

# GAMES OF LOVE AND HATE

Debraj Ray  Rajiv Vohra<sup>†</sup>

September 2018

**Abstract.** A strategic situation with payoff-based externalities is one in which a player’s payoff is a function of her own action and the payoffs of other players. Every action profile therefore induces an interdependent utility system. A strategic situation is *continuous* if each such utility system is continuous. If each utility system is bounded, with a unique payoff solution for every action profile, we call the strategic situation *coherent*, and if the same condition also applies to every subset of players, we call the situation *sub-coherent*. A coherent, sub-coherent and continuous situation generates a standard normal form, referred to as a *game of love and hate*. Our central theorem states that every equilibrium of a game of love and hate is Pareto optimal, in sharp contrast to the general prevalence of inefficient equilibria in the presence of externalities. While externalities 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, sub-coherence and continuity requirements are tight.


## 1. INTRODUCTION

A strategic situation with payoff-based externalities is one 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, Bergstrom 1999, Kockesen, Ok and Sethi 2000, 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, Hori and Kanaya 1989, Bergstrom 1999). This paper directly imposes coherence, as well as coherence on all subsets of agents (“sub-coherence”). We also assume that the payoff of each player is continuous in the payoffs of others. Then the strategic situation can be reduced to a standard game with payoffs defined on action profiles. We call this a game of love and hate. We have just one main result to report:

---

<sup>†</sup>Ray: New York University and University of Warwick, [debraj.ray@nyu.edu](mailto:debraj.ray@nyu.edu); Vohra: Brown University, [rajiv\\_vohra@brown.edu](mailto:rajiv_vohra@brown.edu). Ray acknowledges funding under NSF grant SES-1629370. We thank Dilip Abreu, Ted Bergstrom, Sylvain Chassang, Peter Hammond, George Mailath, David Pearce, Phil Reny and Lones Smith 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.

*Every equilibrium of a game of love and hate is Pareto-optimal.*

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 remarkable result that appears to depend fundamentally — but *only* — on the presumption that all externalities are transmitted via payoffs.

But a bit more is involved. One is naturally drawn to explaining just why games such as the Prisoner’s Dilemma or the Coordination Game, with inefficient equilibria, cannot be written as strategic situations with payoff externalities. The answer is that they *can* be so written (see Theorem 2), but no matter which payoff function we use to represent that conversion, either coherence or continuity must fail. This leads to a new and more subtle interpretation of the coherence restriction.

This discussion should not be taken to mean that we believe Nash equilibria to be efficient, or even efficient “most of the time.” Our result is general, but it is general within the *particular* class of games of love and hate. Such games do have applications (see Section 3), but our aim is not to argue that this class is widespread, or to provide algorithms to verify that a game belongs to this class. What we do find interesting is the fact that equilibria of games of love and hate behave the way they do. In particular, we are 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.

For instance, 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.<sup>1</sup>

## 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  $u_{-i} \equiv \{u_j\}_{j \neq i}$ :

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

A collection  $(N, \{X_i, f_i\})$  is a *strategic situation with payoff externalities* (or a strategic situation, for short). It is *continuous* if for each  $i$  and strategy  $x_i$ ,  $f_i$  is continuous in  $u_{-i}$ . No continuity condition is imposed with respect to  $x_i$ ; in fact, no topological restrictions are placed on the strategy sets.

Define the function  $f : X \times \mathbb{R}^N \rightarrow \mathbb{R}^N$  by:<sup>2</sup>

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

For any  $x$ , the mapping  $f(x, \cdot)$  is an instance of an interdependent utility system, governed by  $n$  component equations  $f_i(x, \cdot)$ , where for every  $i$ , the component  $f_i$  that generates  $u_i$  is defined on the vector of other payoffs  $u_{-i}$ . The strategic situation is said to be *coherent* if for every  $x$ ,

<sup>1</sup>See also Bergstrom (1970) in the special context of a distributive Lindahl equilibrium with non-malevolent agents.

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

- (i) there is  $B(x) < \infty$  such that  $\|f(x, u)\| < \|u\|$  whenever  $\|u\| > B(x)$ , where  $\|\cdot\|$  is the sup norm;  
(ii) the mapping  $f(x, \cdot)$  has a unique fixed point.

We impose coherence for our main result, but explore the implications of this restriction in some detail below. We also ask for “sub-coherence” on every sub-situation generated by 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  $f_i(x_i, u_{N-S-i}, \bar{u}_S)$ , for  $i \in N - S$ .

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:  $f(x, \cdot)$  has a unique fixed point for every strategy profile  $x$ ; call it  $U(x)$ . Pearce (1983), Hori and Kanaya (1989), and Bergstrom (1999) provide sufficient conditions for coherence in specific cases. Vasquez and Weretka (2016) discuss some implications of the lack of coherence (though they impose boundedness).

Given coherence, a strategic situation generates a well-defined normal form with payoffs  $U(x)$ , where  $U(x) = f(x, U(x))$ .<sup>3</sup> We will refer to a normal form game generated by a coherent, sub-coherent and continuous strategic situation as a *game of love and hate*.

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

$$U_i(x^*) \geq U_i(x_i, x_{-i}^*).$$

A strategy profile  $x \in X$  is *Pareto optimal* or *efficient* if there does not exist  $x' \in X$  with  $U(x') > U(x)$ .<sup>4</sup>

### 3. EXAMPLES

Here are four applications that illustrate the concept. To these we could add the obvious examples of altruism or envy that Pearce (1983), Kockesen, Ok and Sethi (2000), and others have studied.

Example 1. *Payoffs as Aspirations*. For some continuous function  $g$ ,

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

where  $x_i$  is an investment with return  $g$ , which is also influenced by the average payoff to others. This is like a reference point or aspiration.<sup>5</sup> 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 each individual’s payoffs are affected by her own actions and the payoffs of others. Of course, assumptions on  $g$ , such as boundedness as well as appropriately bounded responses to its second argument, will need to be placed to guarantee coherence and sub-coherence.

Example 2. *A 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

<sup>3</sup>We use upper case  $U$  to denote the utility function for the normal form game and lower case  $u$  for payoff values.

<sup>4</sup>For vectors  $a$  and  $b$ , “ $a \geq b$ ” means  $a_i \geq b_i$  in every component, “ $a > b$ ” implies  $a \geq b$  and  $a \neq b$ , and “ $a \gg b$ ” means  $a_i > b_i$  in every component.

<sup>5</sup>See, e.g., Ray (2006), Dalton, Ghosal and Mani (2016), Genicot and Ray (2017).

produced by a cottage technique  $y = \ell$ , where  $\ell$  is labor. This technique is available to a competitive fringe. Our firms can also 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 therefore:

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

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 in payoffs. It is easy to verify that coherence and sub-coherence are satisfied.

Example 3. *A 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  $\gamma$  is assumed to be continuous and decreasing. A lower  $\gamma$  increases the chances that a regulation will be placed on the firms, creating a penalty  $\pi(\gamma)$ , where  $\pi$  is continuous. 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  $\gamma$  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. Simple restrictions on these functions will guarantee coherence and sub-coherence.

Example 4. *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 situation of 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 embed this game within a strategic situation with payoff externalities. Consider any bounded continuous  $f_i$  such that for  $i = 1, 2$ ,

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

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 or continuity.

#### 4. MAIN RESULT

Our main result is:

**THEOREM 1.** *Every equilibrium of a game of love and hate 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 sub-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 the payoffs of others.

The context in which this theorem might be easiest to understand is a situation with strategic complementarities. Such is the case with Example 2, 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 2, 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  $U(\bar{x}) = \bar{u}$ ,*

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

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

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

At the same time, because  $x^*$  is an equilibrium, it follows that

$$(4) \quad u_i^* = U_i(x^*) \geq U_i(\bar{x}_i, x_{-i}^*) \text{ for all } i,$$

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

$$(5) \quad u^* = f(x^*, u^*) \geq f(\bar{x}, u^*).$$

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  $n \geq 2$ , 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 generated by some continuous, coherent and sub-coherent strategic situation. 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, sub-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  $x_S$ . Because  $f_i(x_i, \cdot)$  is continuous for all  $i$  and  $x_i$ , the fixed points of each reduced game are upper-hemicontinuous in  $u_S$ . Uniqueness of the fixed point then implies that for each  $x$ ,  $v^S(x, u_S)$  is continuous in  $u_S$ .

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 be 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. It follows from the continuity of  $v^i(x, u_i)$  in  $u_i$  that  $\phi_i(u_i, x)$  is a continuous function of  $u_i$ .

**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  $u'' = U(x'')$ . Then  $u'_i > u''_i$  if and only if  $\phi_i(u'_i, x') > \phi_i(u''_i, x'') = u''_i$ .*

*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  $\tilde{u}_i \neq u_i$ , this contradicts the assumption that there is a unique solution to (1).

(b) Suppose  $\phi_i(u'_i, x') > u''_i$ . Since  $\phi_i(\cdot, x')$  is continuous and  $\phi_i(M, x') < M$  for  $M$  large (by coherence), the intermediate value theorem tells us that 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  $u'_i > u''_i$ . Conversely, if  $\phi_i(u''_i, x') \leq u''_i$ , then using the fact that  $\phi_i(-M, x') \geq -M$  for  $M$  large enough (by coherence), we know that there is  $\tilde{u}_i \leq u''_i$  such that  $\tilde{u}_i = \phi_i(\tilde{u}_i, x')$ . By (a),  $\tilde{u}_i = u'_i$ , which implies  $u'_i \leq u''_i$ . ■

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,

$$(6) \quad U_i(y) = U_i(x_i, x_{-i}^*) > u_i^* \text{ if and only if } \phi_i(u_i^*, y) > u_i^*.$$

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:

$$(7) \quad U_i(y) = U_i(x_i, x_{-i}^*) > u_i^* \text{ if and only if } f_i(x_i, u_{-i}^*) > u_i^*,$$

which establishes the desired result.  $\blacksquare$

*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  $n \geq 2$ . 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.

**Case 1.** There is a player  $j$  with  $\bar{u}_j = u_j^*$ . Define  $S$  by the set of all  $j$  satisfying this equality. Because  $\bar{u} > u^*$ ,  $S$  is a *strict* subset of  $N$ .

**Case 2.**  $\bar{u} \gg u^*$  and there is a player, say  $k$ , such that in the reduced game without  $k$  (and with  $u_k = u_k^*$ ) all the others obtain a higher payoff than  $u_{-k}^*$ , i.e.,

$$(8) \quad v^k(\bar{x}, u_k^*) \gg u_{-k}^*.$$

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

**Case 3.**  $\bar{u} \gg u^*$ , and for every  $k$  (8) fails. Pick any  $k$ . It follows that for some  $j \neq k$ , we have

$$(9) \quad v_j^k(\bar{x}, u_k^*) \leq u_j^*.$$

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

$$(10) \quad u_k^* \leq \hat{u}_k < \bar{u}_k.$$

Because  $v^k(\bar{x}, \cdot)$  is continuous, (10) implies that

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

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

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

In each of the three cases,  $S$  is a strict subset of  $N$ . Consider the reduced game induced on players  $N - S$  by setting  $u_S = u_S^*$ . It has payoff functions

$$f_i^S(x_i, u_{-i}) \equiv f_i(x_i, u_{\{-i, S\}}, u_S^*), \text{ for all } i \in N - S,$$

where we recall that  $u_{-i}$  on the left-hand side excludes all players in  $S$ .

Consider the profile  $\bar{x}$  in the reduced game. In Cases 1 and 2, define a utility profile  $u^r$  on  $N - S$  by  $u^r = v^S(\bar{x}, u_S^*)$ . In Case 3, in which  $u_k = \hat{u}_k$ , define  $u^r$  on  $N - S$  by

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

We claim that in the reduced game,

$$(14) \quad f^S(\bar{x}, u^r) \geq u^r \geq u_{N-S}^* \text{ while } f^S(\bar{x}, u^r) \neq u^r \text{ or } u^r \neq u_{N-S}^*.$$

To establish (14), notice that in Cases 1 and 2,  $v^S(\bar{x}, u_S^*) = u^r$ , so  $f^S(\bar{x}, u^r) = u^r$ . In Case 1,  $u^r = \bar{u}_{N-S} \gg u_{N-S}^*$  and in Case 2, it follows from (8) that  $u^r \gg u_{N-S}^*$ . In both cases, therefore,

$$f^S(\bar{x}, u^r) = u^r \gg u_{N-S}^*,$$

which establishes (14) right away.

It remains to verify (14) in Case 3, where  $k \notin S$ . If there is  $i \in N - S$  with  $i \neq k$ , combine (12) and (13) to see that

$$(15) \quad u_i^r = v_i^k(\bar{x}, \hat{u}_k) > u_i^* \text{ for all } i \in N - S - k.$$

We also claim that

$$(16) \quad u_i^r = f_i^S(\bar{x}, u_{-i}^r) \text{ for all } i \in N - S - k.$$

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:

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

where the first equality is just (13), the second equality comes from  $u_{-k}^r = v_{-k}^k(\bar{x}, \hat{u}_k)$  (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  $k$ , and the last equality comes from  $u_k^r = \hat{u}_k$  (see (13) again), also remembering that  $u^r$  excludes all indices in  $S$ . This yields (16). Combining (15) and (16), we can conclude that

$$(17) \quad f_i^S(x_i, u_{-i}^r) = u_i^r > u_i^* \text{ for all } i \in N - S - k.$$

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

$$(18) \quad \phi_k(\hat{u}_k, \bar{x}) > \hat{u}_k.$$

For if (18) were false, then  $\phi_k(\hat{u}_k, \bar{x}) \leq \hat{u}_k$ . Because  $\phi_k(-M, \bar{x}) \geq -M$  for  $M$  large enough and  $\phi_k$  is continuous, there exists  $u'_k \leq \hat{u}_k$  such that  $\phi_k(u'_k, \bar{x}) = u'_k$ . But that generates a utility solution  $(u'_k, v^k(\bar{x}, u'_k))$  at the action profile  $\bar{x}$  of the original game. It is distinct from  $\bar{u}$  because  $u'_k \leq \hat{u}_k < \bar{u}_k$ , where the last inequality follows from (10). That violates coherence.

Next, observe that

$$(19) \quad \phi_k(\hat{u}_k, \bar{x}) = f_k(\bar{x}, v^k(\bar{x}, \hat{u}_k)) = f_k(\bar{x}, u_{-k}^r, u_S^*) = f_k^S(\bar{x}, u_{-k}^r),$$

where the first equality is just the definition of  $\phi_k$ , the second equality follows from  $u_{-k}^r = v_{-k}^k(\bar{x}, \hat{u}_k)$  (see (13)) and  $u_S^* = v_S^k(\bar{x}, \hat{u}_k)$  (see (11)), and the last equality is just the translation to the reduced game, where  $S$  is excluded with payoff  $u_S^*$ .



Combining (18) and (19) along with  $u_k^r = \hat{u}_k \geq u_k^*$  (the equality is from (13) and the inequality from (10)), we must conclude that

$$(20) \quad f_k^S(\bar{x}_k, u_{-k}^r) > u_k^r \geq u_k^*.$$

Combining (17) and (20), we obtain (14) for the reduced game. This proves that (14) holds in each of the three Cases.

We now use (14) to prove that (2) holds for the reduced game. First, because we maintain that (2) holds for the original game, we have

$$u_i^* \geq f_i(\bar{x}_i, u_{-i}^*) \text{ for all } i \in N - S.$$

Because  $u_i = u_i^*$  for all  $i \in S$ ,  $f_i^S(\bar{x}_i, u_{-i}^*) = f_i(\bar{x}_i, u_{-i}^*)$  for all  $i \in N - S$ . That implies

$$(21) \quad u_i^* \geq f_i^S(\bar{x}, u_{-i}^*) \text{ for all } i \in N - S.$$

Now combine (14) and (21) to see that

$$(22) \quad f^S(\bar{x}, u^r) \geq u^r \geq u_{N-S}^* \geq f^S(\bar{x}, u_{N-S}^*), \text{ with either } f^S(\bar{x}, u^r) \neq u^r \text{ or } u^r \neq u_{N-S}^* \text{ (or both).}$$

To obtain (2) from (22), we claim that  $u^r > u_{N-S}^*$ . If not, then (given that (22) holds) it must be that  $u^r = u_{N-S}^*$ , and so  $f^S(\bar{x}, u^r) > u^r$ . But  $u^r = u_{N-S}^*$ , so that's just the same as saying that  $f^S(\bar{x}, u_{N-S}^*) > u_{N-S}^*$ . That means the very last inequality in (22) cannot hold, a contradiction. So  $u^r > u_{N-S}^*$ , as claimed, and (2) holds for the reduced game.

Now, the reduced game of a coherent and sub-coherent game, with payoff functions continuous in others' payoffs, inherits all these just-named properties. 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. Formally, suppose there is an equilibrium  $x^*$  Pareto dominated by  $\bar{x}$ . Let  $u^* = U(x^*)$  and  $\bar{u} = U(\bar{x})$ . Then

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

where the last inequality makes use of Lemma 2 and the fact that  $x^*$  is an equilibrium. But this implies (2), a contradiction. ■

## 6. DISCUSSION

**6.1. Some Intuition for Differentiable Games of Love and Hate.** 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:

$$(23) \quad \frac{\partial f_i(x)}{\partial x_i} = 0 \text{ for all } i.$$

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

$$\sum_j \lambda_j U_j(x)$$

where  $\lambda \equiv (\lambda_1, \dots, \lambda_n)^\top$  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

$$(24) \quad D_x \lambda = 0,$$

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

$$(25) \quad D_x = F + D_x D_u,$$

where

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

the latter written with the understanding that  $\partial f_i / \partial u_i = 0$  for all  $i$ . Rewriting (25), we see that

$$(26) \quad D_x = F[I - D_u]^{-1},$$

where the presumption that  $I - D_u$  has an inverse is closely connected to coherence; see Pearce (1983). Combining (24) and (26), we must conclude that the first order conditions for a solution to the planner's problem are

$$(27) \quad F[I - D_u]^{-1} \lambda = 0.$$

We can open this out as follows. Let  $b_{ij}$  be a generic entry for the matrix  $[I - D_u]^{-1}$ ; then (27) is equivalent to the condition

$$(28) \quad \left[ \frac{\partial f_i}{\partial x_i} \right] \left[ \sum_{j=1}^n b_{ij} \lambda_j \right] = 0 \text{ for all } i.$$

Using (23), we must conclude that a solution to the equilibrium first order conditions is also a solution to the planner's first-order conditions (28), 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.<sup>6</sup> 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 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 Bergstrom (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 different setting — strategic situations as opposed to competitive equilibrium — and 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.

---

<sup>6</sup>We are grateful to Peter Hammond for alerting us to the existence of this literature.

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 (28) imply (23)?” 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:

$$U_i(x) = h_i(w_1(x_1), \dots, w_n(x_n)).$$

Utilities are not truly interdependent in our sense, and so the coherence issue can be side-stepped.<sup>7</sup> 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), Hori and Kanaya (1989), Bergstrom (1999), 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 statics? (no: find a conditions that guarantee uniqueness — as in Pearce 1983, Hori and Kanaya 1989 or Bergstrom 1999). In short, the coherence of any strategic situation with payoff externalities 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  $U_i : X \rightarrow \mathbb{R}$  — 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:

**THEOREM 2.** *Every regular game with continuous payoffs can be represented as a continuous strategic situation with payoff externalities.*

We relegate the formal proof to the Appendix, but it is easy to see the argument. For player  $i$ , and action  $x_i$ , let  $\mathcal{U}_{-i}$  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  $\mathcal{U}_{-i}$  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  $\mathcal{U}_{-i}$ , and a standard extension argument extends  $f_i$  for every  $i$  and  $x_i$  to all opponent utility profiles on  $\mathbb{R}^{n-1}$ .

<sup>7</sup>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.

But equilibrium inefficiency is rife among games in general. How can Theorem 1 be reconciled with Theorem 2? The answer is that either coherence or sub-coherence must fail for any continuous representation as a strategic situation, whenever the game has an inefficient equilibrium.<sup>8</sup> We alluded to this already in Example 4. To explain further, consider a  $2 \times 2$  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 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

$$(29) \quad c > d > b.$$

Two cases of particular interest for us are (i) a *prisoner's dilemma*, in which  $a > c$ , so that the unique equilibrium is the Pareto-inferior outcome  $x^*$  with payoffs  $(d, d)$ ; and (ii) 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 strategic situations with payoff externalities. Because sub-coherence holds trivially in a strategic situation 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 representation of our two-player game as a strategic situation. Without any loss of generality, we can choose any bounded continuous  $\{f_i\}$  such that for  $i = 1, 2$ ,

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

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 (29)) imply:

$$(30) \quad f_1(\bar{x}_1, d) \leq d.$$

Given the continuity of  $f_1$ , (30) and  $f_1(\bar{x}_1, -M) \geq M$  for large  $M$  (coherence) together imply, by the intermediate value theorem, that there is  $e \leq d$  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  $e \leq d$  and  $c \neq d$ , this implies that  $c < d$ , which contradicts (29). (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 strategic situation with payoff externalities. Because sub-coherence holds trivially for two-person strategic situations, any such transformation must lack coherence.

---

<sup>8</sup>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*.

But the definition of a game of love and hate 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 *sub-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 strategic situation with pure payoff externalities:

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

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

for  $i, j = 1, 2$  and  $j \neq i$ . We now verify that this situation is coherent, which (given just two players) implies that it is also sub-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  $u_j \leq 3$ , 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  $u_j \leq 3$ . By a similar argument,  $u_i \leq 4.5$ . Together, these imply  $u_i = 6 - 2u_j$  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, a property necessitated 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 strategic situation with payoff based externalities. 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. That failure is not a technicality. Indeed, as a parallel to Theorem 2, one could also ask if every regular game has a love-hate representation satisfying coherence and sub-coherence, if one is willing to sacrifice continuity. We do not pursue this question here.

Example 6. *The necessity of sub-coherence.* To show that sub-coherence cannot be dropped from Theorem 1, we construct a continuous and coherent strategic situation with an inefficient equilibrium. Our example will have three players (with two, sub-coherence is satisfied trivially). Let  $X_i = \{x_i^*, \bar{x}_i\}$  for  $i = 1, 2$ ,  $X_3 = \{x_3\}$ . Define the following payoff functions on  $[0, 1]^2$  with range in  $[0, 1]$ :<sup>9</sup>

$$(31) \quad f_1(x_1^*, u_2, u_3) = \begin{cases} 0.95 & \text{if } u_2 \leq 0.6 \\ 95(u_2 - 0.7)^2 & \text{if } 0.6 \leq u_2 \leq 0.7 \\ 0 & \text{if } u_2 \geq 0.7 \end{cases}$$

$$(32) \quad f_1(\bar{x}_1, u_2, u_3) = u_3$$

<sup>9</sup>These functions have trivial bounded extensions to all of  $\mathbb{R}^2$ , changing none of what follows.

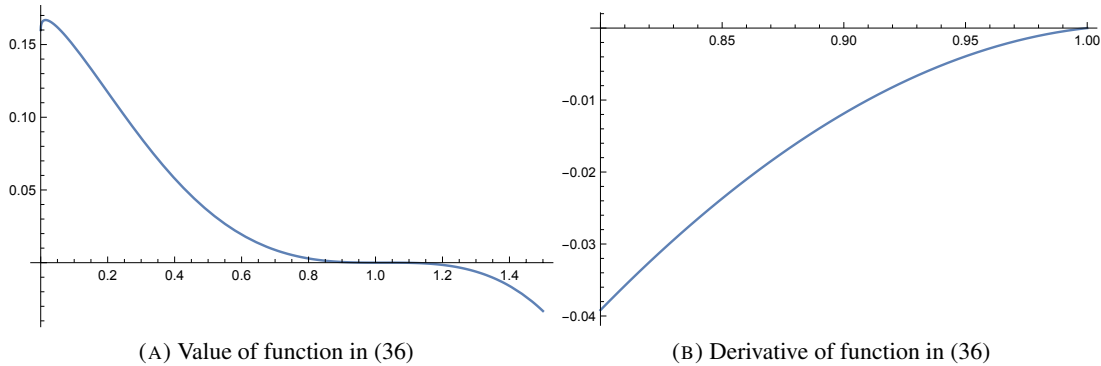


FIGURE 1. Verifying coherence at the strategy profile  $\bar{x}$  in Example 6.

$$(33) \quad f_2(x_2^*, u_1, u_3) = 0.5 \text{ for all } (u_1, u_3)$$

$$(34) \quad f_2(\bar{x}_2, u_1, u_3) = \left[ \frac{u_3^{1/(1+u_1)}}{1 + u_3^{1/(1+u_1)}} + 0.4 + 0.1\sqrt{2u_1 - u_1^2} \right]^{2+2u_1}$$

$$(35) \quad f_3(x_3, u_1, u_2) = u_2^{1+u_1}$$

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 strategic situation is coherent, and payoff functions are continuous in others' payoffs.

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

**Claim 3.** Sub-coherence fails (as implied by Theorem 1 and the previous two Claims).

**Proof of Claim 1.** Continuity is immediate on inspecting (31)–(35). To prove coherence we need to show that for any strategy profile  $x$ ,  $f(x, \cdot)$  has a unique fixed point.

Consider the strategy profile  $x^*$ . In this case  $U(x^*)$  must satisfy (31), (33) and (35). 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  $f(\bar{x}, \cdot)$ . Eliminating  $u_3$  from (32), (34) and (35) 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

$$(36) \quad \left[ \frac{u_1^{1/(1+u_1)}}{1 + u_1^{1/(1+u_1)}} + 0.4 + 0.1\sqrt{2u_1 - u_1^2} \right]^{2+2u_1} - u_1^{1/(1+u_1)} = 0.$$

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

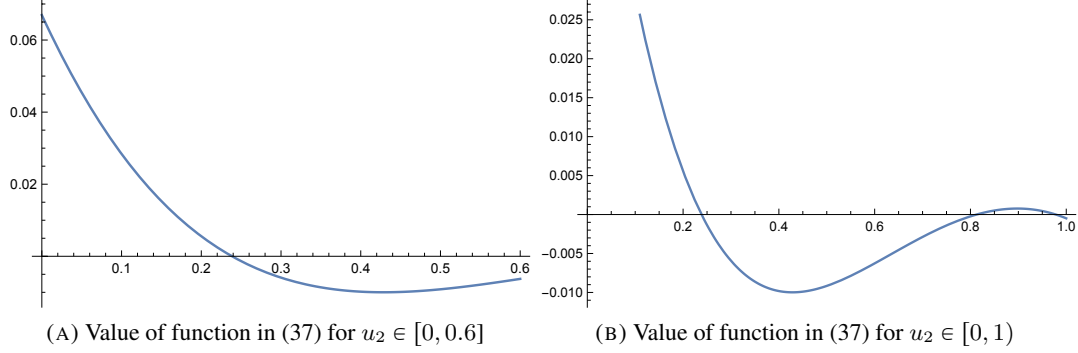


FIGURE 2. More on coherence and sub-coherence in Example 6.

$f(\bar{x}, \cdot)$  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 0 and 1. The accompanying values of  $u_2$  and  $u_3$  are obviously unique.

In the second case,  $x_1 = x_1^*$  and  $x_2 = \bar{x}_2$ . In this case the relevant equations are (31), (34) and (35). Substituting (31) and (35) into (34) 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 (31), 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:

$$(37) \quad \left[ \frac{u_2}{1+u_2} + 0.4 + 0.1\sqrt{1.9 - .95^2} \right]^{3.9} - u_2 = 0 \text{ with } u_2 \leq 0.6$$

Panel A of Figure 2 (again plotted using *Mathematica*) depicts the left hand side of (37). It shows that  $f(x, \cdot)$  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  $U_1(\bar{x}_1, x_2^*) \leq u_1^* = 0.95$  and  $U_2(x_1^*, \bar{x}_2) \leq u_2^* = 0.5$ . The 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 (37). As Panel A of Figure 2 shows, that solution is strictly less than 0.5.

**Proof of Claim 3.** By Theorem 1, sub-coherence fails. Consider the reduced game  $f^1$  with players 2 and 3, and  $u_1 = 0.95$ . Let  $x_2 = \bar{x}_2$ . A fixed point of  $f^1$  is equivalent to a solution of (37), 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. The only difference between the two panels is that in [B], the range of  $u_2$  is  $[0, 1]$  rather than  $[0, 0.6]$ . The piecewise construction of  $f_1(x^*, \cdot)$  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 Theorem 2, 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.<sup>10</sup> In the wider world, replete with inefficient Nash equilibria, coherence is not an appropriate restriction (see the implications of Theorem 2). 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. and A. McLennan (eds), Berlin, Heidelberg: Springer, 2008.

<sup>10</sup>In interactive relationships with altruism, mutually self-reinforcing sympathies may well result in multiple solutions.

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

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):

$$(38) \quad \frac{\partial f_i(x)}{\partial x_i} = 0 \text{ for all } i.$$

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

$$\sum_j \lambda_j U_j(x)$$

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

$$(39) \quad \left[ \frac{\partial f_i}{\partial x_i} \right] \left[ \sum_{j=1}^n b_{ij} \lambda_j \right] = 0 \text{ for all } i.$$

Equation (39) 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 (39) 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 (38), we concluded that the equilibrium conditions are also solutions to the planner's first-order conditions (39), suggesting that equilibria solve necessary conditions for planner optimality. That raises the question:

(a) Are the second order conditions for the planner's problem satisfied, so that (39) characterizes all the (local) optima for the planner's problem?

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

(b) Does (39) imply (38)?

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  $i$ ,  $U_i(\cdot)$  is concave in  $x_i$ . As we shall see in Example A.1 below, for some weights  $\lambda$  it may be convex in some  $x_i$ , which means that (39) 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 (38) implies (39) to argue that an equilibrium is Pareto optimal.

Second, and now moving in the reverse direction, even if (39) holds at a Pareto optimum, it may not imply (38), because it's possible that  $\sum_{j=1}^n b_{ij}\lambda_j = 0$  for some  $i$ . Given our assumption that  $U_i(\cdot)$  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 strategic situation in which  $X_1 = X_2 = [0, 1]$ , 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 situation is coherent (and trivially sub-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:

$$(40) \quad 1 + (\lambda_2 - 2\lambda_1)(0.5 - x_1)^2 + (\lambda_1 - 2\lambda_2)(0.5 - x_2)^2.$$

If  $\lambda \in (1/3, 2/3)$ , the coefficients for  $(0.5 - x_1)^2$  and  $(0.5 - x_2)^2$  are both negative, (40) is strictly concave in  $x$ , and the unique solution to maximizing (40) is  $x^* = (0.5, 0.5)$ . For  $\lambda$  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  $b_{11}\lambda_1 + b_{12}\lambda_2 = 0$  in (39), and the answer to (b) is negative: (39) does not imply (38).<sup>11</sup> Of course, the players' utilities do depend on  $x_1$ . The case  $\lambda_1 = 2/3$  is symmetric.

<sup>11</sup>It can be shown that  $(b_{11}, b_{12}) = (4/3, -2/3)$ .

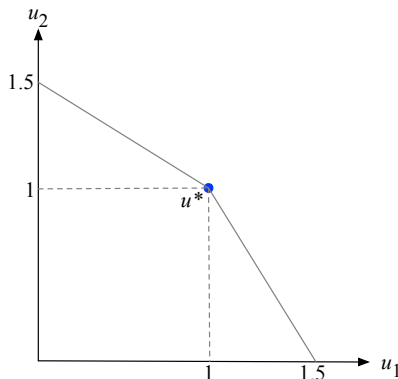


FIGURE 3. Pareto frontier for Example 5.

If  $\lambda_1 < 1/3$ , the planner's objective function, (40), becomes convex in  $x_1$ . If  $\lambda > 2/3$  it becomes convex in  $x_2$ . In either case, (39) is not consistent with the maximization of the planner's objective, (40). 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

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

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  $\lambda$  not in this range, (39) 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 sub-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 (39) imply (38)?" 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:<sup>12</sup> every Pareto optimal allocation is sustainable as a competitive equilibrium with redistribution.<sup>13</sup>

<sup>12</sup>See also Rader (1980) and Parks (1991).

<sup>13</sup>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).

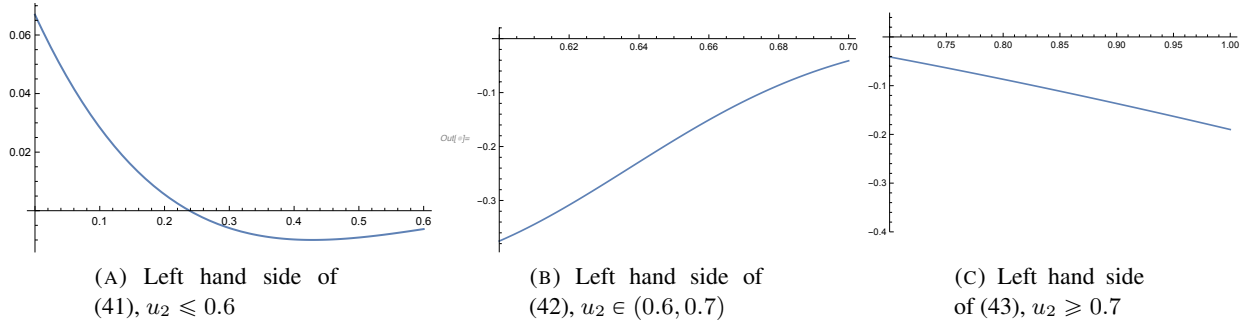


FIGURE 4. Verifying coherence at the strategy profile  $\bar{x}$  in Example 6.

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 Theorem 2.** For each player  $i$ , and action  $x_i$ , define compact  $\mathcal{U}_{-i}$  and  $f_i(x_i, \cdot)$  on  $\mathcal{U}_{-i}$  as in the main text. Let  $u_{-i}^m$  be a sequence of utility profiles in  $\mathcal{U}_{-i}$  converging to some  $u_{-i} \in \mathcal{U}_{-i}$ . Let  $x_{-i}^m$  be some corresponding sequence of action profiles. By compactness, all the limit points of  $x_{-i}^m$  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  $f_i(x_i, u_{-i}^m) \rightarrow f_i(x_i, u_{-i})$ , so  $f_i(x_i, \cdot)$  is continuous on  $\mathcal{U}_{-i}$ . By the Tietze extension theorem (see, for example, Willard 1970, p. 99),  $f_i(x_i, \cdot)$  can be extended to a bounded continuous function on  $\mathbb{R}^{n-1}$ . ■

**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 (31), substituting for  $u_1$  in this equation gives us three distinct possibilities:

$$(41) \quad \left[ \frac{u_2}{1+u_2} + 0.4 + 0.1\sqrt{1.9 - .95^2} \right]^{3.9} - u_2 = 0 \text{ with } u_2 \leq 0.6$$

$$(42) \quad \left[ \frac{u_2}{1+u_2} + 0.4 + 0.1\sqrt{95(u_2 - 0.7)}\sqrt{2 - 95(u_2 - 0.7)^2} \right]^{2+190(u_2-0.7)^2} - u_2 = 0 \text{ with } 0.6 < u_2 < 0.7$$

or

$$(43) \quad \left[ \frac{u_2}{1+u_2} + 0.4 \right]^2 - u_2 = 0 \text{ with } u_2 \geq 0.7$$

Only the graph of the left hand side of (41) was shown in the main text. Figure 4 plots all three equations. Clearly, only (41) has a solution. This shows that  $f(x, \cdot)$  has a unique fixed point and completes the proof of Claim 2.