Ray, Debraj; Vohra, Rajiv

Working Paper

Games of love and hate

Working Paper, No. 2018-8

Provided in Cooperation with:
Department of Economics, Brown University

Suggested Citation: Ray, Debraj; Vohra, Rajiv (2018) : Games of love and hate, Working Paper, No. 2018-8, Brown University, Department of Economics, Providence, RI

This Version is available at:
https://hdl.handle.net/10419/202600

GAMES OF LOVE AND HATE

Debraj Ray ⓡ Rajiv Vohra†

July 2018

Abstract. A game of love and hate is one in which a player's payoff is a function of her own action and the *payoffs* of other players. For each action profile, the associated payoff profile solves an interdependent utility system, and if that solution is bounded and unique for every profile we call the game *coherent*. Coherent games generate a standard normal form. Our central theorem states that every Nash equilibrium of such a game is Pareto optimal, in sharp contrast to the general prevalence of inefficient equilibria in the presence of externalities. While externalities in our model are restricted to flow only through payoffs there are no other constraints: they could be positive or negative, or of varying sign. We further show that our coherence and continuity requirements are tight.

1. INTRODUCTION

A *game with payoff-based externalities*, or more colorfully, a *game of love and hate*, is a strategic setting in which each player's payoff depends on her own action, and the *payoffs* of some or all of the other players. Other actions enter a player's payoff only via the payoffs they generate for other players.

Payoff-based externalities are, of course, natural in situations of altruism or envy (see, for example, Pearce 1983, Ray 1987, Kockesen, Ok and Sethi 2000, Bergstrom 1999, or Vasquez and Weretka 2016). In its purest form, we might derive our happiness or hatred directly from the *extent* to which others are enjoying themselves, and not from *how* they are doing so. But payoff-based externalities also occur in situations in which there is no love or hate as such, but there are pecuniary externalities generated by firm profits, say, via demand (Murphy, Shleifer and Vishny 1989), or in which the payoffs of others serve as reference points or aspirations for an individual (Genicot and Ray 2017).

The interacting cascade generated by interdependent payoff functions can get out of hand, leading to implosions or explosions of utility, or multiple utility solutions for some fixed action profile. Familiar Hawkins-Simon-like conditions guarantee *coherence*; i.e., a bounded utility system with unique solution for every action profile (Pearce 1983, Bergstrom 1999, Hori and Kanaya 1989). This paper directly imposes coherence. Then our setting with payoff-dependent externalities can be reduced to a standard game with payoffs derived from action profiles. We have just one main result to report:

For every coherent game of love and hate satisfying a mild continuity condition on payoff functions, every equilibrium is Pareto-optimal.

†Ray: New York University and University of Warwick, debraj.ray@nyu.edu; Vohra: Brown University, rajiv_vohra@brown.edu. Ray acknowledges funding under NSF grant SES-1629370. We thank Dilip Abreu, Ted Bergstrom, Sylvain Chassang, Peter Hammond, David Pearce and Phil Reny for helpful comments. We are especially grateful to Lucas Pahl for help with Example 6. Names are in random order, following Ray ⓡ Robson (2018). We dedicate this paper to Tapan Mitra — advisor, colleague and dear friend — on the occasion of his 70th birthday. His sense of aesthetics, minimalism and rigor has been an inspiration to us, and we hope he will find some reflection of it in these pages.

The purpose of our paper is to state, prove and discuss this theorem. It is worth mentioning here that this result is independent of the sign of the externalities. "Love" creates full efficiency — despite the fears of a coordination failure, but so does "hate," and so does any mixture of the two — a player could hate some individuals and love others, or indeed could love and hate the same individual at different points on the domain of her payoff function. It is a remarkably general result that appears to depend fundamentally on the presumption that *all* externalities are transmitted via payoffs.

But a bit more is involved. One is naturally drawn to explaining just why a game such as the Prisoner's Dilemma or the Coordination Game, which have inefficient equilibria, cannot be written as a game of love and hate. The answer is that they *can* be so written, but no matter which payoff function we use to convert such a game into a game of love and hate, either the continuity condition or the coherence condition must fail. This leads to a new and more subtle interpretation of the coherence restriction.

Finally, we are also drawn to the philosophical implications of our efficiency theorem, knowing well as we do that Nash equilibria of games with externalities are typically Pareto-suboptimal. A common and obvious criticism of the libertarian doctrine is that when externalities are involved, behavior in accordance with libertarian philosophy can lead to Pareto-inferior outcomes (Sen 1970). Of course, we agree with this position. It is nevertheless of some interest that when all externalities are "non-paternalistic," in the sense of being transmitted entirely via payoffs, a liberal cannot but be a Paretian.[1]

2. THE SETTING

The set of agents is $N = \{1, \dots n\}$. Each agent $i \in N$ has a strategy set X_i. Let $X = \prod_i X_i$. For each i, utility u_i depends on her own action x_i, and on all other utilities $u_{-i} \equiv \{u_j\}_{j \neq i}$:

$$u_i = f_i(x_i, u_{-i}). \tag{1}$$

Define the function $f : X \times \mathbb{R}^N \to \mathbb{R}^N$ by:[2]

$$f(x, u) = f_1(x_1, u_{-1}) \times \ldots \times f_i(x_i, u_{-i}) \times \ldots \times f_n(x_n, u_{-n}).$$

A collection $(N, \{X_i, u_i\})$ where $\{u_i\}$ is a utility system satisfying (1), is a *game of love and hate*. Such a game is *continuous* if f_i is continuous in u_{-i} for all i. Note that no continuity condition is imposed with respect to x_i; in fact, no topological restrictions are placed on the strategy sets.

For any x, the mapping $f(x, \cdot)$ is an instance of an "interdependent utility system," governed by n equations $h = \{h_1, \dots, h_n\}$, where for every i,

$$u_i = h_i(u_{-i})$$

is independent of u_i. Such a system is said to be *coherent* if

(i) there is $B < \infty$ such that $\|h(u)\| < B$ whenever $\|u\| \leqslant B$, where $\|\cdot\|$ is the sup norm;

(ii) the mapping h has a unique fixed point.

We will impose coherence on our model for every x; that is, on every interdependent utility system of the form $h = f(x, \cdot)$. We also ask for "reduced game coherence" on every sub-situation generated by

[1] See also Bergstrom (1970) in the special context of a distributive Lindahl equilibrium with non-malevolent agents.

[2] For $S \subseteq N$, $\mathbb{R}^S$ denotes $|S|$-dimensional Euclidean space with coordinates indexed by members of S.

holding fixed the payoffs to a subset S of players, say at $\bar{u}_S$. That is, for every action profile $\{x_j\}_{j\in N-S}$, we impose coherence on the resulting utility system on player set $N-S$ given by $h_i(u_{N-S-i}) \equiv f_i(x_i, u_{N-S-i}, \bar{u}_S)$, for $i \in N-S$. With coherence in place, we also normalize all payoffs to be nonnegative.

Coherence is our starting point. Without it we are unable to unambiguously assign a utility profile to a profile of actions. With it, we can: $f(x,.)$ has a unique fixed point for every strategy profile x; call it $u(x)$. So the system of utilities generates a well-defined normal form with payoffs $u(x)$, where $u(x) = f(x, u(x))$. Pearce (1983), Bergstrom (1999), and Hori and Kanaya (1989) provide sufficient conditions for coherence in specific cases. Vasquez and Weretka (2016) discuss implications of the lack of coherence (though they impose boundedness).

The following definitions are standard. A strategy profile x^* is a *Nash equilibrium* (or simply *equilibrium*) if for every i and action x_i,

$$u_i(x^*) \geqslant u_i(x_i, x^*_{-i}).$$

A strategy profile $x \in X$ is *Pareto optimal* or *efficient* if there does not exist $x' \in X$ such that $u(x') > u(x)$.[3]

3. EXAMPLES

Here are four examples that illustrate the concept.

Example 1. *A "genuine" game of love and hate.* For some bounded g,

$$u_i = g\left(a_i, \frac{\sum_{j\neq i} u_j}{n-1}\right) - a_i^2,$$

where a_i is an investment with return g, which is also influenced by the average payoff to others. This is like a reference point or aspiration.[4] Such a reference point might serve to inspire or frustrate investment — the cross-partials of g will determine that outcome — but the point is that actions and payoffs are affected by the payoffs of others.

Example 2. *A not-so-genuine game of love and hate.* Consider the Prisoner's Dilemma:

		Player 2	
		$\bar{x}_2$	x_2^*
Player 1	$\bar{x}_1$	c, c	b, a
	x_1^*	a, b	d, d

where $a > c > d > b$. Intuitively, this is not a game of pure payoff externalities. A player's payoff depends on the actions of her opponent and not on the payoff he derives from it. That said, it is mathematically possible to write the game as one of payoff externalities; consider any bounded continuous f_i such that for $i = 1, 2$,

$$f_i(\bar{x}_i, c) = c, f_i(\bar{x}_i, a) = b, f_i(x_i^*, b) = a, f_i(x_i^*, d) = d.$$

[3]For vectors a and b, "$a \geqslant b$" means $a_i \geqslant b_i$ in every component, "$a > b$" implies $a \geqslant b$ and $a \neq b$, and "$a \gg b$" means $a_i > b_i$ in every component.

[4]See, e.g., Ray (2006), Dalton, Ghosal and Mani (2016), Genicot and Ray (2017).

Of course, the function needs to be defined for all utility vectors but that isn't a problem. As we shall see below, though, such a representation *must* fail coherence.

Example 3. *A genuine game of love (without any love in it).* We borrow from the multiple equilibrium notion of industrialization in Rosenstein-Rodan (1943), and specifically invoke the baseline model of Murphy, Shleifer and Vishny (1989). The players are n firms, each producing a distinct good. Each good can be produced by a cottage technique $y = \ell$, where ℓ is labor, and this technique is available to a competitive fringe. The firm can instead choose a "industrial technique" in each sector, where $y = \alpha\ell - F$ for some $\alpha > 1$ and fixed cost $F > 0$. Each firm chooses a binary action: to industrialize or not.

Consumers have a utility function $\sum_i \ln(c_i)$, and so spend their income equally on the n goods. The demand curve for good i is therefore $D_i = Y/np_i$, where Y is national income. National income, in turn, equals wage income plus profit, which generates the payoff externality as follows. If m firms industrialize, each limit-prices the fringe and so:

$$Y(m) = m\left[1 - \frac{1}{\alpha}\right]\frac{Y(m)}{n} - mF + L,$$

where we've normalized wages to 1 and the labor force is L. We thus have the aggregate profits of industrializing firms affecting national income and therefore the profit of every firm, so creating a strategic complementarity.

Example 4. *A genuine game of hate (without any hate in it).* Again there are n firms. Each makes an investment x_i to generate revenue $r(x_i)$ at cost $c(x_i)$. Society (or a collective regulator) receives a payoff $\gamma(u)$ from the vector u of firm payoffs, where γ is assumed decreasing. A lower γ increases the chances that a regulation will be placed on the firms, creating a penalty $\pi(\gamma)$. The payoff for each firm is therefore given by

$$u_i = f_i(x_i, \gamma) = r(x_i) - c(x_i) - \pi(\gamma).$$

The new feature here is that we define γ on the net payoff of each firm, with everything taken into account, including the penalty. Nevertheless, our specification allows the regulator's payoffs to decline with an individual firm's overall fortunes, thereby creating a potential externality imposed by one firm on *all* firms.

4. Main Result

Our main result is:

THEOREM **1.** *Suppose that a game of love and hate is continuous, coherent and reduced-game coherent. Then every equilibrium of that game is Pareto optimal.*

Of course, externalities can result in inefficient outcomes or market failure. Game theory is replete with such examples. It turns out that restricting externalities to be payoff-based, and assuming coherence as well as reduced game coherence, is enough to show that every equilibrium is efficient. Apart from these restrictions, we assume little else. We ask for the continuity of all payoffs in the payoffs of others. We allow for both positive and negative externalities, or indeed both on different sub-regions of the domain. No assumptions are made on payoffs as a function of own actions; indeed, there is no

topological structure on action sets. No assumption is made on the curvature of payoffs as a function of actions and the payoffs of others.

The context in which this theorem might be easiest to understand is a game with strategic complementarities. Such is the case with Example 3, on "industrialization," where the profits of one firm positively affect those of other firms. But even in this "best-case scenario," there may be multiple Pareto-dominated equilibria as in any coordination game. And yet, as noted by Murphy, Shleifer and Vishny (1989), this particular example — or its competitive analogue, to be more exact — has a *unique* equilibrium. The equilibrium is also efficient, which is an implication of our theorem. But our theorem goes way beyond the complementarities in Example 3, and as already mentioned, it is independent of the direction of the externalities.

Perhaps the theorem is best appreciated by reading its proof in detail, but as the argument is long, we provide the reader with an outline. We will establish the following claim, which is a bit stronger than what we need, but is nevertheless the more convenient to prove, as we proceed by induction and will need the additional power to complete the inductive step.

CLAIM. *There is no profile x^* with $u(x^*) = u^*$ such that for some other action profile $\bar{x}$ and utility profile $\bar{u}$,*

$$f(\bar{x}, \bar{u}) \geqslant \bar{u} > u^* \geqslant f(\bar{x}, u^*). \tag{2}$$

Theorem 1 follows from this Claim. Suppose that x^* is an equilibrium, but it is not Pareto-optimal. Then it is Pareto-dominated by some $\bar{x}$ with $\bar{u} = u(\bar{x})$, so that

$$f(\bar{x}, \bar{u}) = \bar{u} > u^*. \tag{3}$$

At the same time, because x^* is an equilibrium, it follows that for every i,

$$u_i^* = u_i(x^*) \geqslant u_i(\bar{x}_i, x_{-i}^*), \tag{4}$$

because a unilateral deviation to $\bar{x}_i$ from x_i^* cannot be profitable for i. A central observation (Lemma 2) proves that the absence of a profitable deviation, as just described in (4), is *equivalent* to the absence of a "naively profitable" deviation, in which player i deviates under the (possibly mistaken) premise that other payoffs will not change — even though they generally will. That is, (4) is equivalent to

$$u_i^* = f_i(x_i^*, u_{-i}^*) \geqslant f_i(\bar{x}_i, u_{-i}^*) \tag{5}$$

for all i. But (3) and (5) together imply (2), which contradicts the Claim.

The remainder of the proof establishes the Claim using induction on n. Specifically, we show that if (2) is true for a game with n players, where $n \geqslant 2$, then we can find a game with a *smaller* number of players where (2) is true as well. But it is very easy to see that for a *single-person* game, (2) must be false. After all, for a one-person game, $f(\bar{x}, \bar{u}) = f(\bar{x}, u^*)$, simply because there are no other players. Echoing the induction upwards as the number of players increases, we see that (2) can never be true.

5. PROOF OF THE MAIN RESULT

Let $(N, \{X_i, u_i\})$ be a game of love and hate that is continuous, coherent and reduced game coherent. We begin with some notation. Consider the reduced game resulting from the removal of some subset S

of players, with their payoffs pegged at u_S. It has player set $N - S$, and payoff functions

$$f_j^S(x_j, u_{-j}) \equiv f_j(x_j, u_{\{-j,S\}}, u_S),$$

for $j \in N - S$, where with some mild abuse of notation, the term u_{-j} on the left-hand side is presumed to exclude all players in S. (Notice that u_{-j} may have no components left; after all, a single player game would be induced if $|S| = n - 1$.) For every action profile x in the original game, reduced game coherence ensures a unique payoff profile in the reduced game; call it $v^S(x, u_S)$. Note that $v^S(x, u_S)$ depends only on x_{N-S} and u_S; it is insensitive to x_S.

A special reduced game is obtained by excluding just one player i with utility u_i. For any action profile x, then, the payoffs to $N - \{i\}$ are given by the vector $v^i(x, u_i)$. It will useful to introduce notation that describes how $v^i(x, u_i)$ maps back to i's payoff in the original game. That is, define

$$\phi_i(u_i, x) = f_i(x_i, v^i(x, u_i)).$$

In words, for a fixed action profile, we consider the reduced utility system that results when player i's utility is pegged at u_i, extract the unique fixed point of that reduced system, and now evaluate player i's utility at her action choice x_i when other players enjoy that fixed point.

LEMMA **1.** (a) *For every action profile x and every i, the well-defined payoff $u_i(x)$ uniquely solves $u_i(x) = \phi_i(u_i(x), x)$.*

(b) *For any pair of action profiles $x', x'' \in X$, let $u' = u(x')$ and $u'' = u(x'')$. The following three statements are equivalent:*

(i) $u_i' > u_i''$,

(ii) $u_i' = \phi_i(u_i', x') > \phi_i(u_i', x'')$,

(iii) $\phi_i(u_i'', x') > \phi_i(u_i'', x'') = u_i''$.

Proof. (a) Let $u(x)$ be the unique solution to (1). Because $v^i(x, u_i(x))$ is the unique solution to the reduced system given x and $u_i(x)$, we have $u_{-i}(x) = v^i(x, u_i(x))$. Therefore $\phi_i(u_i(x), x) = f_i(x_i, v^i(x, u_i(x))) = f_i(x_i, u_{-i}(x)) = u_i(x)$; i.e., $u_i(x)$ is a fixed point of $\phi_i(\cdot, x)$. In fact it is the unique fixed point. For if not, there is $\tilde{u}_i \neq u_i$ with $\tilde{u}_i = \phi_i(\tilde{u}_i, x)$. Let $\tilde{u}_{-i} = v^i(x, \tilde{u}_i)$. Then $\tilde{u}$ satisfies (1), but because $\tilde{u}_i \neq \bar{u}_i$, this contradicts the assumption that there is a unique solution to (1).

(b) [(i) $\Rightarrow$ (ii)] Suppose that (ii) fails, so that $\phi_i(u_i', x'') \geqslant u_i'$. Note that $\phi_i(B, x'') \leqslant B$ for B large enough, by coherence. By the intermediate value theorem, there exists $\tilde{u}_i \geqslant u_i'$ such that $\tilde{u}_i = \phi_i(\tilde{u}_i, x'')$. By part (a), $\tilde{u}_i = u_i''$. But $\tilde{u}_i \geqslant u_i'$, so (i) fails.

[(ii) $\Rightarrow$ (i)] Suppose that (ii) holds, so that $u_i' > \phi_i(u_i', x'')$. Since $\phi_i(0, x'') \geqslant 0$, by the intermediate value theorem there must be $\tilde{u}_i < u_i'$ such that $\tilde{u}_i = \phi_i(\tilde{u}_i, x'')$. By (a), $\tilde{u}_i = u_i''$, so $u_i' > u_i''$, implying (i).

[(i) $\Rightarrow$ (iii)] Suppose that (iii) fails, so that $\phi_i(u_i'', x') \leqslant u_i''$, then because $\phi_i(0, x') \geqslant 0$, by the intermediate value theorem, there is $\tilde{u}_i \leqslant u_i''$ such that $\tilde{u}_i = \phi_i(\tilde{u}_i, x')$. By (a), $\tilde{u}_i = u_i'$, which implies $u_i' \leqslant u_i''$, so (i) fails.

[(iii) ⇒ (i)] Because $\phi_i(u_i'', x') > u_i''$ and $\phi_i(M; x') \leqslant M$ for M large, there is $\tilde{u}_i > u_i''$ with $\tilde{u}_i = \phi_i(\tilde{u}_i, x')$. By (a), $\tilde{u}_i = u_i'$, so $u_i' > u_i''$, implying (i). ∎

A deviation by player i from x^* to x_i is *profitable* if $u_i(x_i, x^*_{-i}) > u_i(x^*)$. It is *naively profitable* if $f_i(x_i, u_{-i}(x^*)) > f_i(x_i^*, u_{-i}(x^*))$, i.e., player i profits under the "naive" presumption that all other utilities will remain unchanged.

LEMMA **2.** *A unilateral deviation is profitable if and only if it is naively profitable.*

Proof. Suppose i deviates from x^* to x_i. Let $u^* = u(x^*)$ and $y = (x_i, x^*_{-i})$. By Lemma 1,

$$u_i(y) = u_i(x_i, x^*_{-i}) > u_i^* \text{ if and only if } \phi_i(u_i^*, y) > u_i^*. \tag{6}$$

Because $v^i(x, u_i)$ is insensitive to x_i and $y_{-i} = x^*_{-i}$, we have $v^i(x, u_i^*) = v^i(y, u_i^*)$, so that

$$\phi_i(u_i^*, y) = f_i(x_i, v^i(y, u_i^*)) = f_i(x_i, v^i(x, u_i^*)) = f_i(x_i, u^*_{-i}).$$

Substituting this in (6) we have:

$$u_i(y) = u_i(x_i, x^*_{-i}) > u_i^* \text{ if and only if } f_i(x_i, u^*_{-i}) > u_i^*, \tag{7}$$

which establishes the desired result. ∎

Proof of Theorem 1. We first prove the Claim described in Section 4 by induction on the number of players. To begin the induction argument, consider any game with a *single* player: 1. Fix any action x_1^* with utility u_1^*. For any other action $\bar{x}$, it is immediate that $f(\bar{x}, \bar{u}) = f(\bar{x}, u^*)$, because there are no other players. So (2) can never hold.

Now for the inductive step. Suppose that the Claim is true of every game with $m < n$ players and satisfying the conditions of the Theorem, where $n \geqslant 2$. Consider a game with player set N, where $|N| = n$. Suppose, contrary to the Claim, that there are profiles x^* and $\bar{x}$ with associated payoff profiles u^* and $\bar{u}$ such that (2) is satisfied. Now we consider the following possibilities. First, if there is j with $\bar{u}_j = u_j^*$, then define S by the set of all j satisfying this equality. By the assertion just proved, S is a *strict* subset of N.

Otherwise, $\bar{u} \gg u^*$. In this case, pick any player, say k, and set $u_k = u_k^*$. If

$$v^k(\bar{x}, u_k^*) \gg u^*_{-k}, \tag{8}$$

then define $S = \{k\}$, which is again a strict subset of N.

Otherwise, (8) fails, so that for some $j \neq k$,

$$v_j^k(\bar{x}, u_k^*) \leqslant u_j^* \tag{9}$$

In this case, pick the *largest* value $\hat{u}_k \in [u_k^*, \bar{u}_k]$ such that (9) holds for some $j \neq k$. Because $\bar{u} \gg u^*$, we have

$$u_k^* \leqslant \hat{u}_k < \bar{u}_k. \tag{10}$$

Since $f_i(\bar{x}, .)$ is continuous for all i, the fixed points of the reduced game are upper-hemicontinuous in u_k. By reduced game coherence, there is a unique fixed point for each u_k, so $v^k(\bar{x}, .)$ is continuous, and

(10) implies

$$v_j^k(\bar{x}, \hat{u}_k) = u_j^* \tag{11}$$

for every $j \neq k$ for which (9) holds at $u_k = \hat{u}_k$. Define S to be this set of agents. Note that $k \notin S$ (so once again S is a strict subset of N), and if some other $i \neq k$ is also not in S, then

$$v_i^k(\bar{x}, \hat{u}_k) > u_i^*. \tag{12}$$

Under each of these three constructions of S, a reduced game is induced on players $N - S$ by setting $u_S = u_S^*$. It has payoff functions

$$f_j^S(x_j, u_{-j}) \equiv f_j(x_j, u_{\{-j,S\}}, u_S^*),$$

for $j \in N - S$, where we recall that u_{-j} on the left-hand side excludes all players in S, and where in all the vectors below that refer to the reduced game, the players in S are similarly excluded.[5]

Consider the profile $\bar{x}$ in the reduced game. In the first two cases above, define a utility profile u^r on $N - S$ by $u^r = v^S(\bar{x}, u_S^*)$. In the third case, in which $u_k = \hat{u}_k$, define u^r on $N - S$ by

$$u_i^r = \begin{cases} \hat{u}_k & \text{for } i = k, \\ v_i^k(\bar{x}, \hat{u}_k) & \text{for } i \neq k, i \in N - S \end{cases} \tag{13}$$

We claim that in the reduced game,

$$f^S(\bar{x}, u^r) \geqslant u^r \geqslant u^* \text{ while } f^S(\bar{x}, u^r) \neq u^r \text{ or } u^r \neq u^*. \tag{14}$$

To establish (14), consider each of the three cases above for which we defined S. In the first, it is easy to see that $v_i^S(\bar{x}, u_S^*) = u_i^r = \bar{u}_i$ for all i, so

$$f_i^S(\bar{x}, u^r) = u_i^r = \bar{u}_i > u_i^* \tag{15}$$

for all i, which establishes (14) right away. In the second case, with $S = \{k\}$, we have $u^r = v^k(\bar{x}, u_k^*)$, which just means that $v^k(\bar{x}, u_k^*) = f^S(\bar{x}, u^r)$. Combining this information with (8), we again obtain (15), which implies (14) as before.

Finally, S is defined as in the third case above, and $k \notin S$. If there is $i \in N - S$ with $i \neq k$, combine (12) and (13) to see that

$$u_i^r = v_i^k(\bar{x}, \hat{u}_k) > u_i^*. \tag{16}$$

We also claim that

$$u_i^r = f_i^S(\bar{x}, u_{-i}^r) \tag{17}$$

for each such $i \in N - S$ with $i \neq k$. To this end, and noting that f^k stands for the payoff function in another reduced game where just player k is removed (with $u_k = \hat{u}_k$), we have:

$$u_i^r = v_i^k(\bar{x}, \hat{u}_k) = f_i^k(\bar{x}, u_{-\{i,k,S\}}^r, u_S^*) = f_i^S(\bar{x}, u_{-\{i,k,S\}}^r, \hat{u}_k) = f_i^S(\bar{x}, u_{-i}^r),$$

where the first equality is just (13), the second equality comes from $u_{-k}^r = v_{-\{k,S\}}^k(\bar{x}, \hat{u}_k)$ (see (13)) and $u_S^* = v_S^k(\bar{x}, \hat{u}_k)$ (see (11)), the third equality switches reduced games to now exclude S but include

[5] As already observed, u_{-j} thus defined may have no components left if $N - S$ is a singleton.

k, and the last equality comes from $u^r_k = \hat{u}_k$ (see (13) again), also remembering that u^r excludes all indices in S. This yields (17). Combining (16) and (17), we can conclude that for $i \in N - S$ with $i \neq k$,

$$u^r_i = f^S_i(x_i, u^r_{-i}) > u^*_i \tag{18}$$

It remains to consider $i = k \in N - S$. We first claim that

$$\phi_k(\bar{x}, u^r_k) > u^r_k. \tag{19}$$

For if (19) were false, then $\phi_k(\bar{x}, u^r_k) \leqslant u^r_k$. Because $\phi_k(\bar{x}, 0) \geqslant 0$ and ϕ_k is continuous, there exists $u'_k \leqslant u^r_k$ such that $\phi_i(\bar{x}, u'_k) = u'_k$. But that generates a new utility solution $(u'_k, v^k(\bar{x}, u'_k))$ at the action profile $\bar{x}$ of the original game, because $u'_k \leqslant u^r_k = \hat{u}_k < \bar{u}_k$, where the equality comes from (13) and the last inequality follows from (10). That violates coherence.

Next, observe that

$$\phi_k(\bar{x}, u^r_k) = f_k(\bar{x}, v^k(\bar{x}, u^r_k)) = f_k(\bar{x}, v^k(\bar{x}, \hat{u}_k)) = f_k(\bar{x}, u^r_{-k}, u^*_S) = f^S_k(\bar{x}, u^r_{-k}), \tag{20}$$

where the first equality is just the definition of ϕ_k, the second equality uses $u^r_k = \hat{u}_k$ (see (13)), the third equality follows from $u^r_{-k} = v^k_{-\{k,S\}}(\bar{x}, \hat{u}_k)$ (see (13)) and $u^*_S = v^k_S(\bar{x}, \hat{u}_k)$ (see (11)), and the last equality simply translates to the reduced game, where S is excluded with payoff u^*_S.

Combining (19) and (20) along with $u^r_k = \hat{u}_k \geqslant u^*_k$ (the equality is from (13) and the inequality from (10)), we must conclude that

$$f^S_k(\bar{x}_k, u^r_{-k}) > u^r_k \geqslant u^*_k. \tag{21}$$

Combining (18) and (21), we obtain (14) for the reduced game.

We next show that for all $i \in N - S$,

$$u^*_i \geqslant f^S(\bar{x}, u^*_{-i}) \tag{22}$$

To establish (22), note that $u_i = u^*_i$ for all $i \in S$, so that

$$f^S_i(x_i, u^*_{-i}) = f_i(x_i, u^*_{\{-i,S\}}, u^*_S) \tag{23}$$

for all $i \in N - S$ and all actions x_i. Because we are presuming (by way of contradiction) that (2) holds for the original game, we have

$$f_i(x^*_i, u^*_{\{-i,S\}}, u^*_S) = u^*_i \geqslant f_i(\bar{x}_i, u^*_{\{-i,S\}}, u^*_S) \tag{24}$$

for all $i \in N - S$. Combining (23) and (24), we must conclude that

$$u^*_i = f^S_i(x^*_i, u^*_{-i}) \geqslant f^S_i(\bar{x}_i, u^*_{-i}),$$

for all $i \in N - S$, which yields (22).

To complete the proof of the Claim, we show that (2) holds for the reduced game. To this end, combine (14) and (22) to see that

$$f(\bar{x}, \bar{u}) \geqslant \bar{u} \geqslant u^* \geqslant f(\bar{x}, u^*), \text{ with either } f(\bar{x}, \bar{u}) \neq \bar{u} \text{ or } \bar{u} \neq u^* \text{ (or both).} \tag{25}$$

To obtain (2) from (25), we claim that $\bar{u} > u^*$. If not, then (given that (25) holds) it must be that $\bar{u} = u^*$, and so $f(\bar{x}, \bar{u}) > \bar{u}$. But $\bar{u} = u^*$, so that's just the same as saying that $f(\bar{x}, u^*) > u^*$. That means the very last inequality in (25) cannot hold, a contradiction. So $\bar{u} > u^*$, as claimed, and (2) holds for the reduced game.

But the reduced game of a coherent and reduced-game-coherent game, with payoff functions continuous in others' payoffs, has all these properties as well. But then, by the the induction hypothesis, (2) cannot hold for that reduced game, a contradiction.

As already noted in Section 4, our Theorem follows from the Claim. We note here formally that (4) implies (5) by way of Lemma 2. Combining (3) and (5), we see that

$$f(\bar{x}, \bar{u}) = \bar{u} > u^* \geqslant f(\bar{x}, u^*),$$

which implies (2). That is a contradiction. ■

6. DISCUSSION

6.1. **Some Intuition for Differentiable Games.** Assuming payoff functions to be differentiable and quasi-concave makes it easier to elicit some intuition about why equilibria might be Pareto optimal. The exposition that follows aims to do this, but is not meant to be rigorous or complete. And by no means is it meant to be a substitute for the proof of our theorem, which follows a completely different approach, relying only on the continuity of the payoff functions in other payoffs and imposing no topological structure on actions.

Suppose that for all i, $u_i(x)$ is continuously differentiable in x and quasi-concave in x_i. An equilibrium can then be characterized in terms of the first order conditions for each player. Using the equivalence of profitable and naively profitable deviations (Lemma 2), these conditions are:

$$\frac{\partial f_i(x)}{\partial x_i} = 0 \text{ for all } i. \tag{26}$$

Now consider the problem of a social planner, who seeks to maximize

$$\sum_j \lambda_j u_j(x)$$

where $\lambda \equiv^i (\lambda_1, \ldots, \lambda_n)^\top$ is a system of nonnegative weights summing to unity. Assuming that the relevant solutions are all interior, the first-order conditions are given by

$$\sum_j \lambda_j \frac{\partial u_j(x)}{\partial x_i} = 0 \text{ for all } i.$$

Collect this in matrix form to write

$$D_x \lambda = 0, \tag{27}$$

where D_x is the matrix of cross-effects

$$D_x = \begin{pmatrix} \frac{\partial u_1(x)}{\partial x_1} & \frac{\partial u_2(x)}{\partial x_1} & \cdots & \frac{\partial u_n(x)}{\partial x_1} \\ \frac{\partial u_1(x)}{\partial x_2} & \frac{\partial u_2(x)}{\partial x_2} & \cdots & \frac{\partial u_n(x)}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial u_1(x)}{\partial x_n} & \frac{\partial u_2(x)}{\partial x_n} & \cdots & \frac{\partial u_n(x)}{\partial x_n} \end{pmatrix}$$

By the chain rule,

$$\frac{\partial u_j(x)}{\partial x_i} = \sum_k \frac{\partial f_j}{\partial u_k}\frac{\partial u_k(x)}{\partial x_i}$$

for $j \neq i$, and for $j = i$:

$$\frac{\partial u_i(x)}{\partial x_i} = \frac{\partial f_i}{\partial x_i} + \sum_k \frac{\partial f_i}{\partial u_k}\frac{\partial u_k(x)}{\partial x_i},$$

so that

$$D_x = F + D_x D_u, \tag{28}$$

where

$$F = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & 0 & \dots & 0 \\ 0 & \frac{\partial f_2(x)}{\partial x_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\partial f_n(x)}{\partial x_n} \end{pmatrix} \text{ and } D_u = \begin{pmatrix} \frac{\partial f_1}{\partial u_1} & \frac{\partial f_2}{\partial u_1} & \dots & \frac{\partial f_n}{\partial u_1} \\ \frac{\partial f_1}{\partial u_2} & \frac{\partial f_2}{\partial u_2} & \dots & \frac{\partial f_n}{\partial u_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial u_n} & \frac{\partial f_2}{\partial u_n} & \dots & \frac{\partial f_n}{\partial u_n} \end{pmatrix},$$

the latter written with the understanding that $\partial f_i/\partial u_i = 0$ for all i. Rewriting (28), we see that

$$D_x = F[I - D_u]^{-1}, \tag{29}$$

where the presumption that $I - D_x$ has an inverse is closely connected to coherence; see Pearce (1983). Combining (27) and (29), we must conclude that the first order conditions for a solution to the planner's problem are

$$F[I - D_u]^{-1}\lambda = 0. \tag{30}$$

We can open this out as follows. Let b_{ij} be a generic entry for the matrix $[I - D_u]^{-1}$; then (30) is equivalent to the condition

$$\left[\frac{\partial f_i}{\partial x_i}\right]\left[\sum_{j=1}^{n} b_{ij}\lambda_j\right] = 0 \text{ for all } i. \tag{31}$$

Using (26), we must conclude that a solution to the equilibrium first order conditions are also solutions to the planner's first-order conditions (31), suggesting that equilibria are Pareto-optimal, or at least solve necessary conditions for planner optimality. We reiterate that this is suggestive but not rigorous (even with the smoothness and curvature assumptions in place). For more discussion, see the Appendix.

6.2. **Connection with the First Theorem of Welfare Economics.** There is a literature that studies the relationship between competitive equilibria and Pareto optimality in the presence of externalities.[6] In general, of course, competitive equilibria need not be Pareto optimal (so the first welfare theorem need not hold) and it may not be possible to decentralize every Pareto optimal allocation as a competitive equilibrium (so the second welfare theorem need not hold either). However, interesting connections can be identified when externalities are payoff-based; see Winter (1969), Ledyard (1971), Osana (1972), Rader (1980) and Parks (1991).

To explore the relationship of this literature to our paper, observe that our main theorem appears similar to the first welfare theorem of competitive equilibrium: it claims that every equilibrium is Pareto

[6]We are grateful to Peter Hammond for alerting us to the existence of this literature.

optimal. However, our conclusion holds regardless of the nature of the externalities, as long as these are purely payoff-based. It doesn't matter whether agents are benevolent or malevolent toward some or all opponents, or indeed whether they are affected in some non-monotone way by the payoffs of others. There is no hope of an analogous first welfare theorem (even for an exchange economy) at this level of generality. The reason is simple. If agents are benevolent they may well wish to allocate a larger part of the resources to some agent(s) than is feasible through the market, given that wealth redistribution is not permitted. In fact, as Winter (1969) and Bergstorm (1970) observe, even allowing agents to *unilaterally* transfer wealth to others may not suffice to restore the first welfare theorem. This literature does look for conditions for the first welfare theorem to hold in the presence of externalities; see, e.g., Ledyard (1971), Osana (1972) and Parks (1991). It identifies a form of "non-benevolence," which is not quite the same as the condition that all externalities are negative, but the point is that no such restriction is needed for the analogous result here.

Theorem 1 is cast in a parallel setting — games as opposed to competitive equilibrium — but the two models are quite distinct. A central difference is that in a game the feasible strategy profiles span the entire set of social outcomes whereas in an exchange economy they don't — specifically, agents cannot alter the wealth distribution. This means that the planner in an exchange economy has an extra instrument compared to the agents, which makes it harder for an equilibrium to satisfy Pareto optimality. While the first welfare theorem tells us that this does not impede Pareto-optimality in the classical setting, it clearly matters when there are externalities. On the other hand, in the game-theoretic model the planner doesn't have the advantage of an extra instrument. The game-theoretic analogue of the classical competitive setting is one in which externalities are central, and efficiency routinely fails. The restriction to the subclass of pure payoff externalities restores that efficiency, no matter what the particular form of those externalities. In this sense, ours is a stronger answer to a weaker question.

Of course, there is also a second welfare theorem for competitive equilibrium, and corresponding to that we have the parallel question for games: might every Pareto optimum be a Nash equilibrium? In terms of our first-order conditions, one might look for the reverse implication: "does (31) imply (26)?" This is not a question we investigate here in any generality, though the Appendix contains a discussion.

Finally, we should note that the literature on welfare theorems with externalities finesses the coherence issue, by postulating private as well as social preferences for each consumer. Consider the case where preferences are represented by utility functions. Suppose each consumer has a "private utility function" $w_i(x_i)$ and a true/social utility function which is Bergsonian:

$$u_i(x) = f^i(w_1(x_1), \ldots, w_n(x_n)).$$

Utilities are not truly interdependent in the sense that we study it, and so the coherence issue can be side-stepped.[7] As we are about to see in Section 6.3, coherence plays a critical role in more ways than one, and its explicit consideration is a central feature of this paper.

6.3. **The Role Played by Coherence.** In Pearce (1983), Bergstrom (1999), Hori and Kanaya 1989, and Vasquez and Weretka (2016), there is a concern with explosive or multiple utility representations. That concern is often at some philosophical level: "should" utility representations explode? (no: bound them — as in Vasquez and Weretka 2016), or: "should" utility representations exhibit the wrong comparative

[7] Parks (1991) shows that under certain assumptions the various non-malevolence and non-benevolence conditions used in this literature imply the Bergsonian form described above.

statics? (no: find a Hawkins-Simon-like condition to obtain uniqueness — as in Pearce 1983, Hori and Kanaya 1989 or Bergstrom 1999). In short, given some game of love and hate, its coherence has intrinsic appeal.

The purpose of this section is to argue that coherence plays a more subtle role, which is related to the intuitive appropriateness of the love-hate representation for certain classes of games. To understand this, begin with a standard game in normal form. We will now assume that the strategy spaces X_i are compact for every i, and that the payoff function $u_i : X \to \mathbb{R}$ — now to be thought of as the primitive — is continuous in the product topology on X. We will say that such a game is *regular* if for every player i and action $x_i \in X_i$, and for every pair of action profiles x_{-i} and x'_{-i} for the other players,

$$u_i(x_i, x_{-i}) \neq u_i(x_i, x'_{-i}) \text{ implies } u_{-i}(x_i, x_{-i}) \neq u_{-i}(x_i, x'_{-i}).$$

This is a mild restriction, stating that if player i is sensitive to some change in the actions of others, then so is at least one other player. Then the following must be true:

OBSERVATION **1.** *Every regular game with continuous payoffs can be represented as a continuous game of love and hate.*

We relegate the formal proof to the Appendix, but it is easy to see the argument. For player i, and action x_i, let U_{-i} be the compact set of utility profiles u_{-i} of the other players, such that $u_{-i} = u_{-i}(x_i, x_{-i})$ for some action profile x_{-i}. Define a function f_i on x_i and this sub-domain U_{-i} by

$$f_i(x_i, u_{-i}) = u_i(x_i, x_{-i}),$$

where x_{-i} is any action profile such that $u_{-i} = u_{-i}(x_i, x_{-i})$ (the exact choice of x_{-i} is unimportant, by regularity). The Appendix verifies the continuity of f_i on U_{-i}, and a standard extension argument extends f_i for every i and x_i to be defined on a common compact cube of utilities.[8]

But we know that the equilibrium inefficiency is rife among games in general. How can Theorem 1 be reconciled with Observation 1? The answer is that either coherence or reduced-game coherence must fail for any continuous love-hate representation, whenever the game has an inefficient equilibrium.

We alluded to this already in Example 2, and now to explain further what we mean, consider the following 2×2 family of regular symmetric games:

		Player 2	
		$\bar{x}_2$	x_2^*
Player 1	$\bar{x}_1$	c, c	b, a
	x_1^*	a, b	d, d

To cut down on the number of cases, suppose that a, b, c, d are all distinct positive numbers. Suppose $x^* = (x_1^*, x_2^*)$ is a Nash equilibrium that is Pareto dominated by $x = (\bar{x}_1, \bar{x}_2)$. This means that

$$c > d > b > 0. \tag{32}$$

[8]Recall that coherence asks for a unique vector of utilities at every action profile, *given the payoff functions* f_i. That is, it is not asking for the demanding — and unreasonable — restriction that there should be just one set of representing payoff functions, but only that there be one set of payoff numbers (per profile), *given* the representation.

Two cases of particular interest for us, depending on whether $a > c$ or $c > a$, are:

A *prisoner's dilemma*, in which $a > c$, so that the unique equilibrium is the Pareto-inferior outcome x^* with payoffs (d, d); and

A *coordination game*, in which $c > a$ so that x^* and x are *both* equilibria, the former Pareto-dominated.

Both cases yield inefficient equilibria. But these games are regular (and trivially continuous), and so have continuous representations as games of love and hate. Because reduced game coherence holds trivially in a game with two players, it follows from Theorem 1 that *no such representation can be coherent*. It is instructive to directly verify this assertion. To this end, let $\{f_1, f_2\}$ be a continuous love-hate representation of our two-player game. Without any loss of generality, we can choose any nonnegative bounded continuous $\{f_i\}$ such that for $i = 1, 2$,

$$f_i(\bar{x}_i, c) = c, f_i(\bar{x}_i, a) = b, f_i(x_i^*, b) = a, \text{ and } f_i(x_i^*, d) = d.$$

Then, even though $f_1(\bar{x}_1, d)$ is not pinned down by the payoff matrix, Lemma 2 and the fact that x^* is an equilibrium (which is implied by $b < d$, as assumed in (32)) imply:

$$f_1(\bar{x}_1, d) \leqslant d. \tag{33}$$

Given the continuity of f_1, (33) and $f_1(\bar{x}_1, 0) \geqslant 0$ together imply, by the intermediate value theorem, that there is $e \in [0, d]$ such that $f_1(\bar{x}_1, e) = e$. By symmetry, $f_2(\bar{x}_2, e) = e$ as well. The uniqueness of the payoffs at $\bar{x}$ must then mean that $e = c$. Because $e \in [0, d]$ and $c \neq d$, this implies that $c < d$, which contradicts (32). (As we shall see in Example 5, however, coherence can be restored if the representing payoff functions are allowed to be discontinuous.)

6.4. **Is Coherence Alone Sufficient for Theorem 1?** We've already seen from the discussion in Section 6.3 that coherence cannot be dropped from the statement of our theorem. For instance, the prisoner's dilemma can be transformed into a game of love and hate. Because reduced game coherence holds trivially for two-person games, any such transformation must lack coherence.

But Theorem 1 relies on two further restrictions. First, it assumes that payoff functions are continuous in the payoffs of others. Second, it assumes that the game in question is not only coherent, it is *reduced-game* coherent. In this Section, we argue that neither restriction can be dropped free of charge.

Example 5. *The need for continuity.* Consider a prisoner's dilemma:

		Player 2	
		$\bar{x}_2$	x_2^*
Player 1	$\bar{x}_1$	$3, 3$	$1, 4$
	x_1^*	$4, 1$	$2, 2$

It is easy to verify that this normal form is generated by the following game of love and hate:

$$f_i(x_i^*, u_2) = \begin{cases} 6 - 2u_j & \text{if } u_j \leqslant 3 \\ 3 & \text{if } u_j > 3 \end{cases}$$

$$f_i(\bar{x}_i, u_j) = \begin{cases} 9 - 2u_j & \text{if } u_j \leqslant 4.5 \\ 4.5 & \text{if } u_j > 4.5 \end{cases}$$

for $i, j = 1, 2$ and $j \neq i$. We now verify that this game is coherent, which (given that it is a two-player game) implies that it is also reduced game coherent. Begin with the profile $x = (x_1^*, x_2^*)$. If $u_j > 3$, then $u_i = f_i(x_i^*, u_j) = 3$. But then $u_j = f_j(x_j^*, u_i) = 0$, a contradiction. Therefore $u_j \leqslant 3$, so that $u_i = 6 - 2u_j$ for $i, j = 1, 2$ and $j \neq i$, the unique solution to which is $u_1 = u_2 = 2$. By a similar argument, $u(\bar{x}_2, \bar{x}_2) = (3, 3)$. Finally, consider $(x_i^*, \bar{x}_j)$. If $u_j > 3$, $u_i = f_i(x_i^*, u_j) = 3$, which implies that $u_j = f_j(\bar{x}_j, , u_i) = 3$, a contradiction. So $u_j \leqslant 3$. By a similar argument, $u_i \leqslant 4.5$. Together, these conditions imply that $u_i = 6 - 2u_j$ and $u_j = 9 - 2u_i$, or $u(x_i^*, \bar{x}_j) = (4, 1)$. That completes the verification of coherence. Of course, these functions are discontinuous. Indeed, a failure of continuity is necessary, by Theorem 1.

While we often view continuity as a mere technical device, here it emerges as having real conceptual power. The prisoner's dilemma is not, intuitively, a game of love and hate. Yet it mathematically can be straitjacketed into one. If we attempt that straitjacketing with continuous payoff functions, then — as already seen — coherence must fail. This example shows that one can *also* impose coherence, but then continuity must fail, and that failure is not a technicality. Indeed, as a parallel to Observation 1, one could also ask if every regular game has a love-hate representation satisfying coherence and reduced-game coherence, if one is willing to sacrifice continuity. We do not pursue this question here.

Example 6. *The necessity of reduced-game coherence.* To show that the reduced game coherence condition cannot be dropped from the statement of Theorem 1, we construct a continuous and coherent game with an inefficient equilibrium. Our example will have three players (we know that with two players reduced game coherence is satisfied trivially). In constructing such an example it is useful to recall the following property from the proof of Theorem 1. Suppose x^* is an equilibrium in a three-person, continuous and coherent game that is Pareto dominated by $\bar{x}$. Then there exists a player, i, such that the reduced game formed by removing i, with utility $u_i(x^*)$, is *not* coherent. Of course, our example also has to simultaneously ensure that the three-player game *is* coherent.

Consider a three-person game with $X_i = \{x_i^*, \bar{x}_i\}$ for $i = 1, 2$, $X_3 = \{x_3\}$ and $u_i \in [0, 1]$ for all i, and with payoff functions:

$$f_1(x_1^*, u_2, u_3) = \begin{cases} 0.95 & \text{if } u_2 \leqslant 0.6 \\ 95(u_2 - 0.7)^2 & \text{if } 0.6 \leqslant u_2 \leqslant 0.7 \\ 0 & \text{if } u_2 \geqslant 0.7 \end{cases} \tag{34}$$

$$f_1(\bar{x}_1, u_2, u_3) = u_3 \tag{35}$$

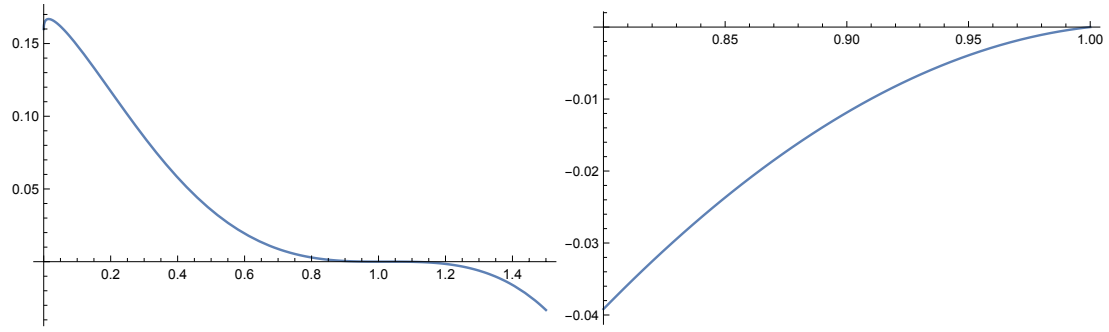
$$f_2(x_2^*, u_1, u_3) = 0.5 \text{ for all } (u_1, u_3) \tag{36}$$

$$f_2(\bar{x}_2, u_1, u_3) = \left[\frac{u_3^{1/(1+u_1)}}{1 + u_3^{1/(1+u_1)}} + 0.4 + 0.1\sqrt{2u_1 - u_1^2} \right]^{2+2u_1} \tag{37}$$

$$f_3(x_3, u_1, u_2) = u_2^{1+u_1} \tag{38}$$

Let $x^* \equiv (x_1^*, x_2^*)$ and $\bar{x} \equiv (\bar{x}_1, \bar{x}_2)$. (Player 3 has only one strategy, so we don't need to take note of this.) We make the following three claims.

Claim 1. The game is coherent, and payoff functions are continuous in others' payoffs.



(A) Value of function in (39) (B) Derivative of function in (39)

FIGURE 1. Verifying coherence at the strategy profile $\bar{x}$ in Example 6.

Claim 2. x^* is an equilibrium that is Pareto dominated by $\bar{x}$: $u(\bar{x}) \gg u(x^*)$.

Claim 3. The game does not satisfy reduced game coherence (as is implied by our Theorem and the previous two Claims).

Proof of Claim 1. Continuity is immediate on inspecting (34)–(38). To prove coherence we need to show that for any strategy profile x, $f(x, .)$ has a unique fixed point.

Consider the strategy profile x^*. In this case $u(x^*)$ must satisfy (34), (36) and (38). It's easy to see that these equations have the unique solution $u(x^*) = (0.95, 0.5, 0.5^{1.95})$.

Next, consider the strategy profile $\bar{x}$. Suppose u is a fixed point of $f(\bar{x}, .)$. Eliminating u_3 from (35), (37) and (38) we have:

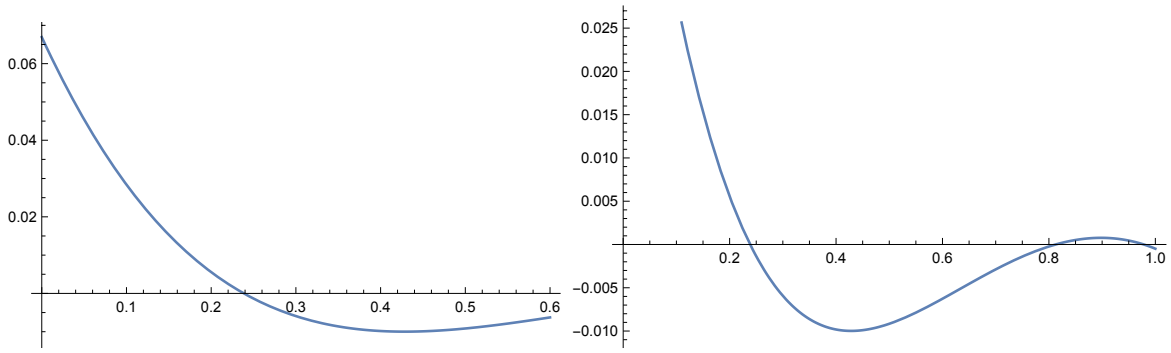
$$u_1 = u_2^{1+u_1} \text{ and } u_2 = \left[\frac{u_1^{1/(1+u_1)}}{1+u_1^{1/(1+u_1)}} + 0.4 + 0.1\sqrt{2u_1 - u_1^2}\right]^{2+2u_1},$$

so that

$$\left[\frac{u_1^{1/(1+u_1)}}{1+u_1^{1/(1+u_1)}} + 0.4 + 0.1\sqrt{2u_1 - u_1^2}\right]^{2+2u_1} - u_1^{1/(1+u_1)} = 0. \tag{39}$$

One solution to this is clearly $u_1 = 1$. Moreover, as Figure 1 (plotted using *Mathematica*) shows, the left hand side of this equation is strictly positive for all $u_1 < 1$; see Panel A. The unique fixed point of $f(\bar{x}, .)$ is therefore $u(\bar{x}) = (1, 1, 1)$. Further verification can be provided by examining the derivative of this function to the left of 1; see Panel B in Figure 1.

There are two remaining cases to consider. In the first of them, $x_1 = \bar{x}_1$ and $x_2 = x_2^*$. Then $u_1 = 0.5^{1+u_1}$. The function $g(u_1) = 0.5^{1+u_1} - u_1$ is strictly decreasing in u_1. Moreover, $g(0) > 0$ and $g(1) < 0$, which implies that $g(u_1) = 0$ has a unique solution strictly between between 0 and 1. The accompanying values of u_2 and u_3 are obviously unique.



(A) Value of function in (40) for $u_2 \in [0, 0.6]$

(B) Value of function in (40) for $u_2 \in [0, 1)$

FIGURE 2. More on coherence and reduced-game coherence in Example 6.

In the second case, $x_1 = x_1^*$ and $x_2 = \bar{x}_2$. In this case the relevant equations are (34), (37) and (38). Substituting (34) and (38) into (37) we have

$$\left[\frac{u_2}{1+u_2} + 0.4 + 0.1\sqrt{2u_1 - u_1^2}\right]^{2+2u_1} - u_2 = 0$$

Given (34), there are three distinct possibilities, depending on whether $u_2 \in [0, 0.6]$, $u_2 \in (0.6.0.7)$ or $u_2 \in [0.7, 1]$. The Appendix shows that the only solution is one that corresponds to the first case:

$$\left[\frac{u_2}{1+u_2} + 0.4 + 0.1\sqrt{1.9 - .95^2}\right]^{3.9} - u_2 = 0 \text{ with } u_2 \leqslant 0.6 \tag{40}$$

Panel A of Figure 2 (again plotted using *Mathematica*) depicts the left hand side of (40). It shows that $f(x, .)$ has a unique fixed point and completes the proof of Claim 1.

Proof of Claim 2. Recall that $u(x^*) = (0.95, 0.5, 0.5^{1.95})$ and $u(\bar{x}) = (1, 1, 1)$. To see that x^* is an equilibrium, we verify that $u_1(\bar{x}_1, x_2^*) \leqslant u_1^* = 0.95$ and $u_2(x_1^*, \bar{x}_2) \leqslant u_2^* = 0.5$. The former inequality follows from the fact that $u_1(\bar{x}_1, x_2^*)$ is the solution to $0.5^{1+u_1} = u_1$, as we saw in the proof of Claim 1. It is easy to see that the solution is strictly less than 0.95. For the latter, observe that $u_2(x_1^*, \bar{x}_2)$ is the (unique) solution to (40). As Panel A of Figure 2 shows, that solution is strictly less than 0.5.

Proof of Claim 3. By Theorem 1, reduced game coherence must fail in this example. This is indeed the case for the reduced game consisting of players 2 and 3 with $u_1 = 0.95$ with $x_2 = \bar{x}_2$. A fixed point of f in this reduced game is equivalent to a solution of (40), but with $u_2 \in [0, 1]$ rather than in $[0, 0.6]$. And there are three solutions to this equation for $u_2 \in [0, 1]$ as Panel B of Figure 2 shows. Note that the only difference between this graph and Panel A is that here the range of u_2 is $[0, 1]$ rather than $[0, 0.6]$. The piecewise construction of $f_1(x^*, .)$ was designed to ensure a unique solution to the utility system at $(x_1^*, \bar{x}_2)$ but not when the *utility* of player 1 is fixed at $u_1(x^*) = 0.95$.

In fact, it follows from the proof of our theorem that if an equilibrium x^* is Pareto dominated by $\bar{x}$, then there must exist a reduced game in which the player(s) that have been removed get $u(x^*)$, the others play $\bar{x}$, and the reduced game is not coherent. In the current example this is the case for the reduced

game with player 1's payoff fixed at u_1^*. We leave it to the reader to verify that in this example this feature does not hold for a reduced game in which one of the other players is removed.

6.5. **Coherence: An Afterword.** Theorem 1, as well as the subsequent discussion centered on Observation 1, tells us that a lot is hidden under the coherence rug. By no means do we suggest that coherence is a universally desirable property. It is desirable only if we believe that the situation at hand is *truly* a game with pure payoff externalities — as in Examples 1, 3 and 4 — and that too, not always.[9] In the wider world, replete with inefficient Nash equilibria, coherence is not an appropriate restriction (see the implications of Observation 1). Whether coherence is a "good" condition or not is deeply contextual.

REFERENCES

Bergstrom, T. (1970), "A 'Scandinavian Consensus' Solution for Efficient Income Distribution Among Nonmalevolent Consumers," *Journal of Economic Theory*, **2**, 383–398.

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

Dalton, P., Ghosal, S. and A. Mani (2016), "Poverty and Aspirations Failure," *Economic Journal*, **126**, 165–188.

Genicot. G and D. Ray (2017), "Aspirations and Inequality," *Econometrica*, **85**, 489–519.

Hori, H. and S. Kanaya (1989), "Utility Functionals with Nonpaternalistic Intergenerational Altruism," *Journal of Economic Theory*, **49**, 241–265.

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

Ledyard, J. (1971), "The Relation of Optima and Market Equilibria with Externalities," *Journal of Economic Theory*, **3**, 54–65.

Murphy, K., A. Shleifer and R. Vishny (1989), "Industrialization and the Big Push," *Journal of Political Economy*, **97**, 1003–1026.

Osana, H. (1972), "Externalities and the Basic Theorems of Welfare Economics," *Journal of Economic Theory*, **4**, 401–414.

Parks, R. (1991), "Pareto Irrelevant Externalities," *Journal of Economic Theory*, **54**, 165–179,

Pearce, D. (1983), "Nonpaternalistic Sympathy and the Inefficiency of Consistent Intertemporal Plans," Ph.D. dissertation, Princeton University, reprinted in *Foundations in Microeconomic Theory*, Jackson M.O., and A. McLennan(eds), Berlin, Heidelberg: Springer, 2008.

Rader, T. (1980), "The Second Theorem of Welfare Economics when Utilities are Interdependent," *Journal of Economic Theory*, **23**, 420–424.

[9] In some situations, for instance, mutually self-reinforcing sympathies or antipathies may well result in multiple solutions; for instance, in interactive human relationships with altruism.

Rosenstein-Rodan, P., (1943), "Problems of Industrialisation of Eastern and South-Eastern Europe," *Economic Journal*, **53**, 202–211.

Ray, D. (1987), "Nonpaternalistic Intergenerational Altruism," *Journal of Economic Theory*, **41**, 112–132.

Ray, D. (2006), "Aspirations, Poverty and Economic Change," in A.Banerjee, R. Bénabou and D. Mookherjee (eds), *What Have We Learnt About Poverty*, Oxford University Press, 409–422.

Ray, D. ⓡ A. Robson (2018), "Certified Random: A New Order for Coauthorship," *American Economic Review*, **108**, 489–520.

Sen, A. (1970), "The Impossibility of a Paretian Liberal," *Journal of Political Economy*, **78**, 152–157.

Vasquez, J. and M. Weretka (2016), "Mutual Empathy in Games," mimeo., Department of Economics, University of Wisconsin, Madison.

Willard, S. (1970), *General Topology*, Reading, MA: Addison-Wesley.

Winter, S. (1969), "On the Second Optimality Theorem of Welfare Economics," *Journal of Economic Theory*, **1**, 99–103.

APPENDIX

A. More on Differentiable Games, Pareto Optima and Equilibria. Section 6.1 of the main text recorded necessary (and under quasi-concavity, sufficient) conditions for a Nash equilibrium, taking advantage of smoothness, as well as the equivalence of profitable and naively profitable deviations (Lemma 2):

$$\frac{\partial f_i(x)}{\partial x_i} = 0 \text{ for all } i. \tag{41}$$

We then considered the problem of a social planner, who seeks to maximize

$$\sum_j \lambda_j u_j(x)$$

where $\lambda \equiv (\lambda_1, \ldots, \lambda_n)^\top$ is a system of nonnegative weights summing to unity. Assuming that the relevant solutions are all interior, the first-order conditions are described as follows. Let b_{ij} be a generic entry for the matrix $[I - D_u]^{-1}$; then:

$$\left[\frac{\partial f_i}{\partial x_i}\right]\left[\sum_{j=1}^{n} b_{ij}\lambda_j\right] = 0 \text{ for all } i. \tag{42}$$

Equation (42) has the flavor of a complementary slackness condition. To understand it, note that b_{ij} can be interpreted as the direct and indirect effects of a change in player i's utility on that of player j, with the direct effects (summarized by $\partial f_j/\partial u_i$) and all indirect effects (echoing through the "utility matrix") factored in. Condition says that as long as this weighted sum of direct and indirect effects is nonzero — as we change the utility of player i by varying her action — we should have player i at a stationary point in her own action at the planner optimum ($\partial f_i/\partial x_i = 0$). On the other hand, if the former weighted sum hits a zero somewhere, the planner might need to prevent player i from maximizing her utility through her own choice of action.

Using (41), we concluded that a solution to the equilibrium first order conditions are also solutions to the planner's first-order conditions (42), suggesting that equilibria are Pareto-optimal, or at least solve necessary conditions for planner optimality. That raises the question:

(a) Are the second order conditions for the planner's problem satisfied, so that (42) characterizes all the (local) optima for the planner's problem?

One might also ask the reverse question: are all Pareto optima in a coherent game of love and hate supportable as equilibria? In terms of first-order conditions, that would be related to:

(b) Does (42) imply (41)?

In general, the answer to both questions could be negative, as our next example illustrates.

First, the answer to (a) may be negative because the planner's objective may not be concave in every x_i even when, for all i, $u_i(.)$ is concave in x_i. As we shall see in Example A.1 below, for some weights λ it may be convex in some x_i, which means that (42) may not even describe a local optimum to the planner's problem. This illustrates the difficulties of a "differential approach" even when the primitive functions are well-behaved. Because the quasiconcavity of the planner's objective function is not guaranteed we cannot use the fact that (41) implies (42) to argue that an equilibrium is Pareto optimal.

Second, and now moving in the reverse direction, even if (42) holds at a Pareto optimum, it may not imply (41), because it's possible that $\sum_{j=1}^{n} b_{ij}\lambda_j = 0$ for some i. Given our assumption that $u_i(.)$ is quasi-concave in x_i for all i, this implies that the Pareto optimum in question is not an equilibrium. In this situation, the optimal x_i imposes a zero marginal effect on the planner's payoff, which could lead to a possible suppression of the best response of agent i.

Example A.1. *Pareto-optima may not be equilibria*. Consider a two-person game where $X_1 = X_2 = [0, 1]$, and each player's payoff is strictly concave in her own action and decreasing in the other player's payoff.

$$\begin{aligned} f_1(x_1, u_2) =& 1.5 - 1.5(0.5 - x_1)^2 - 0.5u_2 \\ f_2(x_2, u_1) =& 1.5 - 1.5(0.5 - x_2)^2 - 0.5u_1. \end{aligned}$$

This game is coherent (and trivially reduced-game coherent). For each $x \in X_1 \times X_2$,

$$\begin{aligned} u_1(x) =& 1 - 2(0.5 - x_1)^2 + (0.5 - x_2)^2 \\ u_2(x) =& 1 - 2(0.5 - x_2)^2 + (0.5 - x_1)^2 \end{aligned}$$

The unique equilibrium is $x^* = (0.5, 0.5)$, with payoff profile $u^* = (1, 1)$. The planner's problem, given $\lambda = (\lambda_1, \lambda_2)$ where $\lambda_i \in [0, 1]$ and $\lambda_1 + \lambda_2 = 1$, is:

$$\max_{x \in X_1 \times X_2} \lambda_1 u_1(x) + \lambda_2 u_2(x).$$

Substituting for $u_i(x)$, the planner objective function is $\lambda_1[1 - 2(0.5 - x_1)^2 + (0.5 - x_2)^2] + \lambda_2[1 - 2(0.5 - x_1)^2 + (0.5 - x_2)^2]$ which can be rewritten as:

$$1 + (\lambda_2 - 2\lambda_1)(0.5 - x_1)^2 + (\lambda_1 - 2\lambda_2)(0.5 - x_2)^2. \tag{43}$$

If $\lambda \in (1/3, 2/3)$, the coefficients for $(0.5 - x_1)^2$ and $(0.5 - x_2)^2$ are both negative, (43) is strictly concave in x, and the unique solution to maximizing (43) is $x^* = (0.5, 0.5)$. For λ in this range the answer to both (a) and (b) is in the affirmative. If $\lambda_1 = 1/3$ the planner's welfare is independent of x_1 and optimality is consistent with any $x_1 \in [0, 1]$, while $x_2 = 0.5$. This corresponds to the case in which $b_{11}\lambda_1 + b_{12}\lambda_2 = 0$ in (42) and the answer

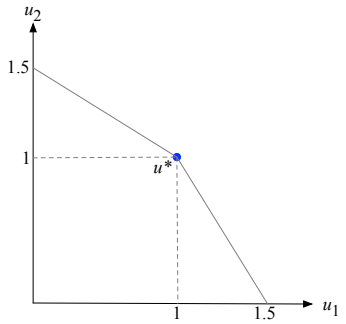


FIGURE 3. Pareto frontier for Example 5.

to (b) is negative: (42) does not imply (41).[10] Of course, the players' utilities do depend on x_1. The case where $\lambda_1 = 2/3$ is symmetric.

If $\lambda_1 < 1/3$, the planner's objective function, (43), becomes convex in x_1. If $\lambda > 2/3$ it becomes convex in x_2. In either case, (42) is not consistent with the maximization of the planner's objective, (43). Of course, x^* continues to satisfy these conditions but is not a solution to the planner's problem for $\lambda_1 \notin [1/3, 2/3]$.

The utility possibility frontier can be shown to have the form

$$u_2 = \begin{cases} 1.5 - 0.5u_1 & \text{if } u_1 \leqslant 1 \\ 3 - 2u_1 & \text{otherwise} \end{cases}$$

and is depicted in Figure 3. There is only one utility profile on the Pareto frontier, u^*, that matches the equilibrium utility profile $u(x^*) = (1, 1)$. It is a solution to the planner's problem for $\lambda \in [1/3, 2/3]$. For λ not in this range, (42) does not describe a solution to the planner's problem. Moreover, every solution to the planner's problem requires that one of the players must be made to choose an action that is sub-optimal.

It should be noted that the situation we have just described cannot happen in a game with complementarities. To see this, recall that for every $i \neq j$, b_{ij} can be viewed as the *full* effect on u_j of changing u_i at any state x; that is,

$$b_{ij} = \frac{\partial v_j^i(x, u_i)}{\partial u_i}.$$

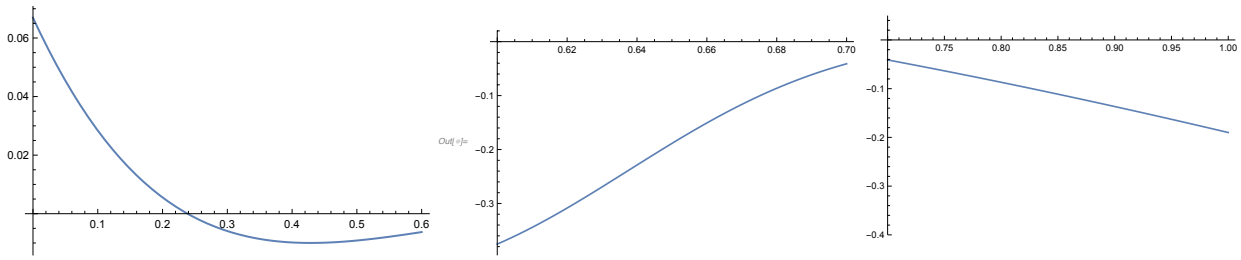
evaluated at $u_i = u_i(x)$, where $v_j^i(x, u_i)$ is given by reduced-game coherence. If these effects are all positive as in a game with payoff complementarities, then $\sum_{j=1}^n b_{ij}\lambda_j$ can never be zero, and so the first-order conditions for a Nash equilibrium must hold throughout. With complementarities, equilibria and Pareto-optima are equivalent.

We return to a discussion of the connections between the welfare theorems of general equilibrium, and our results. This time our focus is on the *second* welfare theorem. That second theorem is related to question (b). With differentiability, it can be phrased as a comparison of two first-order conditions: "does (42) imply (41)?" As we saw in Example A.1, the answer to this can be negative. But that won't happen if all agents are non-malevolent in the sense that their utilities are (weakly) positively related to that of others, or more generally if the game is one of complementarities. Relatedly, Winter (1969) shows that if no consumer is malevolent, then the second welfare theorem holds:[11] every Pareto optimal allocation is sustainable as a competitive equilibrium with redistribution.[12]

[10] It can be shown that $(b_{11}, b_{12}) = (4/3, -2/3)$.

[11] See also Rader (1980) and Parks (1991).

[12] In passing, take note of the tension between the conditions for each welfare theorem. While non-malevolence restores the second welfare theorem, it is non-benevolence that appears to help with the first welfare theorem. Asking for both these conditions to hold is to rule out externalities altogether; see Remark 8 in Parks (1991).



(A) Left hand side of (44), $u_2 \leqslant 0.6$

(B) Left hand side of (45), $u_2 \in (0.6, 0.7)$

(C) Left hand side of (46), $u_2 \geqslant 0.7$

FIGURE 4. Verifying coherence at the strategy profile $\bar{x}$ in Example 6.

It is therefore possible that the second welfare theorem exhibits a closer parallel across games and competitive equilibrium, though this paper is not about the second welfare theorem or its analogue in game theory.

B. Proof of Observation 1. By compactness of the strategy sets and continuity of payoffs, renormalize all utilities if needed so that they lie in some interval $[0, B]$. For each player i, and action x_i, define compact U_{-i} and $f_i(x_i, \cdot)$ on U_{-i} as in the main text. Let u^m_{-i} be a sequence of utility profiles in U_{-i} converging to some $u_{-i} \in U_{-i}$. Let x^m_{-i} be some corresponding sequence of action profiles. By compactness, all the limit points of x^m_{-i} are bonafide action profiles, and by regularity, $u_i(x_i, x_{-i}) = u_i(x_i, x'_{-i})$ for any possible pair of limit points (x_{-i}, x'_{-i}). It follows that $f_i(x_i, u^m_{-i}) \to f_i(x_i, u_{-i})$, so $f_i(x_i, \cdot)$ is continuous on U_{-i}. By the Tietze extension theorem (see, for example, Willard 1970, p. 99), $f_i(x_, \cdot)$ can be continuously extended to $[0, M]^{n-1}$, completing the proof.

C. Missing Details for Claim 2 in Example 6. The only detail for Example 6 that we need to supply is from Claim 2. This is the demonstration that $u(x)$ is unique when $x = (x^*_1, \bar{x}_2)$. As we showed in the main text, this requires us to show that there is a unique solution to:

$$\left[\frac{u_2}{1+u_2} + 0.4 + 0.1\sqrt{2u_1 - u_1^2}\right]^{2+2u_1} - u_2 = 0$$

According to (34), substituting for u_1 in this equation gives us three distinct possibilities:

$$\left[\frac{u_2}{1+u_2} + 0.4 + 0.1\sqrt{1.9 - .95^2}\right]^{3.9} - u_2 = 0 \text{ with } u_2 \leqslant 0.6 \tag{44}$$

$$\left[\frac{u_2}{1+u_2} + 0.4 + 0.1\sqrt{95}(u_2 - 0.7)\sqrt{2 - 95(u_2 - 0.7)^2}\right]^{2+190(u_2-0.7)^2} - u_2 = 0 \text{ with } 0.6 < u_2 < 0.7 \tag{45}$$

or

$$\left[\frac{u_2}{1+u_2} + 0.4\right]^{2} - u_2 = 0 \text{ with } u_2 \geqslant 0.7 \tag{46}$$

Only the graph of the left hand side of (44) was shown in the main text. Figure 4 plots all three equations. Clearly, only (44) has a solution. This shows that $f(x, .)$ has a unique fixed point and completes the proof of Claim 2.