Fair Play Equilibria in Normal Form Games

Eiichi Miyagawa* Columbia University
Ryo-ichi Nagahisa† Kansai University
Koichi Suga‡ Waseda University

June 7, 2005

Abstract

For a social code of conduct to gain universal acceptance in a society, it would have to satisfy minimum requirements of consistency and procedural justice. The so-called universalizability principle in ethics says that any moral judgement made for an action of a person in a situation should be universalizable to other persons' actions in situations that are identical in relevant respects. By adapting standard axioms in social choice theory, we formalize this principle in the framework of normal form games and study its implications on equilibrium outcomes. A social code specifies socially acceptable responses against other individuals' behavior. A fair play equilibrium is an action profile where everyone behaves optimally subject to the social code. We show that for any admissible social code, the set of fair play equilibria coincides with that of Nash equilibria in all games. The result identifies a conflict between the universalizability principle and what a social code can achieve as equilibrium outcomes.

JEL Classification: C70, D63, D71
Keywords: ethics, social norm, universalizability, Hare

*Department of Economics, Columbia University, 420 West 118th Street, New York, NY 10027, U.S.A.

†Department of Economics, Kansai University, 3-3-35, Yamatecho, Suita, Osaka 564-8680, Japan.

‡School of Political Science and Economics, Waseda University, 1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050, Japan.

1 Introduction

In economic theory, social codes of conduct have been studied mainly in terms of the incentives of individuals to follow the social code. Incentive compatibility, however, is not the only characteristic expected for social codes. For a social code to gain universal acceptance in a society, it would have to satisfy minimum conditions of consistency and procedural justice. In this paper, we use ideas in social choice theory to formulate such conditions for social codes and study their implications on equilibrium social behavior.

We use normal form games to describe the basic economic and social interactions among the members of a society. A social code is then a mapping that specifies socially correct responses to each possible action profile of other players, for each player and each normal form game. That is, the set of socially acceptable actions of player i is given by a set $F_i(G, x_{-i})$, which depends on the game G of the society and the action profile x_{-i} of other players.

If a player has more than one socially acceptable response, he can choose one he prefers the most. Given a social code, a *fair play equilibrium* is an action profile where each player chooses a most preferred action within the set of socially correct responses to other players' actions. The equilibrium concept is an application of the social equilibrium of Debreu (1952) when the action set of a player is constrained by the social code.

We formulate a set of axioms that reasonable social codes are expected to satisfy, which can be described briefly as follows.

Anonymity says that all players should be treated in the same way. If a person is allowed to take a certain action in a situation, the same action should be allowed to you in the situation where your position is the same as his in the first situation.

Welfare nondiscrimination says that what matters ultimately for the society is the members' welfare, and therefore a pair of actions or games should be treated in the same way if they are equivalent in terms of welfare. For example, if a person's hairstyle does not affect anyone's welfare, including his own, then the social code should also be indifferent about his choice of hairstyle. We also require *monotonicity*, which ensures that social correctness is associated positively, not negatively, with welfare.

Independence says that $F_i(G, x_{-i})$ is independent of information about action profiles where x_{-i} is not chosen. By definition, $F_i(G, x_{-i})$ is relevant only if other players behave according to x_{-i}. Thus, the axiom says, action profiles in which other players choose $x'_{-i} \neq x_{-i}$ are counter-factual and should be irrelevant for $F_i(G, x_{-i})$. The requirement is natural since what a social code determines is socially correct *responses* to the behavior of other players.

Lastly, *effectiveness* says that it should be physically possible for everyone to follow the social code. That is, for any game, there should be at least one action

profile where no one violates the social code.

An important feature of these axioms is that they do not define social correctness directly. They do not impose a particular ethical view for an individual situation. They rather formulate consistency in what a social code prescribes across situations and players. The basic form of the axioms is "if it is socially (in)correct for player i to choose action A in situation S, then it should be (in)correct for player i' to choose action A' in situation S'."

The axioms of this form capture the principle of *universalizability* in ethics. The principle says that any moral judgement made for an action of a particular person in a particular situation should be universalizable to anyone's action in situations that are equivalent in relevant respects (e.g., Hare, 1963, 1981). Anonymity, welfare nondiscrimination, and independence are applications of this principle, based on specific ideas of what situations are regarded equivalent. The basic idea of universalizability is old and can be traced back to Kant's Categorical Imperative (1785) and the Golden Rule in various religions (e.g., Luke 6:31 in the Christian Bible). The principle is a key element in Hare's ethics. We are interested in its implications on equilibrium outcomes in game-theoretic contexts where agents respond rationally to the social code and each other's behavior.

The axioms, perhaps except for effectiveness, are familiar in social choice theory.[1] A difference is that social choice theory considers axioms for social choice or welfare functions, which directly specify an outcome—an action profile in our context—or social preferences over outcomes. A social code, on the other hand, does not specify an outcome directly. It only specifies a set of socially correct actions for each individual for each possible contingency. The outcome is determined by equilibrium. In this sense, we are concerned with procedural justice.

We show that, under any social code that satisfies the axioms, fair play equilibria are necessarily Nash equilibria in any game. Thus at any fair play equilibrium, the social code is not a binding constraint for any player. We also show that strict Nash equilibria are necessarily fair play equilibria. The social code cannot eliminate any strict Nash equilibrium. These results together imply that if there is any difference between the set of fair play equilibria and that of Nash equilibria, it consists of Nash equilibria where some players have multiple best replies. If the social code also satisfies a mild continuity condition, the difference disappears: the set of fair play equilibria coincides with the set of Nash equilibria.

Since Nash equilibria are not necessarily Pareto efficient, our result identifies a subtle difficulty in achieving socially desirable outcomes by means of a social code. For example, if the game is the Prisoners' Dilemma, our result implies that the unique Nash equilibrium is also the unique fair play equilibrium under any

[1] For introductions to social choice theory, see, e.g., Sen (1986), Moulin (1988), Austen-Smith and Banks (1999), and Arrow *et al.* (2002).

admissible social code. The difficulty is an example of "fallacy of composition" when rational agents have to play fair: the fact that each individual chooses a socially correct action does not necessarily imply that the action profile itself is socially desirable.

The basic intuition of the result is that, if the constraint imposed by the social code is binding for some agent in some game, then we can find a game in which the social code induces a cycle where each agent ought to give an advantage to others and it is physically impossible for all agents to follow the social code.

2 Model

The set of players is fixed and denoted by $N = \{1, 2, \ldots, n\}$ where $n \geq 2$. Let $\mathbb{X}$ be a non-empty infinite set of *potential* actions. A (finite) ***game*** is a list

$$(X, \succcurlyeq) \equiv (\prod_{i \in N} X_i, (\succcurlyeq_i)_{i \in N})$$

where for all $i \in N$, $X_i \subseteq \mathbb{X}$ is a non-empty finite set of actions and $\succcurlyeq_i$ is a complete and transitive preference relation defined over $X \equiv \prod_{i \in N} X_i$. The strict preference and indifference relations associated with $\succcurlyeq_i$ are denoted by $\succ_i$ and $\sim_i$. The class of all games is denoted as $\mathcal{G}$. Let $P(X)$ denote the set of all complete and transitive preference relations defined over X.

We denote $\prod_{j \in N \setminus \{i\}} X_j$ by X_{-i}. A typical element of X_{-i} is denoted by x_{-i}. Let $BR_i(X, \succcurlyeq, x_{-i})$ denote the best replies to x_{-i} for player i: $BR_i(X, \succcurlyeq, x_{-i}) \equiv \{x_i \in X_i : (x_i, x_{-i}) \succcurlyeq_i (y_i, x_{-i}) \text{ for all } y_i \in X_i\}$.

Even when each $\succcurlyeq_i$ is defined over a set $Y \supseteq X$, we write $(X, \succcurlyeq)$ instead of $(X, \succcurlyeq|_X)$. If each $\succcurlyeq_i$ can be represented by a utility function $u_i \colon X \to \mathbb{R}$, we may write the game as (X, u) where $u = (u_i)_{i \in N}$. In what follows, we often present game matrices with specific utility values, but it should be understood that only ordinal preferences induced by those utility values are relevant.

A ***social code*** is a correspondence F that associates with each game $(X, \succcurlyeq) \in \mathcal{G}$, each $i \in N$, and each $x_{-i} \in X_{-i}$ a non-empty subset $F_i(X, \succcurlyeq, x_{-i}) \subseteq X_i$.[2] Here, $F_i(X, \succcurlyeq, x_{-i})$ is the set of i's actions that are considered as fair or socially correct in game $(X, \succcurlyeq)$ when the other players' actions are x_{-i}. We require this set to be non-empty; for each player and each possible situation, there exists at least one socially correct action.

We assume that players respect a given social code; i.e., players do not choose a socially incorrect action although choosing such an action is physically possible. In practice, agents choose to respect the given social code either because a violation of the code is followed by a punishment from other agents or agents have an intrinsic desire to comply with the code (either because they appreciate the

[2]Note that we write $F_i(X, \succcurlyeq, x_{-i})$ instead of $F(i, X, \succcurlyeq, x_{-i})$.

ideas behind the code or they have been educated to have such a desire). Since the incentive issue is not our main concern in this paper, we simply assume that players choose only socially correct actions.

A social code may specify more than one action as socially acceptable and does not necessarily deprive players of their free choice completely. The typical form of a social code is not "one ought to do this" but "one ought not to do this."

When there are multiple actions that are socially correct, we assume that the player chooses an action that is most preferred within the set of socially correct actions. Given a social code F and a game $(X, \succcurlyeq)$, an action profile x is a ***fair play equilibrium*** if, for each player i, x_i is a most preferred action in $F_i(X, \succcurlyeq, x_{-i})$ for $\succcurlyeq_i$. The set of fair play equilibria is denoted $FPE(X, \succcurlyeq, F)$. Thus $FPE(X, \succcurlyeq, F)$ consists of $x \in X$ such that for all $i \in N$, $x_i \in F_i(X, \succcurlyeq, x_{-i})$ and $x_i \succcurlyeq_i y_i$ for all $y_i \in F_i(X, \succcurlyeq, x_{-i})$. At a fair play equilibrium, a player may have better replies but none of them is socially correct.

3 Axioms

We are interested in characterizing fair play equilibria when the social code satisfies the following axioms. The first axiom is *anonymity*, which says that the names of the players should not matter. Formally, given an action profile x and a bijection (permutation) $\pi\colon N \to N$, we define x^π by $x^\pi_{\pi(i)} = x_i$. That is, x^π is the action profile in which player $\pi(i)$'s action coincides with x_i.

Definition. A social code F satisfies ***anonymity*** if for all $(X, \succcurlyeq), (X', \succcurlyeq') \in \mathcal{G}$ and all bijections $\pi\colon N \to N$, if for all $i \in N$ and all $x, y \in X$,

$$X_i = X'_{\pi(i)} \quad \text{and} \quad \left[x \succcurlyeq_i y \iff x^\pi \succcurlyeq'_{\pi(i)} y^\pi\right],$$

then for all $i \in N$ and all $x \in X$,

$$F_i(X, \succcurlyeq, x_{-i}) = F_{\pi(i)}(X', \succcurlyeq', x^\pi_{-\pi(i)}).$$

For example, consider the following games:

		Player 2		
		a	b	c
Player 1	A	$4,4$	$2,5$	$3,7$
	B	$6,0$	$0,9$	$0,7$

		Player 2	
		A	B
Player 1	a	$4,4$	$0,6$
	b	$5,2$	$9,0$
	c	$7,3$	$7,0$

These games are identical except that the players are interchanged. Thus if A is a socially correct response to a in the left game, then anonymity says that the same should be the case in the right game as well.

To state the next axiom, we first introduce a definition. We write $x_i \simeq x_i'$ if these actions are identical in term of welfare for all players, regardless of the actions of $N \setminus \{i\}$. Formally,

Definition. Given $(X, \succcurlyeq) \in \mathcal{G}$ and $i \in N$, actions $x_i \in X_i$ and $x_i' \in X_i$ are ***welfare-equivalent***, denoted $x_i \simeq x_i'$, if for all $x_{-i} \in X_{-i}$ and all $j \in N$, $(x_i, x_{-i}) \sim_j (x_i', x_{-i})$.

Then the next axiom requires that welfare-equivalent actions should be treated as the same action. Formally,

Definition. A social code F satisfies ***welfare nondiscrimination*** if for all $(X, \succcurlyeq) \in \mathcal{G}$, the following conditions are satisfied.

1. For all $x, y \in X$, if $x_i \simeq y_i$ for all $i \in N$ (possibly $x_i = y_i$ for some i), then for all $i \in N$,

$$x_i \in F_i(X, \succcurlyeq, x_{-i}) \iff y_i \in F_i(X, \succcurlyeq, y_{-i}).$$

2. For all $x \in X$, all $i \in N$, and all $y_i \neq x_i$ such that $y_i \simeq x_i$,

$$F_i(X_i \setminus \{y_i\} \times X_{-i}, \succcurlyeq, x_{-i}) = F_i(X, \succcurlyeq, x_{-i}) \setminus \{y_i\},$$
$$F_j(X_i \setminus \{y_i\} \times X_{-i}, \succcurlyeq, x_{-j}) = F_j(X, \succcurlyeq, x_{-j}) \qquad \text{for all } j \neq i.$$

Condition 1 says that for a given game, the social code does not distinguish welfare-equivalent actions. Condition 2 relates two games, saying that if there is a pair of welfare-equivalent actions, then deleting one of those from the game does not change the social code's instructions.

As an illustration, consider the left game.

	A	B
a	$6, 0$	$1, 5$
b	$2, 7$	$3, 4$
c	$2, 7$	$3, 4$

	A	B
a	$6, 0$	$1, 5$
b	$2, 7$	$3, 4$

(1)

Since b and c are welfare-equivalent, welfare nondiscrimination requires that b is socially correct if and only if c is (condition 1 when $i = 1$ and $x_{-i} = y_{-i}$). Welfare nondiscrimination also says that whether player 1 plays b or c does not affect socially correct actions for player 2 (condition 1 when $i = 2$ and $x_i = y_i$).

The game on the right is obtained from the left by deleting c. Since c is a replica of b, there is a sense in which these games are identical. This is why welfare nondiscrimination (condition 2) also requires that at any action profile that exists in both games, a player's action is socially correct in the left game if and only if it is socially correct in the right game. For example, b is a socially

correct response to A in the left game if and only if it is the case in the right game.

Welfare nondiscrimination is a straightforward application of the welfarism principle in social choice theory: what matters is agents' welfare and therefore alternatives should not be distinguished if they have identical welfare consequences. There are two separate issues: how to treat welfare-equivalent actions (condition 1) and how to respond if welfare-equivalent actions are deleted or added (condition 2). Condition 2 ensures a minimum condition of consistency across games that have different numbers of actions. It rules out artificial social codes that give two games different treatments only because they have different numbers of actions.

An important implication of welfare nondiscrimination is what corresponds to *neutrality* in social choice theory. Neutrality, in our context, requires that two games should be treated in the same way if they differ merely in the labeling of actions. Formally,

Definition. A social code F is ***neutral*** if for all $(X, \succcurlyeq), (X', \succcurlyeq') \in \mathcal{G}$, if for all $i \in N$, there exists a bijection $\rho_i \colon X_i \to X'_i$ such that for all $x, y \in X$ and all $i \in N$,

$$x \succcurlyeq_i y \iff (\rho_1(x_1), \ldots, \rho_n(x_n)) \succcurlyeq'_i (\rho_1(y_1), \ldots, \rho_n(y_n)),$$

then for all $x \in X$ and all $i \in N$,

$$x_i \in F_i(X, \succcurlyeq, x_{-i}) \iff \rho_i(x_i) \in F_i(X', \succcurlyeq', \rho_{-i}(x_{-i}))$$

where $\rho_{-i}(x_{-i}) = (\rho_1(x_1), \ldots, \rho_{i-1}(x_{i-1}), \rho_{i+1}(x_{i+1}), \ldots, \rho_n(x_n))$.

Proposition 1. *Welfare nondiscrimination implies neutrality.*

Proof. See Appendix.

Neutrality is strictly weaker than welfare nondiscrimination because neutrality does not relate games that have different numbers of actions: it allows the games in (1) to receive totally different treatments.

The next axiom, *monotonicity*, says that if a player's action is socially correct at an action profile x for a preference profile $\succcurlyeq$, it remains the case for another preference profile $\succcurlyeq'$ if the relative ranking of x in $\succcurlyeq'$ is as high as in $\succcurlyeq$ for every player.

Definition. A social code F is ***monotonic*** if for all $(X, \succcurlyeq) \in \mathcal{G}$, all $x \in X$, all $i \in N$, and all $\succcurlyeq' \in P(X)^N$, if $x_i \in F_i(X, \succcurlyeq, x_{-i})$, and for all $j \in N$ and all

$y \in X$,

$$x \succcurlyeq_j y \Longrightarrow x \succcurlyeq'_j y,$$
$$x \succ_j y \Longrightarrow x \succ'_j y,$$

then $x_i \in F_i(X, \succcurlyeq', x_{-i})$.

This is a natural translation of Arrow's condition of positive association (1963). The condition is also reasonable in the present context since we would like to assume that social correctness is associated positively, not negatively, with welfare.

The next axiom, *independence*, says that whether player i's action is a socially acceptable response to x_{-i} is independent of preferences over action profiles where x_{-i} is not played.

Definition. A social code F satisfies ***independence*** if for all $(X, \succcurlyeq) \in \mathcal{G}$, all $\succcurlyeq' \in P(X)$, all $i \in N$, and all $x_{-i} \in X_{-i}$, if for all $j \in N$, $\succcurlyeq_j$ and $\succcurlyeq'_j$ induce the same preferences over $\{(y_i, x_{-i}) : y_i \in X_i\}$, then $F_i(X, \succcurlyeq, x_{-i}) = F_i(X, \succcurlyeq', x_{-i})$.

This is a natural requirement since what a social code prescribes to a player is conditional on other players' behavior. By definition, $F_i(X, \succcurlyeq, x_{-i})$ is relevant only if other players play x_{-i}. Given x_{-i}, action profiles in $X \backslash \{(y_i, x_{-i}) : y_i \in X_i\}$ are counter-factual and hence deemed immaterial.

For this axiom, it is critical that what a social code decides is whether one's action is a socially correct *response* to others' behavior. Thus a judgement that a player's action is socially incorrect cannot be justified on the ground that it may induce bad behavior of others. Since other players' behavior is held fixed, one can reasonably say that what happens if they behave differently is irrelevant.

The next axiom, *effectiveness*, says that, for any game, there exists at least one action profile where each player plays fair.

Definition. A social code F is ***effective*** if for all $(X, \succcurlyeq) \in \mathcal{G}$, there exists $x \in X$ such that

$$x_i \in F_i(X, \succcurlyeq, x_{-i}) \quad \text{for all } i \in N. \tag{2}$$

Thus a violation of effectiveness means that there exists a game in which it is logically impossible for all players to follow the social code. The axiom represents a basic requirement that a social code's prescriptions to different players should be compatible. Effectiveness is weaker than demanding the existence of a fair play *equilibrium*, since (2) is only a necessary condition for x to be a fair play equilibrium. An action profile that satisfies (2) is called a ***fair play profile***.

4 Main Result

Our main result states that if a social code satisfies all the axioms defined in the previous section, then for any player, there always exists a socially correct action that is also an unconstrained best reply. That is, while the social code may prohibit a player from choosing some of his best replies, it never prohibits all. This result immediately implies that under the social code, all fair play equilibria are necessarily Nash equilibria, and conversely, all strict Nash equilibria are necessarily fair play equilibria.

Formally, let $NE(X, \succcurlyeq)$ and $SNE(X, \succcurlyeq)$ denote the set of pure-strategy Nash equilibria and strict Nash equilibria, respectively. Then

Theorem 1. *If a social code F satisfies anonymity, welfare nondiscrimination, monotonicity, independence, and effectiveness, then for all $(X, \succcurlyeq) \in \mathcal{G}$, all $i \in N$, and all $x_{-i} \in X_{-i}$,*

$$F_i(X, \succcurlyeq, x_{-i}) \cap BR_i(X, \succcurlyeq, x_{-i}) \neq \emptyset. \tag{3}$$

Thus for all $(X, \succcurlyeq) \in \mathcal{G}$,

$$SNE(X, \succcurlyeq) \subseteq FPE(X, \succcurlyeq, F) \subseteq NE(X, \succcurlyeq). \tag{4}$$

(3) indeed implies the first inclusion in (4) since at any strict Nash equilibrium, the best reply is unique and hence must be socially correct. To see that the second inclusion in (4) is also implied by (3), suppose that a fair play equilibrium is not a Nash equilibrium. Then some player is not playing an unconstrained best reply. This is a contradiction with fair play equilibrium since there exists an unconstrained best reply that is also socially correct.

Since the proof is long and involved, we here give a proof for the case of two persons and two actions; the general proof is given in Appendix. Suppose that there are two players (1 and 2) and consider a game $(X, \succcurlyeq)$ where the action set is $X_i = \{a, b\}$ for each player. To prove (3), suppose that it does not hold for some i and x_j $(j \neq i)$. Without loss of generality, assume $(i, j) = (1, 2)$, $x_2 = a$, $F_1(X, \succcurlyeq, a) = \{a\}$, and $BR_1(X, \succcurlyeq, a) = \{b\}$. It is useful to distinguish two cases.

Case 1. $(b, a) \prec_2 (a, a)$. That is, player 1's playing his best reply makes player 2 worse off. Consider the left game:

	a	b
a	$2, 2$	$\cdots$
b	$3, 1$	$\cdots$

	a	b
a	$2, 2$	$\cdots$
b	$3, 1$	$\cdots$
c	$3, 1$	$\cdots$

(5)

where the payoffs in the cells with "$\cdots$" are immaterial. By independence, the

same conclusion continues to hold in the left game: if player 2 plays a, the unique fair play for player 1 is a. Next, we add another action c to the action set of player 1 and consider the right game in (5). By welfare nondiscrimination and independence, action c is regarded identical to b if player 2 plays a. Thus if 2 plays a, the unique fair play for player 1 remains a. Finally, consider the following game:

	a	b
a	$2,2$	$3,1$
b	$3,1$	$2,2$
c	$3,1$	$3,1$

(6)

We now apply the conclusion derived above to this game. Suppose that player 2 plays a. Then the unique fair play for player 1 is a. So now, suppose that player 1 plays a. By anonymity, it is not difficult to show that the unique fair play for player 2 is b. But if player 2 plays b, the unique fair play for 1 is b. If 1 plays b, the unique fair play for 2 is a, and we obtained a cycle. This shows that there exists no fair play profile, a contradiction with effectiveness.

Case 2. $(b,a) \succcurlyeq_2 (a,a)$. Let $\succcurlyeq_2'$ be the preference relation that is obtained from $\succcurlyeq_2$ by moving (b,a) to the bottom of the ordering without changing the other part of the ordering. Then by Case 1, $b \in F_1(X,(\succcurlyeq_1,\succcurlyeq_2'),a)$. Since $\succcurlyeq_2$ is obtained from $\succcurlyeq_2'$ by moving (b,a) upwards, monotonicity implies $b \in F_1(X,\succcurlyeq,a)$, which is a contradiction with our initial assumption. Q.E.D.

In game (6), player 2 wants to choose the same action as player 1 but player 1 does not want to choose the same action as player 2. Since only player 1 can choose c, a Nash equilibrium is where he chooses c. A problem arises if players have to be "kind"—altruistic and self-sacrificing—to each other. To be kind, player 1 ought to choose the same action as player 2 since that is what pleases 2. Player 2, on the other hand, ought to choose the different action to please 1. The problem is that it is impossible for both players to be kind.

Remark 1. The proof for Case 1 can be simplified by using Matching Pennies:

	a	b
a	$2,2$	$3,1$
b	$3,1$	$2,2$

By applying the conclusion from the left game in (5), we can conclude that Matching Pennies has no fair play profile.[3] While this makes the proof simpler,

[3]The simplified proof with Matching Pennies does not change action sets. This implies that, to prove the result for two-person two-action games, we do not need the full force of welfare nondiscrimination: neutrality suffices. However, this conclusion does not generalize beyond the

the added simplicity is obtained in exchange for using a game that has no pure-strategy Nash equilibrium. The theorem itself does not rely on such games. As noted above, the game in (6) has pure-strategy Nash equilibria. The general proof in Appendix also uses only games that have pure-strategy Nash equilibria. Formally, let $\mathcal{G}^{NE} \equiv \{(X, \succcurlyeq) \in \mathcal{G} : NE(X, \succcurlyeq) \neq \emptyset\}$ denote the class of games that have pure-strategy Nash equilibria. Then Theorem 1 holds for any class of games $\mathcal{G}'$ such that $\mathcal{G}' \supseteq \mathcal{G}^{NE}$.

The general proof in Appendix is considerably more complicated. One reason for the complication is that the violation of (3) may be at an action profile where player i has multiple best replies, and other players may have complicated preferences over i's best replies. This complicates the proof since we then need to maintain the same preference structure along each step of a cycle. Another source of complication is that we need to change every player's action to complete a cycle.

The inclusion relations in (4) are tight in the sense that there exist examples where the reverse relations do not hold. The following example shows that $SNE(X, \succcurlyeq) \subsetneq FPE(X, \succcurlyeq, F)$ is possible.

Example 1 (Local weak Pareto code)**.** For all $(X, \succcurlyeq) \in \mathcal{G}$, all $i \in N$, and all $x_{-i} \in X_{-i}$, let $F_i(X, \succcurlyeq, x_{-i})$ be the set of $x_i \in X_i$ such that there exists no $x_i' \in X_i$ such that $(x_i', x_{-i}) \succ_k (x_i, x_{-i})$ for all $k \in N$.

Thus, an action is judged as socially incorrect if there exists an action that makes all players better off. This code is "local" in the sense that x_{-i} is held fixed when Pareto efficiency is invoked. One can easily verify that this code satisfies all the axioms.[4] Apply this code to the following game:

	A	B
A	$2, 2$	$1, 1$
B	$1, 1$	$1, 1$

Then $SNE(X, \succcurlyeq) = \{(A, A)\} \subsetneq \{(A, A), (B, B)\} = FPE(X, \succcurlyeq, F)$.

To show that $FPE(X, \succcurlyeq, F) \subsetneq NE(X, \succcurlyeq)$ is also possible, consider the following social code.

Example 2 (Local strong Pareto code)**.** This is the modified version of Example 1 where strong Pareto efficiency is used. Thus, an action is socially incorrect if there exists an action that makes no player worse off and at least one player strictly better off.

two-person two-action case: a counter-example is Example 7 in Appendix A.3.

[4]This code satisfies effectiveness since any weakly Pareto efficient action profile is a fair play profile.

This code also satisfies all the axioms. Apply this code to the following game:

	A	B
A	$1,1$	$0,1$
B	$1,0$	$0,0$

Then $FPE(X, \succcurlyeq, F) = \{(A, A)\} \subsetneq X = NE(X, \succcurlyeq)$.

The inclusion relations in (4) have a few immediate implications on the existence of fair play equilibria under our axioms. The relation $SNE(X, \succcurlyeq) \subseteq FPE(X, \succcurlyeq, F)$ implies that a fair play equilibrium exists for all games that have strict Nash equilibria. Therefore, for generic games that have pure-strategy Nash equilibria, fair play equilibria exist.

On the other hand, a fair play equilibrium does not necessarily exist in all games. For example, suppose that the social code is the local strong Pareto code (Example 2) and consider

	A	B
A	$1,1$	$1,0$
B	$1,2$	$0,3$

Then $NE(X, \succcurlyeq) = \{(A, A)\}$ but $FPE(X, \succcurlyeq, F) = \emptyset$. The game does have fair play profiles (i.e., (B, A) and (B, B)) but none of them is a fair play equilibrium.

The relation $FPE(X, \succcurlyeq, F) \subseteq NE(X, \succcurlyeq)$ implies that a necessary condition for the existence of a fair play equilibrium is that a pure-strategy Nash equilibrium exists. By definition, effectiveness ensures that fair play profiles do exist. But none of them is a fair play equilibrium if the game has no pure-strategy Nash equilibrium and the social code satisfies the axioms.

For some social codes, the necessary condition for the existence of fair play equilibria is also sufficient. A particularly simple example is the absence of any code:

Example 3 (Empty code)**.** All actions are always socially correct: $F_i(X, \succcurlyeq, x_{-i}) \equiv X_i$.

Under this code, trivially $FPE(X, \succcurlyeq, F) \equiv NE(X, \succcurlyeq)$. Hence, a fair play equilibrium exists if and only if a pure-strategy Nash equilibrium exists. This code is not the only one that has this property. The local weak Pareto code (Example 1) also has $FPE(X, \succcurlyeq, F) \equiv NE(X, \succcurlyeq)$ since any unconstrained best reply is weakly Pareto efficient given x_{-i}.

In Appendix, we show that none of the axioms in Theorem 1 is redundant, by showing that a counter-example can be found if any of the axioms is removed. The examples there also help us see the reasonableness of each axiom.

The relation between Theorem 1 and Arrow's impossibility theorem (1963) is elusive. Our result relies on games with a cyclic nature, just as Arrow's result

relies on Condorcet's voting cycle—the well-known preference profile with 3 voters and 3 alternatives where the majority-rule winner does not exist. It is also interesting to observe that Nash equilibrium can be thought of as a social code in which each player is a *dictator* for his own socially correct actions given the other players' actions. We could easily construct more "democratic" procedures to determine socially correct actions, in such a way that there may not exist an unconditional best response that is also socially correct (e.g., Example 8). However, just in Arrow's theorem, all those procedures violate at least one basic axiom.

The cyclic nature of the game in the proof also suggests a relation to the Liberal Paradox (Sen, 1970). Indeed, a variant of the paradox in Gibbard (1974) is based on Matching Pennies, although the direction of the cycle is opposite, i.e., everyone there tries to increase his own payoff. There is also a similarity in terms of the framework if one interprets $F_i(X, \succcurlyeq, x_{-i})$ as the player's rights. A critical difference is that none of our axioms is about liberty. Each of our axioms permits $F_i(X, \succcurlyeq, x_{-i})$ to be always a singleton, in which case the social code gives no freedom.

The cycle in the game (6) captures situations in which each person ought to give an advantage to others. For example, when everyone insists on paying a bill for the others, it could take long before they reach an agreement. An extreme example of such situations can be found in Japan when a group of business persons go to a restaurant and decide where each person sits. There exists a socially predetermined way of assigning a ranking over the seats around a given table. The socially correct behavior is to offer higher-ranked seats to members with higher social ranks. This social code causes a problem when the social ranking within the group is ambiguous and is a sensitive issue, since then they need to keep standing around a table insisting on giving better seats to each other.

5 Variations

5.1 Full Characterization

Theorem 1 is not a complete characterization of the fair play equilibria, since it fails to determine whether a given $x \in NE(X, \succcurlyeq) \setminus SNE(X, \succcurlyeq)$ is a fair play equilibrium. In this subsection, we show that a complete characterization is obtained if the social code satisfies a continuity condition. We prove that for any social code that satisfies the previous axioms and continuity, the set of fair play equilibria coincides with the set of pure-strategy Nash equilibria.

Given a set of action profiles X, let $U(X)$ denote the set of all utility functions $u_i \colon X \to \mathbb{R}$. Any profile of utility functions $u = (u_i)_{i \in N} \in U(X)^N$ can be represented by a vector of $n|X|$ numbers. Thus $U(X)^N$ can be regarded as $\mathbb{R}^{n|X|}$.

Recall that games $(X, \succcurlyeq)$ and (X, u) are considered to be equivalent if for all i, u_i is a utility-function representation of $\succcurlyeq_i$.

Definition. A social code F satisfies ***continuity*** if for all $(X, u) \in \mathcal{G}$, all $x \in X$, all $i \in N$, and all sequences $(u^k)_{k=1}^{\infty}$ of utility function profiles in $U(X)^N$, if $u^k \to u$ as $k \to \infty$ and $x_i \in F_i(X, u^k, x_{-i})$ for all k, then $x_i \in F_i(X, u, x_{-i})$.

This is the upper-hemi continuity of the correspondence $F_i(X, u, x_{-i})$ with respect to u. It says that for any action x_i, the set of payoff-function profiles in which the action is socially correct is closed. Thus, at the boundary of situations in which x_i is socially correct, the social code takes the permissive side.

Corollary 1. *If a social code F satisfies anonymity, welfare nondiscrimination, monotonicity, independence, effectiveness, and continuity, then for all games $(X, \succcurlyeq) \in \mathcal{G}$,*

$$FPE(X, \succcurlyeq, F) = NE(X, \succcurlyeq).$$

The same conclusion holds on any restricted domain $\mathcal{G}'$ such that $\mathcal{G}' \supseteq \mathcal{G}^{NE}$.

Proof. Take any game $(X, \succcurlyeq)$ in the domain. Let $u \in U(X)^N$ be such that for each i, u_i is a utility-function representation of $\succcurlyeq_i$. By Theorem 1, we only need to prove $NE(X, u) \subseteq FPE(X, u, F)$. So, let $x \in NE(X, u)$. For each $k \in \{1, 2, \ldots\}$ and each $i \in N$, let u_i^k be a utility function defined by $u_i^k(x) = u_i(x) + (1/k)$ and $u_i^k(y) = u_i(y)$ for all $y \neq x$. Then $u_i^k \to u_i$ and $x \in SNE(X, u^k)$ for all k. Theorem 1 then implies $x \in FPE(X, u^k, F)$ for all k, hence $x_i \in F_i(X, u^k, x_{-i})$ for all i and k. Continuity then implies $x_i \in F_i(X, u, x_{-i})$ for all i. Since x is a Nash equilibrium, we obtain $x \in FPE(X, u, F)$. Q.E.D.

Since this result gives a normative axiomatization of Nash equilibria, it is somewhat similar to the result of Peleg and Tijs (1996). They axiomatize Nash equilibria based on the so-called reduced-game property. The property requires consistency across games with different sets of players. Their solution concept, however, chooses action profiles directly, as in social choice theory.

5.2 Domain Restriction

As mentioned above, Theorem 1 holds even if we restrict our attention to games that have pure-strategy Nash equilibria. However, since the theorem mentions strict Nash equilibria, it is of interest to examine what happens if the domain is further restricted to games that have strict Nash equilibria. Under the domain restriction, the proof of Theorem 1 cannot be extended directly. In this section, we give a few results under the domain restriction.

Let $\mathcal{G}^{SNE}$ denote the class of games that have strict Nash equilibria: $\mathcal{G}^{SNE} \equiv \{(X, \succcurlyeq) \in \mathcal{G} : SNE(X, \succcurlyeq) \neq \emptyset\}$. It turns out that on the domain $\mathcal{G}^{SNE}$, the inclusion $SNE \subseteq FPE$ remains true while $FPE \subseteq NE$ does not.

Proposition 2. *If a social code F defined on $\mathcal{G}^{SNE}$ satisfies anonymity, welfare nondiscrimination, monotonicity, independence, and effectiveness, then for all $(X, \succcurlyeq) \in \mathcal{G}^{SNE}$,*

$$SNE(X, \succcurlyeq) \subseteq FPE(X, \succcurlyeq, F).$$

Proposition 3. *On the domain $\mathcal{G}^{SNE}$, there exists a social code F that satisfies anonymity, welfare nondiscrimination, monotonicity, independence, and effectiveness and such that for some game $(X, \succcurlyeq) \in \mathcal{G}^{SNE}$,*

$$FPE(X, \succcurlyeq, F) \nsubseteq NE(X, \succcurlyeq).$$

Proof. Proposition 2 is proved in Appendix (which also contains a corollary). To prove Proposition 3, we exhibit a specific social code. To specify it, consider a game $(X, \succcurlyeq)$ and fix i and x_{-i}. For any subset $X_i' \subseteq X_i$ and any $j \neq i$, let

$$BR_i^j(X, \succcurlyeq, x_{-i}, X_i') \equiv \{x_i \in X_i' : (x_i, x_{-i}) \succcurlyeq_j (x_i', x_{-i})\ \forall x_i' \in X_i'\}.$$

This is the set of i's actions in X_i' that are preferred by j. Given this definition, consider a social code F defined by

$$\begin{aligned} F_i(X, \succcurlyeq, x_{-i}) \equiv & \Big[\bigcup_{j \neq i} BR_i^j(X, \succcurlyeq, x_{-i}, X_i)\Big] \\ & \cup \Big[\bigcap_{j \neq i} BR_i^j\big(X, \succcurlyeq, x_{-i}, BR_i(X, \succcurlyeq, x_{-i})\big)\Big]. \end{aligned} \tag{7}$$

The first line says that if an action x_i is a most preferred choice for at least one $j \neq i$, it is socially correct. The second line says that an action x_i is also socially correct if it is a most preferred choice within i's best replies for all $j \neq i$.

This social code satisfies effectiveness on $\mathcal{G}^{SNE}$. To see this, take any $(X, \succcurlyeq) \in \mathcal{G}^{SNE}$ and let $x \in SNE(X, \succcurlyeq)$. Then, for all i, x_i is the unique best reply and hence it is trivially the preferred choice within i's best replies for all $j \neq i$. Thus, the second line of (7) is $\{x_i\}$ and hence $x_i \in F_i(X, \succcurlyeq, x_{-i})$.

The social code also satisfies monotonicity. To see this, let $x_i \in F_i(X, \succcurlyeq, x_{-i})$ and $\succcurlyeq'$ be as in the definition of monotonicity. First, if $x_i \in BR_i^j(X, \succcurlyeq, x_{-i}, X_i)$ for some $j \neq i$, then by the construction of $\succcurlyeq'$, we have $x_i \in BR_i^j(X, \succcurlyeq', x_{-i}, X_i)$ and hence x_i remains socially correct. Suppose then that x_i belongs to the second line of (7). Then by the construction of $\succcurlyeq'$, x_i remains a best reply for i and the set of i's best replies can only get smaller: $x_i \in BR_i(X, \succcurlyeq', x_{-i}) \subseteq BR_i(X, \succcurlyeq, x_{-i})$. Since x_i was initially a most preferred choice within i's best replies for all $j \neq i$ and the relative ranking of x_i does not go down for any j, it remains a most preferred choice for all j. Hence x_i remains in the second line of (7).

The social code also satisfies welfare nondiscrimination since clearly the code distinguishes actions based only on preferences. It should be also clear that

anonymity and independence are satisfied, too.

To complete the proof, consider the following 3-person game:

		Player 2	
		a	b
Player 1	a	$4,1,4$	$6,6,6$
	b	$4,4,1$	$5,5,5$
	c	$2,6,6$	$1,1,1$

(8)

where player 3 has a single action, say $X_3 = \{a\}$.[5] For this game, $SNE(X, \succcurlyeq) = NE(X, \succcurlyeq) = \{(a, b, a)\}$ while $FPE(X, \succcurlyeq, F) = \{(a, b, a), (c, a, a)\}$. Q.E.D.

The social code used in the proof, (7), does not disprove Theorem 1 because if the domain is enlarged to $\mathcal{G}^{NE}$, the social code does not satisfy effectiveness.

There is a sense in which the counter-example in Proposition 3 is not robust. In game (8), it is critical that player 1 has an indifference over his best replies. If player 2 plays a, player 1 is indifferent between a and b and none of them is unanimously preferred by the other players, which is why player 1 is not allowed to play any of his best replies. If 1's preferences are perturbed so that he has a unique best reply, then he is allowed to play the best reply and thus the problem disappears.

In what follows, we formalize this argument and show that if indifference is ruled out from preferences, Theorem 1 is restored. Since Nash equilibria are all strict under strict preferences, we obtain $FPE = SNE$.

We introduce a few definitions. Let $\mathcal{G}^G$ denote the class of games in which no player has any indifference given the others' actions: $(X, \succcurlyeq) \in \mathcal{G}^G$ if and only if for all $i \in N$ and all $x_{-i} \in X_{-i}$, $\succcurlyeq_i$ has no indifference over $\{(y_i, x_{-i}) : y_i \in X_i\}$. In what follows, we consider $\mathcal{G}^G \cap \mathcal{G}^{SNE}$ as the domain of social codes.

Given a pair of actions $\{x_i, x_i'\}$, we say that x_i ***Pareto dominates*** x_i' if x_i gives a higher payoff than x_i' for every player and for every x_{-i}: i.e., $(x_i, x_{-i}) \succ_j (x_i', x_{-i})$ for all $j \in N$ and all $x_{-i} \in X_{-i}$.

Definition. A social code F satisfies ***Pareto dominance*** if for all $(X, \succcurlyeq) \in \mathcal{G}^G \cap \mathcal{G}^{SNE}$, all $i \in N$, all $x_{-i} \in X_{-i}$, and all pairs $\{x_i, x_i'\} \subseteq X_i$ such that x_i Pareto dominates x_i',

$$x_i' \in F_i(X, \succcurlyeq, x_{-i}) \Longrightarrow x_i \in F_i(X, \succcurlyeq, x_{-i}),$$
$$F_i(X_i \setminus \{x_i'\} \times X_{-i}, \succcurlyeq, x_{-i}) = F_i(X, \succcurlyeq, x_{-i}) \setminus \{x_i'\}.$$

The first line says that if an action is socially correct, then any action that Pareto dominates it is also socially correct. The second line says that deleting

[5] It is easy to modify the game so that player 3 has more than one action.

a Pareto dominated action from the game does not affect whether remaining actions are socially correct.[6]

Definition. A social code F satisfies ***independence**** if for all $(X, \succcurlyeq), (X', \succcurlyeq') \in \mathcal{G}^G \cap \mathcal{G}^{SNE}$ and all $i \in N$ such that $X_i' = X_i$ and $X_{-i} \cap X_{-i}' \neq \emptyset$, and for all $x_{-i} \in X_{-i} \cap X_{-i}'$, if for all $j \in N$, $\succcurlyeq_j$ and $\succcurlyeq_j'$ induce the same preferences over $\{(y_i, x_{-i}) : y_i \in X_i\}$, then $F_i(X, \succcurlyeq, x_{-i}) = F_i(X', \succcurlyeq', x_{-i})$.

This variant of independence is based on the same idea but is technically stronger since it can be applied to pairs of games with different action sets. In the definition, $(X, \succcurlyeq)$ and $(X', \succcurlyeq')$ may have different action sets for $j \neq i$, but these games induce identical preferences over i's actions if other players play x_{-i}.

Proposition 4. *If a social code F defined on $\mathcal{G}^G \cap \mathcal{G}^{SNE}$ satisfies anonymity, neutrality, monotonicity, independence*, effectiveness, and Pareto dominance, then for all $(X, \succcurlyeq)$ in the domain,*

$$FPE(X, \succcurlyeq, F) = SNE(X, \succcurlyeq).$$

Proof. Let F be a social code that satisfies the axioms and consider any game $(X, \succcurlyeq) \in \mathcal{G}^G \cap \mathcal{G}^{SNE}$. It suffices to prove the following: for all i and all x_{-i},

$$BR_i(X, \succcurlyeq, x_{-i}) \subseteq F_i(X, \succcurlyeq, x_{-i}). \tag{9}$$

That is, the (unique) unconstrained best reply is a fair play. It is immediate that this implies the desired result. To prove (9), suppose, by contradiction, that it is false. Thus there exist $y \in X$ and $i \in N$ such that $BR_i(X, \succcurlyeq, y_{-i}) = \{y_i\}$ and $y_i \notin F_i(X, \succcurlyeq, y_{-i})$. Without loss of generality, let $X_i = \{1, 2, \ldots, k\}$ and $y_i = k$, where $k \geq 2$.

If $|X_{-i}| > 1$, choose any action profile $w_{-i} \neq y_{-i}$. If $X_{-i} = \{y_{-i}\}$, let $w_{-i} = y_{-i}$. Let $v\colon X \to \mathbb{R}^N$ be a payoff vector function such that

$$v(x) = \begin{cases} (0, \ldots, 0, x_i, 0, \ldots, 0) & \text{if } x = y, \\ (x_i, \ldots, x_i) & \text{if } x_{-i} = y_{-i} \text{ and } x_i < y_i, \\ (2k, \ldots, 2k) & \text{if } x = (y_i, w_{-i}) \text{ and } w_{-i} \neq y_{-i}, \end{cases}$$

where "x_i" on the first line appears in the ith entry. For all other action profiles x, we can specify $v(x) \in \mathbb{R}^N$ so that (i) $(k, \ldots, k) \ll v(x) \ll (2k, \ldots, 2k)$, (ii) $(X, v) \in \mathcal{G}^G$, and (iii) every $x_i < k - 1$ is Pareto dominated by $k - 1$. Then $(y_i, w_{-i}) \in SNE(X, v)$ and hence $(X, v) \in \mathcal{G}^{SNE} \cap \mathcal{G}^G$. Let $\succcurlyeq'$ denote the preference profile induced by v.

[6]Since the dominated action is not part of any strict Nash equilibrium, the new game remains in the domain: $(X_i \setminus \{x_i'\} \times X_{-i}, \succcurlyeq) \in \mathcal{G}^G \cap \mathcal{G}^{SNE}$.

The construction implies that for all $x \in X$ and all $j \in N$,

$$\begin{aligned} y \succcurlyeq'_j x &\Longrightarrow y \succcurlyeq_j x, \\ y \succ'_j x &\Longrightarrow y \succ_j x. \end{aligned} \tag{10}$$

To see this, note that $y \succcurlyeq'_j x$ holds for $x \neq y$ only if $x_{-i} = y_{-i}$ and $j = i$. Thus we have $y \succ_j x$ since $\{y_i\} = BR_i(X, \succcurlyeq, y_{-i})$ by our initial choice of y. (10) implies that, in the move from $\succcurlyeq$ to $\succcurlyeq'$, the relative ranking of y does not go up for any player. Therefore, by monotonicity and independence, $y_i \notin F_i(X, \succcurlyeq', y_{-i})$. Since $k-1$ Pareto dominates all $x_i < k-1$, we obtain $k-1 \in F_i(X, \succcurlyeq', y_{-i})$.

We now remove all $x_i < k-1$ from X_i. Thus $X'_i \equiv \{k-1, k\}$ is the new action set for i, and $X' \equiv X'_i \times \prod_{j \neq i} X_j$ is the new set of action profiles. Since (k, w_{-i}) remains in the game, $(X', \succcurlyeq') \in \mathcal{G}^{SNE}$. Since the removed actions are all Pareto dominated by $k-1$, the fact that k is not socially correct remains unchanged. Hence $F_i(X', \succcurlyeq', y_{-i}) = \{k-1\}$. Note that given y_{-i}, player i prefers k while all other players prefer $k-1$. Anonymity, neutrality, and independence* thus imply the following: for any game, if a given player j has only two actions a and b and, given x_{-j}, j prefers a and all other players prefer b, then j ought to play b.

To derive a contradiction, we now construct a game that has no fair play profile. If $n = 2$, we can use the following game:

	A	B	C
a	$7,7$	$2,2$	$1,1$
b	$0,0$	$9,3$	$3,9$
c	$5,8$	$4,4$	$6,6$

This game has a strict Nash equilibrium, i.e., (a, A). However, there exists no fair play profile. To see this, suppose that player 2 plays A. We claim that a is not a socially correct reply to A. Note that given A, b is dominated by a and c. Thus, Pareto dominance and independence imply that deleting b from the game does not affect whether remaining actions are socially correct responses to A. If b is removed, the conclusion in the previous paragraph implies that c is the unique fair reply to A, since by playing c, player 1 can make himself worse off and the other player better off. Thus c is a socially correct reply to A, and a is not. Since a dominates b given A, Pareto dominance and independence imply that b is not socially correct (if b is correct, then a should be correct, too). Therefore c is the unique fair reply to A.

So, suppose that player 1 plays c. By the same argument applied to player 2, the unique fair reply for 2 is C (note that B is dominated). But if player 2 plays C, the unique fair play for 1 is b. If 1 plays b, the unique fair play for 2 is B. If 2 plays B, the unique fair play for 1 is c, and we obtained a cycle. Since we went through all actions of player 2, there exists no fair play profile.

If $n \geq 3$, we can use the following game. The action set is $X_i = \{0, 1\}$ for each player. For each $i \in N$, define $t^i \equiv (1, \ldots, 1, 1-n, 1, \ldots, 1) \in \mathbb{R}^N$ where $1-n < 0$ appears in the ith entry. This vector represents a transfer of 1 unit of i's utility to every other player. Let $e^i \equiv (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^N$ be the unit vector for the ith coordinate. Let L and M be two integers such that $L > M > n-1$. Let $\varepsilon \in (0, 1)$. Consider the following payoff vector function $u \colon X \to \mathbb{R}^N$:

$$\begin{aligned} u(x) = &\sum_{i=1}^{n-1} x_i x_n t^i + \sum_{i=1}^{n-1} (1 - x_i)(1 - x_n) L t^i \\ &+ (1 - x_1) x_2 \cdots x_n (1 + L) t^1 \\ &+ x_1 x_2 \cdots x_{n-1} (1 - x_n) M e^n \\ &+ \varepsilon x. \end{aligned} \tag{11}$$

The first term says that if players i and n ($i \neq n$) both choose 1, then one unit of utility is transferred from player i to every other player. The second term says that if i and n both choose 0, then $L > 1$ units of utility are transferred from i to every other player. By the third term, player 1 also makes utility transfers at action profile $(0, 1, \ldots, 1)$. By the fourth term, player n gets a bonus of $M > 0$ at action profile $(1, \ldots, 1, 0)$. By the last term, players who play 1 get a small bonus ε.

In the game (X, u), $(1, \ldots, 1, 0)$ is a strict Nash equilibrium, yielding a payoff vector $(\varepsilon, \ldots, \varepsilon, M)$. To see $(X, u) \in \mathcal{G}^G$, note that the first 3 lines in (11) generate integers for each player. Thus if $x_i = 0$, $u_i(x)$ is an integer. If $x_i = 1$, on the other hand, $u_i(x)$ is not an integer because of $\varepsilon \in (0, 1)$. This implies that players are never indifferent between their actions.

To prove that this game has no fair play profile, suppose, by contradiction, that there exists a fair play profile x. It is useful to distinguish two cases.

Case 1. $x_n = 0$. We first show $x = (0, \ldots, 0)$. To see this, suppose that $x_i = 1$ for some $i \neq n$. Then, by switching to 0, player i can add to the payoff vector either $Lt^i - \varepsilon e^i$ or $Lt^i - \varepsilon e^i - M e^n$. Since $L > M$ and $\varepsilon > 0$, this makes i himself worse off and all the other players better off. Thus, under the social code, i ought to switch to 0, in contradiction with our assumption that x is a fair play profile.

Therefore, if an action profile x such that $x_n = 0$ is a fair play profile, $x = (0, \ldots, 0)$. But then, at this action profile, player n ought to switch to 1, since $u(0, \ldots, 0, 0) = \sum_{i=1}^{n-1} Lt^i = -Lt^n$ while $u(0, \ldots, 0, 1) = \varepsilon e^n$.

Case 2. $x_n = 1$. We first show that $x_i = 1$ for all $1 < i < n$. To see this, suppose that $x_i = 0$ for some $1 < i < n$. This makes the second line of (11) equal to zero. Thus player 1 can affect only the first term and the last term in (11). He ought to play 1 since it allows him to transfer his utility to others. Given $x_1 = 1$, player i can also affect only the first term and the last term in (11). Then by the same

reason, i ought to switch to 1, in contradiction with x being a fair play profile.

Hence $x_i = 1$ for all $i > 1$. This implies that player 1 ought to play 0, since $u(1, 1, \ldots, 1) = \sum_{i=1}^{n-1} t^i + (\varepsilon, \ldots, \varepsilon)$ and $u(0, 1, \ldots, 1) = \sum_{i=2}^{n-1} t^i + (1 + L)t^1 + (0, \varepsilon, \ldots, \varepsilon)$ and hence $u(0, 1, \ldots, 1) - u(1, 1, \ldots, 1) = Lt^1 - \varepsilon e^1$.

Therefore, $x = (0, 1, \ldots, 1)$ is the only remaining possibility. But, at this action profile, player n ought to switch to 0, since $u(0, 1, \ldots, 1, 1) = \sum_{i=2}^{n-1} t^i + (1 + L)t^1 + (0, \varepsilon, \ldots, \varepsilon) = -t^n + Lt^1 + (0, \varepsilon, \ldots, \varepsilon)$ and $u(0, 1, \ldots, 1, 0) = Lt^1 + (0, \varepsilon, \ldots, \varepsilon, 0)$ and hence $u(0, 1, \ldots, 1, 0) - u(0, 1, \ldots, 1, 1) = t^n - \varepsilon e^n$. Q.E.D.

6 Conclusion

Contrary to our result, real-life social codes appear to constrain people's behavior. People often accept unpleasant tasks because of a social code or their own ethical feeling. This is not a contradiction since our result is based on normative properties of social codes and does not necessarily characterize actual social codes. What our result does is to shed light on a tension between the universalizability of a social code and what a social code can achieve as equilibrium social outcomes.

A dilemma is that if one accepts our axioms as ethically desirable, the result says that having a social code that satisfies them has little merit in terms of the induced outcomes. Since an admissible social code does not create new equilibria, it does not do anything for the Prisoners' Dilemma. Since it retains all strict Nash equilibria, its role as a coordination device is also limited: it does not eliminate any equilibrium in coordination games.

The result also shows that a seemingly plausible social code may violate a basic normative requirement. If a social code requires a person to sacrifice his own payoff to make others better off, we know that the social code violates at least one of the axioms. The violation may not be easy to detect since doing so requires one to consider hypothetical situations. This, we believe, gives a useful perspective on real-life social codes and ethical judgements.

Notes

We are grateful to the following people for helpful comments: Nir Dagan, Holger Meinhardt, Tatsuyoshi Saijo, Toyotaka Sakai, Robert Serrano, Tomoichi Shinotsuka, Kotaro Suzumura, Makoto Tanaka, Naoki Yoshihara, Rajiv Vohra, and participants of the Fall 2000 meeting of the Japanese Economic Association and seminars at Brown University and Universitat Autònoma de Barcelona. We are also grateful to Associate Editor in charge and two anonymous referees for very helpful comments. Part of this research was conducted while Miyagawa was visiting the Graduate School of Economics at Kobe University and he thanks the school for its hospitality. Nagahisa thanks Kansai University for financial support. The present paper supersedes Nagahisa and Suga (2000) and Nagahisa (2003).

A Appendix

A.1 Proof of Proposition 1

Suppose that F satisfies welfare nondiscrimination and let $(X, \succcurlyeq), (X', \succcurlyeq') \in \mathcal{G}$ be such that the assumption in the definition of neutrality is satisfied. Assume, without loss of generality, that the games are identical except for the names of player 1's actions: for all $i \neq 1$, $X_i' = X_i$ and $\rho_i(x_i) \equiv x_i$.[7] For example, it may be that these games are both Prisoners' Dilemma but player 1's actions are labeled differently; e.g.,

	A	B
a	$5,5$	$0,7$
b	$7,0$	$1,1$

$(X, \succcurlyeq)$

	A	B
c	$5,5$	$0,7$
d	$7,0$	$1,1$

$(X', \succcurlyeq')$

We further assume $X_1 \cap X_1' = \emptyset$. The case in which $X_1 \cap X_1' \neq \emptyset$ can be proved easily by repeating the argument that follows.[8]

Now, we start with $(X, \succcurlyeq)$ and add X_1' to 1's action set so that x_1 and $\rho_1(x_1)$ are welfare-equivalent for all $x_1 \in X_1$. Denote the constructed game by $((X_1 \cup X_1') \times X_{-1}, \succcurlyeq'')$. For the above example, the game constructed is one where one of the games is "stacked" on top of the other. Welfare nondiscrimination then

[7] The general case can be proved by repeating the argument for the other players.

[8] Specifically, we choose any set $Y \subseteq \mathbb{X}$ such that $|Y| = |X_1| = |X_1'|$, $Y \cap X_1 = \emptyset$, and $Y \cap X_1' = \emptyset$, and then apply our argument first to the pair (X_1, Y) and then to (Y, X_1').

implies that for all $x \in X$,

$$\begin{aligned} x_1 \in F_1(X, \succcurlyeq, x_{-1}) &\iff x_1 \in F_1((X_1 \cup X_1') \times X_{-1}, \succcurlyeq'', x_{-1}) \\ &\iff \rho_1(x_1) \in F_1((X_1 \cup X_1') \times X_{-1}, \succcurlyeq'', x_{-1}) \\ &\iff \rho_1(x_1) \in F_1(X', \succcurlyeq', x_{-1}) \end{aligned}$$

Welfare nondiscrimination also implies that for all $x \in X$ and all $i \in N \setminus \{1\}$,

$$\begin{aligned} F_i(X, \succcurlyeq, (x_1, x_{N\setminus\{1,i\}})) &= F_i((X_1 \cup X_1') \times X_{-1}, \succcurlyeq'', (x_1, x_{N\setminus\{1,i\}})) \\ &= F_i((X_1 \cup X_1') \times X_{-1}, \succcurlyeq'', (\rho_1(x_1), x_{N\setminus\{1,i\}})) \\ &= F_i(X', \succcurlyeq', (\rho_1(x_1), x_{N\setminus\{1,i\}})) \end{aligned}$$

A.2 Proof of Theorem 1

Let F be a social code that satisfies all the axioms. Suppose, by contradiction, that there exist a game $(X, \succcurlyeq) \in \mathcal{G}$, a player $j \in N$, and an action profile $x_{-j}^* \in X_{-j}$ such that

$$F_j(X, \succcurlyeq, x_{-j}^*) \cap BR_j(X, \succcurlyeq, x_{-j}^*) = \emptyset. \tag{12}$$

Let $B \equiv BR_j(X, \succcurlyeq, x_{-j}^*)$. By welfare nondiscrimination, we can assume, without loss of generality, that $|B| \geq 2$. Pick any action $x_j^* \in F_j(X, \succcurlyeq, x_{-j}^*)$. To simplify notation, let $j = 1$, $X_1 = \{1, 2, \ldots, |X_1|\}$, $B = \{1, \ldots, k-1\}$, and $x_1^* = k$ ($k \geq 3$).

Let $\succcurlyeq' \in P(X)^N$ be a preference profile such that for all $y_1, y_1' \in B$, all $z_1, z_1' \in X_1 \setminus B$, and all $i \in N$,

$$(y_1, x_{-1}^*) \succcurlyeq_i' (y_1', x_{-1}^*) \Leftrightarrow (y_1, x_{-1}^*) \succcurlyeq_i (y_1', x_{-1}^*), \tag{13}$$

$$(y_1, x_{-1}^*) \succ_1' (z_1, x_{-1}^*) \sim_1' (z_1', x_{-1}^*), \tag{14}$$

$$(y_1, x_{-1}^*) \prec_i' (z_1, x_{-1}^*) \sim_i' (z_1', x_{-1}^*) \quad \text{if } i \neq 1. \tag{15}$$

Note that these conditions do not specify preferences over x such that $x_{-1} \neq x_{-1}^*$. This does not cause a problem because of independence. The first condition says that we do not change anyone's preferences over B. The indifferences in (14) and (15) say that each player is indifferent over $X_1 \setminus B$. The strict preferences in (14) and (15) say that B remains the set of 1's best replies to x_{-1}^*, while the other players are better off outside of B.

Note that the move from $\succcurlyeq$ to $\succcurlyeq'$ can only shrink players' weak/strict lower-contour sets at action profiles (y_1, x_{-1}^*) with $y_1 \in B$. Thus, monotonicity and independence imply that this change of preferences does not make any action $x_1 \in B$ socially correct against x_{-1}^*, i.e., $F_1(X, \succcurlyeq', x_{-1}^*) \cap B = \emptyset$. Indeed, if there exists $x_1 \in F_1(X, \succcurlyeq', x_{-1}^*) \cap B$, then monotonicity and independence imply that x_1 is also in $F_1(X, \succcurlyeq, x_{-1}^*)$, which is a contradiction.

Now, note that all actions in $X_1 \setminus B$ are indifferent for all players given x^*_{-1}. So, we now remove all these actions except for k: 1's action set is now $B \cup \{k\} = \{1, \dots, k\}$. We keep preferences over the remaining action profiles. Since removed actions are all equivalent to k, welfare nondiscrimination and independence imply that no action in B becomes socially correct:

$$F_1(\{1, \dots, k\} \times X_{-1}, \succcurlyeq', x^*_{-1}) \cap B = \emptyset.$$

Since F is nonempty-valued,

$$F_1(\{1, \dots, k\} \times X_{-1}, \succcurlyeq', x^*_{-1}) = \{k\}. \tag{16}$$

By anonymity and welfare nondiscrimination, this conclusion can be extended to other games and players.

For all $i \in N$, let $u_i \colon X \to \mathbb{R}$ be a utility function that represents $\succcurlyeq'_i$ and satisfies a normalization $u_i(k, x^*_{-1}) = 0$. We define a function $f \colon K \times N \to \mathbb{R}$ by

$$f(x_1, i) = u_i(x_1, x^*_{-1}).$$

This function simply represents the $k \times n$ matrix whose (x_1, i) element denotes the payoffs that i obtains in game $(X, \succcurlyeq')$ at action profile (x_1, x^*_{-1}). Because of the normalization that $u_i(k, x^*_{-1}) = 0$, we have $f(k, i) = 0$ for all $i \in N$.

Define a function $\pi \colon N \times N \to N$ by $\pi(j, i) = i - j + 1 \pmod n$.

For all $i \in N$, let $\alpha_i \colon \{1, \dots, n-1\} \times K \to \mathbb{R}_{++}$ and $\beta_i > 0$. At this point, the values of $\alpha_i(\cdot, \cdot)$ and β_i are arbitrary as long as they are strictly positive. These values are specified at the end of the proof.

We are now ready to construct a game. The action set is $K = \{1, \dots, k\}$ for all players. For each i, the payoff function $v_i \colon K^N \to \mathbb{R}$ is given by

$$v_i(x) = \sum_{j=1}^{n-1} \alpha_i(j, x_n) f(x_j, \pi(j, i)) + \beta_i f(x_n, \pi(n, i)) \qquad \text{if } x_n < k, \tag{17}$$

$$v_i(x) = \sum_{j=1}^{n-1} \alpha_i(j, k) \Big[f(k + 1 - x_j, \pi(j, i)) - f(1, \pi(j, i)) \Big] \qquad \text{if } x_n = k. \tag{18}$$

We claim that game (K^N, v) where $v = (v_i)_{i \in N}$ has no fair play profile. We prove the claim by a series of lemmas.

Lemma 1. For any fair play profile x in (K^N, v), if $x_n < k$, then $x_i = k$ for all $i \neq n$.

Proof. Let x be a fair play profile such that $x_n < k$, and let $i \neq n$. We consider

the set of payoffs that player i can obtain given x_{-i}. Then (17) implies

$$v_i(\cdot, x_{-i}) = \text{constant} + \alpha_i(i, x_n) f(\cdot, 1)$$

where "constant" represents the term that is independent of i's action. We also need to look at the effect of i's action on the other players' welfare. For all $j \in N$,

$$v_j(\cdot, x_{-i}) = \text{constant} + \alpha_j(i, x_n) f(\cdot, \pi(i, j)).$$

These equations imply that i's position when the other players choose x_{-i} is identical to the position of player 1 in the game $(K \times X_{-1}, \succcurlyeq')$ given x^*_{-1}. Player j plays the role that player $\pi(i, j)$ plays in $(K \times X_{-1}, \succcurlyeq')$. Formally, for any pair of actions $a, a' \in K$ and for all $j \in N$,

$$v_j(a', x_{-i}) > v_j(a, x_{-i}) \iff (a', x^*_{-1}) \succ'_{\pi(i,j)} (a, x^*_{-1}).$$

Then, anonymity, independence, welfare nondiscrimination, neutrality, and (16) together imply $F_i(K^N, v, x_{-i}) = \{k\}$. Q.E.D.

Lemma 2. For any fair play profile x in (K^N, v), if $x_i = k$ for all $i \neq n$, then $x_n = k$.

Proof. Let x be a fair play profile such that $x_{-n} = (k, \ldots, k)$. Since $f(k, \cdot) = 0$,

$$v_i(\cdot, x_{-n}) = \beta_i f(\cdot, \pi(n, i)) \qquad \text{for all } i \in N.$$

Since $\pi(n, n) = 1$, this implies that player n's position is identical to that of player 1 in $(K \times X_{-1}, \succcurlyeq')$ given x^*_{-1}. Hence, $F_n(K^N, v, (k, \ldots, k)) = \{k\}$. Q.E.D.

Lemmas 1 and 2 imply that if there exists a fair play profile, $x_n = k$.

Lemma 3. For any fair play profile x in (K^N, v), if $x_n = k$, then $x_i = 1$ for all $i \neq n$.

Proof. Let x be a fair play profile such that $x_n = k$. Let $i \neq n$ and consider the payoff vectors that player $i \neq n$ can induce. The definition of v implies that for all $y_i \in X_i$ and all $j \in N$,

$$v_j(y_i, x_{-i}) = \text{constant} + \alpha_j(i, k) f(k + 1 - y_i, \pi(i, j))$$

where the first term is constant with respect to y_i. This implies that as before, the position of each player $i \neq n$ in (K^N, v, x_{-i}) is identical to that of player 1 in $(K \times X_{-1}, \succcurlyeq', x^*_{-1})$, but this time, action k in $(K \times X_{-1}, \succcurlyeq', x^*_{-1})$ corresponds to action $k + 1 - k = 1$ in (K^N, v, x_{-i}). Thus, the only fair action for player i is 1: $F_i(K^N, v, x_{-i}) = \{1\}$. Q.E.D.

Therefore, if a fair play profile exists, it must be $(1,\dots,1,k)$. However, the next lemma shows that $(1,\dots,1,k)$ is not a fair play profile. This concludes our proof that (K^N,v) has no fair play profile, a desired violation of effectiveness.

Lemma 4. Action profile $(1,\dots,1,k)$ is not a fair play profile in (K^N,v).

Proof. It suffices to prove that for any pair of actions $a,a' \in K$ and for all $i \in N$,

$$v_i(1,\dots,1,a') > v_i(1,\dots,1,a) \iff (k+1-a', x^*_{-1}) \succ'_{\pi(n,i)} (k+1-a, x^*_{-1}) \quad (19)$$

Indeed, if this condition holds, player n' position against $x_{-n} = (1,\dots,1)$ is essentially identical to that of player 1 in $(K \times X_{-1}, \succcurlyeq')$ given x^*_{-1}. The differences are that player i in (K^N,v) corresponds to player $\pi(n,i)$ in $(K \times X_{-1}, \succcurlyeq')$, action a in (K^N,v) corresponds to action $k+1-a$ in $(K \times X_{-1}, \succcurlyeq')$, and players $i \neq 1$ have different numbers of actions in $(K \times X_{-1}, \succcurlyeq')$ and (K^N,v). These differences are, however, not essential under anonymity, welfare nondiscrimination, and independence. Thus (19) implies that $x_n = 1$ is the only fair play for player n in (K^N,v) given $x_{-n} = (1,\dots,1)$: $F_n(K^N,v,(1,\dots,1)) = \{1\}$. Hence $(1,\dots,1,k)$ is not a fair play profile in (K^N,v). The remainder of the proof is devoted to prove (19).

A sufficient condition for (19) is that for all $i \in N$ and all $x_n < k$,

$$v_i(1,\dots,1,x_n) - v_i(1,\dots,1,k) = f(k+1-x_n, \pi(n,i)) - f(1,\pi(n,i)). \quad (20)$$

This implies that the function $v_i(1,\dots,1,\cdot)$ is identical, up to a constant, to $f(\cdot,\pi(n,i))$ except that the ordering in the domain of the function is reversed, i.e., $v_i(1,\dots,1,x_n)$ corresponds to $f(k+1-x_n,\pi(n,i))$.

By (17) and (18), it follows that (20) is equivalent to

$$\begin{aligned} \sum_{j=1}^{n-1} f(1,\pi(j,i))\big[\alpha_i(j,x_n) + \alpha_i(j,k)\big] + \beta_i f(x_n,\pi(n,i)) \\ = f(k+1-x_n,\pi(n,i)) - f(1,\pi(n,i)). \end{aligned} \quad (21)$$

This condition does not necessarily hold if $\alpha_i(\cdot,\cdot)$ and β_i are chosen arbitrarily. However, we claim that for all $i \in N$, there exist $\alpha_i \colon \{1,\dots,n-1\} \times K \to \mathbb{R}_{++}$ and $\beta_i > 0$ such that (21) holds for all $x_n < k$. This is sufficient for our proof since the argument so far does not depend on the values of $\alpha_i(\cdot,\cdot)$ and β_i as long as they are strictly positive.

To prove our claim, fix $i \in N$. An important observation is that for all $x_n < k$,

$$\begin{aligned} f(1,\pi(i,i)) > 0 \quad \text{and} \quad f(x_n,\pi(n,i)) < 0 \quad \text{if } i \neq n, \\ f(1,\pi(1,i)) < 0 \quad \text{and} \quad f(x_n,\pi(n,i)) > 0 \quad \text{if } i = n, \end{aligned}$$

which follows from the definition of $\succcurlyeq'$ and the normalization $f(k,\cdot)=0$. Thus, for all $x_n < k$, the left-hand side of (21) contains a positive term as well as a negative term, so it should be intuitively clear that (21) holds for all $x_n < k$ if we choose the weights on those terms appropriately. It should be noted, however, that β_i is independent of x_n while $\alpha_i(j,x_n)$ can depend on x_n.

A formal proof goes as follows. First, consider the case when $i \neq n$. Then, let $\beta_i > 0$ be sufficiently large so that for all $x_n < k$,

$$\begin{aligned} &\sum_{j=1}^{n-1} f(1,\pi(j,i)) + \beta_i f(x_n,\pi(n,i)) \\ &\quad < f(k+1-x_n,\pi(n,i)) - f(1,\pi(n,i)). \end{aligned} \tag{22}$$

The left-hand side of (22) coincides with that of (21) when $\alpha_i(j,x_n)=1/2$ for all (j,x_n). Since $f(1,\pi(i,i))>0$, the equality of (21) can be attained if we increase $\alpha_i(i,x_n)$.

The case when $i=n$ is similar. First, set $\beta_i>0$ sufficiently large so that for all $x_n<k$, (22) holds with the reverse inequality. The inequality implies that if we set $\alpha_i(j,x_n)=1/2$ for all (j,x_n), then the left-hand side of (21) is larger than the right-hand side. Since $f(1,\pi(1,n))=f(1,n)<0$, the equality in (21) can be attained if we increase $\alpha_n(1,x_n)$.

To sum up, the last two paragraphs prove that for all i, there exist α_i and β_i such that (21) holds for all $x_n<k$, thus (19) holds. As we discussed, (19) implies $F_n(K^N,v,(1,\ldots,1))=\{1\}$, hence $(1,\ldots,1,k)$ is not a fair play profile. Q.E.D.

Remark 2. The proof goes through for any subclass of games $\mathcal{G}'$ such that $\mathcal{G}' \supseteq \mathcal{G}^{NE}$, since only games in $\mathcal{G}^{NE}$ are used. To see this, first note that, for $(X,\succcurlyeq')$ and $(K\times X_{-1},\succcurlyeq')$, the issue is trivial since $\succcurlyeq'$ is specified only for one "column" associated with x^*_{-1}. It is easy to complete the specification of these games so that they belong to $\mathcal{G}^{NE}$. For (K^N,v), since $k>2$, the arguments in the proof reveal that $(1,\ldots,1,2)$ is a Nash equilibrium, hence $(K^N,v)\in\mathcal{G}^{NE}$.

A.3 Independence of the Axioms in Theorem 1

This section shows that none of the axioms in Theorem 1 is redundant. We show that if any of the axioms is removed, there exists a social code that satisfies the remaining axioms and violates the conclusion of the theorem.

Example 4 (Global Pareto code)**.** Let $GP(X,\succcurlyeq)\subseteq X$ denote the set of strongly Pareto efficient action profiles in game $(X,\succcurlyeq)$. Define a social code by

$$F_i(X,\succcurlyeq,x_{-i}) = \{x_i\in X_i : (x_i,x'_{-i})\in GP(X,\succcurlyeq) \text{ for some } x'_{-i}\in X_{-i}\}.$$

Thus, for an action to be acceptable, it suffices that the action achieves a

(global) Pareto efficient outcome for *some* (not necessarily actual) action profile of the other players.

This code satisfies all the axioms except for independence. To see that (4) does not hold, consider

	A	B
A	$1,1$	$1,0$
B	$0,3$	$2,2$

Since the strongly Pareto efficient outcomes are $\{(B,A),(B,B)\}$, the socially correct actions are $\{B\}$ for 1 and $\{A,B\}$ for 2, regardless of each other's action. Then $FPE(X,\succcurlyeq,F)=\{(B,A)\}$, while $NE(X,\succcurlyeq)=SNE(X,\succcurlyeq)=\{(A,A)\}$.

Example 5 (Dictatorship)**.** There exists $k\in N$ such that for all $(X,\succcurlyeq)\in\mathcal{G}$, all $i\in N$, and all $x_{-i}\in X_{-i}$,

$$F_i(X,\succcurlyeq,x_{-i})=\{x_i\in X_i:(x_i,x_{-i})\succcurlyeq_k(y_i,x_{-i})\text{ for all }y_i\in X_i\}.$$

Thus one's action is fair if and only if it is optimal for the dictator. This code satisfies all the axioms except for anonymity. It should be clear that (4) does not hold under this code.

Example 6 (Anti-Pareto code)**.** The code first reverses each player's preference ordering and then applies the local strong Pareto code (Example 2). Formally, given $(X,\succcurlyeq)\in\mathcal{G}$, $x\in X$, and $i\in N$, $x_i\in F_i(X,\succcurlyeq,x_{-i})$ if and only if there exists no $x_i'\in X_i$ such that $(x_i',x_{-i})\preccurlyeq_k(x_i,x_{-i})$ for all $k\in N$ with strict preference holding for some $k\in N$.

This code satisfies all the axioms except for monotonicity. To see that (4) does not hold, consider

	A	B
A	$2,2$	$1,1$
B	$1,1$	$0,0$

Then, $FPE(X,\succcurlyeq,F)=\{(B,B)\}$, while $NE(X,\succcurlyeq)=SNE(X,\succcurlyeq)=\{(A,A)\}$.

Example 7. This code roughly says that an action is not socially correct if it is the unique least preferred action for every other player and the action sets are sufficiently large. Formally, take any game $(X,\succcurlyeq)$. If $\sum_{i\in N}1/|X_i|\geq 1$, then all actions are socially correct for all players in all situations: $F_i(X,\succcurlyeq,x_{-i})=X_i$ for all i and all x_{-i}. On the other hand, if $\sum_{i\in N}1/|X_i|<1$ (i.e., the number of actions is sufficiently large for each player), then for all $x\in X$ and all $i\in N$, $x_i\notin F_i(X,\succcurlyeq,x_{-i})$ if and only if for all $x_i'\in X_i\setminus\{x_i\}$ and all $k\in N\setminus\{i\}$, $(x_i',x_{-i})\succ_k(x_i,x_{-i})$.

This code satisfies all the axioms except for welfare nondiscrimination. To see that effectiveness is satisfied, suppose $\sum_{i\in N} 1/|X_i| < 1$ (otherwise, it is trivial). A key observation is that there exists at most one socially incorrect action for each player given the other players' actions. Thus, the maximum number of action profiles in which at least one player chooses a socially incorrect action is $\sum_{i\in N}\prod_{j\neq i}|X_j|$. Hence, the minimum number of action profiles in which all players play fair is

$$\prod_{i\in N}|X_i| - \sum_{i\in N}\prod_{j\neq i}|X_j| = \big(1 - \sum_{i\in N} 1/|X_i|\big)\prod_{i\in N}|X_i| > 0.$$

To see that this social code violates welfare nondiscrimination, consider

	A	B	C
a	$1,3$	$0,0$	$0,0$
b	$3,1$	$0,0$	$0,0$

	A	B	C
a	$1,3$	$0,0$	$0,0$
b	$3,1$	$0,0$	$0,0$
c	$3,1$	$0,0$	$0,0$

Note that $1/|X_1| + 1/|X_2| < 1$ for either game. Thus if player 2 plays A, the social code states that b is not socially correct in the left game since it gives the *unique* least preferred outcome for player 2. In the right game, however, b is socially correct since c is as bad as b for player 2.[9]

The left game also shows that (4) does not hold for this social code, since $FPE(X, \succcurlyeq, F) = \{(a, A)\}$ while $NE(X, \succcurlyeq) = \{(b, A)\}$. Since this social code satisfies neutrality, this example also shows that Theorem 1 does not hold if welfare nondiscrimination is replaced with neutrality.

Example 8 (Borda code)**.** This social code states that for all $(X, \succcurlyeq) \in \mathcal{G}$, all $x \in X$, and all $i \in N$, $x_i \in F_i(X, \succcurlyeq, x_{-i})$ if and only if (x_i, x_{-i}) is a Borda count winner in $\{(x'_i, x_{-i}) : x'_i \in X_i\}$ when action profiles that are indifferent for all players are viewed as the same alternative.[10]

This code satisfies all the axioms except for effectiveness. To see that effectiveness is not satisfied, consider

	a	b	c
a	$5,3$	$4,2$	$0,1$
b	$1,5$	$3,1$	$2,0$
c	$0,4$	$2,5$	$1,2$

(23)

[9] On the other hand, if we remove the requirement of uniqueness from the definition of the social code, then we lose effectiveness.

[10] That is, if $(x'_i, x_{-i}) \sim_k (x''_i, x_{-i})$ for all $k \in N$, then they are viewed as equivalent, and we apply the Borda count to equivalent classes.

For example, if player 2 plays a, the unique fair play for player 1 is b. Indeed, among the action profiles in the first column, (b, a) ranks first for player 2 and second for 1, while (a, a) ranks first for 1 but third for 2, and hence (b, a) dominates (a, a) under the Border count. This social code is well defined since it specifies a non-empty set of socially correct actions for every contingency. However, under this code, the above game has no fair play profile. Indeed, if 2 plays c, then 1 should play c, in which case 2 should play b, in which case 1 should play a, in which case 2 should play a, in which case 1 should play b, in which case 2 should play b, and we obtained a cycle. Since we went through all actions of player 2, there exists no fair play profile.

The same game also shows that (4) does not hold for this social code, since $FPE(X, \succcurlyeq, F) = \emptyset$ while $NE(X, \succcurlyeq) = SNE(X, \succcurlyeq) = \{(a, a)\}$. To see that $FPE \not\subseteq NE$ is also possible, we can use the first column in (23), viewed as a 3×1 game matrix, where $FPE(X, \succcurlyeq, F) = \{(b, a)\}$ and $NE(X, \succcurlyeq) = \{(a, a)\}$.

A.4 Proof of Proposition 2

We note in advance that the proof will eventually use a part of the proof for Proposition 4 (specifically, we will use game (11)). Let F be a social code that satisfies the axioms and let $(X, \succcurlyeq) \in \mathcal{G}^{SNE}$. It suffices to prove the following:

Lemma 5. For all i and all x_{-i}, if $|BR_i(X, \succcurlyeq, x_{-i})| = 1$, then $BR_i(X, \succcurlyeq, x_{-i}) \subseteq F_i(X, \succcurlyeq, x_{-i})$.

To prove this, suppose otherwise: i.e., there exist $y \in X$ and $i \in N$ such that $BR_i(X, \succcurlyeq, y_{-i}) = \{y_i\}$ but $y_i \notin F_i(X, \succcurlyeq, y_{-i})$. If $|X_{-i}| > 1$, choose any $w_{-i} \in X_{-i}$ such that $w_{-i} \neq y_{-i}$. If $X_{-i} = \{y_{-i}\}$, let $w_{-i} = y_{-i}$. Let $\succcurlyeq'$ be a preference profile represented by the following payoff-vector function $v \colon X \to \mathbb{R}^N$:

$$v(x) = \begin{cases} (1, \dots, 1) & \text{if } x = y, \\ (2, \dots, 2, 0, 2, \dots, 2) & \text{if } x_{-i} = y_{-i} \text{ and } x_i \neq y_i, \\ (3, \dots, 3) & \text{if } x = (y_i, w_{-i}) \text{ and } w_{-i} \neq y_{-i}, \\ (2, \dots, 2) & \text{otherwise,} \end{cases}$$

where "0" on the second line appears in the ith entry. In this game, (y_i, w_{-i}) is a strict Nash equilibrium, hence $(X, \succcurlyeq') \in \mathcal{G}^{SNE}$.

The construction implies that for all $x \in X$ and all $j \in N$,

$$\begin{aligned} y \succcurlyeq'_j x &\Longrightarrow y \succcurlyeq_j x, \\ y \succ'_j x &\Longrightarrow y \succ_j x. \end{aligned} \tag{24}$$

To see this, note that $y \succcurlyeq'_j x$ holds for $x \neq y$ only if $x_{-i} = y_{-i}$ and $j = i$. But then, we have $y \succ_j x$ since $\{y_i\} = BR_i(X, \succcurlyeq, y_{-i})$ by our initial choice of y. (24) implies

that, in the move from $\succcurlyeq$ to $\succcurlyeq'$, the relative ranking of y does not go up for any player. Therefore, by monotonicity and independence, $y_i \notin F_i(X, \succcurlyeq', y_{-i})$. Since all actions $x_i \neq y_i$ are welfare-equivalent given y_{-i}, welfare nondiscrimination and independence imply $F_i(X, \succcurlyeq', y_{-i}) = X_i \setminus \{y_i\}$.

Pick any action $z_i \in X_i \setminus \{y_i\}$. We now remove all actions from X_i except for $\{z_i, y_i\}$. Thus, $X_i' \equiv \{z_i, y_i\}$ is the new action set for i, and $X' \equiv X_i' \times \prod_{j \neq i} X_j$ is the new set of action profiles. Since (y_i, w_{-i}) remains in the game, $(X', \succcurlyeq') \in \mathcal{G}^{SNE}$. By welfare nondiscrimination, $F_i(X', \succcurlyeq', y_{-i}) = \{z_i\}$. Note that given y_{-i}, player i prefers y_i while all the other players prefer z_i. Hence, anonymity, welfare nondiscrimination, and independence imply the following: for any game, if a given player j has only two actions a and b and, given x_{-j}, player j prefers a and all the other players prefer b, then j ought to play b. By welfare nondiscrimination, we can extend this conclusion to the case in which j has more than two actions.

If $n \geq 3$, the remaining argument is the same as that of Proposition 4: use (11). If $n = 2$, the argument in the proof of Proposition 4 does not work here since it relies on Pareto dominance. However, we can use the following game.

	A	B	C
A	$1, 1$	$2, 0$	$2, 0$
B	$0, 2$	$2, 0$	$0, 2$
C	$0, 2$	$0, 2$	$2, 0$

This game has a strict Nash equilibrium, i.e., (A, A). By the conclusion derived in the previous paragraph, if player 2 plays A, player 1 ought to play either B or C since doing so makes himself worse off and the other player better off. Using the same argument, one can easily see that at any action profile, at least one of the players ought to change his action. Q.E.D.

Lemma 5 has another implication: the inclusion $FPE \subseteq NE$ is restored on $\mathcal{G}^{SNE}$ if we limit ourselves to games where each player has two actions.

Corollary 2. *Suppose that $n \geq 3$ and the domain of social codes is the class of games in $\mathcal{G}^{SNE}$ such that $|X_i| = 2$ for all i. Then if a social code F satisfies anonymity, monotonicity, independence, welfare nondiscrimination, and effectiveness, then for all games in the domain,*

$$SNE(X, \succcurlyeq) \subseteq FPE(X, \succcurlyeq, F) \subseteq NE(X, \succcurlyeq).$$

Proof. We only need to prove $FPE(X, \succcurlyeq, F) \subseteq NE(X, \succcurlyeq)$. Thus, suppose that there exists $x \in FPE(X, \succcurlyeq, F)$ such that $x \notin NE(X, \succcurlyeq)$. Let i be such that x_i is not a best reply against x_{-i}. Since there are only two actions, $BR_i(X, \succcurlyeq, x_{-i}) = \{y_i\}$ where $y_i \in X_i \setminus \{x_i\}$. But then, by Lemma 5, $y_i \in F_i(X, \succcurlyeq, x_{-i})$; if $n \geq 3$, the proof of the lemma goes through even if only two-action games are in the domain. Therefore, i is allowed to switch to his best reply, a desired contradiction. Q.E.D.

References

Arrow, J. K. (1963). *Social Choice and Individual Values*, New York: John Wiley, second edition.

Arrow, K. J., Sen, A. K., and Suzumura, K., eds. (2002). *Handbook of Social Choice and Welfare*, volume 1, Amsterdam: North-Holland.

Austen-Smith, D. and Banks, J. S. (1999). *Positive Political Theory I: Collective Preference*, Ann Arbor: University of Michigan Press.

Debreu, G. (1952). "A Social Equilibrium Existence Theorem," *Proceedings of the National Academy of Sciences of the U.S.A.* **38**, 886–893.

Gibbard, A. (1974). "A Pareto-Consistent Libertarian Claim," *Journal of Economic Theory* **7**, 388–410.

Hare, M. R. (1963). *Freedom and Reason*, Oxford: Clarendon Press.

Hare, M. R. (1981). *Moral Thinking: Its Levels, Method, and Point*, Oxford: Clarendon Press.

Kant, I. (1785). *Grundlegung zur Metaphysik der Sitten (Fundamental Principles of the Metaphysic of Morals)*, URL `http://eserver.org/philosophy/kant/metaphys-of-morals.txt`, translated by T. K. Abbott.

Moulin, H. (1988). *Axiom of Cooperative Decision Making*, Cambridge: Cambridge University Press.

Nagahisa, R. (2003). "A Normative Axiomatization of Nash equilibrium by Fair Play Game," mimeo, Kansai University.

Nagahisa, R. and Suga, K. (2000). "Fair Play Equilibria in Normal Form Games: A Normative Justification of Nash Equilibrium and an Impossibility of Procedual Justice," mimeo, Kansai University and Waseda University.

Peleg, B. and Tijs, S. (1996). "The Consistency Principle for Games in Strategic Forms," *International Journal of Game Theory* **25**, 13–34.

Sen, A. K. (1970). "The Impossibility of a Paretian Liberal," *Journal of Political Economy* **78**, 152–7.

Sen, A. K. (1986). "Social Choice Theory," in "Handbook of Mathematical Economics," (K. J. Arrow and M. D. Intriligator, eds.), Amsterdam: North Holland.