

Non-Rationality in non-repeated games

David H. Wolpert,^{1*} Julian Jamison², David Newth³, Michael Harre⁴

¹MS 269-1, NASA Ames Research Center,
Moffett Field, CA, 94035, USA

²USC Brain and Creativity Institute, 3620 McClintock Ave, Suite 250,
Bldg SGM, MC 1061, Los Angeles, CA 90089, USA

³CSIRO Centre for Complex Systems Science
Gungahlin Homestead, Crace, ACT, 2611, Australia,

⁴The Centre for the Mind and The School of Information Technologies,
Sydney University, Australia

*To whom correspondence should be addressed; E-mail: dhw@santafe.edu

A long-standing puzzle in economics and biology is why humans and animals sometimes act non-rationally by seeming not to adopt their optimal strategy when interacting with others (1–3). A particular example of this puzzle is why humans and animals sometimes are altruistic to non-kin and so cooperate with them. Many previous explanations show how non-rationality by a player in an individual game can be optimal for that player if considered in the context of an infinite sequence of identical, repeated instances of that game (1, 4–6, 6–17). Here we introduce a framework that explains non-rationality even in single, non-repeated games. Our framework starts from the observation that an individual i often adopts a “persona” that they signal to others before interacting with them. We formalize such a temporarily adopted persona as a utility function, one that may differ from i ’s true utility function. By changing what

persona they adopt, which may change the behavior of others in the interaction. In particular, we show that sometimes by adopting a “non-rational” persona, an individual i induces behavior by others that increases the value of i 's true utility function. In such cases, it is optimal for i to be “non-rational”. This framework can explain many different instances of non-rationality. To illustrate this breadth, we first derive several quantitative predictions concerning non-rationality in non-repeated play of the Traveler’s Dilemma (TD), all of which agree with experiment. We then show how cooperation can arise even in a non-repeated play of certain versions of the Prisoner’s Dilemma (PD). In addition to explaining cooperation in the PD, our framework reveals an aspect of the PD never previously realized: an unavoidable tradeoff between the robustness of cooperation in the PD and the benefit of the cooperation. On a broader scale, our framework provides a way to formalize the role of non-rationality in “culture gaps”. Finally, adopting an engineering perspective, our framework has implications for how to design mechanisms that regulate groups of interacting individuals, including applications at NASA.

1 Background

A lot of research has been done on explaining non-rational behavior in a game γ by embedding γ in an open-ended sequence of repetitions of γ . The earliest such explanations concerned games where opponents were genetically related (18, 19) or where they never varied across the sequence (4, 20, 21). This work was subsequently expanded into a broader body of work considering the evolution of the strategies of the players of γ across repetitions of γ , with no restriction that opponents be related or fixed throughout the sequence (1, 4–6, 6–14). In this broader work, which we call Evolution Of Strategies (EOS), the strategy of every player in any

particular instance t of γ is fixed. Those fixed strategies jointly determine the payoffs for all the players in the game played at t . As successive versions of γ are played, the probability distribution of the strategies of the players get updated, to reflect each player's payoffs in the preceding instances of γ . Often a strategy that is non-rational for any single instance of γ (i.e., not payoff-maximizing) is actually rational when considered as part of the full sequence of games.

EOS has uncovered and helped analyze many important phenomena related to non-rational behavior, including punishment, "loners", and reputation effects. However EOS has several limitations. The primary one is that EOS is based on having all the individuals in a population repeatedly play the exactly identical game, in a sequence that potentially extends infinitely into the future. In the real world, many games are not infinitely repeated, or at least not in the exact same form. Another limitation is that no formalization of EOS applies to non-rationality in general, but rather each one is tailored to only one type of non-rationality, like altruism. A third limitation is that EOS often requires the players to have some ability to recognize their opponents from one game instance to the next, or to have non-zero probability of encountering the same opponent in multiple game instances. Unfortunately, even given such limitations, the analysis with EOS is often so complicated that computer simulations are needed. (See (22) for discussion of other limitations of EOS.)

There are alternative explanations of apparent non-rationality that apply to non-repeated games, unlike EOS. One of them starts with the observation that real-world social organisms often have different "personas" that they adopt for their interactions with one another. For example, someone might "act dumber than they are" in an interaction. Similarly, often we "act like a different person" when we interact with our boss, our spouse, or a child. In each such instance we act as though we have a different set of preferences and values from our real ones. This phenomenon has even entered common discourse, as illustrated in a recent

newspaper article that said “the workplace is full of chameleons who adopt a different persona each day” (28). In some instances we may choose preferences, signal them to one another, and commit ourselves to them all in an unconscious manner, as “moods” or “emotions” that we signal to one another via tone of voice, body language, and the like. (See the discussion in (23), and of costly signaling in general in (25–27).) In fact, there is reason to believe that unconscious signaling of moods is most pronounced when the signaler is aware that others are watching (24), precisely the context in which game theory considerations come into play.

Adopting a persona that disagrees with one’s true utility would seem to be non-rational. To illustrate how it can actually be rational, say we have two players, Row and Col, each of whom can choose one of two moves (“pure strategies”). We write the sets of pure strategies as (Top, Down) (**T, D**) for Row, and (Left, Right) (**L, R**) for Col. Both players have a “utility function”, which maps any joint move by both players into a real number. As an example, say the utility function pairs (u^R, u^C) for the four possible joint moves can be written as the matrix

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

This matrix says, for instance, that if Row plays **T** while Col plays **L**, then Row’s utility is 6 and Col’s utility is 0. (This game is a particular instance of the broad class of games known as the Prisoner’s Dilemma (PD).)

To play an instance of the game each player $i \in \{\text{Row, Col}\}$ independently chooses a “mixed strategy”, i.e., a probability distribution $P_i(x_i)$ over their set 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. A pair of mixed strategies $(P_{\text{Row}}, P_{\text{Col}})$ is called 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 their mixed strategy, in light of their opponent’s mixed strategy. If either player violates this condition, they are said not to be

“rational”.

For example, in the PD of Table 1, 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 they played **L** with non-zero probability, and given the mixed strategy of Col, Row’s expected utility would decrease if they played **D** with non-zero probability.) Note though that at the (non-NE) joint move (**D, L**), both players have higher expected utility than at the NE. So if they could both be induced to cooperate with one another and choose that move—and in doing so both not be rational—both of the players would benefit.

Now say that rather than being rational in the PD, Col were perfectly **irrational**. That is, they commit to choosing uniformly randomly between their two moves, with no evident concern for the resultant value of their utility function, and therefore no concern for what strategy Row adopts. Given such irrationality of Col, Row would have expected utility of 5 for playing **T**, and of $(5.0 + 0.0)/2 = 2.5$ 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 Col’s expected utility would be 4, being irrational rather than rational would hurt Col in this PD.

Now however modify the PD to have the following utility functions (u^R, u^C) :

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

Again the joint move (**T, 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 their expected utility from 1.0 to 5.5. 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 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 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.

Stated in this informal way, the persona-based explanation of non-rationality predates EOS, going back at least to the 1950's (29–31), and arguably back to antiquity (31). In particular, it played a prominent role in formulation of cold war policies like mutual assured discussion.

2 Evolution of Preferences and Persona Games

We now present our first contribution, a framework that formalizes the persona phenomenon. This framework does not share the limitations of EOS. In particular, it can be applied to non-repeated games. Moreover, as illustrated below, it both explains experimental data concerning non-rational behavior and uncovers novel aspects of such behavior.

To introduce our framework, first note that the EOS equilibrium concept, involving repetitions of a game γ , can be modified to produce the NE concept that concerns a non-repeated instance of γ . In both the EOS equilibrium and the NE, each player i is assigned a utility function u^i over the space X . However the NE concept replaces the infinite game sequence of EOS with a single game and two assumptions. The first is the complete information assumption, common in economics (21). That assumption implies that each player knows X in full, and also knows the utility functions of all the players. The second assumption is that the players have common knowledge and each player i is rational. This means i uses their complete information strategically, to choose the distribution $P(x_i)$ that maximizes their expected utility, given what they think the other players will do. The result of these two assumptions is that the players

jointly play the NE of γ .

We want to formalize the persona framework in an analogous way, i.e., we want to introduce the complete information and rationality assumptions into a repeated game framework, to produce a new framework that applies to non-repeated games. Now note that in the NE players adopt strategies, and the NE is a modification of a repeated game framework where strategies evolve. So by our desired analogy, to build a framework where players adopt personas, we should modify a repeated game framework where personas evolve.

Evolution Of Preferences (EOP) (15–17, 32–34) is such a repeated game framework. EOP can be viewed as a modification of EOS. Like in EOS, in EOP the game γ specifies a space of possible joint moves X . γ also specifies for each player i an associated true, **concrete** utility function u^i defined over X . Unlike in EOS though, rather than fixing the strategy of each player i at the beginning of each instance t of γ , the preference of player i is fixed, i.e., a counterfactual utility function $b_i \in B_i$ is fixed.

In addition to this modification, EOP changes EOS by expanding each instance t of γ into a two-step process. In the first step the players signal their (fixed) preferences to one another. Those signaled preferences are assumed to be binding on the players, in the sense that each player is assumed to act rationally for their signaled preference. This means that in the second step at t , each player i chooses the strategy over X_i that they think will maximize the expected value of their signaled preference b_i , given that the other players will try to maximize their signaled utilities, $\{b_j : j \neq i\}$. So the distribution over X in the game at t is given by a NE of the **realized game** at t specified by that set of signaled preferences at t .

That NE joint strategy over X is evaluated under the concrete utility functions $\{u^i\}$ to get the ultimate expected payoffs to the players for game instance t . Analogously to the evolution of the strategies of the players in EOS, in EOP the fixed preferences the players have before each instance of the game is updated as the sequence of games unfolds. As in EOS, this updating is

based on each player's payoffs in the preceding instances of γ .

Note that in EOP, unlike in EOS, the strategy of player i in game instance t will depend on attributes of the other players in that game instance. Partly as a result, it is often easier to derive results without using computer simulations in EOP than it is in EOS. Another advantage of EOP over EOS is that it does not require any ability of the players to recognize one another from one game to the next. (The signaling in EOP serves the same mathematical purpose as the ability to recognize your opponents does in EOS.)

On the other hand, EOP has some limitations not found in EOS. Much of the formal work in EOP restricts attention to a game (or set of coupled games) with a single symmetric utility function shared by all the players. Such games are rare in the real world. In addition, EOP typically requires that the population be infinite. However when used to evolve a finite population, even a large one, natural selection can result in very different equilibria from when it is used to evolve infinite populations (35, 36). Moreover, abstracting away from real-world geographical constraints, the evolution process assumed in EOP typically requires that *all* individuals in a population interact, even when the population is infinite. Furthermore, for some EOP games, the evolutionary dynamic process has no equilibrium; EOP cannot make predictions for such games. Another limitation is that the results in EOP typically vary with the initial characteristics of the population that the evolution works on. Finally, EOP requires an infinite sequence of exactly identical games, which as mentioned is rare in the real world.

In the same way that the NE is a modification of EOS, we can formalize the persona phenomenon as a modification of EOP. In this formalization, each player i has an associated concrete utility function u^i and a set of possible, counterfactual utility functions over X , B_i . Like in EOP, we assume that every such concrete utility function u_i and set of possible adopted functions B_i is provided exogenously, perhaps through an evolutionary process. (In this paper we restrict attention to B_i 's that seem to be found in the real world.)

Again like in EOP, the concrete utility functions and sets of possible counterfactual utility functions are used to expand the concrete game into two steps. In the first step every player i samples an associated distribution $P(b_i)$ to get a b_i that they will play for the second, realized game, and signals that b_i to the other players. Then in the second step the players play a NE of the realized game specified by those signaled utility functions. That NE sets the expected values of the players' concrete utility functions.

All of this is identical to EOP. Where we differ from EOP is exactly the same place that the NE concept differs from EOS: we replace EOP's repeated games with a single game, and introduce the complete information and common knowledge / full rationality assumptions. In our context, the complete information assumption means that before signaling their counterfactual utility, each player knows their own set B_i , the sets $\{B_j\}$ of the other players, and knows the concrete utility functions of all players. Common knowledge and full rationality means that each player i will use their complete information strategically, in a rational manner, to choose for their self the distribution $P(b_i)$ to be sampled to generate the signaled b_i . More precisely, it means that each player i chooses $P(b_i)$ so as to maximize the associated expected value of their concrete utility, as evaluated under the NE of the realized game. (See (37) for extensions to concrete games of incomplete information.)

This modification means that in the persona framework the distributions $P(b_i)$ themselves are NE, of the full, two-step game. This allows us to bring all the power of techniques for analyzing NE to bear in predicting what the $P(b_i)$'s are. In contrast, in both EOS and EOP the distributions $P(b_i)$ are equilibria of a dynamic evolutionary process, and powerful techniques for analyzing NE usually cannot be applied. This is why it is easier to generate quantitative results without computer simulations in the persona framework than in EOP.

The use of NE techniques also means that every game in the persona framework has an equilibrium, so the persona framework can always make a prediction. This is not the case in

EOP. Furthermore, being based on the NE, no initial characteristics of a population are relevant in the persona framework, and we make no physically impossible assumptions about such a population. (Note that the physical timescales involved in the process that the persona framework models are very different from the timescales of EOP and EOS; the interactions of single individuals versus evolving populations of individuals.)

Perhaps most important, the persona framework offers explanations for apparent non-rationality even in non-repeated games. This corroborates the conclusion in (38) that “cooperation can (be explained), even among non-kin, in situations devoid of repeat interaction”. However the persona framework shows that this conclusion holds even without punishment and genes for non-kin altruism (which have not been found on the human chromosome), which are assumed in (38). Cooperation can exist for purely self-interested reasons.

We refer to the b_i 's determined for a single game via common knowledge as **personas**, to distinguish them from the preferences that are determined in standard EOP over an infinite sequence of games. Accordingly, we refer to each B_i as a **persona set**. Note that the players in the first step can be viewed as playing a game. Their joint move is the joint persona they adopt, b . The utility function of player i in this game is the mapping from all possible b 's to the expected concrete utility of the (NE of the) realized game specified by b . We refer to this game as a persona game. (See the supplemental information for a more detailed definition of persona games, and a discussion of their relation with yet other frameworks, e.g., games involving the signaling of binding contracts.)

3 The Traveler's Dilemma

To illustrate the persona framework, we provide an explanation for some of the experimental data concerning the famous Traveller's Dilemma (TD) (39–44). 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 $\{2, 3, \dots, 101\}$. 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 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. In experiments rich with implications for the sociology of science, it has been found that 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. When describing these results, Basu (39) called for a formalization of “the idea of behavior generated by rationally rejecting rational behavior ... to solve the paradoxes that plague game theory”.

Consider a persona game based on the $R = 2$ TD concrete game. Since it seems that real humans are sometimes fully rational and sometimes irrational, choose those as the possible personas of the players, indicated by $\rho = \infty$ and $\rho = 0$, respectively. When both players are fully rational, the expected utility to both is 2, i.e., $\mathbb{E}(u^i | \rho_1 = \infty, \rho_2 = \infty) = 2$ for both players i . Now say that player i is rational while the other player is irrational. The resultant expected utility $\mathbb{E}(u^i | x_i, \rho_{-i} = 0)$ reaches its (integer) maximum at $x_i \in \{97, 98\}$. Plugging in this value of the full rationality x_i means that $\mathbb{E}(u^i | \rho_i = \infty, \rho_{-i} = 0) \simeq 49.6$ (37). Continuing in this way

gives persona 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)	(3)
$+\infty$	(49.6, 53.3)	(2, 2)	

This persona game has two pure strategy NE, $(\rho_1, \rho_2) = (0, \infty)$ and $(\rho_1, \rho_2) = (\infty, 0)$. The associated distribution $P(x_1)$ for the first of these rationality NE is uniform. The associated $P(x_2)$ instead has half its mass on $x_2 = 97$, and half on $x_2 = 98$. The two distributions for the other pure strategy rationality NE are identical, 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 persona game, at which both rationality players choose $\rho = 0$ 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 $P(\rho_1, \rho_2)$ is not a delta function, $P(x_1, x_2) \neq P(x_1)P(x_2)$.)

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

Uniformly averaging over the three NE of the persona game gives a $P(x)$ that is highly biased to large values of x . This agrees with the experimental data recounted above.

We can do the same analysis for other values of R besides 2. When R grows, the mixed strategy equilibrium of the persona game places more weight on the persona ∞ . This makes $P(x)$ become more weighted towards low values. In fact, when R gets larger than ~ 38.2 , the

two pure strategy NE of the persona 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 (40) on what happens as R changes.

4 Persona Games and the Prisoner's Dilemma

To illustrate the breadth of persona games, we now consider personas for a player that involve the utilities of that player's opponents. Such personas allow us to model "other-regarding preferences", like altruism and 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 elaborate this, let $\{u^j : j = 1, \dots, N\}$ be the utility functions of the original N -player concrete game. Have the persona set of player i be specified by a set of distributions $\{\rho_i\}$, each distribution ρ_i being an N -dimensional vector written as $(\rho_i^1, \rho_i^2, \dots, \rho_i^N)$. 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 the persona $\rho_i^j = \delta_{i,j}$, which equals 1 if $i = j$, 0 otherwise. "Altruism" then is a ρ_i^j that places probability mass on more than one j . ("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 $[(N-1)u^i - \sum_{j \neq i} u^j]^2$.)

As an example, consider the two-player two-move concrete game with the following utility functions:

$$\begin{bmatrix} (2, 0) & (1, 1) \\ (3, 2) & (0, 3) \end{bmatrix} \quad (4)$$

There is one joint pure strategy NE of this game, at (\mathbf{T}, \mathbf{R}) . Say that both players i in the associated persona 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 selfish (\mathcal{E}) and saint (\mathcal{A}), respectively. Under the \mathcal{E} persona, a player acts purely

in their own interests, while under the \mathcal{A} persona, they act purely in their *opponent's* interests.

As an example, if Row chooses \mathcal{E} while Col chooses \mathcal{A} , then the realized game equilibrium for the concrete game in Table 4 is (\mathbf{D}, \mathbf{L}) , since Rows' payoff there is maximal. Note that this joint move gives both players a higher utility (3 and 2, respectively) than at (\mathbf{T}, \mathbf{R}) , the realized game equilibrium when they both adopt the selfish persona. Continuing this way, we get the following pair of utility functions for the possible joint persona choices:

$$\begin{array}{c|cc}
 & \mathbf{Col} \rho & \\
 & \mathcal{E} & \mathcal{A} \\
 \hline
 \mathbf{Row} \rho & & \\
 \mathcal{E} & (1, 1) & (3, 2) \\
 \mathcal{A} & (0, 3) & (3, 2)
 \end{array} \tag{5}$$

The pure strategy NE of this persona game is $(\mathcal{E}, \mathcal{A})$, i.e., the optimal persona for Row to adopt is to be selfish, and for Col is to be a saint. Note that both players benefit by having Col be a saint. One implication is that Row would be willing to pay up to 2.0 to induce Col to be a saint. Perhaps more surprisingly, Col would be willing to pay up to 1.0 to be a saint, i.e., to be allowed to completely ignore their 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 i can choose either the selfish persona, or a "charitable" persona, under which ρ_i is uniform (so that player i has equal concern for their own utility and for their opponent's utility). Then for the PD concrete game in Table 1, the equilibrium of the persona game is for both players to be charitable, a choice that leads them to cooperate in the realized game (see supplemental information). Note that they do this for purely self-centered reasons, in a game they play only once. This result might account for some of the experimental data showing a substantial probability for real-world humans to cooperate in such single-play games (45).

To investigate the breadth of this PD result, consider the fully general, symmetric PD con-

crete game, with utility functions

$$\begin{bmatrix} (\beta, \beta) & (0, \alpha) \\ (\alpha, 0) & (\gamma, \gamma) \end{bmatrix} \quad (6)$$

where (\mathbf{R}, \mathbf{D}) is (defect, defect), so $\alpha > \beta > \gamma > 0$. Also consider the fully general charitable persona, \mathcal{C} , where $\rho_i = s$ for both players i . So \mathcal{E} is $s = 1$, and \mathcal{A} is $s = 0$. We are interested in what happens if the persona sets of both players is augmented beyond the triple {fully rational persona \mathcal{E} , the irrational persona, the anti-rational persona} that was investigated above to also include the \mathcal{C} persona, for some fixed value of s . (See (46) for analysis of the case where the two players have different values of s .)

Working through the algebra (see the supplemental information), we first see that neither the non-rational nor the antirational persona will ever be chosen. We also see that for joint cooperation in the realized game (i.e., (\mathbf{L}, \mathbf{T})) to be a NE under the $(\mathcal{C}, \mathcal{C})$ joint persona choice, we need $R_1 \equiv \beta - s\alpha > 0$ (see the supplemental information). If instead $R_1 < 0$, then under the $(\mathcal{C}, \mathcal{C})$ joint persona either player i would prefer to defect given that $-i$ cooperates. Note that R_1 can be viewed as the robustness of having joint cooperation be the NE when both players are charitable. The larger R_1 is, the larger the noise in utility values, confusion of the players about utility values, or some similar fluctuation would have to be to induce a pair of charitable players not to cooperate.

Given that $R_1 > 0$, we then need $R_2 \equiv \gamma - (1 - s)\alpha > 0$, to ensure that each player prefers the charitable persona to the selfish persona whenever the other player is charitable. R_2 can also be viewed as a form of robustness, this time of the players both wanting to adopt the charitable persona in the first place.

Combining, we see that $(\mathcal{C}, \mathcal{C})$ followed by (\mathbf{L}, \mathbf{T}) is an equilibrium whenever $s \in (1 - \frac{\gamma}{\alpha}, \frac{\beta}{\alpha}]$. For that range on allowed s 's to be non-empty requires that $\gamma > \alpha - \beta$. Intuitively, this means that player i 's defecting in the concrete game provides a larger benefit to i if player $-i$ also

defects than it does if $-i$ cooperates. It is interesting to compare these bounds on α, β and γ to analogous bounds, discussed in (10), that determine when direct reciprocity, group selection, etc., can result in joint cooperation being an equilibrium of the infinitely repeated PD.

At the NE of the concrete game in Table 6, both players defect, and each player's utility is γ . So when we do have $(\mathcal{C}, \mathcal{C})$ followed by (\mathbf{L}, \mathbf{T}) , the benefit to each player of playing the persona game rather than playing the concrete game directly is $B \equiv \beta - \gamma$. Comparing this to the formulas for R_1 and R_2 , we see that $R_1 + R_2 + B \leq 1$. This proves that there are unavoidable tradeoffs between the robustness of cooperation and the potential benefit of cooperation in the PD, whenever (as here) the concrete game matrix is symmetric and both players can either be selfish or charitable for the same value of s .

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

It is interesting to note the implications of this for the "prisoner's dilemma" of a marriage. Having R_2 large means that your spouse must pay significant attention to their own interests as well as yours. It also means that your spouse must benefit substantially by punishing you if you defect. Having R_1 large means that you can't benefit too much by defecting when your spouse cooperates. If both R_1 and R_2 are large, then fluctuations in behavior or perceptions are unlikely to break the joint-cooperation outcome. Our result shows that there is an unavoidable tradeoff between having those values be large and also having there be a large benefit to joint cooperation in the marriage.

5 Discussion

The persona framework goes beyond explaining some types of currently known non-rational behavior, to reveal previously unknown situations in which non-rationality is in a player's best interests. An extreme example is where a player benefits by adopting the persona of being her "own worst enemy" (i.e., by committing to always act to *minimize* utility). There are even simple concrete games where *all* players benefit from adopting this **anti-rational** persona, no matter what persona their opponents adopt. The NE of the associated persona game is for all players to choose to be anti-rational. Furthermore, for some such games, *every* persona player i receives higher utility under that all-anti-rational NE of the persona game than they would if all players instead adopted the persona of full rationality. In such games, every individual would prefer it if everyone (herself included) is their own worst enemy. Translated to the real world, this means that sometimes a governmental regulator should try to induce each player to act precisely against their own interests, since by doing that the player benefits both them and everyone else. An example of such a game is provided in the supplemental information.

As an illustration of a potential practical application of the persona phenomenon, note that many modern engineered systems can be viewed as a distributed set of adaptive, goal-directed subsystems. Often the equilibrium behavior of such a system can be modeled as the NE of a game where the players are those subsystems. Typically in such cases the system designer can set some aspects of the utility functions of the "players" (i.e., some aspects of the goals of the subsystems) 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 (47, 48). In other instances of such engineered systems some of the players are human beings. Examples here

include air-traffic management (49), multi-disciplinary optimization (50, 51), and in a certain sense, much of mechanism design, e.g., design of auctions (4, 52).

The implications of the analysis concerning Table 2 predicts 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). Moreover, the analysis of the game in the supplemental information predicts that some players — perhaps all of them — would sometimes improve their performance if they were induced to always be anti-rational (e.g., by appropriate transformation of their reward signals from their environment).

There are many interesting connections between persona games and real world phenomena. For example, a necessary condition for a real-world player to adopt a persona other than perfect rationality is that they believe that the other players are aware that they can do that. The simple computer programs for maximizing utility that are currently used in game theory experiments do not have such awareness. Accordingly, if a human knows they are playing against such a program, they should always play perfectly rationally, in contrast to their behavior when playing against humans. This distinction between behavior when playing computers and playing humans agrees with much experimental data, e.g., concerning the Ultimatum Game (1, 2, 53).

What happens if the players in a persona game are unfamiliar with the meaning of each others' signals, say due to coming from different cultures? This might lead them to misconstrue the personas (or more generally persona sets) adopted by one another. Intuitively, one would expect that the players would feel frustrated when that happens, since in the realized game they each do what would be optimal if their opponents were using that misconstrued persona — but their opponents aren't doing that. This frustration can be viewed as a rough model of what is colloquially called a “culture gap” (54).

Broadening the discussion beyond humans, note that calculating a persona equilibrium typically involves far more computational work than calculating the equilibria of the associated

concrete game. (Crudely speaking, 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 organisms that can also play persona games. Colloquially speaking, we might characterize a member of such a species who plays persona games well as having “high social intelligence”.

Also for computational reasons, one would expect the persona set of any social animal for any concrete game not to be too large. This is because a large set both increases the computational burden on the player with that set, and on the other players they play against.

Indeed, computational issues might prevent a social animal from calculating the optimal persona from some associated persona set, even a limited persona set, for every concrete game they encounter. (Just think about how many games you play during a typical day, and imagine calculating the precisely optimal persona for every such game.) Rather they might use a simple rule to map any pair {a concrete game, a specification of which player they are in that game} to a persona for that game. As an example, a value for the altruism N -vector ρ can be used to map every N -player concrete game a person might play to a persona for them to adopt for that game. We call such a map a “personality” (see the supplemental information).

Summarizing, persona games provide a very simple justification for irrationality with very broad potential applicability. They also make quantitative predictions that can often be compared with experimental data. (In work currently being written for submission, two of us has found that the predictions of the persona game framework also agree with experimental data for the Ultimatum Game (37).) While here we have only considered personas involving degrees of rationality and degrees of altruism, there is no reason not to expect other kinds of persona sets in the real world. Risk aversion, uncertainty aversion, reflection points, framing effects, and all the other “irrational” aspects of human behavior can often be formulated as personas.

Even so, persona games should not be viewed as a candidate explanation of all non-rational behavior. Rather they are complementary to other explanations, for example those involving sequences of games (like EOS and EOP). Indeed, many phenomena probably involve sequences of persona games (or more generally, personality games). As an illustration, say an individual i repeatedly plays a face-to-face persona game γ involving signaling, persona sets, etc., and adopts persona distribution $P(b_i)$ for those games. By playing all those games i would grow accustomed to adopting $P(b_i)$. Accordingly, if i plays new instances of γ where signaling is prevented, they might at first continue to adopt distribution $P(b_i)$. However as they keep playing signal-free versions of γ , they might realize that $P(b_i)$ no makes sense. This would lead them to adopt the fully rational persona instead. If after doing that they were to play a version of γ where signaling was no longer prevented, they could be expected to return to $P(b_i)$ fairly quickly. This behavior agrees with experimental data (55, 56).

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

References

1. C. Camerer, *Behavioral Game theory: experiments in strategic interaction* (Princeton University Press, 2003).
2. C. Camerer, E. Fehr, *Science* **311**, 47 (2006).
3. D. Kahneman, *American Economic Review* **93**, 1449 (2003).
4. R. B. Myerson, *Game theory: Analysis of Conflict* (Harvard University Press, 1991).

5. D. Fudenberg, D. K. Levine, *The Theory of Learning in Games* (MIT Press, Cambridge, MA, 1998).
6. C. Hauert, A. Traulsen, H. Bradt, M. Nowak, K. Sigmund, *Science* **316**, 1905 (2007).
7. R. Trivers, *Natural Selection and Social Theory: Selected Papers* (Oxford University Press, 2002).
8. S. Bowles, R. Boyd, E. Fehr, H. Gintis, *The Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life* (MIT Press, 2005).
9. O. Gurerk, B. Irlenbusch, B. Rockenbach, *Science* **312**, 108 (2006).
10. M. A. Nowak, *Science* **314**, 1560 (2006).
11. E. Fehr, U. Fischbacher, *Nature* **425**, 785 (2003).
12. L. Keller, H. K. Reeve, *Nature* **394**, 121 (1998).
13. J. McNamara, Z. Barta, L. Fromhage, A. Houston, *Nature* **451**, 189 (2007).
14. A. Dreber, D. Rand, D. Fudenberg, N. M., *Nature* **452**, 348 (2008).
15. S. Huck, J. Oechssler, *Games and economic behavior* **28**, 13 (1996).
16. A. Heifetz, C. Shannon, Y. Spiegel, *Economic Theory* **32**, 251 (2007).
17. L. E. Samuelson, *Journal of Economic Theory* **97** (2001). Special issue on "Evolution of Preferences".
18. J. Haldane, *New Biology* **18**, 34 (1955).
19. W. Hamilton, *American Naturalist* **97**, 354 (1963).

20. D. Kreps, P. Milgrom, J. Roberts, R. Wilson, *J. Econ. Theory* **17**, 245 (1982).
21. D. Fudenberg, J. Tirole, *Game Theory* (MIT Press, Cambridge, MA, 1991).
22. S. von Widekind, *Evolution of non-expected utility preferences* (Springer, 2008).
23. R. Frank, *The American Economic Review* **77**, 593 (1987).
24. C. Frith, *Philosophical Transactions of the Royal Society B* (2008).
Doi:10.1098/rstb.2008.0005.
25. M. Spence, *The quarterly journal of Economics* **87**, 355 (1973).
26. M. Spence, *The review of Economic Studies* **44**, 561 (1977).
27. M. Lachmann, C. Bergstrom, S. Szamado, *Proceedings of the National Academy of Sciences* **98**, 13189 (2001).
28. S. Stern, *Financial Times* (2008). March 17.
29. W. Raub, T. Voss, *Social institutions, their emergence, maintenance and effects*, M. Hechter, K.-D. Opp, R. Wippler, eds. (Walter de Gruyter Inc., 1990).
30. H. Kissinger, *Nuclear Weapons and Foreign Policy* (Harper and Brothers, 1957).
31. T. Schelling, *The strategy of conflict* (Harvard university press, 1960).
32. E. Dekel, J. Ely, Y. O., *Review of Economic Studies* **74**, 685 (2007).
33. W. Guth, M. Yaari, *Internatinal Journal of Game Theory* **24**, 323 (1995).
34. H. Bester, W. Guth, *Journal of Economic Behavior and Organization* **34**, 193 (2000).
35. D. Fogel, G. Fogel, P. Andrews, *BioSystems* **44**, 135 (1997).

36. S. Ficici, O. Melnik, J. Pollack, *IEEE Trans. on Evol. Comp.* **9**, 580 (2005).
37. D. Wolpert, M. Harre, It can be smart to be dumb (2008). In preparation.
38. J. Henrich, et alia, *Science* **312**, 1767 (2006).
39. K. Basu, *Scientific American* (2007).
40. C. M. Capra, J. K. Goeree, R. Gomez, C. H. Holt, *American Economic Review* **19**, 678 (1999).
41. K. Basu, *American Economic Review* **84**, 391 (1994).
42. A. Rubinstein, Instinctive and cognitive reasoning: a study of response times (2004). Ariel-rubinstein.tau.ac.il/papers/Response/pdf.
43. T. Becker, M. Carter, J. Naeve, Experts playing the traveler's dilemma (2005). Universitat Hohenheim Nr. 252/2005.
44. J. K. Goeree, C. A. Holt, *Proceedings National Academy of Sciences* **96**, 10564 (1999).
45. A. Tversky, *Preference, Belief, and Similarity: Selected Writings* (MIT Press, 2004).
46. J. Jamison, D. H. Wolpert, Persona games and altruism (2008). In preparation.
47. J. Ferber, *Foundations of Distributed Artificial Intelligence*, G. O'Hare, N. Jennings, eds. (John Wiley and Sons, 1996), pp. 287–314.
48. J. Shamma, G. Arslan, *IEEE Trans. on Automatic Control* **50**, 312 (2004).
49. H. Hwang, J. Kim, C. Tomlin, *Air Traffic Control Quarterly* (2007). In press.
50. E. Cramer, J. Dennis, et alia, *SIAM J. of Optimization* **4** (1994).

51. S. Choi, J. Alonso, *Proceedings of 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference* (2004). AIAA Paper 2004-4371.
52. N. Nisan, A. Ronen, *Games and Economic Behavior* **35**, 166 (2001).
53. M. Nowak, K. Page, K. Sigmund, *Science* **289.5485**, 1773 (2000).
54. S. Chuah, R. Hoffman, M. Jones, G. Williams, *Journal of Economic Behavior and Organization* **64**, 35 (2007).
55. R. Cooper, D. DeJong, R. Forsythe, T. Ross, *Games and Economic Behavior* **12**, 187 (1996).
56. R. Dawes, R. Thaler, *J. Econ. Perspectives* **2**, 187 (1988).