i} \mid a_{-i}) P'_i(X_i \mid a_i)] \quad > \quad \int da_{-i}\ P_{-i}^A(a_{-i}) a_i [P_{-i}(X_{-i} \mid a_{-i}) P_i(X_i \mid a_i)] \tag{14}$$

[3]There are at least two ways to incorporate refinements of Nash equilibria into the definition of objective games. First, many refinements of objective game equilibria can be expressed simply by modifying the objective functions. For example, if for every player i, $U_i(q) \triangleq \mathbb{E}_q(u_i)$, then we can impose many refinements by modifying $U_i(q)$ to have some value less than $\min_{x \in X}[u_i(x)]$ if q violates the refinement condition. As an alternative, one could define an "equilibrium concept" as a correspondence taking any pair (X, U) to a subset of Δ_X, and then have the specification of the equilibrium concept be part of the specification of an objective game. $\mathcal{E}$ is one possible choice for such a concept, but by choosing other ones, we could have the objective game equilibria correspond be any desired refinement of the Nash equilibria.

So long as all objective functions a_i obey certain benign properties, Kakutani's fixed point theorem can be applied to the associated best-response correspondences in the usual way to establish that there is always such an equilibrium. In particular, if for all players i each $a_i \in A_i$ is the expected value of some utility function over X, then Eq. 14 reduces to the definition of a conventional Bayesian game. So for such sets $\{A_i\}$, there is always an equilibrium of the associated Bayesian objective game.

Persona sample games. The foregoing provides the machinery for us to define and analyze several kinds of persona game. Here we present four such kinds of game. (As usual, which type of game to use to model a particular experiment depends on the precise details of that experiment.)

First, define a **persona world** as any triple

$$(\{X_i : i \in \mathcal{N}\}, \{U_i : i \in \mathcal{N}\}, \{A_i : i \in \mathcal{N}\}) \tag{15}$$

where $\{X_i\}$ is a set of N finite strategy spaces, $\{U_i\}$ is an associated set of N objective functions with domain Δ_X, and each A_i is a set of objective functions with domain Δ_X. 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. Note that any such persona is a mapping from $\Delta_\mathcal{X}$ into $\mathbb{R}$.[4]

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 $P^A(a) = \prod_{i \in \mathcal{N}} P_i^A(a_i)$. We define $\Delta_{\mathcal{A}_{-i}}$ similarly. We also define $\Delta_{\mathcal{X}|\mathcal{A}}$ to mean the members of $\Delta_{X|A}$ that are product distributions, i.e., that are of the form $P(x \mid a) = \prod_{i \in \mathcal{N}} P(x_i \mid a_i)$ for all $a \in A$, and make a similar definition for $\Delta_{\mathcal{X}_{-i}|\mathcal{A}_{-i}}$.

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

Given a persona world and a player $i \in \mathcal{N}$, an associated **summarizer** w_i is a real-valued measure of i's desire for the "realized game" specified by any possible joint persona $a \in A$. In this paper we restrict attention to summarizers w_i that can be expressed as $w_i(a) \triangleq \bar{w}_i(Q(a))$ where Q is a mapping from a into the associated set of equilibria of (X, a) for some pre-fixed equilibrium concept. An example of such a w_i is the uniform average of the values of U_i evaluated at the NE of the realized game (X, a). We refer to such a summarizer as the **equilibrium-averaging** summarizer. A variation, motivated by the entropic prior and its use in predictive game theory (*42*), is to weight the value of U_i at each NE q in this average by $e^{\alpha \mathcal{S}(q)}/Z$, where α is the entropic prior constant and Z is a normalization constant. More generally, $w_i(a)$ could also reflect concerns like the computational cost to i of evaluating those NE. As shorthand, we often abbreviate $(\{X_i : i \in \mathcal{N}\}, \{U_i : i \in \mathcal{N}\}, \{A_i : i \in \mathcal{N}\}, \{w_i : i \in \mathcal{N}\})$ as (X, U, A, w).

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

$$W_i(a) \quad = \quad \bar{w}_i[\mathcal{E}(X, a)]. \tag{16}$$

As an example, say that each $a_i \in A_i$ is the free utility for some associated parameter β_{a_i} and some fixed utility u_i. (So A_i is parameterized by a set of non-zero real numbers $\{\beta_{a_i} : a_i \in A_i\}$). Then the joint persona of the players can be viewed as an N-tuple of logit exponents, in the sense that the equilibrium for the realized game associated with a joint persona a is a QRE with logit exponents given by $\beta_a \equiv (\beta_{a_i} : i \in \mathcal{N})$. Say we also have each each $U_i(q)$ be the expected value under q of a utility function $u_i : X \to \mathbb{R}$. Also have $\bar{w}_i$ be the equilibrium-averaging summarizer. Then $W_i(a)$ is the uniform average of the expected utilities of player i over the

QRE's of the game for the vector of logit exponents β_a.

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

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

Typically a summary sample equilibrium P^A is invariant under any affine transformation of A_i for any $i \in \mathcal{N}$. To illustrate this, write such a transform as $a_i \rightarrow Ca_i + D \ \forall a_i \in A_i$, and consider again the example above where each $a_i \in A_i$ is a free utility. Applying our affine transform to such an A_i is equivalent to multiplying both u_i and $[\beta_{a_i}]^{-1}$ for each $a_i \in A_i$ by C, and then adding D to u_i. Doing this won't affect the value of $e^{\beta_{a_i}[\mathbb{E}_{q_{-i}}(u_i|x_i) - \mathbb{E}_{q_{-i}}(u_i|x_i')}$ for any x_i, x_i', q_{-i}. Accordingly such a transform won't affect the equilibrium q(s) associated with any joint set of free utilities, a, i.e., it won't affect $\mathcal{E}(X, a)$ for any joint persona a. Therefore it won't affect any function W_j, and so won't change the summary sample equilibrium P^A. (The other types of persona equilibria discussed below also typically obey such an invariance under an affine transform of any A_i.)

Just as persona games are a "meta" version of concrete games, so one could have a "meta"

version of a persona game. This would mean no longer assuming that each player is perfectly rational in the persona game, but allowing other personas for the play of that game as well. In particular, if the persona set of each player i is the same size as her move space X_i, then the persona game constitutes a "dual" version of the concrete game. One can investigate how iteratively taking the dual of a dual of such a game moves one through the space of all games.

Say we have an experimental scenario in which, intuitively speaking, a player i is able to adopt a "persona" that consists of an objective function a_i combined with the restriction that i can only choose a strategy from some set $X'_i \subset X_i$, no matter what personas are adopted by the other players. We can model such scenarios as summary persona sample games. To do this, we redefine the value of $a_i(q)$ for all $q \in \Delta_{\mathcal{X}}$ such that $q_i(x_i) \neq 0$ for some $x_i \notin X'_i$. The only thing required of this redefinition is that it ensures that $a_i(q) < a_i(q')$ for every $q' \in \Delta_{\mathcal{X}}$ that obeys $q'_i(x_i) = 0 \ \forall x_i \notin X'_i$. (This redefinition implicitly assumes that the original function a_i is bounded.) Such a redefinition means that persona a_i will place zero probability mass on x_i in the game $(X, (a_i, a_{-i}))$, no matter what distribution q_{-i} the other personas a_{-i} jointly choose.

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

In the second step, in the usual way the N players play a realized objective game with their objective functions set by a': (X, a'). That produces a set of possible equilibrium joint product distributions $\mathcal{E}(X, a')$. Finally, each player i receives "payoff" $\bar{w}_i(\mathcal{E}(X, a')$. Typically, this payoff

is actually an expectation value, over all $q \in \mathcal{E}(X, a')$, of the associated value $U_i(q)$. The goal of each player i in such a persona game is to choose $P_i^A(a_i)$ so that, given the choices $P_{-i}^A(a_{-i})$ of the other players, the expected value of the payoff she receives at the end of the second step is as large as possible.

Persona distribution games. In the version of the two-step process presented above, each player i commits both to adopt her persona choice in the second step, and to communicate that persona choice to all the other players just before the second step. It is presumed that these two commitments are common knowledge. (This common knowledge is in addition to the usual common knowledge of the objective functions and persona sets of all the players.)

However when there is only one equilibrium P^A and it is a pure strategy, these two commitments can be combined into a single, weaker commitment. For such a situation, say that each player i only commits to adopt that persona choice specified in the persona sample equilibrium P^A, *whatever that persona is*. As before, assume that all commitments are common knowledge. Also as before, assume that the objective functions and persona sets of all the players are common knowledge. This means that every player $j \neq i$ can calculate the equilibrium persona a_i of player i. Accordingly, this minimal additional common knowledge beyond the objective functions and persona sets — that i commits to adopt her equilibrium persona — allows the other players to deduce what persona player i will adopt in the second step. Importantly, they get this information of what persona player i will adopt without any explicit communication from player i of what persona she adopts.

The situation when the equilibrium P_i^A is not a pure strategy is more subtle. In such situations, our additional common knowledge would only tell the other players what distribution P_i^A commits to sampling, not what actual sample of that distribution is generated. This means that those other players would have to play a Bayesian best response to P_i^A. We refer to such

a scenario where the players all commit to play the Bayesian objective game with P^A set to an equilibrium of the summary persona sample game as a **summary persona distribution game**. We refer to the Bayesian objective game arising in a summary persona distribution game as a realized (Bayesian) game.

Note that there is no role for a summarizer in the realized Bayesian game that is played after the joint persona distribution is set in a summary persona distribution game. (The summarizer is only used to set persona distributions, with no role once those distributions are set.) Also, note that if the equilibrium P^A of a summary persona distribution equilibrium is a pure strategy, then the realized Bayesian game is just a conventional objective game, where players have fixed objective functions. That means that the entire two-step process of the summary persona distribution game is equivalent to the summary sample game. (So in particular, all the examples of persona games in the text involving pure strategy P^A's are both summary persona sample and summary persona distribution equilibria.) It is only when where there are mixed persona equilibria of a summary persona sample game that there may be a difference in outcomes between that game and the associated summary persona distribution game.

A summary persona distribution game can be viewed as a model of many different physical scenarios. One is where all N players think they are playing a summary persona sample game, and determine their joint distribution P^A accordingly. Each player i then samples P_i^A, to get an associated persona a_i. However at the next, persona revelation step, the players all balk, and refuse to communicate their personas. (Or alternatively, the experimenter prevents them from communicating their personas). Accordingly, all the players have to play the realized Bayesian game given by P^A.

As another example, say that before play each player i communicates to all other players a single bit: whether she will or will not adopt the summary persona sample equilibrium P_i^A. (Implicitly, to not adopt that equilibrium P_i^A means she instead adopts the full rationality per-

sona, U_i.) Say that the players all communicate that they will adopt that equilibrium. Finally, say that the realized game is an anonymous, one-shot game, without any inter-player communication of persona samples before that game. In this scenario, the players are forced to play the realized Bayesian game, using the persona sample equilibrium P^A. So we have a summary persona distribution game.

Note that the communication among the players of the bit that "they will adopt the summary persona sample equilibrium P^A" does not have to be direct communication. For example, say that human beings are pre-disposed to playing persona sample games, and all the players know that. Then those bits would be communicated simply by having every player know that the other players are all human beings. (Of all the models presented in this paper, this one where the players are all told simply that the other players are human beings is perhaps the tightest match to what actually happens in typical real-world experiments involving one-shot anonymous games.)

If human beings do have such a pre-disposition to play persona sample games that disposition may well have a genetic component. (See the section below on communication of personas and natural selection.) Such a pre-disposition may arise even without evolutionary pressures however. Many (if not most) real-world games are not anonymous. Rather they involve face-to-face interaction, in which at least some information concerning persona choices is conveyed. This means that human beings are accustomed to playing persona sample games, not to playing games in which they know nothing of the personas chosen by their opponents. So due to simple "cognitive inertia", one might expect human subjects to be pre-disposed to behave as though they were playing a persona sample game, even if they are actually involved in an anonymous one-shot game experiment.

Note though that if the players are in an experiment in which they repeatedly engage in an anonymous game, then such cognitive inertia might fade, and the players learn that they are

not actually involved in the kind of persona sample game to which they are accustomed. This might lead each human in the experiment to eventually play the concrete game directly, with no concern for personas and the like. This would mean that each player i assumes that the other players will be fully rational in the realized game without any regard for any persona that i might adopt. Such an assumption would lead each player i to be fully rational. (This in turn would validate the assumption of the other players in her rationality, resulting in a sort of equilibrium of assumptions.) Broadly speaking, such a transition to where all subjects play the concrete game directly without any concerns for personas is what is seen in experiments involving repeated anonymous games, where repeated anonymous play leads to more "rational" behavior.

As a technical aside, note that if there are multiple summary persona sample equilibria P^A (pure or mixed), then there must be some way for all the players in the associated realized Bayesian game to agree on one of those equilibrium P^A. Of course, this same requirement arises in conventional noncooperative games with multiple Nash equilibria, where all the players have to agree on which equilibrium joint strategy to adopt. In those conventional games, if no refinement is used to replace the Nash equilibrium concept, the requirement to agree on which NE to adopt is met through some sort of coordination mechanism involving common knowledge, purification, previous interactions, or the like. Without getting into any detailed modeling, we implicitly presume the same sort of mechanism is used with summary persona distribution games. (See (*42*) for a way to avoid all such issues.)

Extensive persona games. A formal shortcoming of summary persona games is the need to specify a summarizer for each persona player. One way to avoid this is to modify the "two-step" process of a summary persona sample game to be an extensive form game that involves two stages. Say we are given an N-player persona world (X, U, A) where all spaces are finite. Consider an associated set of $N + \sum_{i \in \mathcal{N}} |A_i|$ distributions $\{P(A_i) \in \Delta_{A_i}, q(X_i^{a_i}) \in \Delta_{X_i} : i \in \mathcal{N}, a_i \in$

$A_i\}$ where each space $X_i^{a_i}$ 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 $q(X_i^{a_i})$ as the mixed strategy of the realized player who corresponds to persona player i when that persona player adopts persona a_i. Now say that the following two conditions are met:

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

$$U_i\left[\int da_1 da_2 \ldots P'(a_i)P(a_{-i}) \prod_{j\in\mathcal{N}} q(X_j^{a_j})\right] \ > \ U_i\left[\int da_1 da_2 \ldots P(a_i)P(a_{-i}) \prod_{j\in\mathcal{N}} q(X_j^{a_j})\right] \quad (17)$$

and

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

$$a_i\left[q'(X_i^{a_i}) \prod_{j\neq i} q(X_j^{a_j})\right] \ > \ a_i\left[q(X_i^{a_i}) \prod_{j\neq i} q(X_j^{a_j})\right]. \quad (18)$$

Then we say that the distributions $\{P(A_i) \in \Delta_{A_i}, q(X_i^{a_i}) \in \Delta_{X_i} : i \in \mathcal{N}, a_i \in A_i\}$ form an **extensive persona sample equilibrium** for the persona world (X, U, A).

We now prove that any persona world (X, U, A) has an extensive persona distribution 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 Eq.'s 17, 18 as being NE conditions for an objective game involving $N + \sum_{i\in\mathcal{N}} |A_i|$ players. We will see that not only must such an objective game have an NE, it must have an honest NE (*44*).[5]

To begin, letting U_i be the expected value of a utility function u_i, recast the left-hand side in Eq. 17 as

$$\int da' \prod_{j\in\mathcal{N}} (\prod_{a'_j\in A_j} dx_j^{a''_j})\ u_i(x_1^{a'_1}, x_2^{a'_2}, \ldots)P'(a'_i)P(a'_{-i}) \prod_{j\in\mathcal{N}} (\prod_{a''_j\in A_j} q(x_j^{a''_j})) \quad (19)$$

[5]Typically when some objectives are expected utilities and some are free utilities the NE of the objective game are not trembling hand perfect. So some refinement other than perfection must be used to preclude non-credible threats in such scenarios. Honesty is such a refinement.

and similarly for the right-hand side. These integrals are averages over moves by $N + \sum_{i \in \mathcal{N}} |A_i|$ players. To simplify the associated manipulations, write the space of possible joint moves of all $N + \sum_{i \in \mathcal{N}} |A_i|$ players as $\{(a, x_1^{A_1}, x_2^{A_2}, \ldots) : a \in A, X_i^{A_i} \in X_i \forall i\}$. (So each a in any element of this space is an ordered list of N separate personas a_i, one for each $i \in \mathcal{N}$, and similarly each $x_i^{A_i}$ is an ordered list of $|A_i|$ separate elements from X_i, one for each possible persona choice by the i'th persona player.) Given this notation, replace the utility function u_i in Eq. 19 with

$$T_i(\tilde{a}, \tilde{x}_1^{A_1}, \tilde{x}_2^{A_2}, \ldots) \quad \triangleq \quad u_i(\tilde{x}_1^{\tilde{a}_1}, \tilde{x}_2^{\tilde{a}_2}, \ldots), \tag{20}$$

evaluated for the argument $(a', x_1^{A_1}, x_2^{A_2}, \ldots)$.

Having done this, the integral in Eq. 19 becomes the expected value of the function T_i over the joint moves $(a', x_1^{A_1}, x_2^{A_2}, \ldots)$, evaluated under a product distribution over those moves, $P'(a'_i)P(a'_{-i}) \prod_{j \in \mathcal{N}} (\prod_{a''_j \in A_j} q(x_j^{a''_j}))$. Accordingly, the entire expression in that equation is an expected utility for a player setting a mixed strategy $P(A_i)$ for an objective game Ω involving a total of $N + \sum_{i \in \mathcal{N}} |A_i|$ players who have joint move $(a', x_1^{A_1}, x_2^{A_2}, \ldots)$.

Similarly, if a_i is a free utility with logit exponent $\beta_i^{a_i}$ and utility function $u_i^{a_i}$, we can recast the left-hand side in Eq. 18 as

$$\int \prod_{j \in \mathcal{N}} (\prod_{a''_j \in A_j} dx_j^{a''_j}) \; u_i^{a_i}(x_i^{a_i}, x_{-i}^{a_{-i}}) q'(x_i^{a_i}) \prod_{a''_i \neq a_i} q(x_i^{a''_i}) \prod_{j \neq i} (\prod_{a''_j \in A_j} q(x_j^{a''_j})) \; + \; (\beta_{a_i})^{-1} \mathscr{S} \left[q'(X_i^{a_i}) \right] \tag{21}$$

and similarly for the second expression. (Note that many distributions in the integrand marginalize out.) Now replace the utility function $u_i^{a_i}$ in Eq. 21 with

$$V_i^{a_i}(\tilde{a}, \tilde{x}_1^{A_1}, \tilde{x}_2^{A_2}, \ldots) \quad \triangleq \quad u_i^{a_i}(\tilde{x}_1^{a_1}, \tilde{x}_2^{\tilde{a}_2}, \ldots), \tag{22}$$

evaluated for the argument $(a, x_1^{A_1}, x_2^{A_2}, \ldots)$. Having done this, the integral in Eq. 21 becomes an expectation value of the utility function $V_i^{a_i}$ over joint moves $(a', x_1^{A_1}, x_2^{A_2}, \ldots)$, evaluated

under the product distribution $q'(x_i^{a_i}) \prod_{a_i'' \neq a_i} q(x_i^{a_i''}) \prod_{j \neq i} (\prod_{a_j'' \in A_j} q(x_j^{a_j''}))$. Accordingly, the entire expression in that equation is a free utility for a player setting $q(X_i^{a_i})$ for the objective game Ω.

Similar conclusions hold if a_i is an expected utility. Combining and using Coroll. 1 of (*44*), we see that the objective game Ω always has an honest NE, as claimed.

Obvious variations of the type of extensive persona sample game defined here has other kinds of communication between the persona players and the realized game players, along the lines of conventional game theoretic extensive games. Other variations involve objective functions that depend on $P(A) \prod_{i \in \mathcal{N}} (\prod_{a_i \in A_i} q(X_i))$ in other ways. In particular, a persona player i could have U_i be a free utility where the entropy term depends on $P(A_i)$ rather than on $P(X_i) = \int da_i \; P(a_i) q(X_i^{a_i})$.

It is also possible to have a realized Bayesian game be played using the P^A that is an equilibrium of an extensive persona sample game. We refer to such persona games as **extensive persona distribution games**, in direct analogy with summary persona distribution games. In yet another type of persona distribution game, the P^A of the realized Bayesian game is not set by a counterfactual persona sample game, as in summary persona distribution games and extensive persona distribution games. Rather P^A is set so that for all players i, there is no $P_i'^A \neq P_i^A$ such that $U_i(q)$ is higher for the realized Bayesian game based on persona distribution $P_i'^A P_{-i}^A$ than it is for the realized Bayesian game based on persona distribution $P_i^A P_{-i}^A$. Intuitively speaking, such a persona game is a Bayesian game where all players get to set their prior over their types, and the equilibrium is where no player would want to change her prior, given the priors of the other players.

The difference between Persona games and non-credible threats. We now illustrate how the effects arising in persona sample games differ from those of non-credible threats in conventional extensive form games. Have the utility functions in the concrete game of Table 3 give

payoffs in a two-stage perfect information extensive form game where both players can make two moves and Row moves first. So the strategic representation bimatrix is:

$$\begin{array}{lcccc} & \text{player 2:} & [\mathbf{L},\mathbf{L}] & [\mathbf{L},\mathbf{R}] & [\mathbf{R},\mathbf{L}] & [\mathbf{R},\mathbf{R}] \\ \underline{\text{player 1}} & & & & & \\ \mathbf{T} & & (0,0) & (0,0) & (6,1) & (6,1) \\ \mathbf{D} & & (5,5) & (0,6) & (5,5) & (0,6) \end{array} \tag{23}$$

where the 2nd-moving player's strategy is written as [response to move **T** by first-moving player, response to move **D** by first-moving player].

There is a pure strategy expected utility equilibrium at (**D**, [**L**,**R**]). This one is not subgame perfect. (The second player is non-credibly threatening to make move **L** if the first player moves **T**.) There is a second utility equilibrium where the first player moves **T**, and the second player adopts any mixed distribution with support over the last two of her four pure strategies: [**R**, **L**], [**R**, **R**]. The component where the second player makes move [**R**, **R**] is subgame perfect. The other component is not. Summarizing, (**T**, [**R**, **R**]) is the sole subgame perfect utility equilibrium, and is a pure strategy equilibrium. At this equilibrium, the joint payoff is $(6, 1)$.

Now say that we have an associated summary rationality sample game where the first moving player's rationality is fixed to ∞, and the second-moving player can either adopt rationality $\infty, 0$ or $-\infty$. The consequences of the second-moving player adopting rationality ∞ were worked out above. If she instead adopts rationality 0, then the first-moving player will pick **T**, just like when the second-moving player has rationality ∞; the second-moving player gains nothing by reducing her rationality.

Finally, consider what happens if the second-moving player adopts rationality $-\infty$. This is equivalent to her being utility-minimizing rather than utility maximizing, while the first player is still utility-maximizing. Now (**D**, [**L**,**L**]) is the sole subgame perfect equilibrium of the realized game. The associated joint payoff is $(5, 5)$. So the second moving player has benefited in comparison to her utility at the subgame-perfect equilibrium of the concrete game (at which

both players are fully rational). So even if we restrict attention to subgame perfect equilibria of the realized game, the second-moving player benefits by being anti-rational. This illustrates how the potential benefit of being anti-rational is fundamentally different from the potential benefit of making non-credible threats.

Elementary properties of rationality games. Define a **simple** game as a 2-player objective game where both objectives are expected utilities and the game has a single equilibrium, which is a joint pure strategy. Note that both players have rationality ∞ at that joint strategy.

In the text we have considered some two-player summary rationality sample games with averaging summarizers where i) both players could choose rationality $+\infty$; ii) the concrete game was a simple game, and iii) the pure strategy $(+\infty, +\infty)$ is not an equilibrium of the rationality game. In such a game at least one rationality player benefits from choosing a finite (but non-negative) rationality given that the other player is purely rational. (An example is the game in Table 3.)

In addition, as illustrated in our analysis of the TD, there exist such summary rationality sample games based on a simple concrete game that have the following properties: 1) At least one player i has more than two moves in the concrete game; and 2) every rationality player i would benefit, in comparison to the case where both players have infinite rationality, if instead i has a particular finite rationality while her opponent is fully rational.

However there is no such summary rationality sample game that has property (2) and is based on a simple concrete game that has only two possible moves for each player:

Proof: Let the players be Row and Col with moves $\{\mathbf{T}, \mathbf{D}\}$ and $\{\mathbf{L}, \mathbf{R}\}$, as usual. Without loss of generality let the pure strategy objective equilibrium of the simple concrete game be $(\mathbf{T}, \mathbf{R})$, and

have the payoffs there be $(0, 0)$. Accordingly write the utility functions of the concrete game as:

$$\begin{bmatrix} (a,b) & (0,0) \\ (c,d) & (e,f) \end{bmatrix} \tag{24}$$

(under the usual convention that Row's utility is listed first). Since $(\mathbf{T}, \mathbf{R})$ is a NE of the concrete game, $0 \geq b$, and $0 \geq e$. On the other hand, since $(\mathbf{D}, \mathbf{L})$ is not a NE, either $f > d$ and/or $a > c$.

Hypothesize that for both players i, i benefits if she has a particular finite rationality while her opponent is fully rational, in comparison to the case where both i and her opponent are fully rational.

Now the only way that Col would benefit by changing β_{Col} from ∞ to some finite but non-negative value is if her doing so forces Row to be willing to choose **D** rather than **T** (since if Row still plays **T**, then that change to β_{Col} canot improve Col's expected utility). However since $e \leq 0$ and Row is rational, the only way that some finite β_{Col} can induce Row to be willing to play **D** is if it results in $q_{Col}(\mathbf{L}) > 0$ and in addition $c \geq a$. Using similar reasoning with the player roles flipped establishes that $d \geq f$. But we know from our reasoning above that $f > d$ and/or $a > c$. This contradiction establishes the proposition. **QED**.

There are many similar issues that have yet to be addressed. For example, consider an N-player persona sample game where all players other than i are irrational. For such a game, i can always benefit from being more and more rational, since she is effectively playing a game against Nature. So we know that $(0, 0, \ldots)$ cannot be a rationality equilibrium of such a game if any player's persona set includes positive rationalities. It is an open question what the minimal *non-negative* α is such that $(\alpha, \alpha, \ldots)$ is a rationality equilibrium of some game.

Another open formal question is to characterize the conditions under which a rationality game has a joint pure rationality strategy equilibrium in which one of the players is not maximally rational while the others are. In other words, what are necessary and sufficient conditions

so that if $\beta_j = \infty$ for all j other than i, the optimal β_i is finite? Similarly, what are necessary and sufficient conditions so that the NE of the persona game results in a realized game mixed strategy that is Pareto superior to the NE of the concrete game? When such conditions hold, no player is hurt by playing the persona game rather than the associated concrete game.

There are other open issues that address completely different aspects of persona games. For example, can it be the case that a persona equilibrium results in a worse value of the objective function of some player i if she has a persona set A than it does if she instead has a persona set $A' \subset A$? In other words, while in utility games removing pure strategies may help a player, can the analogous phenomenon occur in persona games? If such phenomena can occur, are they at least precluded if A' (and therefore A) contains the purely selfish, full rationality persona?

Empirically distinguishing conventional multistage games from persona sample games. Consider the concrete game in Table 11. As discussed in the text, if the players engage in a summary persona sample game where there are two personas, selfish and saintly, then Col would pick persona saintly with probability 1.0 and Row would pick selfish with probability 1.0. This translates into the realized game joint pure strategy (**L**, **D**).

Now since Row picked the selfish persona with probability 1.0, she is fully rational in the realized game, and so her realized game move is best-response to Col's realized game move. Accordingly, the outcome of the persona game is the same as it would have been if a conventional (non-persona) Stackelberg game were played with Col moving first and picking move **L**. In addition, in both the persona sample game and the Stackelberg game, Col has signaled her choice to Row before Row moves. Accordingly, there is little to distinguish this persona sample game from the associated Stackelberg game in an experimental setting.

Note though that for other concrete games there is not such an equivalence. (For example, there is not such an equivalence for the the concrete game in Table 5.) In addition there is not

this equivalence with persona distribution games, which have no explicit inter-player communication of personas. An important issue for design of persona game experiments is formally characterizing the types of concrete game and types of persona sets for which there is this kind of equivalence with a conventional Stackelberg (or more generally multistage) game.

Communication of personas and natural selection. In much of the previous work on games, be they repeated or one-shot games, it is assumed that before the start of play all players are provided the same common knowledge specifying the game. In particular, it is often assumed that each player knows the specification of the game in full, including the strategy spaces, utility functions and rationalities of all the players, together with the distributions governing all Nature players. For example, this is the case in the QRE and in Bayesian games. Typically such work models how the players use this common knowledge, rather than how they acquired it in the first place and possible errors in that acquisition. (For work that is concerning with the latter, see (*8, 45, 46*).)

In the interests of space, we adopt the same approach here for our consideration of persona sample games. We assume the exact same type of "common knowledge" as in that previous work is communicated to the players before start of the realized game, and assume the same sort of (often ill-specified) mechanism is involved in that communication. So we do not present a detailed model of how in a persona sample game the realized game players acquire the common knowledge of what realized game they are playing (i.e., of what the joint persona sample is). Rather we concentrate on modeling how that realized game is decided in the first place, and how the players behave once they have common knowledge of that game. (The difference from the conventional game theory literature is only that here the realized game is determined endogenously rather than exogenously.)

Nonetheless, we can make some cursory comments on possible mechanisms for transferring

common knowledge of the realized game to the players in a persona sample game. These mechanisms could arise both in persona games where the realized game is determined endogenously and in more conventional scenarios where the "realized game" is determined exogenously.

To be more precise, in persona sample games there must be a mechanism for the choice of each player i in the persona game to be communicated to all the realized game players. There must also be a mechanism that forces i's realized game player to use the persona that i declared. This second mechanism requires there be a cost to a player of changing her persona between when she reveals it and when the realized game gets played.

One example of such a pair of mechanisms is where each player's persona choice is directly signaled to the other players (consciously or otherwise), and there is a cost for signaling one persona while adopting another. Such a cost could manifest itself as an emotional discomfort. For example, it might "hurt" (sic) an academic to act like an airhead in a social situation. Or it might "hurt" (sic) someone who loves opera and loathes punk rock to act as though she likes punk rock. Translated into the language of persona games, that "hurt" is a cost to a player who signals one persona while adopting a different one.

Another example of such a pair of mechanisms arises in repeated non-anonymous instances of the same game (or similar games) being played among a set of players when there is a cost to changing personas between games. Such a cost would induce a player i to stick with the persona that the other players have already encountered, which means that the persona i will use in the current realized game has been honestly revealed to the other players (assuming they can learn the persona in the earlier games from observations of the outcomes of those earlier games (*8*)). Such a cost may be computational, e.g., it "hurts" (sic) to keep changing my persona from one game to the next, since with each change I can no longer behave by force of habit, but have to explicitly think about how I should act.[6]

[6]An extension of this pair of mechanisms would have each player i maintain the same persona distribution

As a more formal model of such a cost, say the players in the repeated games each adopt a set of computational "heuristics" (*47*) to ease their computational burden. Identify each such set of heuristics with a persona. Then maintaining a persona lowers a player's computational burden, while changing a persona raises that burden. In this model, personas arise as a side-effect of the benefit of adopting computational heuristics. In such a situation, the player wants to adopt a persona *anyway*. If faithfully communicating that persona would then cause a further improvement to her utility, no cost is needed to ensure that that communication is honest.

A third type of such a mechanism pair is where there is a trusted, unerring third party who communicates persona choices, perhaps implicitly. An example is where an experimenter tells the subjects of a behavioral game theory experiment that they are playing against human beings rather than against computer programs. Assume the subjects have had sufficient experience interacting with humans in realized games similar to the one at hand to have learned the typical persona distribution of humans in such games. Then the third party comprises a (statistical) way of communicating information about persona distributions to all the players in the current game.

One can imagine that these types of mechanisms and persona sets arise under natural selection processes. For example, say we fix the move spaces of a set of N players, and are given an index set G and then an associated set $R \equiv \{r_1^\gamma, r_2^\gamma, \ldots r_N^\gamma : \gamma \in G\}$ of possible joint utility functions for the N players. So any particular $r \in R$ is a set of N functions, each taking the space of possible joint moves into the reals. Each such r specifies a game. Say we also have a specified set of personas that each player i can adopt for each $r \in R$. (For example, it might be that in all games i is able to adopt either the fully rational or the irrational persona, which would mean that for each γ, her possible personas are the free utility with $\beta_i = 0$ and the expectation of r_i^γ.)

between games rather than the precise same persona.

Given this, we can average over γ the expected utilities that each player achieves after playing the summary persona sample game for concrete game r^γ. Depending on R, the distribution across it, and the persona sets of the players, that average may be higher or lower than the analogous average where all players always adopt full rationality in the concrete game. If the average is higher when the persona games are played, then there might be evolutionary pressure favoring the existence of such a mechanism pair together with the persona sets $\{A_i\}$; it is in the interests of player i to be able to commit to a persona and to be able to honestly communicate that to the other players.

In particular, say the distribution over utility games is symmetric, in that whatever the probability of game r^γ, it equals the probability of every game $r^{\gamma'}$ in which the utility function of some pair of players are interchanged. Physically, this means that each member of the species is *a priori* equally likely to play either role in any game it plays with some conspecific.

Then our analysis reduces to identifying the games $r \in R$ and persona sets A_i such that the average of expected U_i in concrete game r is less than the average of expected U_i in the associated persona game for persona set A_i. So long as the distribution $P(\gamma)$ is restricted to such concrete games, we're guaranteed that all members of the species will, on average, do better by playing persona games.

In this formulation, an equilibrium persona distribution P_i^A is the distribution played by every player of type i in a mix of N separate populations of players. Alternatively, say P_i^A is the relative frequency of the personas of the players in population i. Then having the equilibrium P_i^A be mixed would mean that we need a blend of personas to maximize the average utility of that population. In such a situation that population would **need** a range of personality types in order to function more effectively, including the "psychos", "pushovers" and other non-rational personas.

Consider a situation where repeated instances of the same persona game are played among

the same (or related) sets of players. In such a situation, mechanism pairs like those discussed above may not be needed to implement a persona sample game (and therefore the associated evolutionary forces may not be necessary). More conventional processes might account for each player's declaring her persona and then adopting it in the realized game. Examples of such processes are those that underlie the folk theorems (*1, 2*), and/or those that underlie more recently investigated phenomena associated with reciprocal altruism, like punishment, reputation effects, etc.

As a final comment, note that all of the foregoing considerations about how the players of the realized game are provided with common knowledge of that game also apply to conventional, non-persona games. In those games the common knowledge of the game's specification is determined exogenously rather than endogenously, but it still needs to be revealed to the players as common knowledge, and then faithfully adopted by all the players.

Personalities. In the text, the only persona sets considered are those involving degrees of rationality and degrees of altruism. There is no reason not to consider other kinds of persona sets however. Risk aversion, uncertainty aversion, reflection points, framing effects, and all the other "irrational" aspects of human behavior can often be formulated as personas. Note that often experiments concerning such aspects of human behavior involve games involving a single subject playing against Nature, and the interesting effects of persona games don't apply there. However such aspects of human behavior also affect play in games involving other players (and in fact some of them may "carry over" to games against Nature from games involving other human players.)

More generally, define a **personality** as a mapping taking any pair of a concrete game $(\{X_i : i \in \mathcal{N}\}, \{u_i : i \in \mathcal{N}\})$ and the specification of a particular player $i \in \mathcal{N}$ to a distribution over the set of all possible personas for that player in a realized game based on that concrete

game. (More formally, the image of a personality is the space of probability measures over $\mathbb{R}^{\Delta_{\mathcal{X}}}$, which is the set of possible objective functions for the objective game.) In particular, a deterministic personality is one that maps any concrete game to a single associated persona.

As an example, a rationality value β is a deterministic personality: Given any utility game $(\{X_i : i \in \mathcal{N}\}, \{u_i : i \in \mathcal{N}\})$, it produces a single persona: the free utility $\mathbb{E}_q(u_i)+\beta_i^{-1}\mathcal{S}(q_i)$. Similarly, the parameter ρ specifying a player's degree of altruism is a personality. When modeling how personas might evolve in biological scenarios, it may make sense to consider evolutions of personalities, which are defined for arbitrary sets of games, rather than just the direct evolution of personas.

It would be interesting to investigate whether the real-world distribution over concrete games in human cultures is one in which it is better to play persona games than the associated concrete games. For example, anthropological data may tell tell us what persona sets are in use in a particular culture in the two-move games played by people in that culture. We could then solve for the set of such two-move concrete games such that, for those persona sets and those concrete games, it is advantageous to play persona games rather than the associated cncrete games. Potentially, related anthropological data would allow us to compare that set of concrete games to the actual ones typically arising in that culture. Similarly, we can look for what persona sets are common to all modern cultures. Then by solving as above, we can predict what concrete games are most frequent in the youngest common ancestor of all modern cultures.

Calculations concerning the TD. The expected utilities for rationality pairs (∞, ∞) and $(0, 0)$ are immediate. For a purely irrational player 2 and rational player 1, the player 1 mixed strategies are given by all convex combinations of playing $x_1 \in \{97, 98\}$. The expected utility of

player 2 for those two x_1 values is given by

$$\begin{aligned}
\mathbb{E}(u_2 \mid x_1, \beta_2 = 0) &= \frac{1}{100}\Big(\Big[\sum_{y=2}^{x_1-1}(y+2)\Big] + x_1 + \Big[\sum_{y=x_1+1}^{101}(x_1-2)\Big]\Big) \\
\Rightarrow \quad \mathbb{E}(u_2 \mid x_1 = 97, \beta_2 = 0) &= \frac{2661}{50} \simeq 53.22, \qquad (25) \\
\mathbb{E}(u_2 \mid x_1 = 98, \beta_2 = 0) &= \frac{2665}{50} \simeq 53.30 \qquad (26)
\end{aligned}$$

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

$$\begin{bmatrix} \left(\frac{6967}{200}, \frac{6967}{200}\right) & \left(\frac{s2661+(1-s)2665}{50}, \frac{2479}{50}\right) \\ \left(\frac{2479}{50}, \frac{q2661+(1-q)2665}{50}\right) & (2,2) \end{bmatrix} \qquad (27)$$

Using the uniform averaging summarizers to average over s and q gives Table 9.

As mentioned in the text, there is one mixed NE $P(\beta_1, \beta_2)$ of the rationality game specified in Table 9. To calculate the realized game distribution associated with that mixed rationality NE, write

$$\begin{aligned}
P(x_1, x_2) &= \sum_{\beta_1,\beta_2} P(x_1, x_2 \mid \beta_1, \beta_2) P(\beta_1, \beta_2) \\
&= \sum_{\beta_1,\beta_2} P(x_1, x_2 \mid \beta_1, \beta_2) P(\beta_1) P(\beta_2). \qquad (28)
\end{aligned}$$

where each of the four distributions $P(x_1, x_2 \mid \beta_1, \beta_2)$ is evaluated using the uniform averaging summarizer if there are multiple realized game NE for that (β_1, β_2) pair. For both players i the associated marginal distribution is given by $P(x_i) = \sum_{\beta_1,\beta_2} P(x_i \mid \beta_1, \beta_2) P(\beta_1) P(\beta_2)$. Plugging in gives the values discussed in the text.

If we redo the foregoing analysis for arbitrary (integer) values of the reward / penalty R we find that nothing changes for either the joint rationality ($\beta_1 = 0, \beta_2 = 0$) or ($\beta_1 = \infty, \beta_2 = \infty$). To address the off-diagonal joint-rationalities we solve to find that the integers on either side of $x_1 = \frac{203-4R}{2}$ maximize $\mathbb{E}(u_1 \mid x_1, R, \beta_2 = 0)$. The two associated values of $\mathbb{E}(u_1 \mid x_1, R, \beta_2 = 0)$ both equal $\frac{41201+16(R-50)R}{800}$, and the two associated values of $\mathbb{E}(u_2 \mid x_1, R, \beta_2 = 0)$ are $\frac{1}{50}(2575 + R(49 - 3R))$ and $\frac{1}{50}(2575 + R(51 - 3R))$.

So again using the uniform averaging summarizers, we find that Table 9 changes to

$$\begin{bmatrix} \left(\frac{6967}{200}, \frac{6967}{200}\right) & \left(\frac{2575+R(50-3R)}{50}, \frac{2575-R(50-R)}{50}\right) \\ \left(\frac{2575-R(50-R)}{50}, \frac{2575+R(50-3R)}{50}\right) & (2, 2) \end{bmatrix} \tag{29}$$

This bimatrix will have a mixed NE only for $0 < R < (25 + 5\sqrt{322})/3$; the associated mixed strategy equilibrium is given by:

$$P(\beta_1 = 0) \;=\; P(\beta_2 = 0) \;=\; \frac{4(3R^2 - 50R - 2475)}{8R^2 - 13233} \tag{30}$$

For R outside of this range the two NE at $(\infty, 0)$ and $(0, \infty)$ also vanish, and the only NE is the pure strategy of full rationality (∞, ∞).

Calculations concerning the UG. Formally, the UG concrete game is a perfect and complete information extensive form game with a single move by each of the two players, with B moving first. We take X to be finite, but "almost dense" in $(0.0, 1.0]$, in the sense that $\max_{z \in (0.0,1.0]}[\min_{x \in X}|x - z|]$ is arbitrarily small. Given any differentiable function $M : \mathbb{R} \to \mathbb{R}$, this assumption concerning X allows us to approximate the argmaxima of M over X by solving for the z such that $M'(z) = 0$.

For simplicity, we adopt the multi-agent representation (*36*). So given a value β, at node x of the game tree (i.e., after B has made offer x), S chooses accept with the behavioral strategy

probability $f_\beta(x) \triangleq e^{\beta x}/(1+e^{\beta x})$. So the expected utility of B for making offer x is:

$$\mathbb{E}(u_B \mid x, \beta) \quad = \quad (1-x) f_\beta(x) \tag{31}$$

The associated expected utility for S is $\mathbb{E}(u_S \mid x, \beta) = x f_\beta(x)$.

$\frac{\partial \mathbb{E}(u_B|x,\beta)}{\partial x}$ monotonically decreases from $x = 0$ and is positive for some x only if $\frac{\partial f_\beta(x)}{\partial x}|_{x=0} \geq 1/2$. Since S wants an offer $x > 0$ and S sets β, this means that $\beta \geq 2$.

For a β satisfying that bound, the x that maximizes the expected payoff to B is the value $x(\beta)$ solving $-1 + (\beta - 1)e^{-\beta x(\beta)} - \beta x(\beta) e^{-\beta x(\beta)} \;=\; 0$, i.e.,

$$x(\beta) \quad = \quad \frac{\beta - W(e^{\beta-1}) - 1}{\beta} \tag{32}$$

where $W(x)$ is the Lambert W function. This $x(\beta)$ is the amount that a fully rational B will offer to S if B knows S has rationality β. So the best β that S can adopt is the maximizer of $x(\beta) f_\beta(x(\beta))$. The associated values of β, $x(\beta)$ and $f_\beta(x(\beta))$ are presented in the text.

Phase transitions in persona games. In this section we present examples of summary persona sample games that exhibit "phase transitions". These examples are persona games for two-player, two-move concrete games, where each persona set B_i is the set of all free utilities that have the same utility function u_i and a rationality $\beta_i \in [0, \infty)$. Intuitively, the phase transition point is a pair of rationalities for the players such that if one player reduces her rationality by an infinitesimal amount, the other player says "okay, I've had enough; I'm going to change my choice", and flips her realized game move accordingly.

Since the concrete game is so small, for pedagogical purposes we do all matrix inversions explicitly. We start by writing the utility bimatrix as

$$\begin{bmatrix} (a_1, a'_1) & (a_2, a'_2) \\ (a_3, a'_3) & (a_4, a'_4) \end{bmatrix} \tag{33}$$

where as usual the Row player utility values are the a terms and Col player's utility values are the a' terms. The probability of player i choosing strategy s is written $P_i(s)$, and we define

$P_{Row}(\mathbf{T}) = p_1^{Row}$, $P_{Row}(\mathbf{D}) = p_2^{Row} = 1 - p_1^{Row}$, $P_{Col}(\mathbf{L}) = p_1^{Col}$ and $P_{Col}(\mathbf{R}) = p_2^{Col} = 1 - p_1^{Col}$. The expected utility to player i given that player i adopts strategy s is written $\mathbb{E}(u_i \mid s)$:

$$\mathbb{E}(u_{Row} \mid \mathbf{T}) = p_1^{Col} a_1 + p_2^{Col} a_2 = a_2 + p_1^{Col}(a_1 - a_2), \tag{34}$$

$$\mathbb{E}(u_{Row} \mid \mathbf{D}) = p_1^{Col} a_3 + p_2^{Col} a_4 = a_4 + p_1^{Col}(a_3 - a_4), \tag{35}$$

$$\mathbb{E}(u_{Col} \mid \mathbf{L}) = p_1^{Row} a'_1 + p_2^{Row} a'_3 = a'_3 + p_1^{Row}(a'_1 - a'_3), \tag{36}$$

$$\mathbb{E}(u_{Col} \mid \mathbf{Row}) = p_1^{Row} a'_2 + p_2^{Row} a'_4 = a'_4 + p_1^{Row}(a'_2 - a'_4). \tag{37}$$

So the QRE is given by the quadruple of values $P_i(s)$ solving the following pair of simultaneous equations:

$$p_1^{Row} = \frac{e^{\beta_r \mathbb{E}(u_1|\mathbf{T})}}{\sum_{s \in s_1} e^{\beta_r \mathbb{E}(u_1|s)}}, \tag{38}$$

$$p_1^{Col} = \frac{e^{\beta_c \mathbb{E}(u_2|\mathbf{L})}}{\sum_{s \in s_2} e^{\beta_c \mathbb{E}(u_2|s)}}. \tag{39}$$

We can transform p_i^{Row} and p_i^{Col} into more convenient symmetrized versions by:

$$p_1^{Row} = \frac{1-Y}{2} \text{ and } p_2^{Row} = \frac{1+Y}{2}; \; Y = p_2^{Row} - p_1^{Row} = 1 - 2p_1^{Row}$$
$$p_1^{Col} = \frac{1-X}{2} \text{ and } p_2^{Col} = \frac{1+X}{2}; \; X = p_2^{Col} - p_1^{Col} = 1 - 2p_1^{Col}$$

Under this transformation the equations defining the QRE become

$$Y = f(X, \beta_{Row}) = 1 - \frac{2e^{\beta_{Row}\mathbb{E}(u_{Row}|\mathbf{T})}}{\sum_{s \in s_1} e^{\beta_{Row}\mathbb{E}(u_{Row}|s)}}, \tag{40}$$

$$X = g(Y, \beta_{Col}) = 1 - \frac{2e^{\beta_{Col}\mathbb{E}(u_{Col}|\mathbf{L})}}{\sum_{s \in s_2} e^{\beta_{Col}\mathbb{E}(u_{Col}|s)}}, \tag{41}$$

where the expectations over the utilities in symmetrical variables are:

$$\mathbb{E}(u_{Row} \mid \mathbf{T}) = \frac{1}{2}(a_1 + a_2 + X(a_2 - a_1)), \tag{42}$$

$$\mathbb{E}(u_{Row} \mid \mathbf{D}) = \frac{1}{2}(a_3 + a_4 + X(a_4 - a_3)), \tag{43}$$

$$\mathbb{E}(u_{Col} \mid \mathbf{L}) = \frac{1}{2}(a'_1 + a'_3 + Y(a'_3 - a'_1)), \tag{44}$$

$$\mathbb{E}(u_{Col} \mid \mathbf{Row}) = \frac{1}{2}(a'_2 + a'_4 + Y(a'_4 - a'_2)). \tag{45}$$

Now define $F(Y,\beta_{Row},\beta_{Col}) \triangleq Y - f(X,\beta_{Row}) = Y - f(g(Y,\beta_{Col}),\beta_{Row})$. So a QRE is the Y solution to $F(Y,\beta_{Row},\beta_{Col}) = 0$, with X then set to $g(Y,\beta_{Col})$. In general, there are multiple solutions to $F(Y,\beta_{Row},\beta_{Col}) = 0$ for a given $(\beta_{Row},\beta_{Col})$ pair, call these sets of solutions $\{Y_i\}$ and $\{X_j\}$. Note that not every pair (Y_i, X_j) is a fixed point. In fact, there is a bijection between $\{Y_i\}$ and $\{X_j\}$ that picks out the pairs that are fixed points.

We can now find the rate of change of the fixed point pairs (Y, X) with respect to β_{Row} by holding β_{Col} constant and finding the vector orthogonal to the level set $F = 0$. First we differentiate F with respect to its three arguments, getting

$$F_1(Y,\beta_{Row},\beta_{Col}) = 1 - f_1 g_1, \tag{46}$$

$$F_2(Y,\beta_{Row},\beta_{Col}) = -f_2, \tag{47}$$

$$F_3(Y,\beta_{Row},\beta_{Col}) = -f_1 g_2 \tag{48}$$

where we use subscripts to denote the variable which we are differentiating with respect to and we have dropped the functional dependencies on the right hand side for clarity. The vectors orthogonal to F solve $0 = [r, s, t] \cdot [F_1, F_2, F_3]$, i.e., $r = \frac{-sF_2 - tF_3}{F_1}$, where $r = \delta Y$, $s = \delta\beta_{Row}$, and $t = \delta\beta_{Col}$. Noting that $\frac{r}{s}|_{t=0} = \frac{\delta Y}{\delta\beta_{Row}}|_{t=0} = \frac{\partial Y}{\partial\beta_{Row}}$ we can plug in to get

$$\frac{\partial Y}{\partial\beta_{Row}} = \frac{-F_2}{F_1} = \frac{f_2}{1 - f_1 g_1}. \tag{49}$$

Then using the chain rule on $X = g(Y,\beta_{Col})$ gives

$$\frac{\partial X}{\partial\beta_{Row}} = g_1 \frac{\partial Y}{\partial\beta_{Row}} = \frac{g_1 f_2}{1 - f_1 g_1} \tag{50}$$

where:

$$
\begin{aligned}
f_1(X,\beta_{Row}) &= \frac{\beta_{Row}(1-Y^2)}{4}(a_1-a_2-a_3+a_4), && (51)\\
g_1(Y,\beta_{Col}) &= \frac{\beta_{Col}(1-X^2)}{4}(a'_1-a'_2-a'_3+a'_4), && (52)\\
f_2(X,\beta_{Row}) &= \frac{1-Y^2}{2}(\mathbb{E}(u_{Row}\mid \mathbf{D})-\mathbb{E}(u_{Row}\mid \mathbf{T})),\\
&= \frac{1-Y^2}{4}(-a_1-a_2+a_3+a_4+X(a_1-a_2-a_3+a_4)), && (53)\\
g_2(Y,\beta_{Col}) &= \frac{1-X^2}{2}(\mathbb{E}(u_{Col}\mid \mathbf{Row})-\mathbb{E}(u_{Col}\mid \mathbf{L})),\\
&= \frac{1-X^2}{4}(-a'_1+a'_2-a'_3+a'_4+Y(a'_1-a'_2-a'_3+a'_4)). && (54)
\end{aligned}
$$

Similar calculations give $\frac{\partial Y}{\partial \rho_{Col}}$ and $\frac{\partial X}{\partial \rho_{Col}}$.

We will need to express the rationalities in terms of X and Y. To do that first write

$$
p_1^{Row} = \frac{e^{\beta_{Row}(p_1^{Col}a_1+p_2^{Col}a_2)}}{e^{\beta_{Row}(p_1^{Col}a_1+p_2^{Col}a_2)}+e^{\beta_{Row}(p_1^{Col}a_3+p_2^{Col}a_4)}} \tag{55}
$$

so that

$$
\beta_{Row} = \frac{\ln\left(\frac{1-p_1^{Row}}{p_1^{Row}}\right)}{p_1^{Col}(a_3-a_1)+p_2^{Col}(a_4-a_2)}, \tag{56}
$$

and similarly write

$$
\begin{aligned}
p_1^{Col} &= \frac{e^{\beta_{Col}(p_1^{Row}a'_1+p_2^{Row}a'_3)}}{e^{\beta_{Col}(p_1^{Row}a'_1+p_2^{Row}a'_3)}+e^{\beta_{Col}(p_1^{Row}a'_2+p_2^{Row}a'_4)}}; && (57)\\
\beta_{Col} &= \frac{\ln\left(\frac{1-p_1^{Col}}{p_1^{Col}}\right)}{p_1^{Row}(a'_2-a'_1)+p_2^{Row}(a'_4-a'_3)} && (58)
\end{aligned}
$$

Now substitute in the symmetrical form of the variables:

$$
\begin{aligned}
\beta_{Row} &= \frac{2\ln\left(\frac{1+Y}{1-Y}\right)}{-a_1-a_2+a_3+a_4+X(a_1-a_2-a_3+a_4)} && (59)\\
\beta_{Col} &= \frac{2\ln\left(\frac{1+X}{1-X}\right)}{-a'_1+a'_2-a'_3+a'_4+Y(a'_1-a'_2-a'_3+a'_4)} && (60)
\end{aligned}
$$

We are now ready to demonstrate phase transitions in persona games based on two-player, two-move, simple concrete games. To do this we first define the partition function of a QRE and its derivative in an analogous fashion to that of statistical mechanics. Then we show that the derivative of the partition function can diverge for a particular pair of player rationalities for a particular concrete game.

$$
\begin{aligned}
\ln(Z_{col}) &= \ln\left(e^{\beta_{col}\mathbb{E}(u_{col}|\mathbf{L})} + e^{\beta_{col}\mathbb{E}(u_{col}|\mathbf{R})}\right) \\
\frac{\partial \ln(Z_{col})}{\partial \beta_{col}} &= \frac{1}{Z_{col}}\Big(\mathbb{E}(u_{col} \mid \mathbf{L})e^{\beta_{col}\mathbb{E}(u_{col}|\mathbf{L})} + \mathbb{E}(u_{col} \mid \mathbf{R})e^{\beta_{col}\mathbb{E}(u_{col}|\mathbf{R})}\Big) + \\
&\quad \frac{\beta_{col}}{Z_{col}}\Big(e^{\beta_{col}\mathbb{E}(u_{col}|\mathbf{L})}\frac{\partial \mathbb{E}(u_{col} \mid \mathbf{L})}{\partial \beta_{col}} + e^{\beta_{col}\mathbb{E}(u_{col}|\mathbf{R})}\frac{\partial \mathbb{E}(u_{col} \mid \mathbf{R})}{\partial \beta_{col}}\Big) \\
&= \mathbb{E}(u_{col}) + \beta_{col}\Big(p_{\mathbf{L}}\frac{\partial \mathbb{E}(u_{col} \mid \mathbf{L})}{\partial \beta_{col}} + p_{\mathbf{R}}\frac{\partial \mathbb{E}(u_{col} \mid \mathbf{R})}{\partial \beta_{col}}\Big) \\
&= \mathbb{E}(u_{col}) + \frac{\beta_{col}}{2}\frac{\partial Y}{\partial \beta_{col}}\Big(p_{\mathbf{L}}(a'_3 - a'_1) + p_{\mathbf{R}}(a'_4 - a'_2)\Big) \\
&= \mathbb{E}(u_{col}) + \frac{\beta_{col}}{4}\frac{\partial Y}{\partial \beta_{col}}\Big((1-X)(a'_3 - a'_1) + (1+X)(a'_4 - a'_2)\Big) \\
&= \mathbb{E}(u_{col}) + \frac{\beta_{col}}{4}\frac{\partial Y}{\partial \beta_{col}}\Big(-a'_1 - a'_2 + a'_3 + a'_4 + X(a'_1 - a'_2 - a'_3 + a'_4)\Big) \\
&= \mathbb{E}(u_{col}) + \frac{\beta_{col}g_2 f_1}{4(1 - f_1 g_1)}\Big(-a'_1 - a'_2 + a'_3 + a'_4 + X(a'_1 - a'_2 - a'_3 + a'_4)\Big) \qquad (61)
\end{aligned}
$$

where $p_{\mathbf{L}}$ and $p_{\mathbf{R}}$ are the probabilities of column player choosing strategies **L** and **R** respectively. We now redefine parameters and write

$$a_1 = a_1 - a_2 - a_3 + a_4 \qquad (62)$$

$$a_2 = -a_1 + a_2 - a_3 + a_4 \qquad (63)$$

$$a_3 = -a_1 - a_2 + a_3 + a_4 \qquad (64)$$

$$a'_1 = a'_1 - a'_2 - a'_3 + a'_4 \qquad (65)$$

$$a'_2 = -a'_1 + a'_2 - a'_3 + a'_4 \qquad (66)$$

$$a'_3 = -a'_1 - a'_2 + a'_3 + a'_4 \qquad (67)$$

We also define $c(X,Y) \triangleq \beta_{col}\beta_{row}(1-Y^2)(1-X^2)$. This allows us to write

$$\frac{\partial \ln(Z_{col})}{\partial \beta_{col}} = \mathbb{E}(u_{col}) + \frac{a_1 c(X,Y)(a_3' + Xa_1')(a_2' + Ya_1')}{64 - 4a_1 a_1' c(X,Y)} \tag{68}$$

This diverges if the following constraints hold:

$$c(X,Y) = \frac{16}{a_1 a_1'} \tag{69}$$

$$0 \neq a_1 c(X,Y)(a_3' + Xa_1')(a_2' + Ya_1') \tag{70}$$

Assuming the first constraint holds, the second constraint is guaranteed so long as:

$$a_1 \neq 0, \quad a_2' + Ya_1' \neq 0, \quad a_3' + Xa_1' \neq 0 \tag{71}$$

Now

$$\beta_{col}\beta_{row} = \frac{4\ln\left(\frac{1+Y}{1-Y}\right)\ln\left(\frac{1+X}{1-X}\right)}{(a_3 + Xa_1)(a_2' + Ya_1')} \tag{72}$$

In addition, a pair (X,Y) is guaranteed to be a *QRE* with a unique β_{col} and β_{row} whenever $\mathbb{E}(u_{col} \mid \mathbf{L}) \neq \mathbb{E}(u_{col} \mid \mathbf{R})$ and $\mathbb{E}(u_{row} \mid \mathbf{T}) \neq \mathbb{E}(u_{row} \mid \mathbf{D})$. This in turn means:

$$a_2' + Ya_1' \neq 0 \quad \text{and} \quad a_3 + Xa_1 \neq 0 \tag{73}$$

So for any Y obeying Eq.'s 71 and 73, if we can find an X such that Eq. 70 holds as well as the second constraint in Eq. 73, then we have a phase transition.

An example occurs for the bimatrix

$$\begin{bmatrix} (0,0) & (6,1) \\ (5,5) & (4,6) \end{bmatrix} \tag{74}$$

To verify this write out the redefined parameters:

$$\begin{aligned}
a_1 &= a_1 - a_2 - a_3 + a_4 &= -7 \\
a_2 &= -a_1 + a_2 - a_3 + a_4 &= 5 \\
a_3 &= -a_1 - a_2 + a_3 + a_4 &= 3 \\
a'_1 &= a'_1 - a'_2 - a'_3 + a'_4 &= 0 \\
a'_2 &= -a'_1 + a'_2 - a'_3 + a'_4 &= 2 \\
a'_3 &= -a'_1 - a'_2 + a'_3 + a'_4 &= 10
\end{aligned}$$

Next take the limit as $\beta_{row} \to \infty$, and set β_{col} to the value where the row player decides to flip from one row to the other. This is the limit where $Y \to 0$, and $X \to 3/7$. For convenience, express that as the limit as $z \to 0$ where $Y = z$, and $X = 3/7 - z/7$:

$$\beta_{row} = \frac{2}{7} \ln\left(\frac{1+z}{1-z}\right) \tag{75}$$

$$\beta_{col} = \ln\left(\frac{10-z}{4+z}\right) \tag{76}$$

So we finally get:

$$\lim_{z\to 0} \frac{\partial \ln(Z_{col})}{\partial \beta_{col}} = \lim_{z\to 0} \frac{-2,800 \ \ln(5/2) \ln\left(\frac{1+z}{1-z}\right)}{784\, z} = \infty \tag{77}$$

as claimed.

References

1. D. Fudenberg, J. Tirole, *Game Theory* (MIT Press, Cambridge, MA, 1991).

2. R. B. Myerson, *Game theory: Analysis of Conflict* (Harvard University Press, 1991).

3. J. M. Smith, *Evolution and the Theory of Games* (Cambridge University Press, 1982).

4. C. Camerer, *Behavioral Game theory: experiments in strategic interaction* (Princeton University Press, 2003).

5. C. Starmer, *Journal of Economic Literature* **38**, 332 (2000).

6. C. Camerer, E. Fehr, *Science* **311**, 47 (2006).

7. D. Kahneman, *American Economic Review* **93**, 1449 (2003).

8. D. Fudenberg, D. K. Levine, *The Theory of Learning in Games* (MIT Press, Cambridge, MA, 1998).

9. C. Hauert, A. Traulsen, H. Bradt, M. Nowak, K. Sigmund, *Science* **316**, 1905 (2007).

10. R. Axelrod, D. Dion, *Science* **242**, 1385 (1988).

11. R. Trivers, *Natural Selection and Social Theory: Selected Papers* (Oxford University Press, 2002).

12. S. Bowles, R. Boyd, E. Fehr, H. Gintis, *The Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life* (MIT Press, 2005).

13. M. A. Nowak, K. Sigmund, *Nature* **393**, 573 (1998).

14. A. Rubinstein, *Modeling Bounded Rationality* (MIT press, 1998).

15. R. Frank, *The American Economic Review* **77**, 593 (1987).

16. O. Gurerk, B. Irlenbusch, B. Rockenbach, *Science* **312**, 108 (2006).

17. A. Neyman, *Economics Letters* **19**, 227 (1985).

18. M. Kaminski, *Games Prisoners Play* (Princeton University Press, 2004).

19. J. Ferber, *Foundations of Distributed Artificial Intelligence*, G. O-Hare, N. Jennings, eds. (John Wiley and Sons, 1996), pp. 287–314.

20. J. Shamma, G. Arslan, *IEEE Trans. on Automatic Control* **50**, 312 (2004).

21. A. Schaerf, Y. Shoham, M. Tennenholtz, *Journal of Artificial Intelligence Research* **162**, 475 (1995).

22. S. Bieniawski, I. Kroo, D. H. Wolpert, *Proceedings of 2005 AIAA Guidance, Navigation, and Control Conference, San Francisco, CA* (2005). AIAA Paper 2005-6074.

23. H. Hwang, J. Kim, C. Tomlin, *Air Traffic Control Quarterly* (2007). In press.

24. E. Cramer, J. Dennis, et alia, *SIAM J. of Optimization* **4** (1994).

25. S. Choi, J. Alonso, *Proceedings of 10th AIAA/ISSMO Multidisciplany Analysis and Optimization Conference* (2004). AIAA Paper 2004-4371.

26. N. Nisan, A. Ronen, *Games and Economic Behavior* **35**, 166 (2001).

27. K. Basu, *Scientific American* (2007).

28. C. M. Capra, J. K. Goeree, R. Gomez, C. H. Holt, *American Economic Review* **19**, 678 (1999).

29. K. Basu, *American Economic Review* **84**, 391 (1994).

30. A. Rubinstein, Instinctive and cognitive reasoning: a study of response times (2004). Ariel-rubinstein.tau.ac.il/papers/Response/pdf.

31. T. Becker, M. Carter, J. Naeve, Experts playing the traveler's dilemma (2005). Universitat Hohenheim Nr. 252/2005.

32. J. K. Goeree, C. A. Holt, *Proceedings National Academy of Sciences* **96**, 10564 (1999).

33. R. D. McKelvey, T. R. Palfrey, *Games and Economic Behavior* **10**, 6 (1995).

34. D. H. Wolpert, A predictive theory of games (2006). ArXiv.org/abs/nlin.AO/0512015.

35. J. K. Goeree, C. A. Holt, T. R. Palfrey, *Journal of Economic Theory* **104**, 247 (2002).

36. R. D. McKelvey, T. R. Palfrey, *Experimental Economics* **1**, 9 (1998).

37. A. G. Sanfrey, J. K. Rilling, J. A. Aronson, L. E. Nystrom, J. D. Cohen, *Science* **300.5626**, 1755(4) (2003).

38. M. Nowak, K. Page, K. Sigmund, *Science* **289.5485**, 1773 (2000).

39. D. Talmi, C. Frith, *Nature* **446**, 865 (2007).

40. A. Tversky, *Preference, Belief, and Similarity: Selected Writings* (MIT Press, 2004).

41. J. Jamison, D. H. Wolpert, Persona games and altruism (2007). In preparation.

42. D. H. Wolpert, Predicting the outcome of a game (2007). Submitted. See arXiv.org/abs/nlin.AO/0512015 for an early version.

43. D. Fudenberg, D. Kreps, *Game and Economic Behavior* **5**, 320 (1993).

44. D. H. Wolpert, Honest equilibria (2007). Submitted. Contact author for early version.

45. D. McFadden, *Structural analysis of discrete data and econometric applications* (MIT Press, 1981).

46. R. J. Aumann, *Int. J. Game Theory* **28**, 301 (1999).

47. S. Hart, *Econometrica* **73**, 14011430 (2005).

48. We would like to thank Julian Jamison, Nils Bertschinger, Nihat Ay, and Eckehard Olbrich for helpful discussion.