GAMES OF LOVE AND HATE

Debraj Ray ⓡ Rajiv Vohra[†]

March 2019

Forthcoming, *Journal of Political Economy*

Abstract. A strategic situation with payoff-based externalities is one in which a player's payoff depends on her own action and (continuously) on the payoffs of other players. Every action profile therefore induces an interdependent utility system. 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 and sub-coherent 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 the payoff of each player depends on her own action, and the *payoffs* of some or all of the other players. Others' 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, Kimball 1987, or Vasquez and Weretka 2018). 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:

[†]Ray: NYU 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, George Mailath, David Pearce, Phil Reny, Lones Smith and Yeneng Sun for helpful comments. We are especially grateful to Lucas Pahl for help with Example 7. Names are in random order, following Ray ⓡ Robson (2018). We dedicated 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. Tapan Mitra died February 3, 2019.

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. This result 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-based externalities. The answer is that they *can* be so written (see Theorem 3), but no matter which payoff function we use to represent that conversion, either coherence or continuity must fail. Since the connection between coherence and efficiency is far from obvious, 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 libertarian 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 $u_{-i} \equiv \{u_j\}_{j \neq i}$:

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

A collection $(N, \{X_i, f_i\})$ is a *strategic situation with payoff-based 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}. Except in Section 6, no continuity condition is imposed with respect to x_i; in fact, no topological restrictions are placed on X.

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

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

[1] See also Bergstrom (1970) in the special context of a distributive Lindahl equilibrium with non-malevolent agents and, for a somewhat different perspective, Green (1982), (2003).

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

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}. As Pearce (1983), Kimball (1987), Bergstrom (1999), Vasquez and Weretka (2018) and others have observed, this interdependent system could pose several analytical difficulties.

First, the system may have no solution. To borrow an example from Pearce (1983), suppose that there are two players, and for some x, we have

$$f_i(x, u_{-i}) = 1 + u_{-i} \text{ for } i = 1, 2.$$

Then there is no solution to this interdependent utility system.

It's also possible that the system admits *multiple* solutions. Consider a two-person example from Vasquez and Weretka (2018), where

$$f_i(x, u_{-i}) = \sqrt{u_{-i}} \text{ for } i = 1, 2,$$

so that there are two solutions: $u = (0, 0)$ and $u = (1, 1)$. This reflects a situation in which two people who love each other could both be either happy or unhappy. Or one could generate asymmetric multiplicities; for instance, in situations with intense antipathy. Multiple solutions at some action profile make it impossible to unambiguously pin down the payoffs for that profile.

Finally, even if there is a unique solution, there may be cascades of self-reinforcing echoes that amplify without bound. In other words, the solution may be "unstable" in the sense that a small departure from it will cause payoffs to explode. Drawing on Pearce (1983) and Bergstrom (1989), suppose that there are two players, and

$$f_i(x, u_{-i}) = x_i + 2u_{-i} \text{ for } i = 1, 2.$$

Then

$$u_i = -\frac{x_i + 2x_{-i}}{3} \tag{2}$$

is the unique solution to the system, but it is an absurd solution. In the words of Pearce (1983), slightly adjusted for notation: "although the functions seem to indicate that the individuals enjoy x_1 and x_2 and are mutually sympathetic, the reduced form indicates that both u_1 and u_2 are increased when x_1 or x_2 is decreased! This does not correspond to any plausible dynamic adjustment story, such as the following. If x_1 is reduced by 1 unit, this should 'initially' lower u_1 by 1, which would then cause u_2 to fall 2 units, diminishing u_1 by a further 4 units. This downward spiral does not converge . . . such a counter-intuitive state of affairs might be called instability."

We don't mean to dismiss any of these cases out of hand, and in our discussion below we will argue for an entirely different perspective on such matters. For now, we only wish to consider those strategic situations that are well-behaved in the sense of not being plagued by the problems above. These are situations which can be unambiguously converted into a well-defined *game*, with a unique and "stable" vector of payoffs at every action profile. This is exactly the case considered by Pearce, Bergstrom and others, and as shall see, it yields a remarkable conclusion on the efficiency of equilibria. This result, in turn, will shed some new light on the cases above, once we consider standard normal-form cases and attempt to "work backwards" to their representation as a strategic situation with payoff-based externalities. With these considerations in mind, say that a strategic situation is *coherent* if for every x,

(i) [boundedness] there is $B(x) < \infty$ such that $\|f(x,u)\| < \|u\|$ whenever $\|u\| > B(x)$, where $\|\cdot\|$ is the sup norm;

(ii) [uniqueness] the mapping $f(x,\cdot)$ has a unique fixed point.

Indeed, 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 with player set $N-S$ given by $f_j(x_j, u_{-(S\cup j)}, \bar{u}_S)$, for $j \in N-S$.

Of course, sufficient conditions for coherence can be provided. Pearce (1983), Hori and Kanaya (1989), and Bergstrom (1999) do so. For instance, Pearce (1983) considers the special case of "mutual sympathy," placing bounds on the extent to which u_j affects u_i; these emerge as a Hawkins-Simon condition on the matrix of cross-derivatives. His conditions guarantee both (i) and (ii) above. Or one could presume — mutual sympathy or not — that the f-mapping is a contraction for each action profile. Or assume that for each action profile, the associated directed network of payoff interdependencies is acyclic, and that payoff functions are bounded. That, too, will generate coherence.

Note that coherence is a "robust property": for instance, in the class of all differentiable interdependent utility systems, a perturbation of the Pearce sufficient conditions will still yield coherence. (To be sure, the *lack* of coherence is robust as well; neither coherence nor the lack of it is "generic.") Note, moreover, that any of the above sufficient conditions for coherence will also imply sub-coherence.

"No explosions" and "uniqueness" are independent restrictions. For instance, the former (but not the latter) would follow immediately if we assume that f is bounded, as in Vasquez and Weretka (2018). Our third example above shows that uniqueness can coexist with lack of boundedness.

Coherence is our starting point. Without it we may be 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. Thus a strategic situation generates a well-defined normal form game with payoffs $U(x)$, where $U(x) = f(x, U(x))$.[3] 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,

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

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

3. Applications and Examples

Beginning with Veblen (1899) and Duesenberry (1949), there has been increasing interest in studying economic agents who obtain payoffs (altruistic or invidious) from the well-being of others. Several studies emphasize the relativistic nature of happiness or individual welfare, among them Easterlin (1974), Frank (1985, 1989), Clark and Oswald (1996), Ray and Robson (2012), and many others. These papers typically emphasize invidious comparisons.

[3] We use upper case U to denote the utility function for the normal form game and lower case u for payoff values.

[4] 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.

There is a also large literature on altruism; for a small sample, see, e.g. Strotz (1955), Phelps and Pollak (1968), Ray (1987), Andreoni (1989), Galperti and Strulovici (2017), and of course the references to Pearce, Bergstrom, Kimball and others already cited. One standard formulation presumes that individuals are affected by some measure of the economic standing of others, such as their income, wealth or consumption. But a subset of this literature also emphasizes a "non-paternalistic" formulation in which individuals care — positively or negatively — about the *payoffs* of other individuals. It is common in macroeconomics, for instance, to interpret the value functions of dynamic programming as intergenerational altruism; see, for instance, Barro (1974) or Loury (1981). This is not to suggest that *all* models of interpersonal externalities are profitably written in this way. We may envy or admire the economic positions of others rather than the payoffs that are derived from them. But there are other situations where the non-paternalistic approach is more relevant, and those are the ones we seek to study here.

For instance, and apart from the above interpretation of dynamic programming, the foundations of modern welfare economics laid down by Bergson and Samuelson rely on a non-paternalistic form of interpersonal sympathy. While Bergson's notion of a social welfare function is usually seen as a useful device for studying Pareto optimality in an economy *without* externalities, Samuelson (1981) shows how this idea extends to the case of non-paternalistic agents with sympathy or envy towards others. Suppose each agent has a "private" or "first-level" utility function that depends only on her own consumption of goods. A non-paternalistic agent who does care about others can then be modeled as one having a "final" or "second-level" utility function that is a weakly separable function of *all* the first-level utility functions; see Section 7.2 for further details.

The model of interdependent utilities that we study also connects with the social psychology literature on empathy. Vasquez and Weretka (2018) argue that it captures the psychological phenomenon of affective empathy and emotional contagion, to be contrasted with the notion of cognitive empathy, which works through intentions and beliefs of others, as in Geanakoplos, Pearce and Stacchetti (1989). They also discuss the relationship with "material games," where the interdependence is modeled through material outcomes such as money or consumption.[5]

Here are four examples that illustrate the concept of interdependent utilities. We've deliberately chosen them to illustrate situations other than the obvious ones of altruism or envy.

Example 1. *Aspirations.* For some collection of continuous return and cost functions $\{g_i\}$ and $\{c_i\}$,

$$u_i = g_i\left(x_i, a_i\right) - c_i(x),$$

where a_i is the average value of u_j over some reference group of individuals that influences i. To interpret this, suppose that person i lives in a society in which she is influenced by the average payoff of her reference group. She might observe their lifestyles and economic position, but in the example she is moved, in the end, by how happy they are. The use of perceived or imagined happiness as reference points is, in many cases, just as defensible as the use of "objective economic position."

Think of this average payoff as an "aspiration" for person i, a reference point that affects her not just intrinsically but instrumentally.[6] That is, if x_i is a costly investment in her own life, it will bring a return g that is influenced — positively or negatively — by the aspiration she has. Apart from this "intrinsic effect," there will be an "instrumental effect" on her choice of x_i — her aspirations might serve either to

[5]See Sobel (2005) for a survey.

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

inspire or frustrate investment — the cross-partials of g will determine that outcome. The more general point is that individual payoffs could be affected in this way by own actions and the payoffs of others. Of course, assumptions on g will need to be placed to guarantee coherence and sub-coherence, such as a contraction. Alternatively, if payoff functions are bounded and the associated directed network of payoff interdependencies is acyclic, then coherence will hold automatically by a recursive argument.

Example 2. *Industrialization.* We borrow from the multiple equilibrium theory 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. 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-based 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. *Regulation.* Again there are n firms, but this time the cross-firm externality will be negative. Each firm 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 to be continuous and decreasing. A lower γ increases the chances of social animosity against the firms, and consequently of a regulation placed on the firms, generating a penalty $\pi(\gamma)$, where π is continuous. The payoff for each firm is therefore given by

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

Notice how we define γ on the net payoff of each firm, with everything taken into account, including the penalty. 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 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-based 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-based 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. In other words, although the prisoner's dilemma can be made to fit the definition of a strategic situation with payoff-based externalities, it does *not* result in a game of love and hate.

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 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. (But see Section 7.3 for more on the special case of positive externalities, or "love".)

Perhaps the theorem is best appreciated by reading its proof in detail, but as the argument is long, we provide the reader with a discussion and outline. 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

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

because a unilateral deviation to $\bar{x}_i$ from x_i^* cannot be profitable for i. A central observation (Lemma 1) 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

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

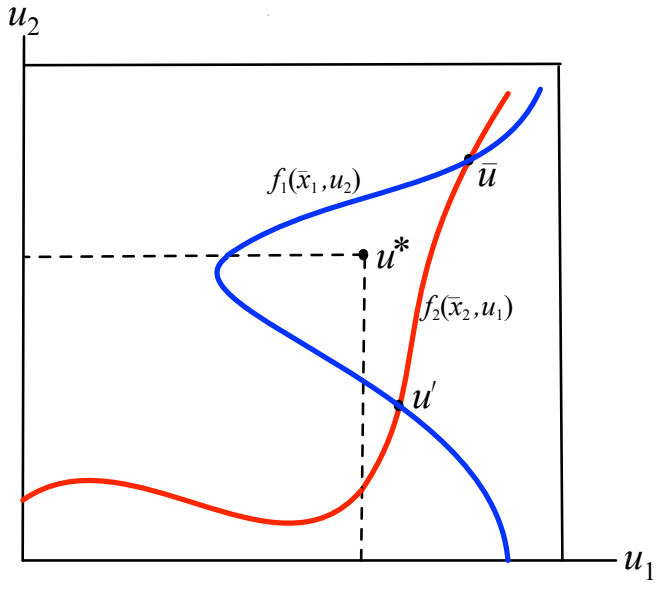


FIGURE 1. Proof of Theorem 1 with Two Players.

With (3) and (5) in hand, we graphically illustrate the proof for $n = 2$. Figure 1 draws the two functions $f_1(\bar{x}_1, u_2)$ and $f_2(\bar{x}_2, u_2)$, where the action vector is fixed throughout at $\bar{x} = (\bar{x}_1, \bar{x}_2)$. By boundedness, all the action takes place in some compact rectangle. The inequality (3) is represented graphically by the location of $\bar{u}$ to the northeast of the utility vector u^*. Next, inequality (5) places one restriction each on the functions f_1 and f_2. The function $f_1(\bar{x}_1, u_2)$ evaluated at $u_2 = u_2^*$ must lie to the "left" of u^*, and the function $f_2(\bar{x}_2, u_1)$ evaluated at $u_1 = u_1^*$ must lie "below" u^*. But the former function continues on and eventually hits the horizontal axis, while the latter must eventually hit the vertical axis. By continuity, then, the two functions must cross at some $u' \neq \bar{u}$. But now we have a contradiction to the uniqueness of the interdependent utility system at $\bar{x}$. The argument generalizes beyond two-person games in other special cases. For instance, if the game in question is one of "love," so that all individual utilities are nondecreasing in the utilities of others, a Tarski-style fixed point theorem reveals the existence of an additional fixed point to the "southwest" of u^*. That, too, leads to a contradiction to the uniqueness of the interdependent utility system at $\bar{x}$. It should be noted that this "special case" is the central case of interest in Pearce (1983) and Bergstrom (1999).

It is also possible to construct a direct argument in the case of "hate."

But no direct analogue of this argument in the general case is possible in three or more dimensions. For instance, with $n = 3$, we can certainly depict inequalities (3) and (5) graphically, as we have just done for two players. But there is no guarantee that an additional fixed point to $f(\bar{x}, .)$, apart from $\bar{u}$, will exist, so that a contradiction can be established. Indeed, Example 7 will show exactly this, and that sub-coherence will need to be invoked to go further. That necessitates a very different argument.

In the general case, we proceed by induction. 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 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 $U(\bar{x}) = \bar{u}$,*

$$f(\bar{x}, \bar{u}) \geqslant \bar{u} > u^* \geqslant f(\bar{x}, u^*). \tag{6}$$

Theorem 1 follows from this Claim. Suppose that x^* is an equilibrium, but it is not Pareto optimal. Then (3) and (5) hold for some $\bar{x} \neq x^*$. But these together imply (6), which contradicts the Claim.

The remainder of the proof establishes the Claim using induction on n. Specifically, we show that if (6) 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 (6) is true as well. But it is very easy to see that for a *single-person* game, (6) 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 (6) can never be true.

This attempt to provide an intuitive argument for Theorem 1 does not adequately highlight the importance of coherence. In fact, as a reading of the formal proof will indicate, both parts of coherence are needed for these arguments. Neither boundedness nor uniqueness can be dropped from Theorem 1 or Lemma 1, which connects profitable and naively profitable deviations. To see this in an explicit example, modify the "unstable equilibrium" from Section 2, described in the discussion leading to (2).

Example 5. *Unstable Hate*. Let $n = 2$, and suppose that both players hate each other:[7]

$$f_i(x, u_{-i}) = x_i - 2u_{-i}, \text{for } i = 1, 2.$$

Let $X_i = [0, 1]$, $i = 1.2$. While boundedness fails, there is a unique solution:

$$u_i = -\frac{1}{3}x_i + \frac{2}{3}x_{-i}, \text{ for } i = 1, 2.$$

Clearly, the unique equilibrium is $x^* = (0, 0)$, with $u^* = u(x^*) = (0, 0)$. The conclusion of Lemma 1 does not hold because at x^* if we keep u_{-i} fixed at 0, player i has a naively profitable deviation to $x_i = 1$. And the conclusion of Theorem 1 does not hold because $x = (1, 1)$ Pareto dominates x^*.

It is easy to see that if we modify this example to impose boundedness we will lose uniqueness. Thus, neither part of coherence can be dropped from our main result. We will return to these matters from a different perspective in Section 7.4.

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

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

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 $u_{-(S \cup j)}$ 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 $U^{-S}(x, u_S)$. Note that $U^{-S}(x, u_S)$ depends only on x_{N-S} and u_S; it is insensitive to x_S. Because $f_i(x_i, .)$ is continuous for all i and x_i, the

[7] We are grateful to Ted Bergstrom for constructing this example and the accompanying story. Hatfield and McCoy hate each other intensely and derive pleasure from their own consumption of whiskey, which alas is bounded above by 1.

fixed points of each reduced game are upper-hemicontinuous in u_S. Uniqueness of the fixed point then implies that for each x, $U^{-S}(x, u_S)$ 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 $U^{-i}(x, u_i)$. It will useful to introduce notation that describes how $U^{-i}(x, u_i)$ maps back to i's payoff in the original game. That is, define

$$\phi_i(u_i, x) = f_i(x_i, U^{-i}(x, u_i)).$$

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 $U^{-i}(x, u_i)$ in u_i that $\phi_i(u_i, x)$ is a continuous function of u_i.

For any x, the fixed point of $f(x, \cdot)$ is closely related to that of $\phi_i(\cdot, x)$. Recall that $U(x)$ is the unique fixed point of $f(x, \cdot)$. Because $U^{-i}(x, U_i(x))$ is the unique solution to the reduced system given x and $U_i(x)$, we have $U_{-i}(x) = U^{-i}(x, U_i(x))$. Therefore $\phi_i(U_i(x), x) = f_i(x_i, U^{-i}(x, U_i(x))) = f_i(x_i, U_{-i}(x)) = U_i(x)$; 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 $\tilde{u}_{-i} = U^{-i}(x, \tilde{u}_i)$. Then $\tilde{u}$ satisfies (1), but because $\tilde{u}_i \neq \bar{u}_i$, this contradicts the uniqueness assumption. Thus, $U(x)$ is a unique fixed point of $f(x, \cdot)$ if and only if $U_i(x)$ is the unique fixed point of $\phi_i(\cdot, x)$ for every i. The function ϕ_i will turn out to be useful in checking equilibrium conditions.

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

Proof. For any pair of action profiles $x', x'' \in X$, let $u' = U(x')$ and $u'' = U(x'')$. We claim that

$$u'_i > u''_i \text{ if and only if } \phi_i(u''_i, x') > \phi_i(u''_i, x'') = u''_i. \tag{7}$$

To see why this is so, suppose $\phi_i(u''_i, x') > u''_i$, as shown in Figure 2A. Since $\phi_i(., x')$ is continuous and $\phi_i(B, x') < B$ for B large (by boundedness), the intermediate value theorem tells us that there is $\tilde{u}_i > u''_i$ with $\tilde{u}_i = \phi_i(\tilde{u}_i, x')$. Since $\phi_i(\cdot, x')$ has a unique fixed point, $\tilde{u}_i = u'_i$, so $u'_i > u''_i$. Conversely, if $\phi_i(u''_i, x') \leqslant u''_i$ (Figure 2B), then using the fact that $\phi_i(-B, x') \geqslant -B$ for B large enough (by boundedness), we know that there is $\tilde{u}_i \leqslant u''_i$ such that $\tilde{u}_i = \phi_i(\tilde{u}_i, x')$. Since $\phi_i(\cdot, x')$ has a unique fixed point, $\tilde{u}_i = u'_i$, which implies $u'_i \leqslant u''_i$, and so establishes (7).

Now suppose that i deviates from x^* to x_i. Let $u^* = U(x^*)$ and $y = (x_i, x^*_{-i})$. By (7),

$$U_i(y) = U_i(x_i, x^*_{-i}) > u^*_i \text{ if and only if } \phi_i(u^*_i, y) > u^*_i. \tag{8}$$

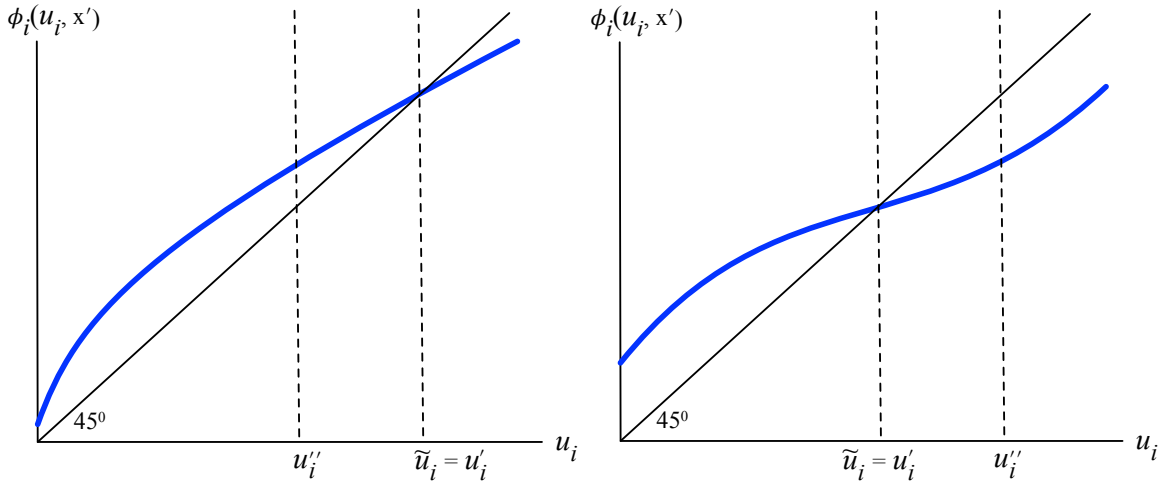
Because $U^{-i}(x, u_i)$ is insensitive to x_i, and $y_{-i} = x^*_{-i}$, we have $U^{-i}(x, u^*_i) = U^{-i}(y, u^*_i)$, so that

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

Substituting this in (8) we have:

$$U_i(x_i, x^*_{-i}) > u^*_i \text{ if and only if } f_i(x_i, u^*_{-i}) > u^*_i, \tag{9}$$

which establishes the desired result. ∎



(A) $\phi_i(u_i'', x') > u_i''$ (B) $\phi_i(u_i'', x') \leqslant u_i''$

FIGURE 2. A Step in the Proof of Lemma 1.

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 (6) can never hold.

Now for the inductive step. Suppose that the Claim is true of every game with $m < n$ players 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 $\bar{x}$, and payoff profiles u^* and $\bar{u}$, such that (6) is satisfied. Now we consider the following two possibilities.

Case 1. There is some player, say k, such that in the reduced game without k (and with $u_k = u_k^*$),

$$U^{-k}(\bar{x}, u_k^*) > u_{-k}^*. \tag{10}$$

In this case, define $S = \{k\}$. In the reduced game with player set $N - S$, define $u^r \equiv U^{-k}(\bar{x}, u_k^*)$. Then we have

$$f^{-k}(\bar{x}, u^r) = u^r > u_{-k}^* \geqslant f^{-k}(\bar{x}, u_{-k}^*) \tag{11}$$

where the first equality is by definition of u^r and the next inequality is from (10). The final weak inequality follows from two observations. First, the last weak inequality of (6) holds, by our contradiction assumption. Second, in the reduced game k has been removed with utility level u_k^*, so $f^{-k}(\bar{x}, u_{-k}^*)$ is $f(\bar{x}, u^*)$ restricted to $N - k$. But the reduced game satisfies all the conditions of the theorem, so that (11) contradicts the induction hypothesis.

Case 2. For every k, (10) fails. Then it must be that $\bar{u} \gg u^*$. (For if not, Case 1 would hold for some k such that $\bar{u}_k = u_k^*$.) Pick any k. Then for for every $j \neq k$,

$$U_j^{-k}(\bar{x}, \bar{u}_k) = \bar{u}_j > u_j^*. \tag{12}$$

Moreover, the fact that (10) fails implies that for some $j \neq k$,

$$U_j^{-k}(\bar{x}, u_k^*) \leqslant u_j^*. \tag{13}$$

Pick the *largest* value $\hat{u}_k \in [u_k^*, \bar{u}_k]$ such that

$$U_j^{-k}(\bar{x}, \hat{u}_k) = u_j^* \text{ for some } j \neq k. \tag{14}$$

Since $U^{-k}(\bar{x}, .)$ is continuous and (13) holds, $\hat{u}_k$ is well defined. Moreover, given (12), it must be that $\bar{u}_k$ is not binding for $\hat{u}_k$:

$$u_k^* \leqslant \hat{u}_k < \bar{u}_k. \tag{15}$$

Let S be the set of agents for whom (14) holds. Note that $k \notin S$ (so S is a strict subset of N), and if some other $i \neq k$ is also not in S, then

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

Consider the reduced game induced on players $N - S$ by setting $u_S = u_S^*$. It has payoff functions

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

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. Define a utility profile u^r on $N - S$ by

$$u_i^r = \begin{cases} \hat{u}_k & \text{for } i = k, \\ U_i^{-k}(\bar{x}, \hat{u}_k) & \text{for } i \neq k, i \in N - S \end{cases} \tag{17}$$

We claim that in the reduced game,

$$f^{-S}(\bar{x}, u^r) \geqslant u^r \geqslant u_{N-S}^* \text{ while } f^{-S}(\bar{x}, u^r) \neq u^r \text{ or } u^r \neq u_{N-S}^*. \tag{18}$$

To this end, first note that if $i \in N - S$ and $i \neq k$, (16) and (17) together imply that:

$$u_i^r = U_i^{-k}(\bar{x}, \hat{u}_k) > u_i^* \text{ for all } i \in N - S - k. \tag{19}$$

Next, by definition, $U^{-k}(\bar{x}, \hat{u}_k)$ solves the interdependent utility system at $\bar{x}$ with k removed, which means in particular that for all $i \in N - S - k$,

$$U_i^{-k}(\bar{x}, \hat{u}_k) = f_i^{-k}(\bar{x}, U_{-i}^{-k}(\bar{x}, \hat{u}_k)) \tag{20}$$

But $U_j^{-k}(\bar{x}, \hat{u}_k) = u_j^*$ for $j \in S$ and $U_j^{-k}(\bar{x}, \hat{u}_k) = u_j^r$ for $j \in N - S - k$, while $u_k^r = \hat{u}_k$; see (17). Putting all this together with (20), we must conclude that

$$u_i^r = f_i^{-S}(\bar{x}, u_{-i}^r) \text{ for all } i \in N - S - k. \tag{21}$$

Combining (19) and (21),

$$f_i^{-S}(x_i, u_{-i}^r) = u_i^r > u_i^* \text{ for all } i \in N - S - k. \tag{22}$$

We now consider individual k. We first claim that

$$\phi_k(\hat{u}_k, \bar{x}) > \hat{u}_k. \tag{23}$$

For if (23) were false, then $\phi_k(\hat{u}_k, \bar{x}) \leqslant \hat{u}_k$. Because $\phi_k(-B, \bar{x}) \geqslant -B$ for B large enough and ϕ_k is continuous, there exists $u_k' \leqslant \hat{u}_k$ such that $\phi_i(u_k', \bar{x}) = u_k'$. But that generates a utility solution $(u_k', U^{-k}(\bar{x}, u_k'))$ at the action profile $\bar{x}$ of the original game. It is distinct from $\bar{u}$ because $u_k' \leqslant \hat{u}_k < \bar{u}_k$, using (15). That violates coherence.

Next, observe that

$$\phi_k(\hat{u}_k, \bar{x}) = f_k(\bar{x}, U^{-k}(\bar{x}, \hat{u}_k)) = f_k(\bar{x}, u^r_{N-S-k}, u^*_S) = f_k^{-S}(\bar{x}, u^r_{-k}), \tag{24}$$

where the first equality is just the definition of ϕ_k, the second equality follows from $u^*_S = U_S^{-k}(\bar{x}, \hat{u}_k)$ (see (14)) and the definition of u^r in (17), and the last equality is just the translation to the reduced game where S is excluded with payoff u^*_S.

Combining (23) and (24) along with $u^r_k = \hat{u}_k \geqslant u^*_k$ (the equality is from (17) and the inequality from (15)), we must conclude that

$$f_k^{-S}(\bar{x}_k, u^r_{-k}) > u^r_k \geqslant u^*_k. \tag{25}$$

Combining (22) and (25), we obtain (18) for the reduced game, as claimed.

We now use (18) to prove that (6) holds for the reduced game. First, recall that by way of contradiction, we've maintained that (6) holds for the original game, so that

$$u^*_i \geqslant f_i(\bar{x}_i, u^*_{-i}) \text{ for all } i \in N - S.$$

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

$$u^*_i \geqslant f_i^{-S}(\bar{x}, u^*_{-i}) \text{ for all } i \in N - S. \tag{26}$$

Now combine (18) and (26) to see that

$$f^{-S}(\bar{x}, u^r) \geqslant u^r \geqslant u^*_{N-S} \geqslant f^{-S}(\bar{x}, u^*_{N-S}), \text{ with } f^{-S}(\bar{x}, u^r) \neq u^r \text{ or } u^r \neq u^*_{N-S} \text{ (or both).} \tag{27}$$

To obtain (6) from (27), we claim that $u^r > u^*_{N-S}$. If not, then (given that (27) 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 (27) cannot hold, a contradiction. So $u^r > u^*_{N-S}$, as claimed, and (6) 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, (6) 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

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

where the last inequality makes use of Lemma 1 and the fact that x^* is an equilibrium. But this implies (6), a contradiction. ∎

6. EXISTENCE

Under what conditions does a game of love and hate possess a Nash equilibrium? One approach to answering this question is to see when standard results apply, once the strategic situation has been converted to the normal form. For instance, if for every i, X_i is non-empty, compact and convex subset of a finite dimensional Euclidean space and $U_i(.) : \prod_j X_j \to \mathbb{R}$ is continuous and quasi-concave in x_i, then the existence of an equilibrium is assured. But this is not satisfactory, because these conditions on U_i should be derived from the primitives of a strategic situation with payoff-based externalities. We

therefore directly examine the strategic situation. It turns out that this approach has the added advantage of yielding an existence result *for pure strategy equilibrium* that is free of any convexity assumptions. We assume that:

(A1.) For all i, X_i is a non-empty, compact subset of a topological space and $f_i : X_i \times \mathbb{R}^N \to \mathbb{R}$ is a continuous function in all its arguments.

We also strengthen boundedness to require a uniform bound for all $x \in X$.

(A.2) There is $B < \infty$ such that for every $x \in X$, $\|f(x,u)\| < \|u\|$ whenever $\|u\| > B$, where $\|\cdot\|$ is the sup norm.

THEOREM **2.** *Suppose the strategic situation* $(N, \{X_i, f_i\})$ *is coherent, sub-coherent and satisfies* (A.1) *and* (A.2)*. Then the induced game of love and hate has an equilibrium in pure strategies.*

Proof. Given B as in (A.2), let $\mathcal{B} = \{u \in \mathbb{R}^N \mid \|u\| \leqslant B\}$. Because f_i is continuous and $X_i \times \mathcal{B}$ is compact, $B_i \equiv \max f_i(x_i, u)$ is well-defined for $(x_i, u) \in X_i \times \mathcal{B}$, for every i. Let $C = \max\{B_1, \ldots B_n, B\}$ and $\mathcal{C} = \{u \in \mathbb{R}^N \mid \|u\| \leqslant C\}$. By construction, the range of f lies in $\mathcal{C}$ for all $u \in \mathcal{B}$. Because $C \geqslant B$, (A.2) implies that the same is true for all $u \in \mathcal{C}$. So for all $(x,u) \in X \times \mathcal{C}$, $f(x,u) \in \mathcal{C}$.

Let $\beta_i : \mathcal{C} \to X_i$ be player i's naive best response correspondence:

$$\beta_i(u) = \{x_i \in X_i \mid f_i(x_i, u_{-i}) \geqslant f_i(x_i', u_{-i}) \text{ for all } x_i' \in X_i\}.$$

Given (A.1), Berge's maximum theorem implies that for every i, β_i is non-empty and upper hemicontinuous and the maximum function, $\gamma_i : \mathcal{C} \to [-C, C]$, is continuous, where

$$\gamma_i(u) = \{f_i(x_i, u_{-i}) \mid x_i \in \beta_i(u_{-i})\}.$$

Since $\gamma = \prod_i \gamma_i : \mathcal{C} \to \mathcal{C}$ is a continuous function, by Brouwer's fixed point theorem, it has a fixed point u^*. For every i, pick any $x_i^* \in \beta_i(u^*)$. We claim that $x^* = (x_i^*)$ is an equilibrium. Clearly, $U(x^*) = f(x^*, u^*) = u^*$ and for every i, x_i^* is a naive best response to u_{-i}^*:

$$f_i(x_i^*, u_{-i}^*) \geqslant f_i(x_i, u_{-i}^*) \text{ for all } x_i \in X_i.$$

From Lemma 1, the lack of a naively profitable deviation implies that no player has a profitable deviation at x^*, i.e., x^* is an equilibrium. ∎

7. DISCUSSION

7.1. **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 curvature condition on payoffs or 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 for a coherent situation (Lemma 1), these conditions are:

$$\frac{\partial f_i(x)}{\partial x_i} = 0 \text{ for all } i. \tag{28}$$

As we saw in Example 5, this condition may not characterize an equilibrium in the absence of coherence. The intuition discussed below relies on (28) and therefore presumes coherence.

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

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

where $\lambda \equiv (\lambda_1, \ldots, \lambda_n)'$ 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

$$D_x \lambda = 0, \tag{29}$$

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

$$D_x = F + D_x D_u, \tag{30}$$

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 (30), we see that

$$D_x = F[I - D_u]^{-1}, \tag{31}$$

where the presumption that $I - D_u$ has an inverse is closely connected to coherence; see Pearce (1983). Combining (29) and (31), we must conclude that the first order conditions for a solution to the planner's problem are

$$F[I - D_u]^{-1}\lambda = 0. \tag{32}$$

We can open this out as follows. Let b_{ij} be a generic entry for the matrix $[I - D_u]^{-1}$; then (32) is equivalent to the condition

$$\left[\frac{\partial f_i}{\partial x_i}\right]\left[\sum_{j=1}^{n} b_{ij}\lambda_j\right] = 0 \text{ for all } i. \tag{33}$$

Using (28), we see that a solution to the equilibrium first order conditions is also a solution to the planner's first order conditions (33), suggesting that equilibria are Pareto optimal, or at least solve necessary conditions for optimality. We reiterate: this is suggestive but incomplete, *even* with the coherence, smoothness and curvature assumptions in place. Moreover, without coherence or sub-coherence, Theorem 1 isn't even generally true; see Examples 5 and 7. For more discussion, see the Appendix.

7.2. **Connection with the First Theorem of Welfare Economics.** There is a parallel between Theorem 1 and the first fundamental theorem of welfare economics. In the latter setting, individual utilities are defined on own consumption, but there is a potential externality: my consumption comes at the expense of yours. That externality is mediated through the price system — an agent who consumes more pays for it at precisely the rate of the marginal deprivation caused to others, and so behaves as if she is socially responsible. The first welfare theorem states that the resulting outcome is efficient. In contrast, there are no prices in Theorem 1, so there is no way to pay for the externalities. But the point is that there are no externalities inflicted locally when an agent hits a maximum, because her payoff is (locally) flat. As indicated in Section 7.1, this argument is far from complete, but in any case, our theorem depends on the vanishing of the externality itself, without the assistance of prices.

These are distinct channels leading to a similar final outcome. To appreciate this distinctness, overlay one model on the other, so that we have a setting with prices *and* with payoff-based externalities. Then the resulting equilibrium is generally not Pareto optimal. 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. Indeed, 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).[8] It identifies a form of "non-benevolence," which is related to though 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 in this paper. (Here, 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.)

Theorem 1 is cast in the setting of strategic situations as opposed to competitive equilibrium. 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

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

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 payoff-based externalities restores that efficiency, irrespective of the particular form of those externalities.

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 (33) imply (28)?" This is not a question we investigate here in any generality, though the Appendix contains a discussion.[9]

Some of this cited literature on welfare theorems avoids the coherence issue via a two-step procedure for defining final payoffs. In Samuelson's (1981) terminology, each consumer i has a "first-order" utility function" $w_i(x_i)$ that depends only on i's commodity bundle. Externalities are introduced through a "second order" Bergsonian utility function:

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

Samuelson (1981) observes that one could consider a "third order" utility function that depends on own consumption and other's second order utility functions, and so on. The logical conclusion of this idea is to consider not some arbitrary, finite iteration but a steady-state in which conjectures about others' utility are accurate.[10] In other words, require U_i to be a function of $w_i(x_i)$ and $U_{j\neq i}$, which is precisely the model of interdependent utilities we have adopted. Under some conditions, Samuelson's formulation can be viewed as a special case of our framework.[11]

7.3. **Love, Hate and the Shape of the Utility Possibility Frontier.** Even though our theorem holds for fully general patterns of love and hate, the special case of positive externalities or "love" has an interesting structure. The corresponding utility possibility frontier has a single payoff profile, which Pareto dominates all other payoff profiles.[12] By our main theorem, every equilibrium picks out this unique payoff profile. (We reiterate that the result continues to be non-trivial even in this special case,

[9]With externalities, 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).

[10]See also the discussion in Vasquez and Weretka (2018). Samuelson doesn't pursue this line on the grounds that in the end, if this process converges, each agent's utility function will be a function of all the first-order functions anyway.

[11]This follows from the proof of the main result in Green (2019), who examines welfare criteria when agents have both subjective preferences and objective interests, which leads naturally to a model of (ordinal) interdependent preferences. An intermediate step in the proof of his main result provides conditions under which such preferences can be represented by means of real-valued functions over the set of alternatives X, as follows. For each agent there is a function $U_i : X \to \mathbb{R}$, representing subjective preferences, and another function $w_i X \to \mathbb{R}$, representing objective interests, where $U_i(x) = \sum_{j=1}^{n} \alpha_{ij} w_j(x) + a_i$. This is his condition (16), which formally corresponds to the additively separable form of the individualistic Bergsonian form stated above. Green shows that under his assumptions, this can be expressed as $U_i(x) = w_i(x) + \sum_{j\neq i} \delta_{ij} U_j(x) + b_i$, which is, of course, the additive form of the model of interdependent preferences we are studying.

[12]This is to be contrasted with the case of an exchange economy, where x_i refers to i's consumption rather than action, and aggregate consumption is required to equal aggregate endowment. In that model, even with "universal love", the utility possibility frontier can have the usual shape; see the Romeo and Juliet example in Bergstrom (1989).

as games of common or even identical interests could exhibit coordination failures in general, whereas here they don't.)

To see why this assertion is true, assume in what follows that for every i, $f_i(x_i, u_{-i})$ is nondecreasing in u_{-i}. Consider any two distinct payoff profiles u and u', and let x and x' be two corresponding action profiles with these payoffs. Define a utility profile $\hat{u}$ by $\hat{u}_i = \max\{u_i, u'_i\}$ for all i. Define an action profile x'' by setting x''_i to i's action under one of the action profiles x or x' that results in payoff $\hat{u}_i$ to i. Now pick any i and suppose without loss of generality that $x''_i = x_i$. Observe that

$$f_i(x''_i, \hat{u}_{-i}) = f_i(x_i, \hat{u}_{-i}) \geqslant f_i(x_i, u_{-i}) = u_i = \max\{u_i, u'_i\} = \hat{u}_i.$$

By boundedness and Tarski's fixed point theorem, the unique fixed point of $f(x'', .)$ — call it u'' — must weakly dominate $\hat{u}$. But $\hat{u}$ in turn weakly dominates both u and u''. This proves that the utility possibility frontier in a strategic situation of love must consist of a unique payoff profile. The Appendix includes a simple example to illustrate this point.

In strategic situations with "universal hate," and *a fortiori* in "mixed situations," the utility possibility frontier looks more conventional, with nontrivial segments. Moreover, as already noted, the set of equilibria may not fully cover such frontiers. The Appendix contains a more detailed discussion.

7.4. **The Role Played by Coherence.** In Pearce (1983), Hori and Kanaya (1989), Bergstrom (1999), and Vasquez and Weretka (2018), 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 2018), 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-based 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 $U_i : X \to \mathbb{R}$ — 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. It is easy to see that if this condition doesn't hold it may be impossible to express a normal form game as a situation with payoff-based externalities. But if it does hold, we have:

THEOREM **3.** *Every regular game with continuous payoffs can be represented as a continuous strategic situation with payoff-based 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, but it should also be clear at this step that regularity *is* needed. 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 $\mathbb{R}^{n-1}$.

But equilibrium inefficiency is rife among games in general. How can Theorem 1 be reconciled with Theorem 3? 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.[13] We alluded to this already in Example 4. To explain further, consider a 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 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

$$c > d > b. \tag{34}$$

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-based 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 1 and the fact that x^* is an equilibrium (which is implied by $b < d$, as assumed in (34)) imply:

$$f_1(\bar{x}_1, d) \leqslant d. \tag{35}$$

Given the continuity of f_1, (35) and $f_1(\bar{x}_1, -m) \geqslant m$ for large m (coherence) together imply, by the intermediate value theorem, that there is $e \leqslant d$ 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 $e \leqslant d$ and $c \neq d$, this implies that $c < d$, which contradicts (34). (As we shall see in Example 6, however, coherence can be restored if the representing payoff functions are allowed to be discontinuous.)

[13] 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.

7.5. **Is Coherence Alone Sufficient for Theorem 1?** Theorem 3 in Section 7.4 shows that coherence cannot be dropped from the statement of Theorem 1. For instance, the prisoner's dilemma can be transformed into a strategic situation with payoff-based externalities. Because sub-coherence holds trivially for two-person strategic situations, any such transformation must lack coherence.

But 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 6. *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-based externalities:

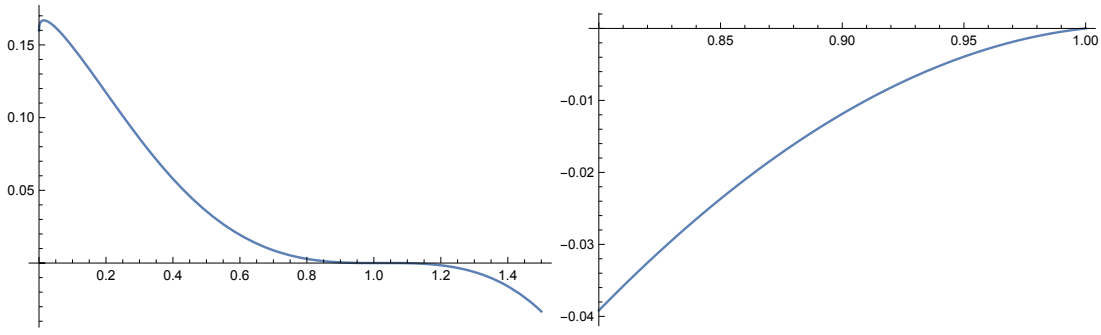
$$f_i(x_i^*, u_j) = \begin{cases} 6 - 2u_j & \text{if } u_j \leqslant 3 \\ 3 & \text{if } u_j > 3 \end{cases}$$

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

for $i, j = 1, 2$ and $j \neq i$. We now verify that this 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 $u_j \leqslant 3$, 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 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 3, 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 7. *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 $X_i = \{x_i^*, \bar{x}_i\}$



(A) Value of function in (41) (B) Derivative of function in (41)

FIGURE 3. Verifying coherence at the strategy profile $\bar{x}$ in Example 7.

for $i = 1, 2$, $X_3 = \{x_3\}$. Define the following payoff functions on $[0,1]^2$ with range in $[0,1]$:[14]

$$f_1(x_1^*, u_2, u_3) = \begin{cases} 0.95 & \text{if } u_2 \leqslant 0.6 \\ 95(u_2 - 0.7)^2 & \text{if } 0.6 \leqslant u_2 \leqslant 0.7 \\ 0 & \text{if } u_2 \geqslant 0.7 \end{cases} \tag{36}$$

$$f_1(\bar{x}_1, u_2, u_3) = u_3 \tag{37}$$

$$f_2(x_2^*, u_1, u_3) = 0.5 \text{ for all } (u_1, u_3) \tag{38}$$

$$f_2(\bar{x}_2, u_1, u_3) = \left[\frac{u_3^{1/(1+u_1)}}{1 + u_3^{1/(1+u_1)}} + 0.4 + 0.1\sqrt{2u_1 - u_1^2} \right]^{2+2u_1} \tag{39}$$

$$f_3(x_3, u_1, u_2) = u_2^{1+u_1} \tag{40}$$

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 (36)–(40). 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 (36), (38) and (40). It's easy to see that these equations have the unique solution $U(x^*) = (0.95, 0.5, 0.5^{1.95})$.

[14]It is trivial to continuously extend these functions to all of $\mathbb{R}^2$, keeping them always in the range $[0,1]$. Therefore our assertions of coherence below will remain unaffected by this extension.

Next, consider the strategy profile $\bar{x}$. Suppose u is a fixed point of $f(\bar{x},.)$. Eliminating u_3 from (37), (39) and (40) 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

$$\left[\frac{u_1^{1/(1+u_1)}}{1+u_1^{1/(1+u_1)}} + 0.4 + 0.1\sqrt{2u_1 - u_1^2}\right]^{2+2u_1} - u_1^{1/(1+u_1)} = 0. \tag{41}$$

One solution to this is clearly $u_1 = 1$. Moreover, as Figure 3 (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 3.

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.

In the second case, $x_1 = x_1^*$ and $x_2 = \bar{x}_2$. In this case the relevant equations are (36), (39) and (40). Substituting (36) and (40) into (39) 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 (36), 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:

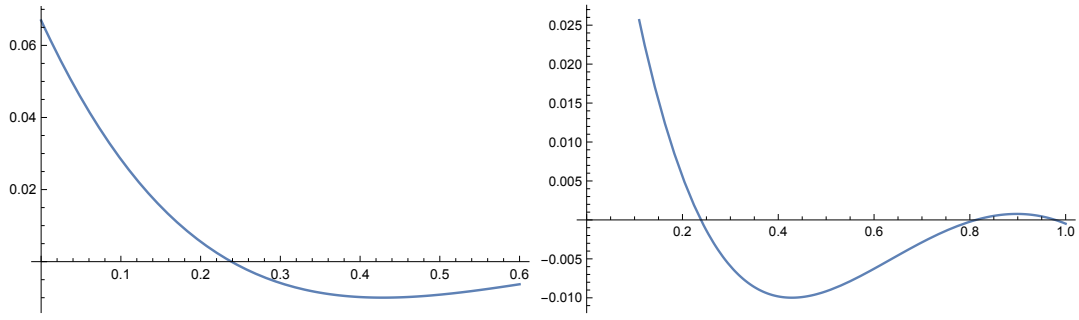
$$\left[\frac{u_2}{1+u_2} + 0.4 + 0.1\sqrt{1.9 - .95^2}\right]^{3.9} - u_2 = 0 \text{ with } u_2 \leqslant 0.6 \tag{42}$$

Panel A of Figure 4 (again plotted using *Mathematica*) depicts the left hand side of (42). 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 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 (42). As Panel A of Figure 4 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 (42), 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 4 shows. The only difference between the two panels is that in [B], the range of u_2 is $[0,1]$ rather than $[0, 0.6]$. There is indeed a unique solution to the full 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



(A) Value of function in (42) for $u_2 \in [0, 0.6]$ (B) Value of function in (42) for $u_2 \in [0, 1)$

FIGURE 4. More on coherence and sub-coherence in Example 7.

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.

7.6. **Coherence: An Afterword.** Theorem 1, as well as the subsequent discussion centered on Theorem 3, 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 payoff-based externalities — as in Examples 1, 2 and 3 — and that too, not always.[15] What we can say is that in the wider world, replete with inefficient Nash equilibria, the imposition of coherence on the payoff-based representation may be inappropriate. In summary, whether coherence is a "good" condition or not is deeply contextual. If our starting point is that the strategic situation is genuinely one of payoff-based externalities, coherence can be defended (as Pearce and Bergstrom do). If, on the other hand, the starting point is a standard normal-form game which has been straitjacketed into a strategic situation with payoff-based externalities — via Theorem 3 — then coherence is a far stronger presumption.

REFERENCES

Andreoni, J. (1989), "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence," *Journal of Political Economy* **97**, 1447–1458.

Barro, R. (1974), "Are Government Bonds Net Wealth?" *Journal of Political Economy* **82**, 1095–1117.

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

Bergstrom, T. (1989), "Love and Spaghetti, the Opportunity Cost of Virtue," *Journal of Economic Perspectives* **3**, 165–173.

[15] After all, such situations may well have no solutions, multiple solutions, or unstable solutions.

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

Clark, A. and A. Oswald (1996), "Satisfaction and Comparison Income," *Journal of Public Economics* **61**, 359–381.

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

Easterlin, R. (1974), "Does Economic Growth Improve the Human Lot? Some Empirical Evidence," in *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*, edited by M. Reder and P. David, New York: Academic Press.

Frank, R. (1985), *Choosing the Right Pond: Human Behavior and the Quest for Status*, New York: Oxford University Press.

Frank, R. (1989), "Frames of Reference and the Quality of Life," *American Economic Review* **79**, 80–85.

Galperti, S. and B. Strulovici (2017), "A Theory of Intergenerational Altruism," *Econometrica* **85**, 1175–1218.

Geanakoplos, J., D. Pearce, E. Stacchetti (1989), "Psychological Games and Sequential Rationality", *Games and Economic Behavior* **1**, 60–79.

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

Green, E. (1982), "Equilibrium and Efficiency under Pure Entitlement Systems," *Public Choice* **39**, 185–212.

Green, E. (2019), "J.S. Mill's Liberal Principle and Unanimity," arXiv:1903.07769 [econ.TH].

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.

Loury, G. (1981), "Intergenerational Transfers and the Distribution of Earnings," *Econometrica* **49**, 843–867.

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*, edited by M. Jackson and A. McLennan, Berlin, Heidelberg: Springer, 2008.

Phelps, E. and R. Pollak (1968), "On Second-Best National Saving and Game-Equilibrium Growth," *Review of Economic Studies* **35**, 185–199.

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 *What Have We Learnt About Poverty*, edited by A.Banerjee, R. Bénabou and D. Mookherjee, Oxford University Press, 409–422.

Ray, D. and A. Robson (2012), "Status, Intertemporal Choice and Risk-Taking,? *Econometrica* **80**, 1505–1531.

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

Samuelson, P. (1981), "'Bergsonian welfare economics" in *Economic welfare and the economics of Soviet socialism: essays in honor of Abram Bergson*, edited by S. Rosefielde, Cambridge University Press, 223–266.

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

Sobel, J. (2005), "Interdependent Preferences and Reciprocity," *Journal of Economic Literature* **4**, 392–436.

Vasquez, J. and M. Weretka (2018), "Affective Empathy in Non-cooperative 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 7.1 of the main text recorded necessary (and under quasi-concavity, sufficient) conditions for equilibrium, taking advantage of smoothness and coherence:

$$\frac{\partial f_i(x)}{\partial x_i} = 0 \text{ for all } i. \tag{43}$$

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

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

where $\lambda \equiv (\lambda_1, \ldots, \lambda_n)'$ 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:

$$\left[\frac{\partial f_i}{\partial x_i}\right]\left[\sum_{j=1}^{n} b_{ij}\lambda_j\right] = 0 \text{ for all } i. \tag{44}$$

Equation (44) 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 (44) 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 (43), we concluded that the equilibrium conditions are also solutions to the planner's first-order conditions (44), 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 (44) 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 (44) imply (43)?

In general, the answer to both questions could be negative (even assuming coherence), 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 (44) 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 (43) implies (44) to argue that an equilibrium is Pareto optimal.

Second, and now moving in the reverse direction, even if (44) holds at a Pareto optimum, it may not imply (43), 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. *A game of love and hate with Pareto optima that are not 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.

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

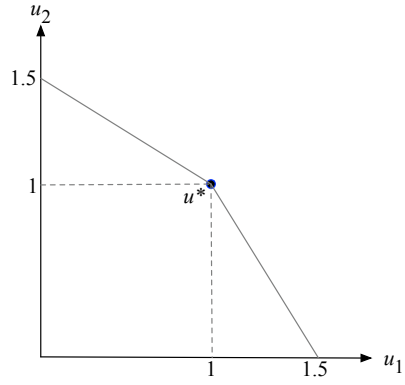


FIGURE 5. Pareto frontier for Example A.1.

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

$$U_1(x) = 1 - 2(0.5 - x_1)^2 + (0.5 - x_2)^2$$
$$U_2(x) = 1 - 2(0.5 - x_2)^2 + (0.5 - x_1)^2$$

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:

$$1 + (\lambda_2 - 2\lambda_1)(0.5 - x_1)^2 + (\lambda_1 - 2\lambda_2)(0.5 - x_2)^2. \tag{45}$$

If $\lambda \in (1/3, 2/3)$, the coefficients for $(0.5 - x_1)^2$ and $(0.5 - x_2)^2$ are both negative, (45) is strictly concave in x, and the unique solution to maximizing (45) 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 $b_{11}\lambda_1 + b_{12}\lambda_2 = 0$ in (44), and the answer to (b) is negative: (44) does not imply (43).[16] Of course, the players' utilities do depend on x_1. The case $\lambda_1 = 2/3$ is symmetric.

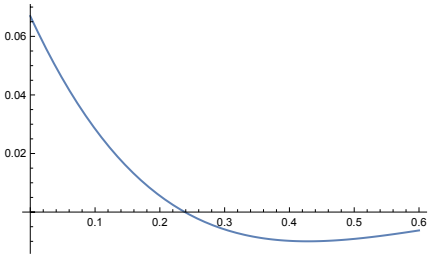
If $\lambda_1 < 1/3$, the planner's objective function, (45), becomes convex in x_1. If $\lambda > 2/3$ it becomes convex in x_2. In either case, (44) is not consistent with the maximization of the planner's objective, (45). Of course, x^* continues to satisfy these conditions but is not a solution to the planner's problem for $\lambda_1 \notin [1/3, 2/3]$.

The utility possibility frontier can be shown to have the form

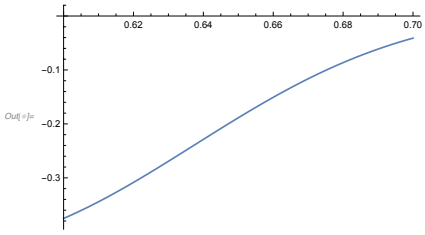
$$u_2 = \begin{cases} 1.5 - 0.5u_1 & \text{if } u_1 \leqslant 1 \\ 3 - 2u_1 & \text{otherwise} \end{cases}$$

and is depicted in Figure 5. 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, (44) 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.

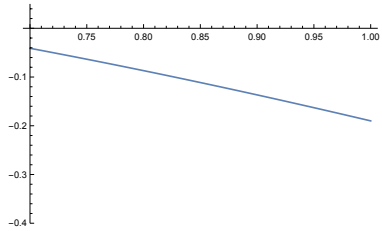
[16] It can be shown that $(b_{11}, b_{12}) = (4/3, -2/3)$.



(A) Left hand side of (46), $u_2 \leqslant 0.6$



(B) Left hand side of (47), $u_2 \in (0.6, 0.7)$



(C) Left hand side of (48), $u_2 \geqslant 0.7$

FIGURE 6. Verifying coherence at the strategy profile $\bar{x}$ in Example 7.

We know from Section 7.3 that this conventional shape of the utility possibility frontier is not possible in the case of positive externalities. To see this, modify Example A.1 as follows:

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

which yields

$$\begin{aligned} U_1(x) =& 3 - 2(0.5 - x_1)^2 - (0.5 - x_2)^2 \\ U_2(x) =& 3 - 2(0.5 - x_2)^2 - (0.5 - x_1)^2 \end{aligned}$$

Now, whatever the weights, λ_1, λ_2, there is a unique solution to the planner's problem, namely $x = (0.5, 0.5)$, and only one point on the utility possibility frontier, $u = (3, 3)$, confirming the general point made in Section 7.3.

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 (44) imply (43)?" As we saw in Example A.1, in general the answer is negative, though not so if all agents are non-malevolent, as we know from Section 7.3. Relatedly, Winter (1969) shows that if no consumer is malevolent, then the second welfare theorem holds:[17] every Pareto optimal allocation is sustainable as a competitive equilibrium with redistribution.[18] 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 3. 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^m_{-i} be a sequence of utility profiles in $\mathcal{U}_{-i}$ converging to some $u_{-i} \in \mathcal{U}_{-i}$. 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 $\mathcal{U}_{-i}$. By the Tietze extension theorem (see, for example, Willard 1970, p. 99), $f_i(x, \cdot)$ can be extended to a bounded continuous function on $\mathbb{R}^{n-1}$. ■

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

[18] 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).

C. Missing Details for Claim 2 in Example 7. The only detail for Example 7 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 (36), substituting for u_1 in this equation gives us three distinct possibilities:

$$\left[\frac{u_2}{1+u_2} + 0.4 + 0.1\sqrt{1.9 - .95^2}\right]^{3.9} - u_2 = 0 \text{ with } u_2 \leqslant 0.6 \tag{46}$$

$$\left[\frac{u_2}{1+u_2} + 0.4 + 0.1\sqrt{95}(u_2 - 0.7)\sqrt{2 - 95(u_2 - 0.7)^2}\right]^{2+190(u_2-0.7)^2} - u_2 = 0 \text{ with } 0.6 < u_2 < 0.7 \tag{47}$$

or

$$\left[\frac{u_2}{1+u_2} + 0.4\right]^2 - u_2 = 0 \text{ with } u_2 \geqslant 0.7 \tag{48}$$

Only the graph of the left hand side of (46) was shown in the main text. Figure 6 plots all three equations. Clearly, only (46) has a solution. This shows that $f(x, .)$ has a unique fixed point and completes the proof of Claim 2.