How bounded rationality can be optimal

David H. Wolpert,[1,4] Julian Jamison[2] and David Newth[3]

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

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

[3]*CSIRO Centre for Complex Systems Science, Gungahlin Homestead, Crace, ACT, 2611, AUSTRALIA*

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

A long-standing puzzle in economics and biology is why humans and animals sometimes exhibit "bounded rationality" by seeming not to adopt their optimal strategy when interacting with others [1–3]. Many previous explanations involve repeated games, combined with mechanisms like reputation effects, punishment, "loners", negotiation and genetic evolution. [1,4–6,6–13]. Here we present a new kind of explanation. We start by observing that often an individual will adopt a counterfactual "persona" for an interaction. We formalize that as the individual's adopting a counterfactual utility function for the interaction. By changing what persona she adopts, the individual may change the behavior of others in the interaction. In particular, we prove that sometimes by adopting a "bounded 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 "bounded rational". Unlike many previous explanations of bounded rationality, ours can be analyzed in closed form without computer simulations. This analysis predicts that in a single, non-repeated play of certain versions of the Prisoner's Dilemma (PD), every individual would benefit by adopting a cooperative persona rather than a rational persona, regardless of the persona adopted by

the individual's opponent. Furthermore, there is an unavoidable tradeoff between the robustness of that cooperation and the benefit of the cooperation. This analysis also predicts how much an individual would be willing to pay others for the right to be altruistic when interacting with them. On a broader scale, our framework provides a way to formalize "culture gaps", and predicts that groups sometimes benefit by having some members be "anti-social". Adopting an engineering perspective, our framework has implications for how to design mechanisms that regulate groups of interacting individuals.

While providing many insights, explanations for bounded rationality based on repeated game have several limitations. Each of them 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. Such sequences do not occur in the real world. Also none of these explanations applies to bounded rationality in general, but rather to only one type of bounded rationality, like altruism. Another limitation is that these explanations often require 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. Even given such limitations though, these explanations are often so complicated that computer simulations are needed to investigate them.

Here we introduce a simple, broadly applicable framework that explains bounded rationality without any of these limitations. Our framework springs from 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". 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, often choosing those apparent preferences semi-consciously.

Why do people do this? To answer this question, say that all players in a game are free to choose such a persona before the start of play. Then as shown in the framework we introduce below, for many games adopting a persona that is bounded rational (acting dumb, or being altruistic for instance) actually results in larger expected utility when the game is played than does adopting a fully rational persona. So "bounded rationality" can actually be utility-maximizing, for a game played anonymously, without concern for possible future games. In particular, [14] is correct that "cooperation can (be explained), even among non-kin, in situations devoid of repeat interaction". However our framework shows that there is not even a need to invoke punishment and genes for altruism (which have not been found on the human chromosome) in such an explanation. Cooperation can exist for purely self-interested reasons.

Our framework goes beyond explaining currently known bounded rational behavior however, to reveal previously unexplored situations in which bounded rationality may be in a player's best interests. An extreme example presented below is where a player benefits by adopting the persona of being her "own worst enemy" (i.e., by committing to always act to *minimize* her utility). There are even simple games where *all* players benefit from adopting this own-worst-enemy persona. In such scenarios, no player would want to "defect" to the rational persona, whatever the personas of her opponents. Moreover, for some such games, each player i would have higher expected utility if everyone adopted the own-worst-enemy persona than i would if everyone were fully rational. Translated to the real world, this means that sometimes a governmental regulator should try to induce each player to act precisely against her own interests, since by doing that the player benefits both herself and everyone else.

To introduce our framework, 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. For example, in one version of 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 is 6 and Col's utility is 0.

To play a game each player $i \in$ {Row, Col} independently chooses a "mixed strategy", i.e., a probability distribution $P_i(x_i)$ over her 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_{Row}, P_{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 her mixed strategy, in light of her opponent's mixed strategy. If either player violates this condition, she is said to exhibit "bounded rationality".

For 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 (**D**, **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.

Now say that rather than being rational in the PD, Col were perfectly irrational. That is,

she commits to choosing uniformly randomly between the two moves, with no evident concern for the resultant value of her 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 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 joint move $(\mathbf{T}, \mathbf{R})$ 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 even more extreme types of real-world bounded rationality. Imagine that Col is a chick in a nest, and Row is Col's parent. Col can either leave the nest to get a seed to eat (**R**) or stay in the nest (**L**). Row can either bring five seeds to the nest (**D**) or none (**T**). Col's utility function is the total number of seeds brought to the nest. In contrast, Row wants Col to have some seeds, but not too many (so that hunger may force Col to fledge and leave the nest for good). This game is captured by the following (u_{Row}, u_{Col}) table:

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

As before, $(\mathbf{T}, \mathbf{R})$ is the sole NE 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, i.e., she always chooses the strategy that *minimizes* $\mathbb{E}(u_c)$ (given the strategy of Row), rather than maximizes it. Since **R** gives Col higher utility for either of of Row's moves, this anti-rationality means that Col always chooses **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).

In this scenario, it benefits Col to develop in her short life what would appear to an outside observer to be a maladaptive fear of seeds in the nest. (Physically, such a fear could manifest itself in many ways, e.g., as blood stress hormone levels.) More precisely, it benefits Col to develop a fear that would not interfere with Col's eating seeds once they are in the nest, but would make her try to minimize the number of seeds in the nest in the first place. Such a fear would prevent Col from bringing seeds to the nest. That would in turn force (rational) Row to bring five seeds to the nest for Col to eat. In contrast, if Col did not develop the fear and was purely rational, Col would choose to bring seeds to the nest, for either of Row's moves.

Knowing that, and being rational, Row would bring no seeds, and so Col would only get a single seed to eat. So it benefits Col not to be rational.

Note that the fear of seeds makes Col less happy the more seeds are brought to the nest. This is the exact negative of Col's true utility function (which increases with the number of seeds), which is why we call the fear 'anti-rational'. This example suggests a way to model persona-based phenomena in general. Say we have an N-player **concrete** game, where the joint pure strategy space is X and the utility functions of the players, defined over X, are $\{u^i : i = 1, \ldots, N\}$. Rather than play the concrete game directly, we hypothesize that the players engage in a sequence of two other, related games, called the **persona** game and **realized game**, respectively. In the first, persona game, every player i chooses what **persona** to adopt, i.e., she chooses a counterfactual utility function over X from her **persona set** of possible such functions. (For example, one persona in the chick's persona set is the negative of her concrete game utility function.) By all choosing a persona, the players jointly construct the realized game: It is the game with joint strategy space X played with the chosen personas (rather than with the concrete game utility functions). So once the personas are all chosen the realized game can be played. The associated NE provides the physically implemented joint strategy of the players, $P(x \in X)$. Finally, the expected utility of each player i of the persona game is given by evaluating the expectation of u^i according to $P(x)$. So the goal of every player i in the persona game is to choose her persona that, when combined with the personas chosen by everyone else, results in a realized game whose NE is as good as possible for i.

As an example of this framework, (see supplementary material §1) we show that for some concrete games *all* the associated persona game players prefer the anti-rational persona over the full rationality persona, no matter what personas are chosen by their opponents. So the NE of

that persona game is for all players to choose to be their own worst enemy. Furthermore, for some such games, *every* persona player *i* receives higher utility under that own-worst-enemy NE of the persona game than she 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.

As an illustration of a potential practical application of the phenomena discussed so far, 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 [15,16]. In other instances of such engineered systems some of the players are human beings. Examples here include air-traffic management [17], multi-disciplinary optimization [18,19], and in a certain sense, much of mechanism design, e.g., design of auctions [4,20].

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). The analysis concerning Table 3 and the game presented in supplementary material §1 predicts that some of the players — perhaps all of them — would sometimes improve their performance if they were induced to always be their own worst enemy (e.g., by appropriate transformation of their reward signals from their environ-

ment).

We now consider personas for a player that involve the utilities of her 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 illustrate this, let $\{u_j : j = 1, \ldots N\}$ 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 $(\rho_i^1, \rho_i^2, \ldots, \rho_i^N)$. 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 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:

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

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 ($\mathscr{E}$) and saint ($\mathscr{A}$), respectively. Under the $\mathscr{E}$ persona, a player acts purely in her own interests, while under the $\mathscr{A}$ persona, she acts purely in her *opponent*'s interests.

As an example, if Row chooses $\mathscr{E}$ while Col chooses $\mathscr{A}$, then the realized game equi-

librium 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:

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

(5)

The pure strategy NE of this persona game is $(\mathscr{E}, \mathscr{A})$, 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 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 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 her own utility and her 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 supplementary material §2). 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 [21].

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

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

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

Working through the algebra, we first see that neither the irrational 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 $(\mathscr{C},\mathscr{C})$ joint persona choice, we need $R_1 \equiv \beta - s\alpha > 0$ (see supplementary material §2). If instead $R_1 < 0$, then under the $(\mathscr{C},\mathscr{C})$ joint persona either player i would prefer to defect given that $-i$ cooperates. 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 $(\mathscr{C}, \mathscr{C})$ 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, both players defect, and each player's utility is γ. So when we do have $(\mathscr{C}, \mathscr{C})$ 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 $R_1 + R_2 + B \leq 1$. So we have proven that when (as here) the concrete game matrix is symmetric and both players can either be selfish or charitable for the same of s, there are unavoidable tradeoffs between the of robustness of cooperation and the potential benefit of cooperation.

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

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 attention to her 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. There is an unavoidable tradeoff between meeting those conditions and having a large benefit

to the marriage of joint cooperation rather than joint defection

There are many connections between persona games and real world phenomena. In the remainder of this paper we sketch a few of them. First, note that a 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 programs for maximizing utility that are currently used in game theory experiments do not have such awareness. Accordingly, if a human knows she is playing against such an program, 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 Ultimatum Game [1,2,22].

When the players know they're all humans, and so play a persona game, it seems that they often play the game semi-consciously. In such scenarios the persona a player adopts is might be what is colloquially called an "emotion", or "mood", with her persona set being the set of all emotions she sometimes adopts [23]. If when playing a particular game she randomly chooses her persona from a distribution over her persona set (i.e., has a mixed strategy in the persona game), then we might say that she is "moody" or "capricious".

However when the players come from different cultures their persona games might involve their cultures at least as much as their emotions. To illustrate this, start with the usual economics presumption that the utility function u_i of any human i is independent of other humans, of where i grew up, and similar factors (it is in essence "pre-fixed", perhaps in i's genes). Now consider a Medieval entrepot, where many individuals play trading games with one another. By our presumption, the distribution of utility functions across those players is independent of where they came from. Yet we might expect that the preferences, values, and bargaining styles guiding

those traders in their interactions will typically cluster into distinct groups, groups that match the society of origin of the traders. After all, typically any particular trader would find it difficult to convincingly adopt a bargaining style not found in her society of origin — such a bargaining style would violate the cultural norms of where she grew up, of "who she is".

Each such cluster is a different persona set, which colloquially speaking corresponds to a "culture". The 'preferences, values, and bargaining styles' of each particular trader at the entrepot is a different persona from her persona set. Such a persona set of a player is made common knowledge with the other traders by factors like her name, dress, language.

In this example, since the traders know so little about one another's cultures, the NE of a persona game may only arise over a long timescale. To illustrate this, say a trader from Fillia plays several games with a trader from Scillia. Over the course of those games the Fillian learns to modify her distribution over bargaining personas, as much as anyone from Fillia easily can (i.e., within the bounds of her persona set), to "optimally match" the bargaining persona the Scillian adopts. Similarly, the Scillian will learn how to modify her distribution over bargaining personas to match the bargaining persona distribution of the Fillian. Once those learning processes are carried to completion, so that the Scillian and Fillian have a final (distribution of) personas from their persona sets, the persona players have found a NE.

Given that persona game NE, just before the Fillian and Scillian start a new bargaining game each of them (perhaps semi-consciously) samples her persona distribution, and signals the resultant persona sample to her opponent. This signaling occurs as the traders greet each other, on a comparatively short timescale, via phenomena like body language and language inflection. Next, based on the signaled persona of her opponent together with her own persona, each of the players uses the usual game-theoretic reasoning to determine what strategy to use in the

(realized version of the) bargaining game, which is then finally played.

What happens if the players in the entrepot persona game 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" [24].

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 she plays 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 she encounters. (Just think about how many games you play during a typical day, and imagine calculating the precisely optimal persona for every such game.) Rather she might use a simple

rule to map any pair {a concrete game, a specification of which player she is in that game} to a persona for that game. As an example, a value for the altruism N-vector ρ can be used to map every N-player concrete game a person might play to a persona for her to adopt for that game. We call such a map a "personality".

This definition of personalities can be used to make quantitative predictions about many aspects of human societies. It should be possible to compare those predictions to anthropological data, e.g., to an extended version of the data in [14,25]. One intriguing prediction is presented in supplementary material §3. It suggests that sometimes social welfare is higher if there are many personalities found in the members of a society. In such a situation, the society as a whole would *need* a range of personalities — potentially including anti-social, psychopathic and other "dysfunctional" bounded rational personalities — in order to function most effectively.

Ideas related to the persona game concept have been discussed at least since the 1950's [26–28], and may even extend back to antiquity [28]. Unfortunately this earlier work is too informal and stylized to provide quantitative predictions. (See the end of supplementary material §1 for a brief summary of what issues a fully formal approach must address.) Much of this work also implicitly imposes a huge computational burden on the players by allowing infinite persona sets, which rules it out as an explanation of real-world behavior. There are other more formal studies that appear related to the persona concept [23,29–31]. However in truth each of these studies applies a model of very limited scope to a restricted (and often rather complicated) scenario (e.g., [23,29]). For example, many of these studies focus so narrowly on the PD that their results do not apply to other types of bounded rationality. As a result, these studies do not concern the persona concept in full generality.

Perhaps the closest any formal earlier work comes to the persona concept is the work on

"evolution of preferences" [32–34]. However the evolution of preferences work invokes repetitions of the exact same game, extending potentially infinitely into the future, a type of sequence that does not occur in the real world. Also, much of the evolution of preferences work restricts attention to a game (or set of coupled games) with a single symmetric utility function shared by all the players, which is very rare in the real world. In addition, heedless of real-world geographical constraints, this work typically requires that *all* individuals in a population interact, even when the population is infinite. Furthermore, some evolution of preferences games have no equilibria. Another problem is that the results in this work typically vary with the initial characteristics of the population. None of these difficulties apply to the persona games framework (which does not require a population of players).

Summarizing, persona games provide a very simple justification for bounded rationality that is potentially applicable to any type of bounded rationality. They also make quantitative predictions that can often be compared with experimental data. (In work currently being written for submission, one of us has found that the predictions of the persona game framework agree with experimental data both for the Ultimatum Game and for the Traveler's Dilemma [35,36]. 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 bounded rational phenomena. Rather they are complementary to explanations based on mechanisms like kin selection and reciprocal altruism.

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

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

2. Camerer, C. & Fehr, E. When does economic man dominate social behavior? Science **311**, 47–52 (2006).

3. Kahneman, D. Maps of bounded rationality: Psychology of behavioral economics. American Economic Review **93**, 1449–1475 (2003).

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

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

6. Hauert, C., Traulsen, A., Bradt, H., Nowak, M. & Sigmund, K. Via freedom to coercion: the emergence of costly punishment. Science **316**, 1905–1907 (2007).

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

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

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

10. Nowak, M. A. Five rules for the evolution of cooperation. Science **314**, 1560–1563 (2006).

11. Fehr, E. & Fischbacher, U. The nature of human altruism. Nature **425**, 785–791 (2003).

12. Keller, L. & Reeve, H. K. Familiarity breeds cooperation. Nature **394**, 121–122 (1998).

13. McNamara, J., Barta, Z., Fromhage, L. & Houston, A. The coevolution of choosiness and cooperation. Nature **451**, 189–192 (2007).

14. Henrich, J. & et alia. Costly punishment across human societies. Science **312**, 1767 – 1770 (2006).

15. Ferber, J. Reactive distributed artificial intelligence: Principles and applications. In O-Hare, G. & Jennings, N. (eds.) Foundations of Distributed Artificial Intelligence, 287–314 (John Wiley and Sons, 1996).

16. Shamma, J. & Arslan, G. Dynamic fictitious play, dynamic gradient play, and distributed convergence to nash equilibria. IEEE Trans. on Automatic Control **50**, 312–327 (2004).

17. Hwang, H., Kim, J. & Tomlin, C. Protocol-based conflict resolution for air traffic control. Air Traffic Control Quarterly (2007). In press.

18. Cramer, E., Dennis, J. & et alia. Problem formulation for multidisciplinary optimization. SIAM J. of Optimization **4** (1994).

19. Choi, S. & Alonso, J. Multi-fidelity design optimization of low-boom supersonic business jet. In Proceedings of 10th AIAA/ISSMO Multidisciplany Analysis and Optimization Conference (2004). AIAA Paper 2004-4371.

20. Nisan, N. & Ronen, A. Algorithmic mechanism design. Games and Economic Behavior **35**, 166–196 (2001).

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

22. Nowak, M., Page, K. & Sigmund, K. Fairness versus reason in the ultimatum game. Science **289.5485**, 1773 (2000).

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

24. Chuah, S., Hoffman, R., Jones, M. & Williams, G. Do cultures clash? evidence from cross-national ultimatum game experiments. Journal of Economic Behavior and Organization **64**, 35–48 (2007).

25. Henrich, J. et al. Cooperation, reciprocity and punishment in fifteen small-scale societies. American Economic Review **91** (2001).

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

27. Kissinger, H. Nuclear Weapons and Foreign Policy (Harper and Brothers, 1957).

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

29. Israeli, E. Sowing doubt optimally in two-person repeated games. Games and Economic Behavior **28**, 203–216 (1996).

30. Becker, G. Altruism, egoism and genetic fitness: economics and sociobiology. J. Econ. Lit **14**, 817 (1976).

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

32. Huck, S. & Oechssler, J. Games and economic behavior **28**, 13–24 (1996).

33. Heifetz, A., Shannon, C. & Spiegel, Y. The dynamic evolution of preferences. Economic Theory **32**, 251–286 (2007).

34. Samuelson, L. E. Journal of Economic Theory **97** (2001). Special issue on ”Evolution of Preferences”.

35. Capra, C. M., Goeree, J. K., Gomez, R. & Holt, C. H. Anomalous behavior in a traveler's dilemma game. American Economic Review **19**, 678–690 (1999).

36. Goeree, J. K. & Holt, C. A. Stochastic game theory: for playing games, not just doing theory. Proceedings National Academy of Sciences **96**, 10564–10567 (1999).

How bounded rationality can be optimal: Supporting Material

David H. Wolpert,[1*] Julian Jamison[2] and David Newth[3]

[1]MS 269-1, NASA Ames Research Center,
Moffett Field, CA, 94035, USA
[2]USC Brain and Creativity Institute, 3620 McClintock Ave, Suite 250,
Bldg SGM, MC 1061, Los Angeles, CA 90089
[3]CSIRO Centre for Complex Systems Science
Gungahlin Homestead, Crace, ACT, 2611
[*]To whom correspondence should be addressed; E-mail: david.h.wolpert@nasa.gov

1 Anti-rational persona

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

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

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 persona game:

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

(2)

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

The joint persona $(-\infty, -\infty)$ 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 persona, neither player would benefit from changing to rational behavior, no matter which persona her opponent adopted. Note in particular that $(+\infty, +\infty)$ is not a NE of the persona game. So if the players are sophisticated enough to play the persona game with each other rather than the concrete game, they will both act anti-rationally, and will thereby both benefit.

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

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

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

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

(4)

This pair of utility matrices is just the PD (up to irrelevant rescalings, etc.) with "defect-defect" identified as $(-\infty, -\infty)$, and "cooperate-cooperate" as $(+\infty, +\infty)$. So for this 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.

It is important to realize that in both of these persona games the benefit to a player of being anti-rational is not the same as the benefit that "non-credible threats" can provide in certain extensive form games [1]. A player making a non-credible threat says to her opponent, "If you do α, I'll do something that hurts me — but also hurts you. So you must not do α, and I will exploit that". In contrast, an anti-rational player says "*No matter what you do*, I'll do something that will hurt me".

To properly analyze such distinctions a fully formal definition of persona games is needed. There are many subtleties that such a formalism must address. For example, for some concrete games and persona sets, some joint personas result in a realized game that has more than one NE. To define the associated utility functions in the persona game, either those multiple NE must be somehow summarized or one of them must be somehow selected. Another subtlety arises if we try to specify a counterfactual utility function that player i can adopt such that at the associated NE of the realized game player i is irrational, playing all her moves with equal probability: in general there is *no* such counterfactual utility function. (In particular, the uniform utility function allows any realized game mixed strategy, rather than picking out the uniform one.) This means that to include the possibility of the irrational persona, we must define personas using a more flexible type of preference order than utility functions. A third subtlety is that often some type of enforcement mechanism is needed to force a real world player to actually adopt the persona she signals to the others, i.e., some sort of costly signaling is needed. This enforcement mechanism will distort the persona game NE in general. These issues are addressed in various ways, and the existence of associated persona game equilibria guaranteed, in [2, 3].

2 Prisoner's Dilemma

We now consider the general Prisoner's Dilemma (PD) concrete game, parameterized as

$$\begin{bmatrix} (\beta,\beta) & (0,\alpha) \\ (\alpha,0) & (\gamma,\gamma) \end{bmatrix} \tag{5}$$

with $\alpha > \beta > \gamma > 0$. Thus each player's first strategy is cooperation and second strategy is defection. We will explore what outcomes are possible in the corresponding persona game, where we consider persona sets that include charitable personas in addition to rational, irrational, and/or anti-rational ones. For simplicity in the analysis, if there are multiple Nash equilibria of the realized game, we presume that each player is individually "optimistic" and considers only the NE outcome that is best for her. Furthermore, we restrict attention (whenever possible) to NE of the realized game that are in pure strategies for both players. These assumptions are relaxed in [3], and more involved calculations are given in [2].

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

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

$$\begin{bmatrix} (\beta, \beta) & ((1-s)\alpha, s\alpha) \\ (s\alpha, (1-s)\alpha) & (\gamma, \gamma) \end{bmatrix} \tag{6}$$

For mutual cooperation to be a NE here, obviously we need that $R_1 \equiv \beta - s\alpha \geqq 0$. Meanwhile, if Row chooses $\mathscr{E}$ while Col chooses $\mathscr{C}$, we end up in the following realized game:

$$\begin{bmatrix} (\beta, \beta) & (0, s\alpha) \\ (\alpha, (1-s)\alpha) & (\gamma, \gamma) \end{bmatrix} \tag{7}$$

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

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

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

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

$$\begin{bmatrix} (\beta,\beta) & ((1-s_1)\alpha, s_2\alpha) \\ (s_1\alpha,(1-s_2)\alpha) & (\gamma,\gamma) \end{bmatrix} \tag{8}$$

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

3 Personality Games

We can define "personality games" in a manner similar to "persona games". The major difference is that in personality games there are multiple concrete games rather than just one. This makes the definition intrinsically more complicated.

Say that a society consists of a set of N individuals, labeled $i \in \{1, \ldots, N\}$. Call a **game scenario** the combination of an N-player concrete game Γ and a mapping f assigning each of the N individuals to be a separate one of the N players in that game. So for individual i, a personality is a mapping that takes any game scenario (Γ, f) to a persona for i to adopt when playing as player $f(i)$ in a persona game associated with concrete game Γ. (At the expense of more notation, we can also allow for game scenarios involving games with fewer than N players.)

Say that in their daily lives people in a certain society are involved in game scenarios g that are distributed according to some $P(g)$. So for example if every individual in that society is equally likely to be any of the players in concrete game Γ, then $P(\Gamma, f) = P(\Gamma, f')$ for all f and f'.

Given all this, consider a product distribution $q(b) = \prod_{i=1}^{N} q_i(b_i)$ in which each individual i randomly chooses a personality $b_i \in B_i$ according to the distribution $q_i(b_i)$. Define $b(g)$ as the joint persona for game scenario g and joint personality b. Next define $u_i^*(b, g)$ as the expected utility for individual i at the realized game equilibrium specified by joint persona $b(g)$ and game scenario g. So any pair of distributions $q(b)$ and $P(g)$ fix a value for $\mathbb{E}_{q,P}(u_i^*)$, the average of each individual i's expected utilities at the realized game equilibria specified by game scenarios g and joint personalities b.

Finally, call $q(b)$ a "personality equilibrium" for distribution $P(g)$ if no individual i can unilaterally raise $\mathbb{E}_{q,P}(u_i^*)$ by changing $q_i(b_i)$. Intuitively, this is the Nash equilibrium in personalities, given the concrete game distribution $P(g)$ and personality sets $\{B_i\}$. At a personality equilibrium, no individual in the society would have an incentive to change her personality distribution, given the personality distribution of the other individuals in the society.

Using $P(g)$, one can calculate the expected utility of all individuals at an associated personality equilibrium $q(b)$. One can also calculate an analogous average where the individuals play the concrete games specified in each g directly, without any personas. (Or equivalently, where each B_i is replaced with a set that only contains the fully rational persona.) For some pairs $(P(g), \{B_i\})$ it will be the case that the average expected utility for playing the persona games is higher than the average for playing the concrete games directly. (For example, this is the case for the persona set and single game discussed in §1.)

In such situations, society as a whole will benefit if its members play persona games with personality sets $\{B_i\}$ rather than concrete games. One might expect the $(P(g), \{B_i\})$ pair of a typical real human society is "matched" this way, so that it benefits those societies to have its members play persona games. In particular, it might be that for certain $P(g)$'s, this matching arises for sets $\{B_i\}$ where each B_i is spread over many personalities rather than few. In extreme versions of such a situation, it would benefit society to contain some individuals whose personality set is so large that it includes "anti-social" personalities.

References

[1] Myerson, R. B. Game theory: Analysis of Conflict (Harvard University Press, 1991).

[2] Wolpert, D. & Harre, M. It can be smart to be dumb (2008). In preparation.

[3] Jamison, J. & Wolpert, D. H. Persona games and altruism (2008). In preparation.