An Axiomatic Approach to the Design of Moral Codes*

Eiichi Miyagawa† Ryo-ichi Nagahisa‡
Kobe University Kansai University

Koichi Suga§
Waseda University

June 18, 2015

Abstract

For a moral code of conduct to gain universal acceptance in society, it would have to satisfy minimum requirements of consistency and procedural justice. The so-called prescriptivity and universalizability principles in ethics together say that any moral judgement prescribes

*This paper was presented at 14th SAET Conference, 2014, Tokyo. We thank all of the participants for their valuable comments. A very early version of the paper superseded Nagahisa, presented with the different title, "fair play equilibria with mixed strategies," at 8th International Meeting of the Society of Social Choice and Welfare, 2006, Istanbul, although the proof of the main result turned out to be incomplete. Cooperation with Miyagawa and Suga has finally completed the paper. We also thank Prof. Kohei Kamaga (Sophia University) and Prof. Tsuyoshi Adachi (Takasaki City University) for their helpful comments.

†Department of Economics, Kobe University, 2-1, Rokkodaicho Kobe Hyogo 657-8501 Japan

‡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.

a person what to do in a situation and that the prescription should be universalizable to apply to other persons' actions in situations that are identical in relevant respects. By adapting standard axioms in social choice theory, we formalize these principles in the framework of normal form games and study the implications on equilibrium outcomes. A moral code specifies socially acceptable responses against other individuals' behavior. A fair play equilibrium is a strategy profile where everyone behaves optimally subject to the moral code. We show that for any admissible moral code, the set of fair play equilibria coincides with that of Nash equilibria in all games. The result identifies a conflict between the two principles of moral judgements and what a moral code can achieve as equilibrium outcomes.

JEL Classification: *D63, D71*

Keywords: *moral code, prescriptivity, universalizability, Hare, fair play equilibrium, Nash equilibrium, Arrow's impossibility theorem*

1 Introduction

This paper develops a new axiomatic approach to the design of moral codes, motivated by the moral philosophy of Hare (1952, 1963, 1981). Richard Mervyn Hare (1919 – 2002), a leading English moral philosopher in the twentieth century, appeals to two logical properties of moral judgements: prescriptivity and universalizability. Prescriptivity states that moral judgements entail imperatives and normally lead to action. It does not just describe or evaluate decisions; "You ought to do this" (used evaluatively) logically entails the imperative "Do this."

Those imperatives must be universalizable in the sense that if it is right for a particular person P to do an action A, then the same action must likewise be right for any person exactly like P, or like P in the relevant respects. Furthermore, if P is right in doing A in a situation, then it must be right for

the person to do the same in other relevantly similar situations. This is the essence of universalizability[1].

Universalizability has its roots in a wide range of world cultures. For example, the so called Golden Rule of the gospel is one of them: "Do to others as you would have them do to you." Kant is the first moral philosopher to appreciate the concept of universalizability, and formulates it as his categorical imperative, which states that the only morally acceptable maxims of our actions are those that could rationally be willed to be universal law (Kant, 1785). Sidgwick (1907) is also a stout defender of universalizability, which underlies his principle of "equity" or "fairness."

...whatever action any of us judges to be right for himself, he implicitly judges to be right for all similar persons in similar circumstances. Or, as we may otherwise put it, 'if a kind of conduct that is right (or wrong) for me is not right (or wrong) for some one else, it must be on the ground of some difference between the two cases, other than the fact that I and he are different persons.' (Sidgwick (1907), Book III, Chap. XIII, p.379.)

Although prescriptivity and universalizability are certainly fundamental and not without practical importance, a question still remains in our mind: What is the outcome if all the members in a society make moral judgments according to prescriptivity and universalizability and they choose their courses of action independently ? For example, if it is right for a person P to do an action A, then prescriptivity makes this moral judgement an imperative that orders P to do A. Universalizability applies this imperative to any other person who is exactly like P, or like P in the relevant respects: it orders a person Q, who is similar to P in the relevant respects, to do A. As a result of the applications of prescriptivity and universalizability, every imperative that is applied to a particular person in a particular circumstance induces a large number of imperatives, which are applied to similar persons in similar

[1]The explanation of universalizability presented here is partly indebt to Chapter 9 in Sen (1970b), which gives an excellent review of universalizability.

circumstances. Furthermore, it is noted that those imperatives are strategically related: a person cannot make a moral judgment without considering what actions other persons choose following their moral judgements. It is therefore not clear what outcome results from those imperatives, and it is also doubtful if those imperatives, which are strategically interdependent, are compatible[2].

By applying technique familiar from social choice theory, this paper explores the meaning of prescriptivity and universalizability and clarifies its implications on social behavior. We use a normal-form game specifying a set of feasible actions and preferences for each member. We are interested in what we call a *moral code*, which is defined as a rule that specifies whether a given strategy, an assignment of a probability to each action, is socially acceptable for a given member in a given situation. Thus, whether it is all right for you to choose a certain strategy depends on the situation. In our context, the "situation" consists of the game and other members' strategies. We denote this as $m_i \in F_i(G, m_{-i})$, which says that m_i is a socially acceptable strategy for individual i if the game is G and the vector of other individuals' strategy is $m_{-i} = (m_j)_{j \neq i}$. The entry of m_{-i} in the formula comes from the observation that the social acceptability of a behavior often depends on the behavior of other individuals. Whether it is all right to drive at 50mph on a highway depends on the average speed on the road in the particular situation. While legal, driving at 50mph may not be considered appropriate by other drivers if they are all driving at 70mph. Similarly, whether it is all right to spend six months to write a referee report for a paper depends on the average in the field. How much you should spend for a gift depends on the amount spent by other people in your circle. Since G and m_{-i} enter the formula, a moral code as defined here is a generator of social judgements, generating instructions for each possible situation. Therefore, the social norm for drivers

[2]Hare argues, as Harsanyi (1955) does, that the combination of universalizability and prescriptivity leads to a certain form of utilitarianism, namely, preference utilitarianism (Hare 1981). But his argument ignores strategic interdependency of moral judgements.

on highways and the social norm for referees may be generated by a single moral code. But a single moral code may not govern all situations. The moral code may differ across different regions, organizations, and circles.

We assume that if a player has more than one socially acceptable strategies, he choose one that he prefers the most. This suggests the following equilibrium concept. Given a moral code, a *fair play equilibrium* is a strategy profile where each player chooses his most preferred socially acceptable strategy given the other players' strategy profile. The equilibrium concept is an application of the social equilibrium of Debreu (1952) when a player's strategy set is constrained by the moral code. It is noted that different moral codes may generate different fair play equilibria in the same game.

The two concepts, moral code and fair play equilibrium, capture prescriptivity because a moral code prescribes players to choose from the set of socially acceptable strategies. Universalizability is captured by axioms that reasonable moral 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 strategy in a situation, the same strategy should be allowed to you in the situation where your position is the same as the person's in the previous situation.

Welfare nondiscrimination says that what matters ultimately for the society is the members' welfare, and therefore a pair of strategies 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 moral 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 $F_i(G, m_{-i})$ is independent of the payoffs at strategy profiles where m_{-i} is not chosen. By definition, $F_i(G, m_{-i})$ is relevant only if other players choose m_{-i}. Thus, the axiom says, situations where

the other players' strategy profile is not m_{-i} are counter-factual and should be immaterial for $F_i(G, m_{-i})$. The requirement is natural in our framework since what the moral code determines is whether one's strategy is a socially acceptable *response* to other players' strategies[3].

Lastly, effectiveness says that it should be feasible for all players to follow the moral code simultaneously. That is, for any game, there should exist a pure strategy profile where no one violates the moral code.

An important feature of these axioms is that they do not define social correctness directly. They rather formulate consistency in what a moral code prescribes across situations and players. The basic form of the axioms is "if it is socially acceptable for player P to choose strategy A in situation S, then it should be acceptable for player P' to choose action A' in situation S'," as in the quotation from Sidgwick (1907). Hare (1981) seems to be using universalizability in the sense to refer to anonymity, not referring to our other axioms. But we interpret here universalizability in a broader sense, as a consistency requirement in moral judgements across situations and players.

The axioms, perhaps except for effectiveness, are familiar in social choice theory[4]. A difference is that social choice theory is concerned with mappings specifying the outcome or orderings over possible outcomes. We are, on the other hand, concerned with mappings specifying the set of permitted strategies for each player. In particular, our mappings, i.e., social codes, do not specify the outcome directly. The outcome is determined indirectly as an equilibrium.

We have three theorems, Theorems 1-3. The first two demonstrate a close association of the fair play equilibria with Nash equilibria and the third shows a welfare property of fair play equilibrium. The first two are the most

[3] Although the axiom resembles Arrow's independence of Irrelevant Alternatives (1963), it is not a requirement of informational economization.

[4] For an introduction to social choice theory, see, e.g., Sen (1986), Moulin (1988), Austen-Smith and Banks (1999), and Campbell and Kelly (2002).

important so that we explain the details here. Theorem 1 shows that, under any moral code that satisfies the axioms, fair play equilibria are necessarily Nash equilibria in any game. Thus at any fair play equilibrium, the moral code is never binding for any player. We also show that strict Nash equilibria are all fair play equilibria. Thus a moral 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 moral code satisfies a mild continuity condition, the difference disappears: the set of fair play equilibria coincides with the set of Nash equilibria (Theorem 2).

The basic intuition of the result is as follows. If a moral code does not lead to a Nash equilibrium, our axioms (in particular, independence and monotonicity) imply that there must be a game where some player is required to sacrifice his payoffs for the sake of other players' payoffs. Then we can find a game in which the moral code induces a cycle where each player is required to sacrifice his own payoffs and it is not possible for all players to follow the moral code simultaneously, which is a violation of effectiveness axiom. This observation therefore leads to a conclusion that an admissible moral code permits players to take at least one of his best replies, which in turn implies that any strict Nash equilibrium is a fair play equilibrium and any fair play equilibrium is a Nash equilibrium. The proof therefore resembles that of Arrow's impossibility theorem (Arrow 1963), where if a voter is decisive over a single pair of alternatives, Arrow's axioms imply that the voter is actually decisive over all pairs of alternatives.

A way of escaping from the negative results of Theorems 1 and 2 is to abandon independence or the combination of weak Independence with the equivalent utility representation axiom that prevents any kind of interpersonal welfare comparison. Abandoning these axioms makes way for two admissible moral codes: One is called the Utilitarian code, based on the same idea as that in utilitarian rules. The other is called the Lexi-min code, base on

the same idea as that in Lexi-min rule due to Sen (1970b), a lexicographic completion of Rawls'(1972) difference principle. We show that these code work: both the codes have fair play equilibria for any game. Although axiomatic characterization of the codes is the issue that remains in future[5], we stress here that our approach is open to the subject of seeking the possibility of admissible codes that make interpersonal welfare comparison possible.

These informational bases are often called 'welfarism' whether cardinality and interpersonal comparability of utilities are allowed. In other words, social decision based on welfarism makes use of utility information and excludes others. In the literature we find other types of informational bases, that is, non-utility informational bases such as rights and procedures. The most important contribution to these bases was introduced by Sen (1970a) to show impossibility of a Paretian liberal. Subsequent discussions have been dedicated to the definition of rights. In the debate there are two main streams, one based on the social choice theoretic framework and the other on game theoretic one à la Nozick (1974)[6]. In the real world rights of others are often considered to give constraints on our behavioral decisions. We are not free to give risks and fears to others if we accept the value of democracy. Procedures are also important bases to restrict our decision and judgment in relation to our behavior. Our behavior is often prohibited from the reason not to observe rules and procedures. We will incorporate such informational bases into our framework and explore the effects on the judgment of our behavior in our future work.

Lastly, we mention a few papers that are closely related to ours. Peña (2003) develops a different axiomatic analysis of moral codes. Moral code there does not restrict actions that are available to players. It rather cre-

[5] A significant amount of literature deals with the axiomatizations of the utilitarian and lexi-min rules. d'Aspremont (1985), Bossert and Weymark (2004), d'Aspremont and Gevers (2002), and Sen (2011) are excellent surveys.

[6] See for example, Gaertner, Pattanaik and Suzumura (1992), Sen (1992, 1996, 2011), Suzumura (1996, 2000, 2011)

ates moral values that influence players' payoffs and change rational choices: When a player chooses a good action he receives moral rewards, and when he chooses a bad action he suffers from moral penalties.

Since our result gives an axiomatization of Nash equilibria, it is similar to the results of Peleg and Tijis (1996), Peleg et.al. (1996) and Salonen (1992). The first two use a reduced game property with a variable number of agents, while the last one uses axioms in axiomatic bargaining theory with a fixed number of agents. Our paper differs from these papers in at least two respects. First, we do not use a variable population setting. Second, their solution concept chooses social outcomes directly as in social choice theory

This paper is organized as follows. Section 2 provides notation and definitions, where we give the definition of moral code and fair play equilibrium. Section 3 defines our axioms. Section 4 states and proves a few preliminary results concerning the axioms. Section 5 states our main results and illustrates their proofs using a simple case with two players and two actions. Examples in Section 6 serve the interpretation of the main results. Section 7 introduce the Utilitarian and the Lexi-min codes. Section 8 is the conclusion. Section 9 proves the main results.

2 Model

The set of players is fixed and denoted by $N = \{1, 2, ..., n\}$ where $n \geq 2$. Let Ω be an infinite set of potential actions (pure strategies). A (finite) game is a list

$$(X, u) := (\prod_{i \in N} X_i, (u_i)_{i \in N})$$

where for all $i \in N$, $X_i \subset \Omega$ is a non-empty finite set of actions and u_i is a utility function defined over X. An element of X is called an action profile and denoted $x = (x_1, ..., x_n) \in X$. We call $(u_i)_{i \in N}$ a utility profile. The class of all games is denoted as Γ.

Let a game $(X, u) \in \Gamma$ be given. A (mixed) strategy of player i, denoted by m_i, is a probability distribution over i's actions. Thus m_i is a function from X_i to $[0, 1]$ that associates with each $x_i \in X_i$ a probability $m_i(x_i) \in [0, 1]$ such that $\sum_{x_i \in X_i} m_i(x_i) = 1$. We write $m_i = (m_i(x_i))_{x_i \in X_i}$. Let $M(X_i)$ be the set of mixed strategies generated from X_i, which is simply written as M_i wherever X_i is apparent in the context. We let $M = \prod_{i \in N} M_i$. A typical element of M is written as $m = (m_1, ..., m_n)$ or $m = (m_i)_{i \in N}$, and called a (mixed) strategy profile[7]. Each player i has an expected utility function defined over M given by

$$v_i(m) := \sum_{x \in X} \prod_{j \in N} m_j(x_j) u_i(x) \text{ for all } m \in M$$

where $\prod_{j \in N} m_j(x_j)$ is the probability that m assigns to $x = (x_1, ..., x_n) \in X$. When the utility function over pure strategies is denoted u_i', the associated utility function over mixed strategies is denoted v_i', and so on. We denote $\prod_{j \in N \setminus \{i\}} X_j$ and $\prod_{j \in N \setminus \{i\}} M_j$ by X_{-j} and M_{-j} respectively. Typical elements of X_{-j} and M_{-j} are denoted by x_{-j} and m_{-j}.

Given a subset $Y_i \subset X_i$ for all i and $Y := \prod_{i \in N} Y_i$, let $u_{i|Y}$ denote the restriction of u_i to Y. We can then define a profile $u_{|Y} = (u_{1|Y}, ..., u_{n|Y})$ and a game $(Y, u_{|Y})$. We simply write (Y, u) instead of $(Y, u_{|Y})$ if there is no risk of confusion.

Given $(X, u) \in \Gamma$, $i \in N$ and $m_{-i} \in M_{-i}$, let $BR_i(X, u_i, m_{-i})$ denote the set of best replies to m_{-i} for player i: $BR_i(X, u_i, m_{-i}) := \{m_i \in M_i : v_i(m_i, m_{-i}) \geq v_i(m_i', m_{-i}) \ \forall m_i' \in M_i\}$.

Definition A *moral code* is a correspondence F that associates with each game $(X, u) \in \Gamma$, each $i \in N$, and each $m_{-i} \in M_{-i}$ a non-empty subset $F_i(X, u, m_{-i}) \subset M_i$.

[7] By definition, an action profile is also a strategy profile.

Here $F_i(X, u, m_{-i})$ is the set of i's strategies that are considered as fair or socially acceptable in game (X, u) when the other players' strategies are m_{-i}. We require this set to be non-empty; for each player and each possible situation, there exists at least one socially acceptable strategies.

We assume that players respect a given moral code; i.e., players do not choose a socially unacceptable strategy, although choosing such a strategy is physically possible. In practice, people choose to respect the given moral code either because a violation of the code results in a punishment from other people, or people have an intrinsic desire to comply with the code (either because they appreciate the ideas behind the code or they have been educated to have such a desire). Since the issue of these incentives is not our main concern in this paper, we simply assume that players choose only socially acceptable strategies.

A special case is where a moral code specifies only a set of pure strategies and allows any randomization among those pure strategies. That is,

$$F_i(X, u, m_{-i}) = \{m_i \in M_i : \text{supp}(m_i) \subset F_i^*(X, u, m_{-i})\}$$

where $\text{supp}(m_i)$ denotes the support of m_i and $F_i^*(X, u, m_{-i}) \subset X_i$ is the set of pure strategies that are fair replies to m_{-i}. Our results are valid for, but not limited to, this subclass of moral codes.

A moral code may specify more than one strategy as socially acceptable and does not necessarily deprive players of their free choice completely. The typical form of a moral code is not "one ought to do this" but "one ought not to do these." When there are multiple strategies that are socially acceptable, we assume that the player chooses a strategy that is most preferred within the set of socially acceptable strategies. This consideration suggests the following equilibrium concept.

Definition Given a moral code F and a game (X, u), a strategy profile m is a *fair play equilibrium* if and only if, for each player i, m_i is a most preferred

strategy in $F_i(X, u, x_i)$ for v_i.

The set of fair play equilibria is denoted $FPE(X, u, F)$. Thus $FPE(X, u, F)$ consists of $m \in M$ such that for all $i \in N$, $m_i \in F_i(X, u, m_{-i})$ and $v_i(m_i, m_{-i}) \geq v_i(m'_i, m_{-i})$ for all $m'_i \in F_i(X, u, m_{-i})$.

At a fair play equilibrium, a player may have better replies, but none of which is socially acceptable. It is also noted that different moral codes may generate different fair play equilibria in the same game.

3 Axioms

We are interested in characterizing fair play equilibria when the moral code satisfies the following axioms.

The first axiom is *anonymity*, which says that the name of the players should not matter. Let a permutation be $\pi : N \longrightarrow N$. Given an action profile x, 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. Similarly, given a mixed strategy profile m, we define m^π by $m^\pi_{\pi(i)} := m_i$.

Definition A moral code F satisfies *anonymity* if for all $(X, u), (X', u') \in \Gamma$ and all permutations $\pi : N \longrightarrow N$, if

$$X'_{\pi(i)} = X_i \text{ and } u'_{\pi(i)}(x^\pi) = u_i(x) \ \forall i \in N, \forall x \in X \tag{1}$$

then for all $i \in N$ and all $m \in M$, $F_i(X, u, m_{-i}) = F_{\pi(i)}(X', u', m^\pi_{-\pi(i)})$.

It can be easily verified that (1) implies $v'_{\pi(i)}(m^\pi) = v_i(m)$ for all $m \in M$. Thus (X', u') is the game generated from (X, u) by renaming player i as $\pi(i)$.

For example, consider the following pair of games:

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

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

These games are identical except that the players are interchanged; player $i = 1, 2$ in the left game is identical to player $j \neq i$ in the right game. Suppose that, in the left game, a mixed strategy $(A, \frac{1}{3}, B, \frac{2}{3})$ of player 1 is a socially acceptable response to 2's strategy $(a, \frac{1}{6}, b, \frac{1}{3}, c, \frac{1}{2})$. Then anonymity says that, in the right game, the same strategy $(A, \frac{1}{3}, B, \frac{2}{3})$, now defined for player 2, is a socially acceptable response to player 1's strategy $(a, \frac{1}{6}, b, \frac{1}{3}, c, \frac{1}{2})$.

To state the next axiom, we first introduce a definition. We write $m_i \simeq m_i'$ if these strategies are identical in term of welfare for all players, regardless of the strategies of $j \neq i$. Formally,

Definition Given $(X, u) \in \Gamma$ and $i \in N$, strategies $m_i \in M_i$ and $m_i' \in M_i$ are *welfare-equivalent*, denoted $m_i \simeq m_i'$, if $v_j(m_i, m_{-i}) = v_j(m_i', m_{-i})$ for all $j \in N$ and all $m_{-i} \in M_{-i}$.

The next axiom says that welfare-equivalent strategies should be treated equally.

Definition A moral code F satisfies *welfare nondiscrimination* if for all $(X, u) \in \Gamma$, the following conditions are satisfied.

1. For all $m, m' \in M$ such that $m_i \simeq m_i'$ for all $i \in N$ (possibly $m_i = m_i'$ for some i), then for all $i \in N$, $m_i \in F_i(X, u, m_{-i}) \iff m_i' \in F_i(X, u, m_{-i}')$.

2. For all $m \in M$, all $i \in N$, and all $y_i \in X_i$ such that $m_i \simeq y_i$ and $m_i(y_i) = 0$,

2.1. $F_i(X_i \backslash \{y_i\} \times X_{-i}, u, m_{-i}) = F_i(X, u, m_{-i}) \backslash M_i(y_i)$, where $M_i(y_i) = \{m_i \in M_i : m_i(y_i) > 0\}$,and

2.2. $F_j(X_i \backslash \{y_i\} \times X_{-i}, u, m_{-j}) = F_j(X, u, m_{-j})$ for all $j \neq i$.

Condition 1 says that the moral code does not distinguish welfare-equivalent strategies. Condition 2 considers a case where there is a pure strategy y_i that has a welfare-equivalent substitute strategy m_i whose support does not include y_i. In this case, y_i is redundant in the game in the sense that for any mixed strategy whose support includes y_i, there is a welfare-equivalent substitute whose support does not include y_i. Condition 2 says that, in this case, deleting y_i from the game makes no essential difference on the set of fair plays for any player. For player i, the set of fair plays is reduced by simply removing those strategies whose support includes y_i. For players $j \neq i$, the set of fair plays does not change.

By applying condition 1, condition 2.2 is generalized as follows:

2.2'. $F_j(X_i \backslash \{y_i\} \times X_{-i}, u, m_{-j}) = F_j(X, u, m_{-\{i,j\}}, m'_i)$ for all $j \neq i$ and all $m'_i \simeq m_i$ in (X, u).

As an illustration, consider the following pair of games.

		Player 2	
		a	b
Player 1	A	$6, 0$	$1, 6$
	B	$2, 8$	$3, 4$
	C	$4, 4$	$2, 5$

		Player 2	
		a	b
Player 1	A	$6, 0$	$1, 6$
	B	$2, 8$	$3, 4$

(2)

It follows from Lemma 1 (see Section 9) that $(A, \frac{1}{2}, B, \frac{1}{2})$ and C are welfare-equivalent. Welfare nondiscrimination requires that $(A, \frac{1}{2}, B, \frac{1}{2})$ is socially acceptable if and only if C is (condition 1 when $i = 1$ and $m_{-i} = m'_{-i}$). Welfare nondiscrimination also says that whether player 1 plays $(A, \frac{1}{2}, B, \frac{1}{2})$ or C does not affect socially acceptable strategies for player 2 (condition 1 when $i = 2$ and $m_i = m'_i$).

The game on the right is obtained from the left by deleting C. Since C is a replica of $(A, \frac{1}{2}, B, \frac{1}{2})$, there is a sense in which these games are identical. This is why welfare nondiscrimination (condition 2) also requires that at

any strategy profile that exists in both games, a player's strategy is socially acceptable in the left game if and only if it is socially acceptable in the right game. For example, $(A, \frac{1}{2}, B, \frac{1}{2})$ is a socially acceptable response to $(a, \frac{1}{2}, b, \frac{1}{2})$ 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 individuals' welfare and therefore alternatives should not be distinguished if they have identical welfare consequences. There are two separate issues: how to treat welfare-equivalent strategies (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 moral codes that treat two games differently only because they have different numbers of actions.

The next axiom, independence, says that whether player i's strategy is a socially acceptable response to m_{-i} is independent of preferences over strategy profiles where m_{-i} is not played. Let a game $(X, u) \in \Gamma$ be given. For each player i, the preference $\succcurlyeq_{v_i}$ induced from i's utility function v_i is a binary relation on M given by

$$m \succcurlyeq_{v_i} m' \Longleftrightarrow v_i(m) \geq v_i(m') \ \forall m, m' \in M.$$

Definition A moral code F satisfies *independence* if for all $(X, u), (X, u') \in \Gamma$, all $i \in N$, and all $m_{-i} \in M_{-i}$, if for all $j \in N$, $\succcurlyeq_{v_j}$ and $\succcurlyeq_{v'_j}$ are identical on $\{(m_i, m_{-i}) : m_i \in M_i\}$, then $F_i(X, u, m_{-i}) = F_i(X, u', m_{-i})$.

This is a natural requirement since what a moral code prescribes to a player is conditional on other players' behavior. By definition, $F_i(X, u, m_{-i})$ is relevant only if other players play m_{-i}. Given m_{-i}, strategy profiles in $M \backslash \{(m_i, m_{-i}) : m_i \in M_i\}$ are counter-factual and hence deemed immaterial.

For this axiom, it is critical that what a moral code decides is whether one's strategy is a socially acceptable *response* to others' behavior. Thus a

judgement that a player's strategy is socially unacceptable 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.

A weaker version of independence is defined by replacing "$\succcurlyeq_{v_j}$ and $\succcurlyeq_{v'_j}$" with "v_j and v'_j" and thus restricting its scope to the case where cardinal utilities are also identical over the relevant set.

Definition A moral code F satisfies *weak independence* if for all $(X,u),(X,u') \in \Gamma$, all $i \in N$, and all $m_{-i} \in M_{-i}$, if for all $j \in N$, v_j and v'_j are identical on $\{(m_i, m_{-i}) : m_i \in M_i\}$, then $F_i(X,u,m_{-i}) = F_i(X,u',m_{-i})$.

As we show in the next section (Proposition 3), the two versions are equivalent under the next axiom.

Definition A moral code F satisfies *invariance to equivalent utility representations* if for all $(X,u),(X,u') \in \Gamma$, if for all $j \in N$, there are some real values, $\alpha_j > 0$ and β_j, such that $u'_j(x) = \alpha_j u_j(x) + \beta_j$ for all $x \in X$, then for all $i \in N$, and all $m_{-i} \in M_{-i}$, $F_i(X,u,m_{-i}) = F_i(X,u',m_{-i})$.

Obviously we have $v'_j(m) = \alpha_j v_j(m) + \beta_j$ for all $m \in M$. This axiom says that a judgement that a player's strategy is socially acceptable is not influenced by any positive affine transformations of utility functions.

The next axiom, monotonicity, says that if a player's strategy is socially acceptable at a strategy profile m under a utility profile u, it remains so under a utility profile u' if the change from u to u' only decreases utilities at strategy profiles $m' \neq m$.

Definition *A* moral code F satisfies *monotonicity* if for all $(X,u),(X,u') \in \Gamma$, all $m \in M$, and all $i \in N$, if $m_i \in F_i(X,u,m_{-i})$ and

$$v'_j(m) = v_j(m) \text{ and } v'_j(m') \leq v_j(m') \quad \forall m' \neq m, \forall j \in N, \tag{3}$$

then $m_i \in F_i(X,u',m_{-i})$.

Remark 1 From the perspective of social choice theory, it might be more natural to replace (3) with the following:

$$v_j(m) \geq v_j(m') \Longrightarrow v'_j(m) \geq v'_j(m'); \tag{4}$$

$$v_j(m) > v_j(m') \Longrightarrow v'_j(m) > v'_j(m') \tag{5}$$

The condition (4)-(5) says that the change from u to u' only raises the relative ranking of m in individuals' preferences. The monotonicity axiom with (3) being replaced with (4)–(5) is a natural translation of Arrow's condition of positive association (Arrow, 1963). Since (4)–(5) is weaker than (3), the axiom of monotonicity is stronger with (4)–(5) than with (3). Our main results remain valid with the stronger version of monotonicity.

The next axiom, effectiveness, says that, for any game, there exists at least one pure strategy profile where all players play fair.

Definition A moral code F satisfies *effectiveness* if for all $(X, u) \in \Gamma$, there exists $x \in X$ such that

$$x_i \in F_i(X, u, x_{-i}) \quad \forall i \in N. \tag{6}$$

Thus a violation of effectiveness means that there exists a game where there is no pure strategy profile where all players follow the moral code. The axiom represents a basic requirement that a moral code's prescriptions to different players should be compatible. An action profile that satisfies (6) is called a *fair play profile*. Effectiveness is weaker than demanding the existence of a pure-strategy fair play *equilibrium*, since (6) is only a necessary condition for x to be a pure-strategy fair play equilibrium. If there exists no pure-strategy fair play profile in a game $(X, u) \in \Gamma$, then mixed-strategy profiles $m \in M$ such that $m_i \in F_i(X, u, m_{-i})$ for all $i \in N$, even if they exist, are not grounded on moral rightness, because none of the pure-strategy

profiles x with $\prod_{i \in N} m_i(x_i) > 0$ is a fair play profile; m never leads players to an action profile at which they all take socially acceptable actions.

The last axiom is continuity. Given a set of action profiles X, let $U(X)$ denote the set of all utility functions defined on X. 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 $\mathbb{R}^{n|X|}$.

Definition A moral code F satisfies *continuity* if for all $(X, u) \in \Gamma$, all $m \in M$, all $i \in N$, and all sequences $(u^\nu)_{\nu=1}^\infty$ from $U(X)^N$, if $m_i \in F_i(X, u^\nu, m_{-i})$ for all ν and $u^\nu \to u$ as $\nu \to \infty$, then $m_i \in F_i(X, u, m_{-i})$.

This is the upper-hemi continuity of the correspondence $F_i(X, u, m_{-i})$ with respect to u. It says that for any strategy m_i, the set of utility-function profiles for which the strategy is socially acceptable is a closed set. Thus, at the boundary of the area where m_i is socially acceptable, the moral code takes the permissive side.

There exist moral codes satisfying all the axioms: see Examples 1 and 2 in Section 6.

4 Preliminary Results

We state here four preliminary results concerning the axioms. The first two propositions, Propositions 1 and 2, are concerned with welfare nondiscrimination. 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 moral code F is *neutral* if for all (X, u), $(X', u') \in \Gamma$, if there exists a bijection $\rho_i \colon X_i \to X'_i$ for all $i \in N$ such that $u'_i(\rho_1(x_1), \ldots, \rho_n(x_n)) =$

$u_i(x)$ for all $x \in X$,[8] then for all $m \in M$ and all $i \in N$,

$$m_i \in F_i(X, u, m_{-i}) \iff \rho_i(m_i) \in F_i(X', u', \rho(m)_{-i}),$$

where $\rho_i(m_i)$ is a mixed strategy that plays $\rho_i(x_i) \in X_i'$ with probability $m_i(x_i)$ for each x_i, i.e., the composition of $\rho_i^{-1}\colon X_i' \to X_i$ and $m_i\colon X_i \to [0,1]$; and $\rho(m)_{-i}$ denotes the profile of $\rho_j(m_j)$ for all $j \neq i$.

Neutrality is strictly weaker than welfare nondiscrimination because neutrality does not relate games with different numbers of actions: neutrality allows the pair of games in (2) to receive totally different treatments.

Proposition 1 *Welfare nondiscrimination implies neutrality.*

Proof. Let F be a moral code satisfying welfare nondiscrimination, and let $(X, u), (X', u') \in \Gamma$ 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 $\rho_i(x_i) = x_i$.[9] We further assume $X_1 \cap X_1' = \emptyset$. The case of $X_1 \cap X_1' \neq \emptyset$ can be proved by repeating the argument that follows.[10]

Now we start with (X, u) 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 $((X_1 \cup X_1') \times X_{-1}, u'')$. Welfare nondiscrimination implies that for all $m \in M$,

$$\begin{aligned} m_1 \in F_1(X, u, m_{-1}) &\iff m_1 \in F_1((X_1 \cup X_1') \times X_{-1}, u'', m_{-1}) \\ &\iff \rho(m_1) \in F_1((X_1 \cup X_1') \times X_{-1}, u'', m_{-1}) \\ &\iff \rho(m_1) \in F_1(X', u', m_{-1}). \end{aligned}$$

For all $i \neq 1$,

[8] While this is stated for pure-strategy profiles, it implies $v_i'(\rho_1(m_1), \ldots, \rho_n(m_n)) = v_i(m)$ for all $m \in M$.

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

[10] Specifically, we choose any set $Y \subset \Omega$ such that $|Y| = |X_1| = |X_2|$, $Y \cap X_1 = \emptyset$, and $Y \cap X_2 = \emptyset$, and then apply our argument first to the pair (X_1, Y) and (Y, X_2).

$$\begin{aligned} F_i(X,u,(m_1,m_{N\setminus\{1,i\}})) &= F_i((X_1\cup X_1')\times X_{-1},u'',(m_1,m_{N\setminus\{1,i\}})) \\ &= F_i((X_1\cup X_1')\times X_{-1},u'',(\rho_1(m_1),m_{N\setminus\{1,i\}})) \\ &= F_i(X',u',(\rho_1(m_1)),m_{N\setminus\{1,i\}})). \ \blacksquare \end{aligned}$$

Proposition 2 *Let $G=(X,u)\in\Gamma$ and $i\in N$ be given. Suppose that there exist $m_i^*\in M_i$ and $y_i\in X_i$ such that $m_i^*\simeq y_i$ and $m_i^*(y_i)=0$. Then for all $m_i\in M_i$, there exists $m_i'\in M_i$ such that $m_i'(y_i)=0$ and $m_i'\simeq m_i$.*

Proof. Let m_i' be defined by $m_i'(y_i)=0$ and

$m_i'(x_i)=m_i(x_i)+m_i^*(x_i)m_i(y_i)\ \forall x_i\neq y_i$.

Then for all $m_{-i}\in M_{-i}$ and all j,

$$\begin{aligned} v_j(m_i',m_{-i}) &= \sum_{x_i\neq y_i}\left[m_i(x_i)+m_i^*(x_i)m_i(y_i)\right]v_j(x_i,m_{-i}) \\ &= \sum_{x_i\neq y_i} m_i(x_i)v_j(x_i,m_{-i})+m_i(y_i)v_j(m_i^*,m_{-i}) \\ &= \sum_{x_i\neq y_i} m_i(x_i)v_j(x_i,m_{-i})+m_i(y_i)v_j(y_i,m_{-i})=v_j(m_i,m_{-i}). \ \blacksquare \end{aligned}$$

The proposition says that for every player, deleting an action y_i that is welfare-equivalent to a strategy whose support does not include y_i does not make any change as far as we are only concerned with utility levels players enjoy, although it changes the set of strategies. This proposition makes welfare-nondiscrimination more appealing.

The next two propositions, Propositions 3 and 4, are concerned with logical relationship between independence, weak independence, and invariance to equivalent utility representations.

Proposition 3 *For any $(X,u),(X,u')\in\Gamma$, any $i,j\in N$, and any $m_{-i}\in M_{-i}$, if $\succcurlyeq_{v_j}$ and $\succcurlyeq_{v_j'}$ are identical on $\{(m_i,m_{-i}):m_i\in M_i\}$, then there are some constants $\alpha_j>0$ and β_j such that for all $m_i\in M_i$, $v_j'(m_i,m_{-i})=\alpha_j v_j(m_i,m_{-i})+\beta_j$.*

Proof. This follows from Mas-Colell et. al.'s Proposition 6.B.2 (p173) applied to $\{(m_i,m_{-i}):m_i\in M_i\}$ $\blacksquare$

Proposition 4 *A moral code satisfies independence if and only if it satisfies weak independence and invariance to equivalent utility representations.*

Proof. The "only if" part is obvious, so we only prove the "if" part. Let $(X,u),(X,u') \in \Gamma, i \in N$, and $m_{-i} \in M_{-i}$ be as in the assumption in the definition of independence. By Proposition 3, for all $j \in N$, there exist $\alpha_j > 0$ and β_j such that for all $m_i \in M_i$, $v'_j(m_i, m_{-i}) = \alpha_j v_j(m_i, m_{-i}) + \beta_j$. For all $j \in N$, define v''_j by $v''_j(m) = \alpha_j v_j(m) + \beta_j$ for all $m \in M$. Then invariance to equivalent utility representations and weak independence imply

$$F_i(X,u,m_{-i}) \overset{\text{invari.}}{=} F_i(X,u'',m_{-i}) \overset{\text{weak.I.}}{=} F_i(X,u',m_{-i}). \quad \blacksquare$$

5 Main Results

We have three main results. The first one states that if a moral code satisfies all the axioms defined in Section 3, except for continuity, then for any player, there always exists a socially acceptable strategy that is also an unconstrained best reply. That is, while the moral code may prohibit a player from choosing some of his best replies, it never prohibits all. This result immediately implies that under the moral code, all fair play equilibria are necessarily Nash equilibria, and conversely, all strict Nash equilibria are necessarily fair play equilibria. Formally, let $NE(X,u)$ and $SNE(X,u)$ denote the set of Nash equilibria and strict Nash equilibria, respectively. Then

Theorem 1 *If a moral code F satisfies anonymity, welfare nondiscrimination, weak independence*[11]*, monotonicity, and effectiveness, then for all $(X,u) \in \Gamma$, all $i \in N$ and all $m_{-i} \in M_{-i}$,*

$$F_i(X,u,m_{-i}) \cap BR_i(X,u,m_{-i}) \neq \phi. \tag{7}$$

[11] By Proposition 4, independence can be replaced with weak independence and invariance to equivalent utility representations.

This implies that for all $(X, u) \in \Gamma$*,*

$$SNE(X, u) \subset FPE(X, u, F) \subset NE(X, u). \tag{8}$$

(7) indeed implies the first inclusion in (8) since at any strict Nash equilibrium, a best reply is unique and hence must be socially acceptable. To see that the second inclusion in (8) is also implied by (7), 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 acceptable.

If a moral code satisfies not only the axioms of Theorem 1 but also continuity, the set of fair play equilibria coincides with the set of Nash equilibria, which is our main result.

Theorem 2 *If a moral code* F *satisfies continuity and all the axioms in Theorem 1, then for all* $(X, u) \in \Gamma$*,*

$$FPE(X, u, F) = NE(X, u).$$

Once Theorem 1 is proved, the proof of Theorem 2 is not difficult. The rough sketch of the proof is as follows: Let $m \in NE(X, u)$. It is possible to construct a sequence of games (X, u^ν) such that for every i, m_i is a unique best reply to m_{-i}, and $(X, u^\nu) \longrightarrow (X, u)$. Then we have $m \in SNE(X, u^\nu)$ for all ν and hence Theorem 1 means $m \in FPE(X, u^\nu, F)$ for all ν. Continuity then yields $m \in FPE(X, u, F)$, which shows $NE(X, u) \subset FPE(X, u, F)$. See Section 9 for the detail.

Since the proof of Theorem 1 is long and involved, we here give a proof for the case of two players with two actions; the general proof is given in Section 9. Suppose that there are two players (1 and 2) and consider a game (X, u) where the action set is $X_1 = \{A, B\}$ and $X_2 = \{a, b\}$. The proof of (7) goes on with three steps.

Step 1: (7) holds when m_{-i} consists of pure strategies.

By noting anonymity and neutrality, it suffices to show $F_1(X,u,a) \cap BR_1(X,u,a) \neq \emptyset$. Suppose on the contrary that

$$F_1(X,u,a) \cap BR_1(X,u,a) = \emptyset \tag{9}$$

This implies $BR_1(X,u,a)$ is $\{A\}$ or $\{B\}$ (otherwise, $BR_1(X,u,a) = M_1$). Without loss of generality, assume that it is $\{A\}$, i.e., $u_1(A,a) > u_1(B,a)$. Then (9) implies that

$$A \notin F_1(X,u,a) \tag{10}$$

That is, for all $m_1 \in F_1(X,u,a), m_1(A) < 1$.

By invariance to equivalent utility representations, we can normalize each player's payoffs by setting $u_1(B,a) = 0$ and $u_2(A,a) = 0$. By independence, the values $u_i(\cdot, b)$ do not affect (9). Altogether, we can assume that the payoff matrix is given by

	a	b
A	$1,0$	$0,y$
B	$0,y$	$1,0$

(11)

where $y \in \mathbb{R}$. There are two cases.

Case 1: $y \leq 0$. Then, let us replace y with $y' > 0$ and denote the resulting utility profile by u'. Then (10) remains true under u': i.e., $A \notin F_1(X,u',a)$. By way of contradiction, suppose that $A \in F_1(X,u',a)$. Since the change from u' to u only decreases payoffs at $m \neq (A,a)$, monotonicity implies $A \in F_1(X,u,a)$, a contradiction, which shows that $A \notin F_1(X,u',a)$. Since $y' > 0$, the result shows that it suffices to consider the next case.

Case 2: $y > 0$. By invariance to equivalent utility representations, we can set $y = 1$. Since $A \notin F_1(X,u,a)$, (A,a) is not a fair play profile. By

neutrality, (B, b) is not a fair play profile either. By anonymity, neither (A, b) nor (B, a) is a fair play profile. Thus there exists no action profile that is a fair play profile, which is a contradiction with effectiveness.

Step 2: The result of Step 1 extends to games where an action $c \notin \{a, b\}$ is added to X_2.

Consider a game $(\widetilde{X}, \widetilde{u}) \in \Gamma$ with $\widetilde{X} = X_1 \times (X_2 \cup \{c\})$. We show that $F_1(\widetilde{X}, \widetilde{u}, a) \cap BR_1(\widetilde{X}, \widetilde{u}_1, a) \neq \emptyset$.

Let u' be a utility profile that is identical to $\widetilde{u}$ on X and such that $u'_i(x_1, c) = \widetilde{u}_i(x_1, a)$ for all x_1 and all i. Under u', we have $c \simeq a$. Independence and welfare nondiscrimination imply that

$$F_1(\widetilde{X}, \widetilde{u}, a) \overset{\text{Indep.}}{=} F_1(\widetilde{X}, u', a) \overset{\text{welfare.non.}}{=} F_1(X, u'|_X, a).$$

By Step 1, $F_1(X, u'|_X, a) \cap BR_1(X, u'_1|_X, a) \neq \emptyset$. Since $BR_1(X, u'_1|_X, a) = BR_1(\widetilde{X}, \widetilde{u}_1, a)$, we obtain $F_1(\widetilde{X}, \widetilde{u}, a) \cap BR_1(\widetilde{X}, \widetilde{u}_1, a) \neq \emptyset$, as desired.

Step 3: (7) holds when m_{-i} contains non-pure strategies.

Given a non-pure strategy m_2, we show that $F_1(X, u, m_2) \cap BR_1(X, u_1, m_2) \neq \emptyset$. Take an action $c \notin \{a, b\}$ and define a game (X', u') by $X' = X_1 \times (X_2 \cup \{c\})$ and, for all $(x_1, x_2) \in X'$ and for all $i = 1, 2$,

$$u'_i(x_1, x_2) = \begin{cases} u_i(x_1, x_2) & \text{if} \quad x_2 \neq c, \\ v_i(x_1, m_2) & \text{if} \quad x_2 = c. \end{cases}$$

For this game, $c \simeq m_2$ (see Lemma 1). Step 2 implies $F_1(X', u', c) \cap BR_1(X', u'_1, c) \neq \emptyset$. This together with welfare nondiscrimination implies $F_1(X', u', m_2) \cap BR_1(X', u'_1, m_2) \neq \emptyset$. Deleting c and invoking welfare nondiscrimination, we obtain $F_1(X, u, m_2) \cap BR_1(X, u_1, m_2) \neq \emptyset$.

The general proof in Section 9, in particular that of Lemma 3, is considerably more complicated. One reason for the complication is that the violation of (7) may be at an action profile where player i has multiple best replies, and other players may have complex preferences over i's best replies. This complicates the proof since we then need to maintain the same preference structure along a cycle that shows the non-existence of fair play profile. An-

other source of complication is that we need to change every player's action to complete a cycle.

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 moral code in which each player is a dictator for his own socially acceptable actions given the other players' actions. We could easily construct more "democratic" procedures to determine socially acceptable actions, in such a way that there may not exist an unconditional best response that is also socially acceptable (e.g., Example 7 in section 6). However, as 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 $F_i(X, u, m_{-i})$ as the player's rights. A critical difference is that none of our axioms is about liberty. Each of our axioms permits $F_i(X, u, m_{-i})$ to be always a singleton, in which case the moral code gives no freedom.

The cycle in the game (11) captures situations in which everyone ought to give an advantage to others. Such situations are often observed when there is an unpleasant job to be done by someone and everyone wants to be nice: everyone insists on taking the job. We also often observe situations where everyone insists on shouldering responsibility, paying a bill, going after others, etc.

Our previous paper (Miyagawa, Nagahisa, and Suga, 2005) proves Theorems 1 and 2 in the case where only pure strategies are available to players. Their results can be derived from our present proof. In particular they are

from direct consequences of Lemma 3 in Section 9.

We turn to the third result concerning a welfare property of fair play equilibrium.

Definition Let a game (X, u) be given. A profile $m \in M$ is *locally weak Pareto efficient* if for all $i \in N$, m is weak Pareto efficient on $\{(m_i', m_{-i}) : m_i' \in M_i\}$.

It directly follows from Theorem 1 that a fair play equilibrium is locally weak Pareto efficient. This still holds if we relax the assumptions.

Theorem 3 *If a code F satisfies independence, monotonicity and welfare nondiscrimination, then any fair play equilibrium is locally weak Pareto efficient.*

Proof. It suffices to show that

For any $(X, u) \in \Gamma$, any $i, j \in N$, and any $m_i, m_i' \in M_i$,

if $m_i \in F_i(X, u, m_{-i})$ and $v_j(m_i', m_{-i}) > v_j(m_i, m_{-i})$ for all $j \in N$

then $m_i' \in F_i(X, u, m_{-i})$.

Let us take $a_k \notin X_k$ for every $k \in N$ arbitrarily. Lemma 2 assures a game $(X^{(1)}, u^{(1)}) \in \Lambda$ such that

$X_k^{(1)} = X_k \cup \{a_k\}$ and $a_k \simeq m_k$ for all $k \in N$; and

$u^{(1)}|_X = u$.

Since $m_i \in F_i(X, u, m_{-i})$, welfare nondiscrimination implies

$$a_i \in F_i(X^{(1)}, u^{(1)}, a_{-i}). \tag{12}$$

Next we define a game $(X^{(1)}, u^{(2)})$ such that for any $k \in N$ and for any $x \in X^{(1)}$,

$$u_k^{(2)}(x) = \begin{cases} v_k^{(1)}(m_i', m_{-i}) & \text{if } x = (a_k)_{k \in N} \\ u_k^{(1)}(x) & \text{otherwise} \end{cases}$$

By (12) and monotonicity, we have

$$a_i \in F_i(X^{(1)}, u^{(2)}, a_{-i}). \tag{13}$$

Lemma 2 assures a game $(X^{(1)}, u^{(3)}) \in \Lambda$ such that
$a_j \simeq m_j$ for all $j \neq i$ and $a_i \simeq m'_i$;
$u^{(3)}|_X = u$.

A simple computation leads us to that for all $k \in N$, $v^{(2)}$ and $v^{(3)}$ induce the same utility function on the set $\{(m''_i, a_{-i}) : m''_i \in M_i^{(2)} (= M_i^{(3)})\}$.

Thus (13) and independence imply

$$a_i \in F_i(X^{(1)}, u^{(3)}, a_{-i}). \tag{14}$$

Welfare nondiscrimination is applied to (14), we have

$$m'_i \in F_i(X^{(1)}, u^{(3)}, m_{-i}). \tag{15}$$

By deleting all a_k, the game $(X^{(1)}, u^{(3)})$ reduces to the original (X, u). (15) and welfare nondiscrimination imply $m'_i \in F_i(X, u, m_{-i})$, the desired result. ■

6 Examples

This section gives a few examples of moral codes and games. The following two moral codes satisfy all the axioms in Theorem 2.

Example 1 (Amoral code)

All strategies are always socially acceptable: $F_i(X, u, m_{-i}) := M_i$.

Example 2 (Local weak Pareto code)

For all $(X, u) \in \Gamma$, all $i \in N$ and all $m_{-i} \in M_{-i}$, let $F_i(X, u, m_{-i})$ be the set of strategies $m_i \in M_i$ such that there exists no $m'_i \in M_i$ such

that $v_j(m_i', m_{-i}) > v_j(m_i, m_{-i})$ for all $j \in N$. In other words, a strategy is judged as socially unacceptable if there exists a strategy that makes all players strictly better off. This code is "local" in the sense that m_{-i} is held fixed when Pareto efficiency is invoked. One can easily verify that this code also satisfies all the axioms.

Although these two codes look very different, since they both satisfy all the axioms in Theorem 2, they satisfy $FPE(X, u, F) = NE(X, u)$ for all games.

For these codes, it is possible that $\emptyset \neq SNE(X, u) \subsetneq FPE(X, u, F)$ for some game. For example, consider the following game:

	a	b
A	$2, 2$	$1, 1$
B	$1, 1$	$1, 1$

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

The converse of Theorem 2 is false. That is, there exists a moral code that satisfies $FPE(X, u, F) = NE(X, u)$ for all games but does not satisfy all the axioms of Theorem 2. A simple example of proving this is the following code.

Example 3 (Best Response code)

For all $(X, u) \in \Gamma$, all $i \in N$ and all $m_{-i} \in M_{-i}$, $F_i(X, u, m_{-i}) := BR_i(X, u, m_{-i})$.

Under this code, players are allowed to play any best reply and hence fair play equilibrium is equivalent to Nash equilibrium. This code, however, violates effectiveness since a pure-strategy Nash equilibrium may not exist.

If a moral code satisfies all the axioms of Theorem 2, then a fair play equilibrium exists because Theorem 2 implies that the existence of fair play equilibrium is reduced to the existence of Nash equilibrium. The same cannot be said for moral codes that satisfy only the axioms of Theorem 1. As the

next example shows, if a moral code lacks continuity, fair play equilibria may not exist.

Example 4 (Local strong Pareto code)

For all $(X, u) \in \Gamma$, all $i \in N$, and all $m \in M$, $m_i \in F_i(X, u, m_{-i})$ if and only if there exists no $m_i' \in M_i$ such that $v_j(m_i', m_{-i}) \geq v_j(m_i, m_{-i})$ for all $j \in N$ and $v_j(m_i', m_{-i}) > v_j(m_i, m_{-i})$ for some $j \in N$.

This code satisfies all the axioms except for continuity. Under this code, a fair play equilibrium does not exist for some game. An example is the following game.

$$\begin{array}{c|c|c} & a & b \\ \hline A & 1,0 & 2,0 \\ \hline B & 0,2 & 2,1 \end{array} \tag{16}$$

The set of fair plays are

$$F_1(X, u, m_{-1}) = \begin{cases} \{B\} & \text{if } m_2(b) = 1 \\ M_1 & \text{otherwise} \end{cases}$$

$$F_2(X, u, m_{-2}) = \begin{cases} \{b\} & \text{if } m_1(A) = 1 \\ M_2 & \text{otherwise.} \end{cases}$$

A player can choose any strategy in all but one case: Player 1 is constrained to B if player 2 chooses b, and player 2 is constrained to b if player 1 chooses A. This game has no fair play equilibrium. (Proof: If player 2 chooses b with probability one, player 1 chooses B, but then player 2 is free to choose anything and thus chooses a, a contradiction. If player 2 does not choose b with probability one, player 1 is free to choose anything and thus chooses A, which constrains player 2 to b, which is a contradiction.) On the other hand, as long as player 1 chooses A, a Nash equilibrium results independently of player 2's choice.

Theorem 1 holds true for this code, but it contains the trivial case where $SNE(X, u) \subset FPE(X, u, F) = \emptyset \subset NE(X, u)$ as is the case for (16). The

game below shows a case of $\emptyset \neq FPE(X, u, F) \subsetneq NE(X, u)$ for this code.

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

In this game, (A, a) is a fair play equilibrium and hence $FPE(X, u, F) \neq \emptyset$. (A, a) is also a Nash equilibrium, and so is (B, b). However, (B, b) is not a fair play equilibrium, since it is Pareto dominated by (B, a) and (A, b).

The remaining examples show that none of the axioms in Theorems 1 and 2 is redundant. We show that if any of the axioms is removed, there exists a moral code that satisfies the remaining axioms but violates the inclusive relation of Theorem 1 (and hence Theorem 2).

Example 5 (Anonymity: Dictatorial code)

There exists a player $k \in N$ such that for all $(X, u) \in \Gamma$, all $i \in N$, and all $m_{-i} \in M_{-i}$,

$F_i(X, u, m_{-i}) := \{m_i \in M_i : v_k(m_i, m_{-i}) \geq v_k(m'_i, m_{-i}) \text{ for all } m'_i \in M_i\}$.

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. For this code, the conclusion of Theorem 1 (and hence that of Theorem 2) is false. Indeed, if there exists a unique action profile $x \in X$ that maximizes the dictator's payoff $u_k(x)$, it is a fair play equilibrium under this code but it may not be a Nash equilibrium.

Example 6 (Weak independence: Global weak Pareto code)

Let $WP(X, u) \subset M$ denote the set of weakly Pareto efficient strategy profiles in the game (X, u). Define a moral code F by

$$F_i(X, u, m_{-i}) := \{m_i \in M_i : \exists m'_{-i} \in M_{-i} s.t. (m_i, m'_{-i}) \in WP(X, u)\}.$$

Thus, for a strategy to be acceptable, it suffices that the strategy yields a weakly (and globally) Pareto efficient outcome if the strategy is combined with some (not necessarily actual) strategy profile of the other players. This code satisfies all the axioms except for independence. Independence is violated since $F_i(X, u, m_{-i})$ depends on the utility values $v_j(m')$ for strategy profiles $m'_{-i} \neq m_{-i}$. To see that this code satisfies continuity, consider a sequence $\{u^\nu\}$ such that $u^\nu \to u$ and suppose that $m_i \in F_i(X, u^\nu, m_{-i})$ for all ν. The definition of F implies that, for all ν, there exists $m^\nu_{-i} \in M_{-i}$ such that $(m_i, m^\nu_{-i}) \in WP(X, u^\nu)$. Since M_{-i} is compact, we can let $m^\nu_{-i} \to m^0_{-i}$ without loss of generality. Then $(m_i, m^\nu_{-i}) \in WP(X, u^\nu)$ for all ν implies $(m_i, m^0_{-i}) \in WP(X, u)$[12], i.e., $m_i \in F_i(X, u, m_{-i})$, as desired.

To see that Theorem 1 does not hold without weak independence, consider the following game.

To see that Theorem 1 does not hold without independence, consider the following game.

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

Then (a, a) is a fair play equilibrium under this code; b is not a fair play since it is not a part of any Pareto efficient outcome. Since (a, a) is not a Nash equilibrium, we have $FPE(X, u, F) \not\subset NE(X, u)$. On the other hand, (b, b) is a strict Nash equilibrium but not a fair play equilibrium, which shows that $SNE(X, u) \not\subset FPE(X, u, F)$.

Example 7 (Monotonicity: Anti-Pareto code)

[12]If not, then there exists $m' \in M$ such that $v_j(m') > v_j(m_i, m^0_{-i})$ for all j. Since $m^\nu_{-i} \to m^0_{-i}$ and $u^\nu \to u$, there exists ν' such that for all $\nu \geq \nu'$, $v^\nu_j(m') > v^\nu_j(m_i, m^\nu_{-i})$ for all j, which is a contradiction since $(m_i, m^\nu_{-i}) \in WP(X, u^\nu)$.

The code first reverses each player's preference ordering and then applies the local weak Pareto code (Example 2). Formally, given $(X, u) \in \Gamma$, $i \in N$, $m_{-i} \in M_{-i}$, $m_i \in F_i(X, u, m_{-i})$ if and only if there is no $m_i' \in M_i$ such that $-v_j(m_i', m_{-i}) > -v_j(m_i, m_{-i})$ for all $j \in N$.

This code satisfies all the axioms except for monotonicity. For this code, the conclusion of Theorem 1 (and hence that of Theorem 2) does not hold. Indeed, if a game has two action profiles $x, y \in X$ such that $u_i(x) > u_i(z) > u_i(y)$ for all $z \in X \setminus \{x, y\}$ and all i, then x is a strict Nash equilibrium but not a fair play equilibrium, whereas y is a fair play equilibrium but not a Nash equilibrium.

Example 8 (Effectiveness: Altruistic code)

Under this code, a strategy is acceptable if it maximizes someone else's utility: given $(X, u) \in \Gamma$, $i \in N$, and $m_{-i} \in M_{-i}$,

$$F_i(X, u, m_{-i}) = \bigcup_{j \neq i} \{m_i \in M_i : v_j(m_i, m_{-i}) \geq v_j(m_i', m_{-i}) \ \forall m_i' \in M_i\}.$$

This code satisfies all the axioms except for effectiveness. To see that this code lacks effectiveness, it suffices to consider matching pennies. The following game shows that the conclusion of Theorem 1 (and hence Theorem 2) does not hold for this code.

	a	b
A	$1, 1$	$2, 0$
B	$0, 2$	$1, 1$

(A, a) is a strict Nash equilibrium, but not a fair play equilibrium, whereas (B, b) is a fair play equilibrium, but not a Nash equilibrium.

Example 9 (Welfare nondiscrimination: Aviod-the worst-of-many-for-others-code)

Under this code, a strategy is unacceptable socially if it is the unique minimizer of everyone else's utility given m_{-i}, provided that the number of actions is sufficiently large. That is, if

$$\sum_{i \in N} \frac{1}{|X_i|} \geq 1, \tag{17}$$

no restriction is imposed: $F_i(X, u, m_{-i}) = M_i$ for all i and all m_{-i}. If (17) is violated, which is the case if the number of actions is large for all players, then for all $m \in M$ and all $i \in N$, $m_i \notin F_i(X, u, m_{-i})$ if and only if $(m_i, m_{-i}) \prec_j (m'_i, m_{-i})$ for all $m'_i \neq m_i$ and all $j \neq i$.

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 unacceptable 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 unacceptable 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

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

To see that this moral 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 moral code states that B is not socially acceptable in the left game since it gives the unique least preferred outcome for player 2. In the right game,

however, B is socially acceptable since C is as bad as B for player 2.(On the other hand, if we remove the requirement of uniqueness from the definition of the moral code, then we lose effectiveness.) The left game also shows that the conclusion of Theorem 1(and hence that of Theorem 2) does not hold for this moral code, since $FPE(X, u, F) = \{(A, a)\}$ while $NE(X, u) = SNE(X, u) = \{(B, a)\}$. Since this moral code satisfies neutrality, this example also shows that Theorem 1 does not hold if welfare nondiscrimination is replaced with neutrality.

7 The Utilitarian and Lexi–min codes

Independence or the combination of weak independence with invariance to equivalent utility representations makes the informational base in moral judgement impoverished because these axioms never allow any kind of interpersonal welfare comparison.

This section defines two moral codes both of which violate independence, but have rich informational bases that make interpersonal welfare comparison possible. One is called Local Utilitarian Code, based on the same idea as that in Utilitarian rule. The other is called Local Lexi-min Code, based on the same idea as that in Lexi-min rule due to Sen (1970), a lexicographic completion of Rawls (1972)' difference principle. We show that these codes work: both codes have fair play equilibria for any game.

Definition The *Local Utilitarian Code* is defined as follows: Given exogenously specified weights λ_j, $j \in N$, such that $\lambda_j > 0$ for all j and $\sum_{j \in N} \lambda_j = 1$,

$F_i(X, u, m_{-i}) := \{m_i \in M_i : \sum_{j \in N} \lambda_j v_j(m_i, m_{-i}) \geq \sum_{j \in N} \lambda_j v_j(m_i', m_{-i})$ for all $m_i' \in M_i\}$

Different weights generate a different Local Utilitarian code. Every Local Utilitarian code satisfies all the axioms except anonymity and independence. If the weights are all equal to $1/n$, we call it Local Pure Utilitarian Code,

which is the only code satisfying anonymity in the family of Local Utilitarian codes. It is easy to see that every Local Utilitarian code satisfies welfare nondiscrimination, monotonicity, and continuity.

Proposition 5 *For any game (X, u), there exists a fair play equilibrium under any Local Utilitarian code.*

Proof. First we define the following codes, called ε-codes.

For any $\varepsilon \in (0, 1)$, ε-code is defined by:

for all $(X, u) \in \Gamma$, all $i \in N$, and all $m_{-i} \in M_{-i}$,

$F_i^{\varepsilon}(X, u, m_{-i}) := \{m_i \in M_i : \sum_{j \in N} \lambda_j v_j(m_i, m_{-i}) \geq (1 - \varepsilon) Max(m_{-i})\}$,

where $Max(m_{-i}) = \max_{m_i \in M_i} \sum_{j \in N} \lambda_j v_j(m_i, m_{-i})$.

Under ε-code, a strategy m_i given m_{-i} is a fair play if and only if it brings about at least as much as $1 - \varepsilon$ portion of the maximum of weighted utility sum. If there exists a fair play equilibrium m^{ε} for any ε-code, then there exists a fair play equilibrium for Local Utilitarian code: As $\varepsilon \longrightarrow 0$, we can let $m^{\varepsilon} \longrightarrow m^0$, the limit of the sequence of fair play equilibria under ε code, which turns out to be a fair play equilibrium for Local Utilitarian code.

Therefore the proof reduces to show the existence of fair play equilibrium for ε-code, which follows Kakutani's fixed point theorem argument.

Given a game (X, u), we define a correspondence $\Psi : M \to M$ as follows:

$\Psi(m) = \prod_{i \in N} \{m_i^* \in M : m_i^* \in F_i^{\varepsilon}(X, u, m_{-i})$ and $v_i(m_i^*, m_{-i}) \geq v_i(m_i', m_{-i})$ for all $m_i' \in F_i^{\varepsilon}(X, u, m_{-i})\}$.

It is easy to see that Ψ is convex valued. To see that Ψ is continuous, the only difficult part to prove is that F_i^{ε} is lower hemi continuous. (Since it is easy to see that F_i^{ε} is upper hemi continuous, F_i^{ε} is then continuous. The Berge maximum theorem is applied so that Ψ is continuous.)

Let $m_i^0 \in F_i^{\varepsilon}(X, u, m_{-i}^0)$ and $m_{-i}^{\nu} \longrightarrow m_{-i}^0$. It suffices to show the existence of a sequence such that $m_i^{\nu} \in F_i^{\varepsilon}(X, u, m_{-i}^{\nu})$ and $m_i^{\nu} \longrightarrow m_i^0$. By

definition of F_i^{ε},

$$\sum_{j\in N}\lambda_j v_j(m_i^0, m_{-i}^0) \geq (1-\varepsilon)Max(m_{-i}^0).$$

Let $\widetilde{m}_i^0$ be a strategy that defines $Max(m_{-i}^0)$, i.e.,

$$\sum_{j\in N}\lambda_j v_j(\widetilde{m}_i^0, m_{-i}^0) > (1-\varepsilon)Max(m_{-i}^0).$$

For any ν, we have

$$\left(1-\frac{1}{\nu}\right)\sum_{j\in N}\lambda_j v_j(m_i^0, m_{-i}^0) + \frac{1}{\nu}\sum_{j\in N}\lambda_j v_j(\widetilde{m}_i^0, m_{-i}^0) > (1-\varepsilon)Max(m_{-i}^0).$$

Since $Max(m_{-i}^{\nu}) \longrightarrow Max(m_{-i}^0)$ as $m_{-i}^{\nu} \longrightarrow m_{-i}^0$, there exists a number ν^* such that

for all $\nu \geq \nu^*$,

$$\left(1-\frac{1}{\nu}\right)\sum_{j\in N}\lambda_j v_j(m_i^0, m_{-i}^{\nu}) + \frac{1}{\nu}\sum_{j\in N}\lambda_j v_j(\widetilde{m}_i^0, m_{-i}^{\nu}) > (1-\varepsilon)Max(m_{-i}^{\nu}).$$

That is,

$$\sum_{j\in N}\lambda_j v_j\left(\left(1-\frac{1}{\nu}\right)m_i^0 + \frac{1}{\nu}\widetilde{m}_i^0, m_{-i}^{\nu}\right) > (1-\varepsilon)Max(m_{-i}^{\nu}).$$

Thus the desired sequence m_i^{ν} is given by

$$m_i^{\nu} = \left(1-\frac{1}{\nu}\right)m_i^0 + \frac{1}{\nu}\widetilde{m}_i^0 \ \forall \nu \geq \nu^* ,$$

where m_i^{ν} is arbitrarily taken when $\nu < \nu^*$. ■

We now turn to define the Lexi-min code. For every point $a \in R^n$, i.e., a point in the $n-$dimensional Euclidean space, let $i^*(a)$ denote the index of its

i-th smallest component. Two binary relations $>_L^*$ and $=_L^*$ on R^n are defined as follows:

$$a >_L^* b \Longleftrightarrow \exists r \in N : \begin{cases} \forall i \in \{1, ..., r-1\} : a_{i^*(a)} = b_{i^*(b)} \\ a_{r^*(a)} > b_{r^*(b)} \end{cases}$$

and

$a =_L^* b \Longleftrightarrow \forall i \in N : a_{i^*(a)} = b_{i^*(b)}$.

Let $a \geq_L^* b \Longleftrightarrow a >_L^* b \vee a =_L^* b$.

Definition The *Local Lexi-min Code* is defined as follows:

Let $(X, u) \in \Gamma$ *be given. For any* $i \in N$, *any* $m_i \in M_i$, *and any* $m_{-i} \in M_{-i}$,

$F_i(X, u, m_{-i}) := \{m_i \in M_i : v(m_i, m_{-i}) \geq_L^* v(m_i', m_{-i})$ *for all* $m_i' \in M_i\}$,

where $v(m_i, m_{-i}) = (v_1(m_i, m_{-i}), ..., v_n(m_i, m_{-i}))$.

It is easy to see that this code satisfies anonymity, welfare nondiscrimination, monotonicity and effectiveness, and that it violates independence and continuity.

We show the existence of a fair play equilibrium, not depending on the fixed point argument.

Proposition 6 *For the Local Lexi-min code, any fair play profile is a fair play equilibrium, and hence there exists a fair play equilibrium for any game.*

Proof. Let x be a fair play profile of (X, u). Let $i \in N$ be given.

Suppose $m_i \in F_i(X, u, x_{-i})$. By definition, we have $v(x_i, x_{-i}) =_L^* v(m_i, x_{-i})$.

Let S be the set of players who are the smallest components of $v(x_i, x_{-i})$, and similarly let T be the set of players who are the smallest components of $v(m_i, x_{-i})$. It follows from $v(x_i, x_{-i}) =_L^* v(m_i, x_{-i})$ that $|S| = |T|$. We show $S = T$. Suppose not. Then there are subsets S' of S and T' of T such that $|S'| = |T'|$ and any $i \in S'$ is the smallest in $v(x_i, x_{-i})$ but not in $v(m_i, x_{-i})$ whereas any $i \in T'$ is the smallest in $v(m_i, x_{-i})$ but not in $v(x_i, x_{-i})$. Thus for a mixed strategy m_i' of x_i and m_i, we see that $v(m_i', x_{-i}) >_L^* v(x_i, x_{-i}) =_L^*$

$v(m_i, x_{-i})$, which is a contradiction. Thus we have $S = T$, i.e., the set of players who are the smallest components of $v(x_i, x_{-i})$ coincides with the set of players who are the smallest components of $v(m_i, x_{-i})$. This argument can be applied to the second smallest components, the third smallest components, ..., and the n-th smallest components so that we conclude that i is indifferent between (x_i, x_{-i}) and (m_i, x_{-i}) and that x_i is i's best reply to x_{-i} in the situation (X, u, x_{-i}). This holds for all $i \in N$, and hence x is a fair play equilibrium. ■

8 Conclusion

In economic theory, moral codes have been studied mainly in terms of the incentives of individuals to follow the moral code. The theory of repeated games has shown that a society can sustain almost any outcome as an equilibrium by designing a punishment scheme. Incentive compatibility, however, is not the only characteristic expected for moral codes. moral codes are also expected to be fair and consistent in their instructions.

Contrary to our result, real-life moral codes appear to constrain people's behavior. People often accept unpleasant tasks because of a moral code or their own ethical feeling. This is not a contradiction since our result is based on normative properties of moral codes and does not necessarily characterize actual moral codes. What our result does is to shed light on a tension between prescriptivity and universalizability of a moral code and what a moral 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 moral code that satisfies them has little merit in terms of the induced outcomes.

The result also shows that a seemingly plausible moral code may violate a basic normative requirement. If a moral code requires a person to sacrifice his own payoff to make others better off, we know that the moral 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 may give a useful perspective on real-life moral codes and ethical judgements.

9 Appendix

Lemma 1 *For all* $(X,u)\in\Gamma$*, all* $i\in N$*, and all* $m_i,m_i'\in M_i (m_i\neq m_i')$*,if*

$$v_j(m_i,z_{-i})=v_j(m_i',z_{-i})\quad\forall j\in N,\ \forall z_{-i}\in X_{-i} \tag{18}$$

then $m_i\simeq m_i'$.

Proof. For all $m_{-i}\in M_{-i}$,

$$\begin{aligned} v_j(m_i,m_{-i}) &= \sum_{z_{-i}\in X_{-i}} v_j(m_i,z_{-i})\prod_{k\neq i} m_k(z_k) \\ &\overset{(18)}{=} \sum_{z_{-i}\in X_{-i}} v_j(m_i',z_{-i})\prod_{k\neq i} m_k(z_k) = v_j(m_i',m_{-i}), \end{aligned}$$

where z_k is the $k-$th component of z_{-i}. Since this holds true for all $j\in N$, $m_i\simeq m_i'$. ■

Lemma 2 *For all* $(X,u)\in\Gamma$*, all* $m\in M$*, and all* $a\in\Omega^N\backslash X$*,*

there exists a game $(X',u')\in\Gamma$ *satisfying the following conditions:*

1. $X_i'=X_i\cup\{a_i\}$ *for all* $i\in N$,
2. $u'|_X=u$*,and*
3. $a_i\simeq m_i$ *for all* $i\in N$.

Proof. We construct a game (X',u') recursively as follows. First of all, we define $(X^1,u^1)\in\Gamma$ by setting $X_1^1=X_1\cup\{a_1\}$, and $X_j^1=X_j$ for all $j\neq 1$, and

$$u_j^1(x_1,x_{-1})=\begin{cases} v_j(m_1,x_{-1}) & \text{if } x_1=a_1 \\ u_j(x_1,x_{-1}) & \text{if } x_1\neq a_1 \end{cases}$$

for all $j\in N$ and all $(x_1,x_{-1})\in X^1$. Since u^1 differs from u only when $x_1=a_1$, we have $u^1|_X=u$. We have $a_1\simeq m_1$ in the game (X^1,u^1). Indeed,

the definition of u_j^1 implies

$$u_j^1(a_1, x_{-1}) = v_j(m_1, x_{-1}) = v_j^1(m_1, x_{-1}) \tag{19}$$

since the support of m_1 does not include a_1. (19) and Lemma 1 imply $a_1 \simeq m_1$.

Next, we define $(X^2, u^2) \in \Gamma$ by setting $X_2^2 = X_2 \cup \{a_2\}$, $X_j^2 = X_j^1$ for all $j \neq 2$, and

$$u_j^2(x_2, x_{-2}) = \begin{cases} v_j^1(m_2, x_{-2}) & \text{if } x_2 = a_2 \\ u_j^1(x_2, x_{-2}) & \text{if } x_2 \neq a_2 \end{cases}$$

for all $j \in N$ and all $(x_2, x_{-2}) \in X^2$. Since u^2 differs from u only when either $x_1 = a_1$ or $x_2 = a_2$, we have $u^2|_X = u$. As in the previous step, $a_2 \simeq m_2$ in (X^2, u^2). We also have $a_1 \simeq m_1$ in (X^2, u^2). Here is the proof: If 2's component of x_{-1} is not a_2, then

$v_j^2(m_1, x_{-1}) = v_j^1(m_1, x_{-1})$,

$u_j^2(a_1, x_{-1}) = u_j^1(a_1, x_{-1})$.

These two values are also equal since $a_1 \simeq m_1$ in (X^1, u^1). On the other hand, if player 2's component of x_{-2} is a_2, then

$v_j^2(m_1, x_{-1}) = v_j^2(m_1, a_2, x_{-\{1,2\}}) = v_j^1(m_1, m_2, x_{-\{1,2\}})$

$u_j^2(a_1, x_{-1}) = u_j^2(a_1, a_2, x_{-\{1,2\}}) = u_j^1(a_1, m_2, x_{-\{1,2\}})$.

These two values are also equal since $a_1 \simeq m_1$ in (X^1, u^1). Thus, in both case, $v_j^2(m_1, x_{-1}) = u_j^2(a_1, x_{-1})$, which implies that $a_1 \simeq m_1$ in (X^2, u^2).

Repeating the similar argument for each action $a_3, ..., a_n$, we obtain games $(X^3, u^3), ..., (X^n, u^n)$. The last game (X^n, u^n) is the desired one. ■

If a moral code F satisfies welfare nondiscrimination, then for the game just constructed above, we can say that $F_i(X, u, m_{-i}) = F_i(X', u', a_{-i}) \setminus M_i(a_i)$, where $M_i(a_i) = \{m_i \in M_i : m_i(a_i) > 0\}$.

To prove the next lemma, we need to consider a framework where only pure strategies are used. Moral codes in this framework are denoted as X-codes:

Definition An *X-code* is a correspondence G that associates with each

game $(X,u) \in \Gamma$, each $i \in N$, and each $x_{-i} \in X_{-i}$ a non-empty subset $G_i(X,u,x_{-i}) \subset X_i$.

Thus, an X-code takes a pure-strategy profile of other players and gives a subset of pure strategies. Our axioms can be redefined for X-codes in a straightforward way.

Definition (Axioms of X-code)

An X-code G satisfies *X-anonymity* if for all $(X,u),(X',u') \in \Gamma$ and all permutations $\pi\colon N \to N$, if for all $i \in N$ and all $x \in X$, $X'_{\pi(i)} = X_i$ and $u'_{\pi(i)}(x^\pi) = u_i(x)$, then for all $i \in N$ and all $x \in X$, $G_{\pi(i)}(X',u',x^\pi_{-\pi(i)}) = G_i(X,u,x_{-i})$.

An X-code G satisfies *X-welfare nondiscrimination* if for all $(X,u) \in \Gamma$, the following conditions are satisfied.

(i) 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 $i \in N$, $x_i \in G_i(X,u,x_{-i}) \Longleftrightarrow y_i \in G_i(X,u,y_{-i})$;

(ii) For all $x \in X$, all $i \in N$, and all $y_i \in X_i$, if $x_i \simeq y_i$ and $x_i \neq y_i$, then $G_i(X_i \setminus \{y_i\} \times X_{-i},u,x_{-i}) = G_i(X,u,x_{-i}) \setminus \{y_i\}$ and $G_j(X_i \setminus \{y_i\} \times X_{-i},u,x_{-j}) = G_j(X,u,x_{-j})$ for all $j \neq i$.

An X-code G satisfies *X-monotonicity* if for all $(X,u),(X,u') \in \Gamma$, all $x \in X$, and all $i \in N$, if $x_i \in G_i(X,u,x_{-i})$ and

$$u'_j(x) = u_j(x) \text{ and } u'_j(y) \leq u_j(y) \quad \forall y \neq x, \forall j \in N,$$

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

An X-code G satisfies *X-weak independence* if for all $(X,u),(X,u') \in \Gamma$, all $i \in N$, and all $x_{-i} \in X_{-i}$, if, for all $j \in N$, u_j and u'_j are identical on $\{(y_i,x_{-i}) : y_i \in X_i\}$, then $G_i(X,u,x_{-i}) = G_i(X,u',x_{-i})$.

An X-code G satisfies *X-invariance* to equivalent utility representations if for all $(X,u),(X,u') \in \Gamma$, if, for all $j \in N$, there exist $\alpha_j > 0$ and $\beta_j \in R$ such that $u'_j(x) = \alpha_j u_j(x) + \beta_j$ for all $x \in X$, then for all $i \in N$ and all $x_{-i} \in X_{-i}$, $G_i(X,u',x_{-i}) = G_i(X,u,x_{-i})$.

An X-code G satisfies *X-effectiveness* if for all $(X,u) \in \Gamma$, there exists $x \in X$ such that for all $i \in N$, $x_i \in G_i(X,u,x_{-i})$. We call this x an *X-fair play profile.*

As for moral codes, X-welfare nondiscrimination implies X-neutrality, which is defined as follows: For all $(X,u),(X',u') \in \Gamma$, if, for all $i \in N$, there exists a bijection $\rho_i \colon X_i \to X'_i$ such that for all $x \in X$, $u'_i(x) = u_i(\rho_1(x_1),\rho_2(x_2),\ldots,\rho_n(x_n))$, then for all $x \in X$ and all $i \in N$, $x_i \in G_i(X,u,x_{-i}) \Longleftrightarrow \rho_i(x_i) \in G_i(X',u',\rho(x)_{-i})$.

Lemma 3 *If an X-code G satisfies X-anonymity, X-welfare nondiscrimination, X-monotonicity, X-weak independence, X-invariance to equivalent utility representations, and X-effectiveness, then for all $(X,u) \in \Gamma$, all $i \in N$ and all $x_{-i} \in X_{-i}$, $G_i(X,u,x_{-i}) \cap BR_i(X,u_i,x_{-i}) \neq \phi$*

Proof. Let G be a X-code satisfying all the axioms. Suppose on the contrary that there exist $(X,u^*) \in \Gamma$, $i \in N$ and $x^*_{-i} \in X_{-i}$ such that $G_i(X,u^*,x^*_{-i}) \cap BR_i(X,u^*_i,x^*_{-i}) = \emptyset$. Let $B := BR_i(X,u^*_i,x^*_{-i}) \cap X_i$. To simplify notation, let $i = 1$, $X_1 = \{1,2,...,|X_1|\}$, $B = \{1,...,k-1\}$, where $k \geq 2$. Let $K = B \cup \{k\}$. By X-neutrality, we can assume $x^*_{-1} = (k,...,k)$.

By using X-weak independence and X-welfare nondiscrimination, it is further assumed to be $X = X_1 \times K^{N-1}$. Let us show the details. First we delete all actions in $X_j (j \neq 1)$ except for k and make a game $(X_1 \times \{k\}^{N-1}, u')$ in the following way:

Let u' be an utility profile such that

$u'_i(x_1,x_{-1}) = u^*_i(x_1,x^*_{-1})\ \forall i \in N, \forall x_1 \in X_1, \forall x_{-1} \in X_{-1}$.

It is easy to see that $x_j \simeq k$ for all $j \neq 1$, all $x_j \in X_j$. By X-weak independence and X-welfare nondiscrimination, we have $G_i(X_1 \times \{k\}^{N-1}, u', x^*_{-i}) = G_i(X,u^*,x^*_{-i})$.

Next we add all actions in B to this game, and make a game $(X_1 \times K^{N-1}, u'')$ in the following way:

Let u'' be an utility profile such that

$u_i''(x_1, x_{-1}) = u_i'(x_1, x_{-1}^*)(= u_i^*(x_1, x_{-1}^*))\ \forall i \in N, \forall x_1 \in X_1, \forall x_{-1} \in X_{-1}$.

It is easy to see that $b \simeq k$ for all $j \neq 1$, all $b \in K_j$. By X-weak independence and X-welfare nondiscrimination, we have $G_i(X_1 \times K^{N-1}, u'', x_{-i}^*) = G_i(X, u^*, x_{-i}^*)$.

With out loss of generality we denote $u'' = u^*$.

We now modify u^* to define a new profile u satisfying the following conditions:

$u_i(c, x_{-1}^*) = u_i(c', x_{-1}^*)\ \forall c, c' \notin B, \forall i \in N$,

$u_1(b, x_{-1}^*) = u_1^*(b, x_{-1}^*) > u_1(c, x_{-1}^*) \geq u_1^*(c, x_{-1}^*)\ \forall b \in B, \forall c \notin B$,

$u_j(b, x_{-1}^*) = u_j^*(b, x_{-1}^*) < u_j(c, x_{-1}^*) \geq u_j^*(c, x_{-1}^*)\ \forall b \in B, \forall c \notin B, \forall j \neq 1$

Thus we do not change anyone's utility for action profiles $B^* := \{(b, x_{-1}^*) : b \in B\}$. We increase everyone's utility over the action profiles $C^* := \{(c, x_{-1}^*) : c \in X_1 \backslash B\}$ so that (i) everyone is indifferent within C^*, (ii) player 1 prefers B^* to C^*(which is possible since B^* gives strictly higher utilities than C^* under u_1^*), and (iii) all players $j \neq 1$ prefer C^* to B^*. The utilities at the other action profiles in $X_1 \times K^{N-1}$ are arbitrary.

Suppose that, under the new profile u, there exists $b_1 \in B$ such that $b_1 \in G_1(X_1 \times K^{N-1}, u, x_{-1}^*)$. Since u gives higher utilities than the original u^* over C^*, X-monotonicity (together with X-weak independence) implies $b_1 \in G_1(X_1 \times K^{N-1}, u^*, x_{-1}^*)$, which is a contradiction. Thus $G_1(X_1 \times K^{N-1}, u, x_{-1}^*) \cap B = \emptyset$.

We define a profile $\widetilde{u}$ such that for all $i \in N$,

$$\begin{cases} \widetilde{u}_i(c, x_{-1}) = u_i(c, x_{-1}^*) & \text{if } c \notin B \text{ and } x_{-1} \neq x_{-1}^* \\ \widetilde{u}_i = u_i & \text{otherwise} \end{cases}$$

By X-weak independence, $G_1(X_1 \times K^{N-1}, \widetilde{u}, x_{-1}^*) \cap B = G_1(X_1 \times K^{N-1}, u, x_{-1}^*) \cap B = \emptyset$.

Since all actions in $X_1 \backslash B$ are welfare equivalent, this implies that $G_1(X_1 \times K^{N-1}, \widetilde{u}, x_{-1}^*) = X_1 \backslash B$. If we delete all the actions in $X_i \backslash B$ except for k, then X-welfare nondiscrimination implies

$$G_1(K^N, \widetilde{u}, x_{-1}^*) = \{k\}.$$

Thus, player 1 is instructed to choose k, which is what the other players want him to do but the worst action for player 1 himself. By X-anonymity and X-neutrality, everyone is instructed to act in the same way in the same situation. In what follows, we show that this X-code violates X-effectiveness by constructing a game where the X-code has no X-fair action profile.

For all $i \in N$, let $u_i : X \longrightarrow \mathbb{R}$ be the utility function obtained from u_i by adding a constant so that $u_i(k, x^*_{-1}) = 0$. By X-invariance to equivalent utility representations,

$$G_1(K^N, u^*, x^*_{-1}) = \{k\}. \tag{20}$$

Define a function $\pi : N \times N \longrightarrow N$ by $\pi(i, j) := j - i + 1 \ (mod\ n)$. For all $i \in N$, let $\alpha_i : \{1, ..., n-1\} \times K \longrightarrow \mathbb{R}_{++}$ 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 where the X-code has no fair action profile. The action set is $K = \{1, ..., k\}$ for all players, as before. For each i, the utility function $w_i : K^N \longrightarrow \mathbb{R}$ is given by

$$w_i(x) := \sum_{j=1}^{n-1} \alpha_i(j, x_n) u_{\pi(j,i)}(x_j, x^*_{-1}) + \beta_i u_{\pi(n,i)}(x_n, x^*_{-1}) \text{ if } x_n < k, \tag{21}$$

$$w_i(x) := \sum_{j=1}^{n-1} \alpha_i(j, k)[u_{\pi(j,i)}(k + 1 - x_j, x^*_{-1}) - u_{\pi(j,i)}(1, x^*_{-1})] \text{ if } x_n = k. \tag{22}$$

We prove that game (K^N, w), where $w = (w_i)_{i \in N}$, has no X-fair play profile. The proof goes on with four steps.

Step 1. For any X-fair play profile x in (K^N, w), if $x_n < k$, then $x_j = k$ for all $j \neq n$.

Let x be an X-fair play profile such that $x_n < k$. Let $j \neq n$. (21) implies $w_i(\cdot, x_{-j}) =$constant$+\alpha_i(j, x_n) u_{\pi(j,i)}(\cdot, x^*_{-1})\ \forall i \in N$,

where "constant" represents the term that is independent of j's action x_j, and the dot $\cdot$ represents x_j. This equation for $i = j$ implies that player j's position in (K^N, w) given x_{-j} is identical to the position of player $\pi(j, j) = 1$ in (K^N, u) given x^*_{-1}. Similarly, player i's position in (K^N, w) given x_{-j} is identical to the position of player $\pi(j, i)$ in (K^N, u) given x^*_{-1}. Then our axioms and (20) imply that $G_j(K^N, w, x_{-j}) = \{k\}$, which completes the proof of Step 1.

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

Let x be an X-fair play profile such that $x_{-n} = (k, ..., k)$. Since (21)-(22) and $u_i(k, x^*_{-1}) = 0$ for all i, we have $w_i(\cdot, x_{-n}) = \beta_i u_{\pi(n,i)}(\cdot, x^*_{-1})$ for all $i \in N$. This equation for $i = n$ implies that player n's position in (K^N, w) given x_{-n} is identical to that of player $\pi(n, n) = 1$ in (K^N, u) given x^*_{-1}. Similarly, player $i's$ position in (K^N, w) given x_{-n} is identical to that of player $\pi(n, i)$ in (K^N, u) given x^*_{-1}. Hence $G_n(K^N, w, x_{-n}) = G_1(K^N, u, x^*_{-1}) = \{k\}$, which completes the proof of Step 2.

Steps 1 and 2 imply that if there exists an X-fair play profile x, then $x_n = k$.

Step 3. For any X-fair play profile x in (K^N, w), if $x_n = k$, then $x_j = 1$ for all $j \neq n$.

Let x be an X-fair play profile with $x_n = k$. Let $j \neq n$ be given and consider the utility vectors he can induce. The definition of u implies that

$w_i(y_j, x_{-j}) =$constant$+\alpha_i(j, k)u_{\pi(j,i)}(k + 1 - y_j, x^*_{-1})$ $\forall y_j \in K$, $\forall i \in N$,

where the first term is "constant" with respect to y_j. As before, the equation for $i = j$ implies that the position of player j in (K^N, w) given x_{-j} is identical to that of player $\pi(j, j) = 1$ in (K^N, u) given x^*_{-1}. Similarly, the position of player i in (K^N, w) given x_{-j} is identical to that of player $\pi(j, i)$ in (K^N, u) given x^*_{-1}. But this time, action k given x^*_{-1} corresponds to action $k + 1 - k = 1$ given x_{-j}. Thus, the only fair play for player j is 1: $G_j(K^N, w, x_{-j}) = \{1\}$. This completes the proof of Step 3.

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

Step 4. $x = (1, ..., 1, k)$ is not an X-fair play profile in (K^N, w).

It suffices to show that for all $a \in K$ and all $i \in N$,

$$w_i(1, ..., 1, a) = u_{\pi(n,i)}(k + 1 - a, x^*_{-1}) + \text{constant}. \tag{23}$$

where the term "constant" is constant with respect to a. Indeed, if this condition holds, the condition for $i = n$ implies that the position of player n in (K^N, w) given $x_{-n} = (1, ..., 1)$ is identical to the position of player $\pi(n, n) = 1$ in (K^N, u) given x^*_{-1}. Similarly, the position of player i in (K^N, w) given $x_{-n} = (1, ..., 1)$ is identical to that of player $\pi(n, i)$ in (K^N, u) given x^*_{-1}. And player n's action $a = 1$ in (K^N, w) given x_{-n} corresponds to player 1's action $k + 1 - 1 = k$ in (K^N, u) given x^*_{-1}. Thus (20) implies $G_n(K^N, w, x_{-n}) = \{1\}$. This implies that $(1, ..., 1, k)$ is not an X-fair play profile in (K^N, w). The reminder of the proof is devoted to the proof of (23).

A sufficient condition for (23) is that for all $i \in N$ and all $x_n < k$,

$$w_i(1, ..., 1, x_n) - w_i(1, ..., 1, k) = u_{\pi(n,i)}(k+1-x_n, x^*_{-1}) - u_{\pi(n,i)}(1, x^*_{-1}). \tag{24}$$

Proof of (24)$\Longrightarrow$(23): Assume (24) and let $a \in K$. It is obvious when $a = k$. Suppose $a < k$. Then substituting $x_n = a$ in (24) yields

$$w_i(1, ..., 1, a) = u_{\pi(n,i)}(k + 1 - a, x^*_{-1}) + \text{constant} \quad \forall a < k \tag{25}$$

where

$$\text{constant} = w_i(1, ..., 1, k) - u_{\pi(n,i)}(1, x^*_{-1}). \tag{26}$$

But rewriting (26) yields $w_i(1, ..., 1, k) = u_{\pi(n,i)}(1, x^*_{-1})$+constant. This and (25) imply (23).

The definition of w_i implies that (24) is equivalent to the following: for all i

$$\sum_{j=1}^{n-1} u_{\pi(j,i)}(1, x^*_{-1})[\alpha_i(j, x_n) + \alpha_i(j, k)] + \beta_i u_{\pi(n,i)}(x_n, x^*_{-1})$$

$$= u_{\pi(n,i)}(k + 1 - x_n, x^*_{-1}) - u_{\pi(n,i)}(1, x^*_{-1}) \ \forall x_n < k \tag{27}$$

This condition may not hold if $\alpha_i(\cdot, \cdot)$ and $\beta_i(\cdot, \cdot)$ are chosen arbitrarily. However, we claim that for all $i \in N$, there exists $\alpha_i : \{1, ..., n-1\} \times K \to \mathbb{R}_{++}$ and $\beta_i > 0$ such that (27) holds. This is sufficient for our proof since the argument so far does not depend on the values of $\alpha_i(\cdot, \cdot)$ and $\beta_i(\cdot, \cdot)$ as long as they are strictly positive.

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

$u_1(1, x^*_{-1}) > 0$ and $u_{\pi(n,i)}(x_n, x^*_{-1}) < 0$ if $i \neq n$,

$u_{\pi(1,i)}(1, x^*_{-1}) < 0$ and $u_1(x_n, x^*_{-1}) > 0$ if $i = n$,

which follows from the definition of u and the normalization $u_i(k, x^*_{-1}) = 0$. Thus, for all $x_n < k$, the left-hand side of (27) contains a positive term as well as a negative term, so it should be intuitively clear that (27) 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.

The 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 $x_n < k$,

$$\sum_{j=1}^{n-1} u_{\pi(j,i)}(1, x^*_{-1}) + \beta_i u_{\pi(n,i)}(x_n, x^*_{-1}) < u_{\pi(n,i)}(k+1-x_n, x^*_{-1}) - u_{\pi(n,i)}(1, x^*_{-1}). \tag{28}$$

The left-hand side of (28) coincides with that of (27) when $\alpha_i(j, x_n) = \alpha_i(j, k) = 1/2$ for all (j, x_n) and all (j, k). Since $u_{\pi(i,i)}(1, x^*_{-1}) > 0$, the equality of (27) can be attained by increasing $\alpha_i(i, x_n)$.

The case when $i = n$ is similar. First, set $\beta_i > 0$ sufficiently large so

that for all $x_n < k$, (28) holds with the reverse inequality. This inequality implies that if we set $\alpha_i(j, x_n) = \alpha_i(j, k) = 1/2$ for all (j, x_n) and all (j, k), then the left-hand side of (28) is larger than the right-hand side. Since $u_{\pi(1,n)}(1, x^*_{-1}) = u_n(1, x^*_{-1}) < 0$, the equality in (27) 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 (27) holds for all $x_n < k$, thus (24) holds. As we discussed, (24) implies $G_n(K^N, w, (1, ..., 1)) = \{1\}$, hence $(1, ..., 1, k)$ is not a X-fair play profile. ■

The next lemma is a weak version of Theorem 1 restricting attention to the case where other players use pure strategies.

Lemma 4 *If a moral code F satisfies anonymity, welfare nondiscrimination, weak independence, monotonicity, and effectiveness, then for all $(X, u) \in \Gamma$, all $i \in N$, and all $x_{-i} \in X_{-i}$, $F_i(X, u, x_{-i}) \cap BR_i(X, u, x_{-i}) \neq \phi$.*

Proof. Consider the following particular X-code G.

$G_i(X, u, x_{-i}) := [F_i(X, u, x_{-i}) \cap X_i] \cup [\bigcup_{j \neq i} \{x_i \in X_i : u_j(x_i, x_{-i}) \geq u_j(x'_i, x_{-i})$ $\forall x'_i \in X_i\}]$.

The second term, i.e., $\bigcup_{j \neq i} \{\cdots\}$, ensures the non-emptyness of $G_i(X, u, x_{-i})$. Since F satisfies our original axioms for moral code, G satisfies all the axioms redefined for X-codes.

Hence Lemma 3 implies that for all $(X, u) \in \Gamma$, all $i \in N$, and all $x_{-i} \in X_{-i}$,

$$G_i(X, u, x_{-i}) \cap BR_i(X, u_i, x_{-i}) \neq \phi. \tag{29}$$

The final step is to go back to F and prove that for all $(X, u) \in \Gamma$, all $i \in N$, and all $x_{-i} \in X_{-i}$,

$$F_i(X, u, x_{-i}) \cap BR_i(X, u_i, x_{-i}) \neq \phi. \tag{30}$$

If $BR_i(X, u_i, x_{-i}) \cap X_i = X_i$, (30) is obvious because of $BR_i(X, u_i, x_{-i}) =$

M_i. Thus suppose $BR_i(X, u_i, x_{-i}) \cap X_i \subsetneq X_i$.

For all $j \neq i$, let u'_j be a utility function obtained from u_j by increasing the utilities at action profiles (x_i, x_{-i}) where $x_i \notin BR_i(X, u_i, x_{-i}) \cap X_i$. That is,

$u'_j(b_i, x_{-i}) = u_j(b_i, x_{-i}) \ \ \forall b_i \in BR_i(X, u_i, x_{-i}) \cap X_i$,
$u'_j(c_i, z_{-i}) = u_j(c_i, z_{-i}) + L \ \ \forall c_i \in X_i \backslash BR_i(X, u_i, x_{-i})$,
where $L > 0$ is set large so that for all $j \neq i$,

$$max\{u'_j(x_i, x_{-i}) : x_i \in X_i\} > max\{u'_j(b_i, x_{-i}) : b_i \in BR_i(X, u_i, x_{-i}) \cap X_i\}. \tag{31}$$

Thus, if player i does not play his best reply against x_{-i}, the other players obtain a large additional payoff L. Since the other players want this additional payoff, they do not want i to play his best reply. Since (29) holds for all utility profiles including $(u_i, (u'_j)_{j \neq i})$, we obtain

$$G_i(X, (u_i, (u'_j)_{j \neq i}), x_{-i}) \cap BR_i(X, u_i, x_{-i}) \cap X_i \neq \phi.$$

By definition of G, this implies either

$$F_i(X, (u_i, (u'_j)_{j \neq i}), x_{-i}) \cap BR_i(X, u_i, x_{-i}) \neq \phi. \tag{32}$$

or

$$[\bigcup_{j \neq i} \{x_i \in X_i : u'_j(x_i, x_{-i}) \geq u_j(x'_i, x_{-i}) \ \forall x'_i \in X_i\}] \cap BR_i(X, u_i, x_{-i}) \neq \phi. \tag{33}$$

Since u'_j was defined to satisfy (31), (33) does not hold. Hence we obtain (32).

Let $x_i \in X_i$ be any element of the intersection in (32). We now go back to the utility profile u. Since the move from $(u_i, (u'_j)_{j \neq i})$ to u changes utilities only downwards, but does not change the utilities at (x_i, x_{-i}), the monotonicity of F implies $x_i \in F_i(X, u, x_{-i})$. Thus we obtain $F_i(X, u, x_{-i}) \cap$

$BR_i(X, u_i, x_{-i}) \neq \phi$. ■

The proof of Theorems 1 and 2 goes on as follows:

Proof. Theorem 1: Consider any $(X, u), i, m_i$. Take $a_j \in \Omega \backslash X_j$ for each $j \neq i$ to construct a game $(X', u') \in \Gamma$ as in Lemma 2. Lemma 4 says $F_i(X', u', a_{-i}) \cap BR_i(X', u', a_{-i}) \neq \phi$. This and welfare nondiscrimination imply $F_i(X', u', m_{-i}) \cap BR_i(X', u', m_{-i}) \neq \phi$. Deleting all a_j from (X', u'), and applying welfare nondiscrimination again, we obtain the desired result.

Theorem 2: It suffices to show that $BR_i(X, u, m_{-i}) \subset F_i(X, u, m_{-i})$. Indeed, this implies $NE(X, u) \subset FPE(X, u, F)$, which together with Theorem 1 completes the proof.

So, let $m_i \in BR_i(X, u, m_{-i})$. Let $a_i \in \Omega \backslash X_i$ and define $X_i^0 = X_i \cup \{a_i\}$ and $X_j^0 = X_j$ for all $j \neq i$. Consider a sequence of utility functions $\{u_k^\nu\}_{\nu=1}^\infty$ by

$u_k^\nu(a_i, x_{-i}) := v_k(m_i, x_{-i}) + \frac{1}{\nu}$,

$u_k^\nu(x_i, x_{-i}) := u_k(x_i, x_{-i})$

for all $k \in N$, all $x_i \in X_i \backslash \{a_i\}$, and all $x_{-i} \in X_{-i}$. Let u_j^0 be the limit of u_j^ν as $\nu \longrightarrow \infty$. Then, by Lemma 1, $a_i \simeq m_i$ in (X^0, u^0).

We now show that $BR_i(X^\nu, u^\nu, m_{-i}) = \{a_i\}$ for all ν. Indeed, for any $m_i' \in M_i$ (i.e., strategy whose support does not include a_i),

$$v_i^\nu(m_i', m_{-i}) = v_i(m_i', m_{-i}) \leq v_i(m_i, m_{-i}) < v_i(m_i, m_{-i}) + \frac{1}{\nu} = v_i^\nu(a_i, m_{-i}). \tag{34}$$

For any $m_i' \in M_i'$ (i.e., strategy whose support includes a_i) such that $m_i' \neq a_i$,

$v_i^\nu(m_i', m_{-i}) = m_i'(a_i) v_i^\nu(a_i, m_{-i}) + \sum_{x_i \neq a_i} m_i'(x_i) v_i^\nu(x_i, m_{-i})$

$< m_i'(a_i) v_i^\nu(a_i, m_{-i}) + \sum_{x_i \neq a_i} m_i'(x_i) v_i^\nu(a_i, m_{-i}) = v_i^\nu(a_i, m_{-i})$,

where the inequality follows from (34) and $x_i \neq a_i$.

This proves $BR_i(X^\nu, u^\nu, m_{-i}) = \{a_i\}$ for all ν.

Theorem 1 then implies that $a_i \in F_i(X^0, u^\nu, m_{-i})$ for all ν. The continuity of F implies $a_i \in F_i(X^0, u^0, m_{-i})$. Since $a_i \simeq m_i$ in (X^0, u^0), welfare

nondiscrimination implies $m_i \in F_i(X^0, u^0, m_{-i})$. Deleting a_i and applying welfare nondiscrimination implies $m_i \in F_i(X, u, m_{-i})$. ■

References

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

[2] Ausitin-Smith, D. and Banks, J.S. (1999). Positive Political Economy I: Collective Preference, Ann Arbor:University of Michigan Press.

[3] Bossert, W. and Weymark, J. A. (2004). Utility in Social Choice. In: Barbera, S., Hammond, P. J. and Seidl, C. (eds) Handbook of Utility Theory, Vol. 2, Chapter 20, Kluwer Academic Press.

[4] Campbell, D.E. Kelly J.S. (2002). Impossibility Theorems in the Arrovian Framework. In:Arrow KJ, Sen AK, Suzumura K (eds) Handbook of social choice and welfare Vol.1

[5] d'Aspremont, R. (1985). "Axioms for Social Welfare Orderings," in Hurwicz, L., Schmeidler, D., and H. Sonnenschein eds. Social Goals and Social Organization, Cambridge: Cambridge University Press.

[6] d'Aspremont, R., and Gevers, L. (2002). "Social Welfare Functionals and Interpersonal Comparability. " In: Arrow, J. K., Sen, A. K. and Suzumura, K. (eds) Handbook of Social Choice and Welfare, Vol. 1, Chapter 10, Elsevier B. V..

[7] Debreu, G. (1952). A Social Equilibrium Existence Theorem. In: Proceedings of the national academy of sciences of the U.S.A. 38:886-893.

[8] Gaertner, W., Pattanaik, P. K., and Suzumura, K. (1992). "Individual Rights revised," Economica 59, 161-177.

[9] Gibbard, A. (1974). “A Pareto-Consistent Libertarian Claim,” Journal of Economic Theory 7, 388–410.

[10] Hare, R. M. (1952). The Language of Morals, Oxford: Clarendon Press.

[11] Hare, R. M. (1963). Freedom and Reason, Oxford: Clarendon Press.

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

[13] Harsanyi, J.C. (1955). "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," Journal of Political Economy 63, 309-21.

[14] 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.

[15] Mas-Colell, A. Whinston, MD. and Green,JR (1995) Micro Economic Theory, Oxford Student edition.

[16] Miyagawa, E. Nagahisa, R. and Suga, K. (2005) "Fair Play Equilibria in Normal Form Games," Mimeo.

[17] Moulin, H. (1988). Axioms of Cooperative Decision Making. Cambridge University Press

[18] Nozick, R. (1974). Anarchy, State, and Utopia, Basic Books:New York.

[19] Peleg, B. and Tijs, S. (1996). “The Consistency Principle for Games in Strategic Forms,” International Journal of Game Theory 25, 13–34.

[20] Peleg, B. Potters, J.A.M. and Tijs, S. (1996). “Minimality of the Consistent Solutions for Strategic Games, in Particular for Potential Games," Economic Theory 7, 81–93.

[21] Peña,JP. (2003). "Ethical implementation and the creation of moral values," centrA: Fundación Centro de Estudios Andaluces, Documento de Trabajo Serie Economia E2003/25 1-34

[22] Rawls, J. (1971, 2nd ed.1999) A Theory of Justice, Harvard University Press.

[23] Salonen, H. (1992). "An axiomatic analysis of the Nash equilibrium concept," Theory and Decision 33, 177-189

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

[25] Sen, A. K. (1970b). Collective Choice and Social Welfare, SanFrancisco: Holden-Day.

[26] Sen, A. K. (1986). "Social Choice Theory," In: Arrow KJ, Intriligator MD (eds) Handbook of mathematical economics Vol.3 North Holland.

[27] Sen, A. K. (1992). "Minimal liberty." Economica, 59, 139-159.

[28] Sen, A. K. (1996). "Rights: formulation and consequences. " Analyse & Kritik, 18, 153-170.

[29] Sen, A. K. (2011). "The Informational Basis of Social Choice." In: Arrow, J. K., A. K. Sen and K. Suzumura (eds) Handbook of Social Choice and Welfare, Vol. 2, Chapter 14, Elsevier B. V.

[30] Sidgwick, H. (1907). The Method of Ethics, Macmillan, London.

[31] Suzumura, K. (1996). "Welfare, rights, and social choice procedure: a perspective." Analyse & Kritik, 18, 153-170.

[32] Suzumura, K. (2000). "Welfare economics beyond welfarist-consequentialism." Japanese Economic Review, 50, 1-32.

[33] Suzumura, K. (2011). "Welfarism, Individual Rights, and Procedural Fairness." In: Arrow, J. K., A. K. Sen and K. Suzumura (eds) Handbook of Social Choice and Welfare, Vol. 2, Chapter 23, Elsevier B. V.