Evolution of payoff-dependent preferences *

Tatsuhiro Shichijo
Osaka Prefecture University

First version: August 2008
This version: June 2009

Abstract

We consider a situation where each player's preferences are "payoff-dependent" in the sense that the utility of the payoff-dependent preferences is determined by the payoff of each person and those of others. In this paper, we consider the evolution of such payoff-dependent preferences using an "indirect evolutionary approach"(Güth and Yaari (1992)). Our model is essentially the same model as the one studied in Dekel, Ely and Yilankaya (2007) except that they consider a larger set of possible preferences set. It turns out that if an outcome is stable in our setting, then it is also stable in the setting in Dekel et al. (2007). That is, our model can check the robustness of the result about stable outcome in Dekel et al. (2007). We obtain a sufficient condition for a stable outcome in our setting which is less restrictive than the condition in Dekel et al. (2007). When we consider a game with roles, we find that an important criterion for stability is whether each role obtains equal payoff. For example, in an ultimatum game, an equal division is stable because each role obtains equal payoff. We find that the necessary conditions for instability include unequal payoff for each role. For example, in the ultimatum game, unequal division with pure strategy is unstable.

1 Introduction

Many experimental results show that human beings are not simple material payoff maximizers. For instance, subjects sometimes reciprocate or punish others, even though it reduces their own rewards. The indirect evolutionary approach, first proposed by Güth and Yaari (1992), is one theory that attempts to explain these anomalies.

As in standard evolutionary game theory, the indirect evolutionary approach assumes that a population of individuals is randomly matched in a two-player game. The special feature of this approach is that it considers (subjective) preferences that may differ from (material) payoffs and studies the evolution of preferences. Players choose their actions according to their own preferences and information about the opponents' preferences. The distribution of actions then determines the average (material) payoff of each preference. Finally, those preferences that have earned larger payoffs increase. As discussed in Samuelson (2001), subjective preferences are effective as a commitment device if they are observable by others. As a result, the possibility exists that irrational preferences, which are inconsistent with material payoffs, can survive. This approach has two useful features. First, it can explain a result that is not a Nash equilibrium. For example, Dekel et al. (2007)(hereafter DEY) show that a cooperative state can be stable in the prisoner's dilemma game. Second, this approach can be used for further refinement of the Nash equilibrium. For example, Güth (1995) construct a model that explains fair allocation in the ultimatum game.

Samuelson (2001) pointed out two important questions about this approach. First, "applications typically consider only a handful of possible preferences, including true preferences as well as commitment types that are carefully tailored to the particular game of interest, $\cdots$ What would happen if one allowed the entire collection of possible preferences to compete?" Second, "the commitment type's advantage arises out of the fact that his opponents can observe his commitment $\cdots$. Is it reasonable to assume that preferences can be observed, and what happens if they cannot?"

Recently, DEY made an important contribution to the literature on the indirect evolutionary approach by responding to the questions posed by Samuelson (2001). At first, DEY considered various degrees of observability. That is, they considered not only the perfectly observable case, but also the partially observable and unobservable

*I am grateful to Munetomo Ando, Michihiro Kandori and seminar participants at Osaka University for helpful comments on an earlier version of this paper. All remaining errors are mine. This research is partially supported by a Japan Society for the Promotion of Science Grant-in-Aid for Young Scientists (B).

cases. Moreover, DEY check the robustness of extreme cases. Second, DEY allow for all possible preferences based on action profiles[1]. That is, the possible preferences are all functions from pure strategy profiles to $[0, 1]$ in the DEY model.

Our model is essentially the same model as the one studied in DEY except that they consider a larger set of possible preferences set. DEY considered a large preference set that includes counterintuitive preferences. For example, consider the game with a material payoff matrix as in Table 1. θ^{DEY} in Table 2 shows the possible preferences in DEY's model. θ^{DEY}'s utility is 1 when both of players choose a_1 and obtain material payoff 10. On the other hand, θ^{DEY}'s utility is 0 when both of players choose a_2 and obtain material payoff 10. The possibility exists that these counterintuitive preferences cause problems. Lemma 1 in this paper shows that the larger the set of possible preferences, the looser the stability condition. That is, a state that is stable in DEY's model can be unstable in a model with a smaller set of possible preferences.

We provide an intuitive exposition here and the formal proof in Lemma 1 in Section 2. Consider two sets of preferences, say Θ^S, Θ^L. We suppose that Θ^S is a subset of Θ^L. We now consider two models depending on the set of preferences, a model with Θ^S and a model with Θ^L. We then consider which model requires a more restrictive condition for stability. When we consider stability, we should pay heed to *indifferent preferences* that are indifferent between strategies for any strategies of the opponent (see θ^0 in Table 2 for an example and Definition 6 for a formal definition). Now we assume that Θ^S includes indifferent type.

The size of the possible preference set has two effects on stability. First, the model with Θ^L considers more mutants than the model with Θ^S. From this point of view, it appears to be more difficult for an outcome to be stable in the model with a larger preference set, Θ^L. We should, however, notice that Θ^S includes indifferent types. Any strategy is a best response strategy against any strategy for the indifferent preferences and there is no restriction on the equilibrium play of these types. That is, this type can mimic any other type. If a mutant can destabilize a population, then the indifferent type, by copying its strategy, can also destabilize the population. Thus, the problem of stability reduces to whether the indifferent type can destabilize the population. Because both Θ^L and Θ^S include the indifferent type, both models have the same restriction. Second, the model with Θ^L considers more incumbents than the model with Θ^S. That is, a preference in Θ^L but not Θ^S can compose a stable state as the incumbent. From this viewpoint, the larger preference set, Θ^L, makes the stable condition less restrictive.

Therefore, it is worthwhile studying the robustness of DEY's results under a smaller set of possible preferences as some biological constraints may influence the evolution of the brain. In this paper, we consider a model with subjective utilities that depend only on the payoff profiles. That is, we consider preferences that have the same utility level for the same payoff profile. We refer to these functions as *payoff-dependent preferences*. (See θ^π in Table 2 for payoff-dependent preferences in games with the material payoffs shown in Table 1.) Theorem 2 shows that a strategy profile is stable in a model with payoff-dependent preferences for any level of observability if the profile consists of a pure, strong, neutrally stable strategy, where the strong neutral stability is defined in Section 2. As a strong neutral stability is the weaker requirement of DEY's propositions, and the smaller set of possible preferences used in the theorem, we make two contributions using it. First, we confirm that DEY's results are robust in a model with payoff-dependent preferences. Second, we obtain a weaker sufficient condition for stability.

In Section 3, we consider two-player games with two roles. In these games, roles are determined at random before players choose an action. For example, the ultimatum game is a two-player game with two roles. Note that even if a game has a symmetric payoff matrix, it may still have roles, such as a "Battle of the Sexes" game. If we consider a model with all preferences based on an action profile as in DEY, then the profile of the pure, strong, neutrally stable strategy is stable in games with roles (see Theorem 4). However, if we consider payoff-dependent preferences, whether each role obtains the same payoff becomes an additional, critical condition for stability. If each role obtains the same payoff in an action profile, and the strategy that induces the action profile is a strong, stable strategy, then the profile of the strategy is stable in a model with payoff-dependent preferences (see Theorem 5). On the other hand, if each role obtains a different payoff, and some additional conditions are satisfied, then the strategy profile cannot be stable (see Theorem 6). That is, the evolution of payoff-dependent preferences has a tendency for a fair allocation.

[1]They did not consider preferences that care about another's preference, as discussed in some of the literature, e.g., Bester and Güth (1998), Sethi and Somanathan (2001). Herold and Kuzmics (2008) study a general model that considered these preferences.

	a_1	a_2
a_1	10,10	0,0
a_2	0,0	10,10

Table 1: Material payoff

	a_1	a_2
a_1	1	1
a_2	0	0

θ^{DEY}

	a_1	a_2
a_1	α	β
a_2	β	α

θ^{π}

	a_1	a_2
a_1	0	0
a_2	0	0

θ^{0}

Table 2: Examples of subjective utility over Table 1

2 The basic model

2.1 The environment

We borrow the notations and the stability concept from DEY. In this section, we consider a two-player symmetric normal-form game $G = (A, \pi, \{1, 2\})$, where $A = \{a_1, a_2, \ldots, a_n\}$ is a finite action set and $\pi : A \times A \to \mathbb{R}$ is a (material) payoff function. As in standard indirect evolutionary theory, we interpret (material) payoffs as "fitness". As we consider the symmetric game, a_i is the action of player i, $\pi(a_1, a_2)$ is the payoff to player 1 and $\pi(a_2, a_1)$ is the payoff to player 2. We denote the set of mixed actions over A by Δ. The payoff function π extends naturally to $\Delta \times \Delta$. As in standard evolutionary game theory, we suppose that individuals are repeatedly drawn at random from a large population and match to play G. A (subjective) preference is a function from $A \times A$ to $[0, 1]$. Note that preferences can be different from the payoff function. $\theta(a, a')$ is the (subjective) utility of preference θ when a player plays $a \in A$ and her opponent plays $a' \in A$. Let Θ be the set of possible preferences, which is a variable depending on the model. We refer to the unit of evolution as "type". In this paper, the unit of evolution is preference, hence we often refer to $\theta \in \Theta$ as a "type". We denote with $\theta(\sigma, \sigma')$ the expected utility of type θ when a player with θ adopts σ and her opponent adopts σ'. That is, $\theta(\sigma, \sigma') = \sum_{a_i, a_j \in A} \theta(a_i, a_j)\sigma(a_i)\sigma'(a_j)$ where $\sigma(a_i)$ or $\sigma'(a_j)$ is the probability that strategy σ or σ' plays action a_i or a_j, respectively. In DEY's model, the set of possible preferences is assumed to be the set of all possible functions from $A \times A$ to $[0, 1]$, denoted by Θ^{DEY}. In this paper, we pay attention to payoff-dependent preferences that have the same utility level for the action profiles whose payoff profile is the same. We write Θ^{π} for the set of all payoff-dependent preferences. Formally,

$$\Theta^{\pi} \equiv \{\theta \in \Theta^{DEY} : \theta(\alpha, \beta) = \theta(\alpha', \beta') \text{ if } (\pi(\alpha, \beta), \pi(\beta, \alpha)) = (\pi(\alpha', \beta'), \pi(\beta', \alpha'))\}.$$

We assume that the population is large but finite and hence consider only the case where finite types exist in a population. We denote the set of possible finite support probability distributions on Θ by $\mathcal{P}(\Theta)$. $C(\mu)$ denotes the support of distribution $\mu \in \mathcal{P}(\Theta)$. Early indirect evolutionary approaches assumed that each player could observe her opponent's preference. According to DEY and other more recent literature, we consider not only observable preference cases but also unobservable and intermediate preference cases. We assume that players independently observe their opponents' preferences with probability $p \in [0, 1]$, and with probability $1 - p$ players do not have any information about their opponents. We consider all cases of $p \in [0, 1]$. Given μ and p, we have an incomplete information game, denoted by $\Gamma_p(\mu)$.

2.2 The solution concept

We consider, as in the standard indirect evolutionary literature, a static two-step solution concept. First, it is assumed that players optimize their strategies to obtain better (subjective) utility and hence learn to play a Bayesian–Nash equilibrium of the incomplete information game defined by the environment, $\Gamma_p(\mu)$. Second, using the expected fitness of each type determined by the Bayesian–Nash equilibrium, we define the static evolutionary stability. As in DEY, we consider that a population is stable only if any mutants cannot invade the population, i.e., mutants cannot earn greater fitness than the incumbents.

A strategy for type θ is a function $b_\theta : C(\mu) \cup \{\phi\} \to \Delta$, where $b_\theta(\theta')$ is the strategy of type θ when she observes the opponent's type is $\theta' \in \Theta$ and $b_\theta(\phi)$ is the strategy when she does not observe any information.
When a player with θ observes that the opponent's type is θ', the expected strategy of the opponent is

$$pb_{\theta'}(\theta) + (1 - p)b_{\theta'}(\phi),$$

as the opponent plays $b_{\theta'}(\theta)$ with probability p and $b_{\theta'}(\phi)$ with probability $1-p$. When a player with θ cannot observe information about the opponent, the expected strategy of the opponent is

$$\sum_{\theta' \in C(\mu)} \left[p b_{\theta'}(\theta) + (1-p) b_{\theta'}(\phi) \right] \mu(\theta')$$

where $\mu(\theta')$ is the population share of θ'. Let $BR_\theta(\sigma)$ be the set of best response strategies of type θ against σ. That is, $BR_\theta(\sigma) \equiv \text{argmax}_{\sigma^* \in \Delta} \theta(\sigma^*, \sigma)$. We assume that each player chooses a best response strategy for her type and that the strategy profile b is a Bayesian–Nash equilibrium of $\Gamma_p(\mu)$. In other words, each player adjusts her beliefs from her experience, and the time taken for belief adjustment is so short when compared with the length of her life that we can neglect it when we consider the evolution of preferences. Therefore, we assume that each b_θ is a best response strategy given $\mu \in \mathcal{P}(\Theta)$ and the others' strategies. Formally, for all $\theta', \theta \in C(\mu)$,

$$b_\theta(\theta') \in BR_\theta(p b_{\theta'}(\theta) + (1-p) b_{\theta'}(\phi))$$

$$b_\theta(\phi) \in BR_\theta(\sum_{\theta' \in C(\mu)} \left[p b_{\theta'}(\theta) + (1-p) b_{\theta'}(\phi) \right] \mu(\theta'))$$

Let $B_p(\mu)$ denote the set of such all Bayesian–Nash equilibria of the game $\Gamma_p(\mu)$. The average payoff of type $\theta \in C(\mu)$ is denoted by $\Pi_\theta(\mu \mid b)$ and given by the following equation:

$$\begin{aligned} \Pi_\theta(\mu \mid b) \quad = \quad & \sum_{\theta' \in C(\mu)} [p^2 \pi(b_\theta(\theta'), b_{\theta'}(\theta)) + p(1-p)\pi(b_\theta(\theta'), b_{\theta'}(\phi)) \\ & + [p(1-p)\pi(b_\theta(\phi), b_{\theta'}(\theta)) + (1-p)^2 \pi(b_\theta(\phi), b_{\theta'}(\phi))] \mu(\theta') \end{aligned} \tag{1}$$

The pairing of the distribution μ and the strategy profile b determines the outcome of the game. We refer to the combination of (μ, b) as *configuration*, where $b \in B_p(\mu)$. We are interested in the strategy profile resulting from a configuration. We say a configuration (μ, b) induces (σ^*, σ^*) if the aggregated strategy of the configuration is σ^*, i.e.,

$$\sum_{\theta, \theta' \in C(\mu)} \left[p b_\theta(\theta') + (1-p) b_\theta(\phi) \right] \mu(\theta) \mu(\theta') = \sigma^*.$$

Note that when σ^* is a pure strategy, (μ, b) induces (σ^*, σ^*) if and only if $b_\theta(\theta') = b_\theta(\phi) = \sigma^*$ for all $\theta, \theta' \in C(\mu)$. We use a stability concept that parallels DEY. Given the set of possible preferences, Θ, a configuration is stable if it satisfies four conditions. The first condition is trivial: the concerned configuration, (μ, b), must have support in the given preference set, Θ, i.e., $C(\mu) \subset \Theta$. The second condition concerns the incumbent's payoff. That is, all types present must earn the same payoff. If there is a type that earns a higher payoff than ther types do, then the population share of this type will increase and the configuration will change. The third condition requires that the mutant must not earn a larger payoff than the incumbents in a post-entry Bayesian–Nash equilibria achieved after the mutant has entered. If a mutant earns a larger payoff, then the mutant can invade the population and increase its share. As there are usually many equilibria, we must consider which post-entry equilibrium is relevant.

It is not natural that after the mutant has entered, all players can change their strategy at the same time and move to totally different equilibria from the original, even if they can remain at an equilibrium that is close to the original. When players change their strategy greatly, they should usually pay some cost, e.g., a mental cost for getting used to the new strategy. Thus, we assume that players choose a post-entry equilibrium that is close to the original if a Bayesian–Nash equilibrium exists. Therefore, we require that the mutant must not earn a larger payoff than the incumbents in any "nearby" post-entry equilibria close to the original. The final condition concerns the existence of the "nearby" post-entry equilibria. We define the distance of the post-entry equilibria from the original and require that, for any $\delta > 0$, there exists a post-entry equilibrium that is δ away from the original if the mutant share is sufficiently small. DEY requires a similar condition for stability. Moreover, we require an additional condition that does not appear in DEY because we consider a smaller set of possible preferences. The following example shows the necessity of the additional restriction.

Example 1. Consider the case where only two preferences $\Theta = \{\theta, \theta^0\}$ exist, as defined by the Table 3. That is, $\theta(a_1, a_1) = \theta(a_1, a_2) = \theta(a_2, a_1) = 0, \theta(a_2, a_2) = 1$ and $theta^0(a_i, a_j) = 0$ for all $i, j \in \{1, 2\}$. We assume that in a configuration (μ, b), all players have preferences θ and use strategy a_1, i.e., $\mu(\theta) = 1$, $b_\theta(\theta) = b_\theta(\phi) = a_1$. In this case, there exists the following post-entry equilibrium, $\tilde{b}$, for any share of mutants: $\tilde{b}_\theta(\theta) = \tilde{b}_{\theta^0}(\theta) = \tilde{b}_{\theta^0}(\phi) = \tilde{b}_\theta(\phi) = a_1$. We can say that this equilibrium is close to the original because incumbents use exactly the same strategies. That is, a nearby post-entry equilibrium exists for any mutant. On the other hand, if the mutant θ^0 chooses a_2 for any

	a_1	a_2
a_1	0	0
a_2	0	1

type θ

	a_1	a_2
a_1	0	0
a_2	0	0

type θ^0

Table 3: Subjective utility of each type

information, i.e., $\tilde{b}_{\theta^0}(\theta) = \tilde{b}_{\theta^0}(\phi) = a_2$, then $\tilde{b}_\theta(\theta) = \tilde{b}_\theta(\phi) = a_2$ is the unique best response strategy of θ for any positive mutant share. As a_2 is always best response strategy for θ^0, it is natural that θ^0 keeps adopting a_2 and greatly changes the Bayesian–Nash equilibrium from the original. Thus, we would not say that (μ, b) is stable[2]. **End of example**

The DEY model considers all preferences. Thus, it considers the preference for each strategy, say β in Example 2.2, that is strictly dominant. Therefore, while DEY need not consider the problem in Example 2.2, we must because we consider a smaller set of possible preferences. When mutants have some dominant strategy, we also consider the post-entry equilibrium where mutants adopt a fixed dominant strategy and require that some of the post-entry equilibria are close to the original [3].

To develop a formal discussion, we introduce some notation.

We denote by $N_\epsilon(\mu, \tilde{\theta})$ the set of preference distributions where the original population is μ and the proportion of mutants $\tilde{\theta}$ is at most ϵ. Formally, $N_\epsilon(\mu, \tilde{\theta}) = \{\tilde{\mu} : \tilde{\mu} = (1-\epsilon')\mu + \epsilon'\tilde{\theta}, \epsilon' < \epsilon\}$

We define the distance between post-entry and original equilibrium as follows:

Definition 1. A strategy profile $\tilde{b}$ is δ from the b on μ if

$$\delta = \max_{\theta \in C(\mu), \theta' \in C(\mu) \cup \{\phi\}} |\tilde{b}_\theta(\theta') - b_\theta(\theta')|$$

Note that we focus only on the difference of strategies over $C(\mu)$.We denote the set of "nearby" Bayesian–Nash equilibria where the distance of the incumbent's strategy between the original and post-entry equilibria is at most $\delta \in [0, 1)$ by $B_p^\delta(\tilde{\mu} \mid \mu, b)$. Formally[4]

$$B_p^\delta(\tilde{\mu} \mid \mu, b) = \{\tilde{b} \in B_p(\tilde{\mu}) : |\tilde{b}_\theta(\phi) - b_\theta(\phi)| \leq \delta, |\tilde{b}_\theta(\theta') - b_\theta(\theta')| \leq \delta \text{ for } \forall \theta, \theta' \in C(\mu)\}$$

where , for any $\tilde{\sigma}, \sigma \in \Delta$, $|\tilde{\sigma} - \sigma| = \sum_{a \in A} |\tilde{\sigma}(a) - \sigma(a)|^2$.

Definition 2. We say a strategy $\sigma^* \in \Delta$ is *a dominant strategy* for θ if $\theta(\sigma^*, \sigma) \geq \theta(\sigma', \sigma)$ for $\forall \sigma', \forall \sigma \in \Delta$.

Let $D(\theta)$ be the set of dominant strategies for θ, i.e., $D(\theta) = \{\sigma \in \Delta : \theta(\sigma, \sigma') \geq \theta(\sigma'', \sigma') \text{ for } \forall \sigma', \forall \sigma'' \in \Delta\}$. We consider the subset of $B_p^\delta(\tilde{\mu} \mid \mu, b)$ where mutants adopt the fixed strategy determined by the given strategy profile, $\{\sigma_\theta\}_{\theta \in C(\mu) \cup \phi}$. Formally:

$$B_p^\delta(\tilde{\mu} \mid \mu, b, \{\sigma_\theta\}_{\theta \in C(\mu) \cup \phi}) = \{\tilde{b} \in B_p^\delta(\tilde{\mu} \mid \mu, b) : \forall \theta \in C(\mu), \tilde{b}_{\tilde{\theta}}(\theta) = \sigma_\theta, \tilde{b}_{\tilde{\theta}}(\phi) = \sigma_\phi\} \tag{2}$$

The condition about the existence of nearby equilibria is defined as follows: [5]

Definition 3. (μ, b) is δ-robust in Θ if there exists $\epsilon > 0$ such that $\forall \tilde{\theta} \in \Theta \setminus C(\mu)$, $\forall \tilde{\mu} \in N_\epsilon(\mu, \tilde{\theta})$:

(i) $B_p^\delta(\tilde{\mu} \mid \mu, b) \neq \emptyset$,

(ii) If $D(\tilde{\theta}) \neq \emptyset$, then $B_p^\delta(\tilde{\mu} \mid \mu, b, \{\sigma_\theta\}_{\theta \in C(\mu) \cup \phi}) \neq \emptyset$ for any dominant strategy profiles $\{\sigma_\theta\}_{\theta \in C(\mu) \cup \phi}$ such that $\sigma_\theta \in D(\tilde{\theta})$ for all $\theta \in C(\mu) \cup \{\phi\}$.

[2]As we define formally below, we consider a strategy to be dominant if its (subjective) utility is larger than or equal to the utility from any other strategy.

[3]We only consider the case where the mutant uses dominant strategies. This is enough because, as shown in the following, we only consider the set of possible preferences that include an infinite number of indifferent type. As any strategies are dominant for indifferent type, this requirement is sufficiently strong.

[4] $B_p^0(\tilde{\mu} \mid \mu, b)$ is the same as the set of "focal" equilibria in DEY. Note that the definition of "nearby" equilibrium is slightly different from DEY.

[5]In the definition, we require that the type of mutant is not the same as the incumbent's, i.e., $\tilde{\theta} \in \Theta \setminus C(\mu)$. This requirement comes from the assumption that the strategy of preference θ is b_θ. That is, we assume every player uses the same strategy as a player with the same preference. This means that we do not consider the case where a mutant has the same preference as an incumbent and adopts a different strategy to the incumbent. We can change this assumption, but we retain the assumption, as in DEY, for simplicity.

Obviously, if (μ, b) is δ-robust, then for all $\delta' \in (\delta, \infty)$, (μ, b) is δ'-robust.

The formal definition of stability is as follows [6]:

Definition 4. A configuration (μ, b) is *stable* in Θ if:

(1) $C(\mu) \subset \Theta$,

(2) $\Pi_\theta(\mu \mid b) = \Pi_{\theta'}(\mu \mid b)$ for all $\theta, \theta' \in C(\mu)$,

(3) There exists $\epsilon > 0$ such that $\forall \tilde{\theta} \in \Theta$, $\forall \tilde{\mu} \in N_\epsilon(\mu, \tilde{\theta})$,

  (a) If (μ, b) is 0-robust in Θ then, $\forall \tilde{b} \in B_p^0(\tilde{\mu} \mid \mu, b)$, $\forall \theta \in C(\mu)$, $\Pi_\theta(\tilde{\mu} \mid \tilde{b}) \geq \Pi_{\tilde{\theta}}(\tilde{\mu} \mid \tilde{b})$.

  (b) If (μ, b) is not 0-robust in Θ, then (μ, b) is δ-robust in Θ for all $\delta > 0$ and $\forall \tilde{b} \in B_p(\tilde{\mu})$, $\forall \theta \in C(\mu)$, $\Pi_\theta(\tilde{\mu} \mid \tilde{b}) \geq \Pi_{\tilde{\theta}}(\tilde{\mu} \mid \tilde{b})$.

If a configuration is stable in Θ^{DEY}, then it is stable in the definition of DEY. We say that a strategy profile (σ^*, σ^*) is stable if there exists a stable configuration that induces (σ^*, σ^*). If a stable configuration exists that induces (σ^*, σ^*) as the dominant strategy of their preferences, we can guess that the strategy profile (σ^*, σ^*) is easily materialized. Moreover, the strategy profile is stable, even if we consider a more restrictive solution concept for $\Gamma_p(\mu)$ than Bayesian–Nash equilibrium, say, trembling perfect equilibrium.

Definition 5. A strategy profile (σ^*, σ^*) is *stable with dominant strategy* in Θ if there exists a configuration (μ, b) such that (μ, b) is stable in Θ and $b_\theta(\theta')$ are dominant for any $\theta \in C(\mu)$ and any $\theta' \in C(\mu) \cup \{\phi\}$.

A strategy profile (σ^*, σ^*) is *unstable* in Θ if there exists no configuration (μ, b) such that (μ, b) is stable in Θ and induce (σ^*, σ^*).

Throughout the paper, we use the following conservative policy. When we find that something is stable, we consider that it is stable with a dominant strategy. When we obtain the result that a strategy profile is unstable, we use ordinal stability.

2.3 The results of the basic model

As discussed, indifferent type plays an important role in our discussion. The formal definition of an indifferent type is as follows:

Definition 6. We say a preference, $\theta^0 \in \Theta$, is *indifferent type* if $\theta^0(a_i, a_j) = \theta^0(a'_k, a'_\ell)$ for all $a_i, a_j, a'_k, a'_\ell \in A$.

For example, if $\theta(a_1, a_2) = 1$ for all $a_1, a_2 \in A$, then θ is an indifferent type. If $\theta'(a_1, a_2) = 0.5$ for all $a_1, a_2 \in A$, then θ' is another indifferent type.

For the indifferent type, any strategies are best-response strategies. That is, the type can take any strategies in post-entry Bayesian–Nash equilibrium. Thus, this type is the hardest mutant to remove. That is, if an indifferent type cannot destabilize a configuration, then any mutants also cannot. Thus, we need only consider an indifferent type as a mutant if it can be a mutant. In this discussion, we require that the indifferent type can be a mutant.

Formally, we have the following theorem .

Lemma 1. *Suppose that $\Theta^S \setminus C(\mu)$ includes an indifferent type θ^0 and that $C(\mu) \subset \Theta^S \subset \Theta^L$. Then a configuration (μ, b) is stable in Θ^L if it is stable in Θ^S.*

From Lemma 1, we can easily obtain the following result.

Theorem 1. *Let $\Theta^\pi \subseteq \Theta^S \subseteq \Theta^L \subseteq \Theta^{DEY}$. Then a strategy profile (σ^*, σ^*) is stable in Θ^L if the profile is stable in Θ^S.*

The following efficiency of the strategy profile plays an important role in both DEY and our paper.

Definition 7. (σ^*, σ^*) is *efficient* if $\pi(\sigma^*, \sigma^*) \geq \pi(\sigma, \sigma)$ for all $\sigma \in \Delta$.

The following criterion is the central concept of this paper.

[6] In the definition of (3)-(b), we can replace $\forall \tilde{b} \in B_p(\tilde{\mu})$ by $\forall \tilde{b} \in B_p^\delta(\tilde{\mu} \mid \mu, b)$ without any change in our results. However, we use it in order for the definition to be parallel to DEY's stability. (!!In the DEY, istead of (3)-(b) they use the following definition: If $B_p^0(\tilde{\mu} \mid \mu, b) \neq \emptyset$, then $B_p^\delta(\tilde{\mu} \mid \mu, b) \neq \emptyset$ for all $\delta > 0$ and $\Pi_\theta(\tilde{\mu} \mid \tilde{b}) \geq \Pi_{\tilde{\theta}}(\tilde{\mu} \mid \tilde{b})$, for all $\theta \in C(\mu)$, for all $\tilde{b} \in B_p(\tilde{\mu})$!!!)

Definition 8. We say σ^* is a *strong, neutrally stable strategy* (hereafter, strong NSS) if σ^* satisfies the following two properties:

(1) $\pi(\sigma^*, \sigma^*) \geq \pi(\sigma, \sigma^*)$ for all $\sigma \in \Delta$,

(2) if $\pi(\sigma^*, \sigma^*) = \pi(\sigma, \sigma^*)$, then $\pi(\sigma^*, \sigma) \geq \pi(\sigma', \sigma')$ for all $\sigma' \in \Delta$.

If σ^* is a strong NSS, then σ^* is a neutrally stable strategy and (σ^*, σ^*) is a Nash equilibrium. Moreover, inserting $\sigma = \sigma^*$ into condition (2) of the above definition, we have $\pi(\sigma^*, \sigma^*) \geq \pi(\sigma', \sigma')$ for all $\sigma' \in \Delta$. That is, if σ^* is a strong NSS, then (σ^*, σ^*), is efficient.

We now present an intuitive interpretation of a strong NSS, assuming $p = 1$ for simplicity. Suppose that the incumbents adopt strategy σ^* whatever they observe, and that mutants adopt strategy σ against incumbents and strategy σ' against mutants. In this case, the expected payoff to incumbents is $(1-\epsilon)\pi(\sigma^*, \sigma^*) + \epsilon\pi(\sigma^*, \sigma)$ where ϵ is the population share of mutants. On the other hand, the expected payoff to mutants is $(1-\epsilon)\pi(\sigma, \sigma^*) + \epsilon\pi(\sigma', \sigma')$. If σ^* is a strong NSS and ϵ is small enough, the expected payoff to incumbents becomes larger than or equal to that for mutants. Thus, if a type exists for which a strong NSS is dominant, the profile of the strong NSS appears to be stable. We add the existence of a uniform invasion barrier ϵ and extend the discussion to any observability and payoff-dependent preference in the following Theorem.

Theorem 2. *(a^*, a^*) is stable with a dominant strategy in Θ^π for any $p \in [0, 1]$ if $a^* \in A$ is a pure strong NSS.*

Using Theorem 1, we can directly extend Theorem 2 to Θ^{DEY}.

Corollary 1. *Let $\Theta^\pi \subseteq \Theta \subseteq \Theta^{DEY}$. (a^*, a^*) is stable with a dominant strategy in Θ for any $p \in [0, 1]$ if $a^* \in A$ is a pure strong NSS.*

In Propositions 1 and 6 in DEY, they use the condition "strict Nash and efficient". If (σ^*, σ^*) is strict Nash and efficient, then σ^* is a strong NSS. That is, strong NSS is a less restrictive condition than "strict Nash and efficient". Thus, the corollary above corresponds to Propositions 1 and 6 in DEY and implies the robustness of their results.

DEY obtained two important conditions necessary for stability. When p is high enough, a pure strategy profile is stable only if it is efficient. When p is low enough, a pure strategy profile is stable only if it is a Nash equilibrium.

Theorem 3. *(Dekel et al. (2007))*

(i) *If a pure strategy profile (a, a) is not efficient, then there exists $\bar{p} \in (0, 1)$ such that it is not stable in Θ^{DEY} for any $p \in (\bar{p}, 1]$.*

(ii) *If a pure strategy profile is not a Nash equilibrium of G, then there exists $\underline{p} \in (0, 1)$ such that it is not stable in Θ^{DEY} for any $p \in [0, \underline{p})$.*

See Dekel et al. (2007) for the proof[7]. This theorem can be extended to Θ^π easily, by combining Theorem 3 and the contraposition of Theorem 1.

Corollary 2. *Suppose that $\Theta^\pi \subseteq \Theta \subseteq \Theta^{DEY}$.*

(i) *If a pure strategy profile (a, a) is not efficient, then there exists $\bar{p} \in (0, 1)$ such that it is not stable in Θ for any $p \in (\bar{p}, 1]$.*

(ii) *If a pure strategy profile is not a Nash equilibrium of G, then there exists $\underline{p} \in (0, 1)$ such that it is not stable in Θ for any $p \in [0, \underline{p})$.*

Note that strong NSS can be weakly dominated strategy. The following example illustrates the case that weakly dominated strategy can be strong NSS.

Example 2. Consider the following two stage game. At the first stage, two players play the prisoner's dilemma game where there is two choices "C"(corporate) or "D" (defeat). At the second stage, a player can choose either "P" or "N" only if he/she chooses "C" and the opponent chooses "D" at the first stage. "P" stands for punishing the opponent and "N" stands for not punishing. If he/she chooses "P", then he/she should pay some cost in order to punish the opponent and the payoffs of both players decrease. If he/she chooses "N", both of payoff

[7]There is a slight difference between our notion of stability and that in DEY. However, the same logic in Dekel et al. (2007) is useful for proving the theorem.

do no change from the first stage. The game tree is in Figure 1 where the dotted lines describe the information sets and circled numbers indicate the players. The payoff matrix of the strategic form of the game is in Table 4. "CP" is the strategy where the player chooses "C" at the first stage and chooses "P" at the second stage if he/she is defeated. "CN" is the strategy where the player chooses "C" at the first stage and chooses "N" at the second stage. "D" is the strategy where the player chooses "D" at the first stage and does not have alternatives at the second stage.

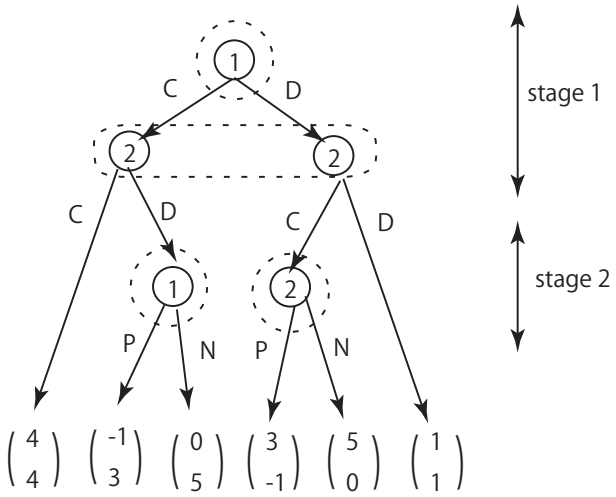

Figure 1: Prisoner's dilemma Game with punishment

In the game, "CP" is strong NSS. Thus, the outcome (CP,CP) is stable with dominant strategy in Θ^{π} for any observability from Theorem 2. Note that (CP,CP) can be (subjective) subgame perfect equilibrium from the point of view of incumbents even though it consists of (materially) weakly dominated strategy. On the other hand, (D,D) is not efficient, thus it is not stable in Θ^{DEY} from Theorem 3 (i) when p is large enough. Moreover, (CN,CN) is not Nash equilibrium, thus it is not stable in Θ^{DEY} from Theorem 3 (ii) when p is small enough.

	CN	D	CP
CN	4,4	0,5	4,4
D	5,0	1,1	3,-1
CP	4,4	-1,3	4,4

Table 4: Payoff matrix of prisoner's dilemma game with punishment

End of example

3 Two-player games with two roles

In this section, we consider two-player games with two roles. For example, the ultimatum game has two roles: a responder and a proposer.

Before players choose an action, nature determines the role of each player at random and lets them know their role. Players choose an action in the asymmetric games, depending on their roles. We have two interpretations of role assignment. First, at the beginning of each game, nature assigns roles. That is, they are assigned new roles when they play a new game. Second, when a player is born, the player is assigned a fixed role. This role does not change for their whole life, but the roles are independent of their parents' roles. For example, "male" and "female" are roles that do not change and are determined at random. Both interpretations are consistent with our model, as the evolution of preferences depends only on the average payoff of each type. As we see in Theorem 4, the profiles of a pure strong NSS are stable in Θ^{DEY}. However, the profiles of a pure strong NSS can be unstable if each role obtains a different payoff from the other. Let us denote a role by $\ell \in \{1, 2\}$. We denote the counterpart's role

against role ℓ by $-\ell$. That is, $-\ell = 1$ if $\ell = 2$ and $-\ell = 2$ if $\ell = 1$. The set of possible actions to role ℓ is A^ℓ. A pure strategy is an action plan for each role, i.e., (a^1, a^2) is a pure strategy for $a^1 \in A^1, a^2 \in A^2$. The set of possible pure strategies is a Cartesian product $A = A^1 \times A^2$. The set of mixed actions is $\Delta(A^\ell)$. The material payoff to role ℓ is $\pi^\ell(a^\ell, a^{-\ell})$ where a^ℓ and $a^{-\ell}$ are the actions of role ℓ and role $-\ell$, respectively. The expected payoff to $a = (a^1, a^2) \in A$ against $\tilde{a} = (\tilde{a}^1, \tilde{a}^2) \in A$ is $\pi(a, \tilde{a}) = (\pi^1(a^1, \tilde{a}^2) + \pi^2(a^2, \tilde{a}^1))/2$.

Consider the case where the realized outcome of a game is (a^1, a^2) where $a^1 \in A^1$ and $a^2 \in A^2$. Here, the utility of type θ is $\theta(a^1, a^2)$ if the player's role is 1, and it is $\theta(a^2, a^1)$ if the player's role is 2. That is, a preference θ is a function from $(A^1 \times A^2) \cup (A^2 \times A^1)$ to $[0, 1]$. In this section, Θ^{DEY} is the set of all of such functions. We also focus on the payoff-dependent preferences in this section. Given that the utility of these preferences is decided by the payoff profile of their own and the opponent's payoff, utility levels for the same payoff profiles are fixed, even if the role is different. Formally, the set of payoff-dependent preferences in this section is $\Theta^\pi \equiv \{\theta \in \Theta^{DEY} \mid \theta(\alpha, \beta) = \theta(\alpha', \beta')$ for $(\alpha, \beta) \in (A^\ell \times A^{-\ell}), (\alpha', \beta') \in (A^r \times A^{-r})$ if $\pi^\ell(\alpha, \beta) = \pi^r(\alpha', \beta')$ and $\pi^{-\ell}(\beta, \alpha) = \pi^{-r}(\beta', \alpha'),$ for all$\ell, r \in \{1, 2\}\}$. Note that role ℓ may differ from role r in the definition of Θ^π.

Example 3. We consider a simplified ultimatum game. In this game, there are two roles, a *proposer* and a *responder*, who must divide four resources. Table 5 shows the payoff matrix of the game. The column player, the proposer, decides how much of the resources he will offer to the other player, the responder. He has three options, "E", "F", and "G". "E" is the egoistic offer that is "I'll take 3, so you only obtain 1". "F" is the fair offer that is "Both of us obtain 2". "G" is the generous offer that is "I'll only take 1 and you obtain 3". The row player, the responder, has three options, "a", "r_E", and "r_F". "a" means that "I accept any offers". "r_E" means that "I reject the egoistic offer, but accept other offers". "r_F" means that "I reject both the fair offer and the egoistic offer, but accept the generous offer". That is, $A^{prop} = \{\text{E,F,G}\}$ and $A^{res} = \{\text{a}, r_E, r_F\}$ where "prop" and "res" are the proposer and responder, respectively.

If the responder accepts the proposer's offer, then they get exactly the same payoff as in the offer. If the responder rejects the proposer's offer, then both get nothing. For example, $\pi^{res}(a, E) = 1$ and $\pi^{prop}(E, a) = 3$. If θ is a payoff-dependent preference, $\theta(\text{a,E}) = \theta(\text{G,a})$ as both action profiles, she obtains 1 and her opponent obtains 3. **End of example**

We denote the strategy of type θ for role ℓ by b^ℓ_θ. A strategy of type θ is a pair of strategies for each role, denoted by $b_\theta = (b^1_\theta, b^2_\theta)$. In the same way as Section 2, we assume that b is a Bayesian–Nash equilibrium. That is, it satisfies for all $\theta, \theta' \in \Theta$ and $\ell \in \{1, 2\}$

$$b^\ell_\theta(\theta') \in BR_\theta \left(p b^{-\ell}_{\theta'}(\theta) + (1-p) b^{-\ell}_{\theta'}(\phi) \right)$$

$$b^\ell_\theta(\phi) \in BR_\theta \left(\sum_{\theta' \in C(\mu)} [p b^{-\ell}_{\theta'}(\theta) + (1-p) b^{-\ell}_{\theta'}(\phi)] \mu(\theta') \right)$$

where $\mu(\theta')$ is the population share of θ'. The definitions of $\Pi_\theta(\mu \mid b)$, $B_p(\mu)$, $B^\delta_p(\tilde{\mu} \mid \mu, b)$ and $N_\epsilon(\mu, \tilde{\theta})$ do not change from the model in Section 2.

When the set of possible preferences is Θ^{DEY}, we have a similar result to Theorem 2.

Theorem 4. *If $a^* \in A$ is a pure, strong NSS in a two-player game with two roles, then (a^*, a^*) is stable with a dominant strategy in Θ^{DEY} for any $p \in [0, 1]$.*

Note the last part of the above theorem. If the game concerned is a strategy expression for an extensive form, (a^*, a^*) can be a (subjective) subgame perfect equilibrium for incumbents because a^* is dominant for the incumbents. Moreover, as a strong NSS can be an (objective) weakly dominated strategy based on the payoff, a (objective) weakly dominated strategy can compose a (subjective) subgame pefect equilibrium.

Example 4. We consider the same simplified ultimatum game as Example 3. As discussed, an action plan for each role is a pure strategy in the game. For example, "Choose 'a' when I am a responder and choose 'E' when I am a proposer" is a strategy, denoted by "aE". In the same way, we denote each strategy with two letters of actions, e.g., "aF", "r_EF". Table 6 shows the payoff matrix of pure strategies. There are three strong NSS: "aE", "r_EF", and "r_FG". From Theorem 4, all are stable in Θ^{DEY} for any $p \in [0, 1]$. **End of example**

In Example 3, each role player obtains the same payoff if both players choose "r_EF". If the material payoff to each role is likewise the same, then a strong NSS is sufficient for a strategy profile to be stable in Θ^π.

	E	F	G
a	1,3	2,2	3,1
r_E	0,0	2,2	3,1
r_F	0,0	0,0	3,1

Table 5: Simplified ultimatum game

	aE	aF	aG	r_EE	r_EF	r_EG	r_FE	r_FF	r_FG
aE	2	2.5	3	0.5	1	1.5	0.5	1	1.5
aF	1.5	2	2.5	1.5	2	2.5	0.5	1	1.5
aG	1	1.5	2	1	1.5	2	1	1.5	2
r_EE	1.5	2.5	3	0	1	1.5	0	1	1.5
r_EF	1	2	2.5	1	2	2.5	0	1	1.5
r_EG	0.5	1.5	2	0.5	1.5	2	0.5	1.5	2
r_FE	1.5	1.5	3	0	0	1.5	0	0	1.5
r_FF	1	1	2.5	1	1	2.5	0	0	1.5
r_FG	0.5	0.5	2	0.5	0.5	2	0.5	0.5	2

Table 6: Payoff matrix of simplified ultimatum game

Theorem 5. *Let $a^* = (a^{1*}, a^{2*})$ be a pure, strong NSS. If $\pi^1(a^{1*}, a^{2*}) = \pi^2(a^{2*}, a^{1*})$, then (a^*, a^*) is stable with a dominant strategy in Θ^π.*

Example 5. We consider the same game as Example 3. As discussed, there are three strong NSS: "aE", "r_EF", and "r_FG". When both players choose "r_EF", both get payoff 2. From Theorem 5, (r_EF,r_EF) is stable in Θ^π. In the profiles of the other two strong NSS, each role obtains a different payoff from the other. Thus, we cannot identify from the theorem whether (aE,aE) and (r_FG,r_FG) are stable. **End of example**

When we consider payoff-dependent preferences, the role changes the stability condition drastically because the payoff dependency of preferences requires some constraint about the strategy of each role. We provide an example here. Consider the payoff matrix in Table 5. If (aE,aE) is stable in Θ^π with configuration (μ, b), then $\theta(\text{a}, \text{E}) \geq \theta(r_E, \text{E})$ for all $\theta \in C(\mu)$. As we assume θ is a payoff-dependent preference, it implies that the player with θ prefers the payoff profile (1,3) to (0,0) where the first coordinate of the payoff profile shows the payoff to the concerned player. This implies that $\theta(\text{G}, r_F) \geq \theta(\text{F}, r_F)$ and $\theta(\text{G}, r_F) \geq \theta(\text{E}, r_F)$. Consider a mutant with indifferent type θ^0. Suppose that the mutant chooses "E" when she is a proposer and r_F when she is a responder. If a player with θ knew that the opponent's preference is θ^0, then "aG" is the best response strategy. In this case, the expected material payoff to type θ^0 is asymptotic to $3p/2 + 3/2$ as the population share of mutants goes to zero. On the other hand, the expected payoff to incumbents is asymptotic to $3/2 + 1/2 = 2$. Thus, if $p > 2/3$, then the mutant can invade the population. In our discussion, the kind of payoff profile possible is important. We denote the set of possible material payoff profiles for the opponent player with role $-\ell$ chooses $a^{-\ell}$ by $\Xi^\ell(a^{-\ell}) \equiv \{(\pi^\ell(a^\ell, a^{-\ell}), \pi^{-\ell}(a^{-\ell}, a^\ell)) | a^\ell \in A^\ell\}$

The following theorem shows conditions for instability.

Theorem 6. *Consider a pure strategy $\alpha = (\alpha^1, \alpha^2) \in A^1 \times A^2$. There exists $\bar{p} \in (0, 1)$ such that the strategy profile (α, α) cannot be stable in Θ^π for any $p \in [\bar{p}, 1]$ if the following conditions are satisfied for $\ell \in \{1, 2\}$:*

(1) $\pi^\ell(\alpha^\ell, \alpha^{-\ell}) < \pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell)$,

(2) *there exists a strategy profile $(\beta^\ell, \beta^{-\ell})$ such that $\pi^\ell(\alpha^\ell, \alpha^{-\ell}) = \pi^{-\ell}(\beta^{-\ell}, \beta^\ell)$, $\pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell) = \pi^\ell(\beta^\ell, \beta^{-\ell})$*

(3) *If $\Xi^{-\ell}(\beta^\ell) \setminus (\Xi^\ell(\alpha^{-\ell})) \neq \emptyset$, then $\pi^\ell(\alpha^\ell, \alpha^{-\ell}) < \tilde{\pi}^\ell$ for all $(\tilde{\pi}^{-\ell}, \tilde{\pi}^\ell) \in \Xi^{-\ell}(\beta^\ell) \setminus \Xi^\ell(\alpha^{-\ell})$.*

Example 6. We consider the same game as Example 3.

We consider the stability of the pure strategy "aE". In the action profile (a,E), each role obtains a different payoff from the other. Consider that $\ell =$ "res", $-\ell =$ "pro" , $\alpha^\ell =$ a and $\alpha^{-\ell} =$ E. Then we have $\pi^\ell(\alpha^\ell, \alpha^{-\ell}) = \pi^{res}(\text{a}, \text{E}) = 1$ and $\pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell) = \pi^{prop}(\text{E}, \text{a}) = 3$. Thus, condition (1) in Theorem 6 is satisfied. Moreover, if

we put $\beta^{-\ell} = \mathrm{G}$ and $\beta^{\ell} = r_F$, then we have $\pi^{-\ell}(\beta^{-\ell}, \beta^{\ell}) = \pi^{prop}(\mathrm{G}, r_F) = 1 = \pi^{\ell}(\alpha^{\ell}, \alpha^{-\ell})$ and $\pi^{\ell}(\beta^{\ell}, \beta^{-\ell}) = \pi^{res}(r_F, \mathrm{G}) = 3 = \pi^{-\ell}(\alpha^{-\ell}, \alpha^{\ell})$. Thus, condition (2) in Theorem 6 is satisfied. Furthermore, $\Xi^{-\ell}(\beta^{\ell}) = \Xi^{prop}(r_F)$ is the set of the payoff profile the responder chooses r_F where the first coordinate of each element of the set is the proposer's payoff and the second coordinate is the responder's payoff. As shown in the row of r_E in Table 5, $\Xi^{-\ell}(\beta^{\ell}) = \Xi^{prop}(r_F) = \{(0,0),(1,3)\}$. On the other hand, $\Xi^{\ell}(\alpha^{-\ell}) = \Xi^{res}(\mathrm{E}) = \{(1,3),(0,0)\}$. Therefore, we have $\Xi^{-\ell}(\beta^{\ell}) \setminus \Xi^{\ell}(\alpha^{-\ell}) = \emptyset$. That is, condition (3) in Theorem 6 is satisfied. As shown, the strategy profile (aE,aE) satisfies all conditions of Theorem 6 and hence is not stable. In the same way, we obtain that (r_FG,r_FG) is not stable. **End of example**

4 Discussion

We do not confirm the robustness of the model with payoff-dependent preferences. We can extend this in two ways. First, we could place some function restrictions on the payoff profiles such that the function would be smooth and continuous. Second, we could consider a middle case between this paper and DEY. Most interest lies in the payoff profile, but perhaps both players can respond to the label of the actions.

5 Appendix

Lemma 1. *Suppose that $\Theta^S \setminus C(\mu)$ includes an indifferent type θ^0 and that $C(\mu) \subset \Theta^S \subset \Theta^L$. Then a configuration (μ, b) is stable in Θ^L if it is stable in Θ^S.*

Proof. We show that a configuration (μ, b) is not stable in Θ^S if a configuration (μ, b) is not stable in Θ^L. If $\Pi_{\theta}(\mu \mid b) \neq \Pi_{\theta'}(\mu \mid b)$ for $\exists \theta, \theta' \in C(\mu)$, then (μ, b) is not stable Θ^S. Because $\Theta^S \setminus C(\mu)$ includes an indifferent type, Θ^S (μ, b) is δ-robust in Θ^S if and only if it is δ-robust in Θ^L.

If a mutant $\tilde{\theta} \in \Theta^L$ invade into the population, then θ^0 can invade into the population copying the strategy. We show the detail. Consider an mutant $\tilde{\theta} \in \Theta^L$ which earn larger payoff than incumbents in the configuration $(\tilde{\mu}, \tilde{b})$ where $\tilde{\mu} \in N_{\epsilon}(\mu, \tilde{\theta})$ and $\tilde{b} \in B_p^0(\mu \mid b, \mu)$ or $\tilde{b} \in B_p^{\delta}(\tilde{\mu} \mid b, \mu)$.

Consider a configuration on Θ^S, (μ', b'), where mutants are indifferent type θ^0 and mimic the strategy of $\tilde{\theta}$ and incumbents use the same strategy in the $\tilde{b}$. Formally, (1)θ^0 mimic the strategy of $\tilde{\theta}$ in $\tilde{b}$ i.e. $\tilde{b}_{\tilde{\theta}}(\theta) = b'_{\theta^0}(\theta)$ for any $\theta \in C(\mu) \cup \{\phi\}$, (2)The other types adopt the same strategy, i.e. for all $\theta \in C(\mu)$, $\tilde{b}_{\theta}(\tilde{\theta}) = b'_{\theta}(\theta^0)$, $\tilde{b}_{\theta}(\theta') = b'_{\theta}(\theta')$ for $\theta' \in C(\mu) \cup \{\phi\}$ (3)The proportion of θ^0 in μ' is the same as the proportion of $\tilde{\theta}$ in μ, i.e. $\tilde{\mu}(\tilde{\theta}) = \mu'(\theta^0)$ and (4) The proportion of the other types in μ' is the same as the proportion in $\tilde{\mu}$, i.e. $\tilde{\mu}(\theta) = \mu'(\theta)$ for all $\theta \in C(\mu)$.

Since any strategies are best response strategies for indifferent type θ^0, b' defined above is Bayesian equilibrium in μ'. That is, if $B_p(\mu' \mid b, \mu) \neq \emptyset$ or $B_p(\mu' \mid b, \mu) = \emptyset$, then $b' \in B_p(\mu' \mid b, \mu)$ or $b' \in B_p^{\delta}(\mu' \mid b, \mu)$ respectively. Moreover, $\mu' \in N_{\epsilon}(\mu, \theta_0)$. Therefore, if $\tilde{\theta}$ earns larger payoff than incumbents, then θ^0 does and (μ, b) is not stable in Θ^S □

Theorem 1. *Let $\Theta^{\pi} \subseteq \Theta^S \subseteq \Theta^L \subseteq \Theta^{DEY}$. Then a strategy profile (σ^*, σ^*) is stable in Θ^L if the profile is stable in Θ^S.*

Proof. Θ^{π} includes infinite number of indifferent strategies. If (μ, b) is stable in Θ^S, then there exists a indifferent type $\theta^0 \in \Theta^S \setminus C(\mu)$. Thus, (μ, b) is stable in Θ^L from the Lemma 1 □

Lemma 2. *Let a^* be a pure, strong NSS. Then there exists $\bar{\epsilon}$ such that for any $\epsilon \in (0, \bar{\epsilon})$ and for any $\sigma, \sigma' \in \Delta$, the following inequality holds:*

$$(1-\epsilon)\{\pi(a^*, a^*) - \pi(\sigma, a^*)\} + \epsilon\{\pi(a^*, \sigma) - \pi(\sigma', \sigma')\} \geq 0$$

Proof. We denote the left-hand side of the above inequality by $L(\sigma, \sigma', \epsilon)$. That is,

$$L(\sigma, \sigma', \epsilon) = (1-\epsilon)\{\pi(a^*, a^*) - \pi(\sigma, a^*)\} + \epsilon\{\pi(a^*, \sigma) - \pi(\sigma', \sigma')\}$$

First, we consider the case where σ is a pure strategy. Let $\sigma = a \in A$.

If $\pi(a^*, a^*) - \pi(a, a^*) = 0$, then $\pi(a^*, a) - \pi(\sigma', \sigma') \geq 0$, i.e., $L(a, \sigma', \epsilon) \geq 0$, because a^* is a strong NSS. Moreover, we have $\pi(a^*, a^*) - \pi(a, a^*) \geq 0$ from the definition of a strong NSS. As a result, we only need to consider the case $\pi(a^*, a^*) - \pi(\sigma, a^*) > 0$. We denote by A' the set of actions whose payoff against a^* is less than $\pi(a^*, a^*)$. $A' = \{a \in A \mid \pi(a^*, a^*) > \pi(a, a^*)\}$. Let $d_1 = \min_{a \in A'} [\pi(a^*, a^*) - \pi(a, a^*)]$ and $d_2 = \min_{a \in A', \sigma' \in \Delta} [\pi(a^*, a) - \pi(\sigma', \sigma')]$.

From the definition of A', we have $d_1 > 0$. On the other hand, as action space is finite, we have $-\infty < d_2$. Thus, there exists $\bar{\epsilon}$ such that $(1-\epsilon)d_1 + \epsilon d_2 \geq 0$ for any $\epsilon < \bar{\epsilon}$. On the other hand, $L(a, \sigma', \epsilon) \geq (1-\epsilon)d_1 + \epsilon d_2$ for all $a \in A'$. Therefore, we have $L(a, \sigma', \epsilon) \geq 0$ for any $a \in A$.

We next consider the case where σ is a mixed strategy. Let $\sigma = \sum_i \sigma_i a_i$.

$$\begin{aligned} L(\sigma, \sigma', \epsilon) &= (1-\epsilon)\{\pi(a^*, a^*) - \sum_i \sigma_i \pi(a_i, a^*)\} + \epsilon\{\sum_i \sigma_i \pi(a^*, a_i) - \pi(\sigma', \sigma')\} \\ &= \sum_i \sigma_i \left[(1-\epsilon)\{\pi(a^*, a^*) - \pi(a_i, a^*)\} + \epsilon\{\pi(a^*, a_i) - \pi(\sigma', \sigma')\}\right] \end{aligned}$$

As we showed, $L(a_i, \sigma', \epsilon) = (1-\epsilon)\{\pi(a^*, a^*) - \pi(a_i, a^*)\} + \epsilon\{\pi(a^*, a_i) - \pi(\sigma', \sigma')\} \geq 0$. Therefore, we have $L(\sigma, \sigma', \epsilon) \geq 0$ for any $\epsilon < \bar{\epsilon}$. □

Theorem 2. *(a^*, a^*) is stable with a dominant strategy in Θ^π for any $p \in [0,1]$ if $a^* \in A$ is a pure, strong NSS.*

Proof. Suppose that for any $a, a' \in A$, (1) $\theta^*(a, a') = 1$ if there exists $\hat{a} \in A$ such that the payoff profile from (a, a') is the same as that from $(a^*, \hat{a})$, i.e., $\pi(a, a') = \pi(a^*, \hat{a})$ and $\pi(a', a) = \pi(\hat{a}, a^*)$ and (2) $\theta^*(a, a') = 0$ otherwise. In this case, θ^* is a payoff-dependent preference and a^* is a dominant strategy for θ^*.

Consider a configuration (μ, b) such that $C(\mu) = \{\theta^*\}$, and $b_{\theta^*}(\theta^*) = b_{\theta^*}(\phi) = a^*$. As we see in Lemma 1, we can assume that the mutant is of the indifferent type. Suppose that an indifferent mutant $\tilde{\theta}$ enters and the distribution of preferences becomes $\tilde{\mu} = (1-\epsilon)\mu + \epsilon\tilde{\theta}$. Given that a^* is dominant for incumbents, $B_p^0(\tilde{\mu} \mid \mu, b) \neq \emptyset$ and $B_p^0(\tilde{\mu} \mid \mu, b, \{\sigma_\theta\}_{\theta \in C(\mu) \cup \phi}) \neq \emptyset$. As $B_p^0(\tilde{\mu} \mid \mu, b) \subset B_p^\delta(\tilde{\mu} \mid \mu, b)$ and $B_p^0(\tilde{\mu} \mid \mu, b, \{\sigma_\theta\}_{\theta \in C(\mu) \cup \phi}) \subset B_p^\delta(\tilde{\mu} \mid \mu, b, \{\sigma_\theta\}_{\theta \in C(\mu) \cup \phi})$, we have $B_p^\delta(\tilde{\mu} \mid \mu, b) \neq \emptyset$ and $B_p^\delta(\tilde{\mu} \mid \mu, b, \{\sigma_\theta\}_{\theta \in C(\mu) \cup \phi}) \neq \emptyset$ for any δ.

For any $\tilde{b} \in B_p^0(\tilde{\mu} \mid \mu, b)$, we have $\tilde{b}_{\theta^*}(\theta^*) = \tilde{b}_{\theta^*}(\phi) = a^*$. Let the strategy of θ^* against $\tilde{\theta}$ be $s^*_{\tilde{\theta}} = \tilde{b}_{\theta^*}(\tilde{\theta})$, the aggregated strategy of $\tilde{\theta}$ against θ^* be $\tilde{\sigma}_{\theta^*} = p\tilde{b}_{\tilde{\theta}}(\theta^*) + (1-p)\tilde{b}_{\tilde{\theta}}(\phi)$ and the aggregated strategy of $\tilde{\theta}$ against themselves $\tilde{\sigma}_{\tilde{\theta}} = p\tilde{b}_{\tilde{\theta}}(\tilde{\theta}) + (1-p)\tilde{b}_{\tilde{\theta}}(\phi)$.

From the definition of θ^*, we have $\theta^*(a^*, \tilde{\sigma}_{\theta^*}) = 1$. Because $s^*_{\tilde{\theta}}$ is a best response strategy against $\tilde{\sigma}_{\theta^*}$, we also have $\theta^*(s^*_{\tilde{\theta}}, \tilde{\sigma}_{\theta^*}) = 1$. Let the support of $s^*_{\tilde{\theta}}$ and $\tilde{\sigma}_{\theta^*}$ be A' and $\tilde{A}$, respectively. Then we obtain $\theta^*(a', \tilde{a}) = 1$ for all $a' \in A', \tilde{a} \in \tilde{A}$. From the definition of θ^*, for each $a' \in A'$ and $\tilde{a} \in \tilde{A}$, there exists $a \in A$ such that $\pi(a', \tilde{a}) = \pi(a^*, a)$ and $\pi(\tilde{a}, a') = \pi(a, a^*)$. Thus, there exists $\hat{\sigma} \in \Delta$ such that $\pi(s^*_{\tilde{\theta}}, \tilde{\sigma}_{\theta^*}) = \pi(a^*, \hat{\sigma})$, $\pi(\tilde{\sigma}_{\theta^*}, s^*_{\tilde{\theta}}) = \pi(\hat{\sigma}, a^*)$.

$\Pi_\theta^p(\tilde{\mu} \mid \tilde{b}) = (1-\epsilon)\pi(a^*, a^*) + \epsilon\pi(ps^*_{\tilde{\theta}} + (1-p)a^*, \tilde{\sigma}_{\theta^*})$

$\Pi_{\tilde{\theta}}^p(\tilde{\mu} \mid \tilde{b}) = (1-\epsilon)\pi(\tilde{\sigma}_{\theta^*}, ps^*_{\tilde{\theta}} + (1-p)a^*) + \epsilon\pi(\tilde{\sigma}_{\tilde{\theta}}, \tilde{\sigma}_{\tilde{\theta}})$

$$\begin{aligned} \pi(ps^*_{\tilde{\theta}} + (1-p)a^*, \tilde{\sigma}_{\theta^*}) &= p\pi(s^*_{\tilde{\theta}}, \tilde{\sigma}_{\theta^*}) + (1-p)\pi(a^*, \tilde{\sigma}_{\theta^*}) && (3) \\ &= p\pi(a^*, \hat{\sigma}) + (1-p)\pi(a^*, \tilde{\sigma}_{\theta^*}) && (4) \\ &= \pi\left(a^*, p\hat{\sigma} + (1-p)\tilde{\sigma}_{\theta^*}\right) && (5) \end{aligned}$$

Let $\tilde{\sigma}^* = p\hat{\sigma} + (1-p)\tilde{\sigma}_{\theta^*}$. Then we have $\pi(ps^*_{\tilde{\theta}} + (1-p)a^*, \tilde{\sigma}_{\theta^*}) = \pi(a^*, \tilde{\sigma}^*)$. Similarly, we have $\pi(\tilde{\sigma}_{\theta^*}, ps^*_{\tilde{\theta}} + (1-p)a^*) = \pi(\tilde{\sigma}^*, a^*)$

$$\begin{aligned} &\Pi_\theta^p(\tilde{\mu} \mid \tilde{b}) - \Pi_{\tilde{\theta}}^p(\tilde{\mu} \mid \tilde{b}) \\ &\quad = (1-\epsilon)\left\{\pi(a^*, a^*) - \pi(\tilde{\sigma}^*, a^*)\right\} \\ &\qquad + \epsilon\left\{\pi(a^*, \tilde{\sigma}^*) - \pi(\tilde{\sigma}_{\tilde{\theta}}, \tilde{\sigma}_{\tilde{\theta}})\right\} \end{aligned}$$

From Lemma 2, we have $\bar{\epsilon}$ such that for any $\epsilon < \bar{\epsilon}$, $\Pi_\theta^p(\tilde{\mu} \mid \tilde{b}) \geq \Pi_{\tilde{\theta}}^p(\tilde{\mu} \mid \tilde{b})$. □

Theorem 4. *If $a^* \in A$ is a pure, strong NSS in a two-player game with two roles, then (a^*, a^*) is stable with a dominant strategy in Θ^{DEY} for any $p \in [0,1]$.*

Proof. $a^* = (a^{1*}, a^{2*})$ is the action profile of each role. Suppose that $\theta^*(a^{1*}, a^2) = \theta^*(a^{2*}, a^1) = 1$ for all $a^1 \in A^1$, $a^2 \in A^2$ and that $\theta^*(a^1, a^2) = 0$, $\theta^*(a^2, a^1) = 0$ otherwise.

In this case, a^{1*} and a^{2*} are (strictly) dominant strategies for type θ^*.

Consider a configuration (μ, b) such that $C(\mu) = \{\theta^*\}$, and $b_{\theta^*}(\theta^*) = b_{\theta^*}(\phi) = a^*$. Suppose that mutant $\tilde{\theta}$ enters and the distribution of preferences becomes $\tilde{\mu} = (1-\epsilon)\mu + \epsilon\tilde{\theta}$. For any $\tilde{b} \in B_p^0(\tilde{\mu} \mid \mu, b)$, we have $\tilde{b}_{\theta^*}(\theta^*) = \tilde{b}_{\theta^*}(\phi) = \tilde{b}_{\theta^*}(\tilde{\theta}) = a^*$. Let $\tilde{\sigma}_{\theta^*} = p\tilde{b}_{\tilde{\theta}}(\theta^*) + (1-p)\tilde{b}_{\tilde{\theta}}(\phi)$ and $\tilde{\sigma}_{\tilde{\theta}} = p\tilde{b}_{\tilde{\theta}}(\tilde{\theta}) + (1-p)\tilde{b}_{\tilde{\theta}}(\phi)$.

$\Pi_\theta^p(\tilde{\mu} \mid \tilde{b}) = (1-\epsilon)\pi(a^*, a^*) + \epsilon\pi(a^*, \tilde{\sigma}_{\theta^*})$

$$\Pi^p_{\tilde{\theta}}(\tilde{\mu} \mid \tilde{b}) = (1-\epsilon)\pi(\tilde{\sigma}_{\theta^*}, a^*) + \epsilon\pi(\tilde{\sigma}_{\tilde{\theta}}, \tilde{\sigma}_{\tilde{\theta}})$$

$$\begin{aligned}&\Pi^p_{\theta}(\tilde{\mu} \mid \tilde{b}) - \Pi^p_{\tilde{\theta}}(\tilde{\mu} \mid \tilde{b})\\ &\quad=(1-\epsilon)\left\{\pi(a^*, a^*) - \pi(\tilde{\sigma}_{\theta^*}, a^*)\right\} + \epsilon\left\{\pi(a^*, \tilde{\sigma}_{\theta^*}) - \pi(\tilde{\sigma}', \tilde{\sigma}')\right\}\end{aligned}$$

From Lemma 2, we have $\bar{\epsilon}$ such that for any $\epsilon < \bar{\epsilon}$, $\Pi^p_{\theta}(\tilde{\mu} \mid \tilde{b}) \geq \Pi^p_{\tilde{\theta}}(\tilde{\mu} \mid \tilde{b})$.

□

Lemma 3. *If σ^* is a strong NSS in a two-player game with two roles, then we have the following:*

(i) $\pi^1(\sigma^{1*}, \sigma^{2*}) \geq \pi^1(\sigma^1, \sigma^{2*})$ *for all* $\sigma^1 \in \Delta(A^1)$
$\pi^2(\sigma^{2*}, \sigma^{1*}) \geq \pi^2(\sigma^2, \sigma^{1*})$ *for all* $\sigma^2 \in \Delta(A^2)$.

(ii) $\pi^1(\sigma^{1*}, \sigma^2) \geq \pi^1(\sigma^{1*}, \sigma^{2*})$ *if* $\pi^2(\sigma^{2*}, \sigma^{1*}) = \pi^2(\sigma^2, \sigma^{1*})$
$\pi^2(\sigma^{2*}, \sigma^1) \geq \pi^2(\sigma^{2*}, \sigma^{1*})$ *if* $\pi^1(\sigma^{1*}, \sigma^{2*}) = \pi^1(\sigma^1, \sigma^{2*})$.

Proof. proof of (i)

As σ^* is a strong NSS, we have $\pi(\sigma^*, \sigma^*) \geq \pi(\sigma, \sigma^*)$ for all $\sigma \in A$. Inserting $\sigma = (\sigma^1, \sigma^{2*})$ into the inequality, we have the first line of (i). Similarly, inserting $\sigma = (\sigma^{1*}, \sigma^2)$ into the inequality, we have the second line of (i).

proof of (ii)

As σ^* is a strong NSS, we have the statement $\pi(\sigma^*, \sigma) \geq \pi(\sigma^*, \sigma^*)$ if $\pi(\sigma, \sigma^*) = \pi(\sigma^*, \sigma^*)$. Inserting $\sigma = (\sigma^{1*}, \sigma^2)$ into the statement, we have the first line of (ii). Similarly, inserting $\sigma = (\sigma^1, \sigma^{2*})$ into the statement, we have the second line of (ii).

□

Theorem 5. *Let $a^* = (a^{1*}, a^{2*})$ be a pure, strong NSS. If $\pi^1(a^{1*}, a^{2*}) = \pi^2(a^{2*}, a^{1*})$, then (a^*, a^*) is stable with a dominant strategy in Θ^π.*

Proof. Consider a type θ^* for which $a^* = (a^{1*}, a^{2*})$ is dominant. For example, θ^* satisfies $\theta^*(a^{1*}, a^2) = \theta^*(a^{2*}, a^1) = 1$ for all $a^1 \in A^1$, $a^2 \in A^2$. If a strategy profile has the same payoff profile as that where a player adopts the dominant strategy, then the (subjective) utility from the strategy profile should be equal to 1 from the property of payoff-dependent preferences. Because of this, we have $\theta^*(a^1, a^2) = 1$ if $\pi^1(a^1, a^2) = \pi^1(a^{1*}, \hat{a}^2)$ and $\pi^2(a^2, a^1) = \pi^2(\hat{a}^2, a^{1*})$ for $\exists \hat{a}^2 \in A^2$. Similarly, we have $\theta^*(a^2, a^1) = 1$ if $\pi^2(a^2, a^1) = \pi^2(a^{2*}, \hat{a}^1)$ and $\pi^1(a^1, a^2) = \pi^1(\hat{a}^1, a^{2*})$ for $\exists \hat{a}^1 \in A^1$. Moreover, we have the case where a payoff profile with one role assignment is the same as the other payoff profile with a different role assignment. That is, we have $\theta^*(a^1, a^2) = 1$ if $\pi^1(a^1, a^2) = \pi^2(a^{2*}, \hat{a}^1)$ and $\pi^2(a^2, a^1) = \pi^1(\hat{a}^2, a^{2*})$ for $\exists \hat{a}^1 \in A^1$. Similarly, we have $\theta^*(a^2, a^1) = 1$ if $\pi^2(a^2, a^1) = \pi^1(a^{1*}, \hat{a}^2)$ and $\pi^1(a^1, a^2) = \pi^2(\hat{a}^2, a^{1*})$ for $\exists \hat{a}^2 \in A^2$. We put zero for the other utility of profiles that do not satisfy the above conditions.

Consider a configuration (μ, b) such that $C(\mu) = \{\theta^*\}$, and $b_{\theta^*}(\theta^*) = b_{\theta^*}(\phi) = a^*$. We show that (μ, b) is stable in Θ^π. Obviously, the configuration satisfies stability conditions (1) and (2). Moreover, after any mutant entry, there exists a post-entry equilibrium where the incumbent θ^* keeps adopting a^* as a^* is dominant for θ^*. That is, (μ, b) is 0-robust. The remaining problem is whether the configuration satisfies stability condition (3).

Suppose that mutant $\tilde{\theta}$ enters and that the post-entry distribution of preferences is $\tilde{\mu} = (1-\epsilon)\mu + \epsilon\tilde{\theta}$. We now pay attention to the post-entry equilibrium $\tilde{b} \in B^0_p(\tilde{\mu} \mid \mu, b)$. For notational simplicity, let $\hat{\sigma} = \tilde{b}_{\theta^*}(\tilde{\theta})$, $\tilde{\sigma} = p\tilde{b}_{\tilde{\theta}}(\theta^*) + (1-p)\tilde{b}_{\tilde{\theta}}(\phi)$ and $\tilde{\sigma}' = p\tilde{b}_{\tilde{\theta}}(\tilde{\theta}) + (1-p)\tilde{b}_{\tilde{\theta}}(\phi)$.

We have:

$$\Pi^p_{\theta}(\tilde{\mu} \mid \tilde{b}) = (1-\epsilon)\pi(a^*, a^*) + \epsilon\pi(p\hat{\sigma} + (1-p)a^*, \tilde{\sigma}) \tag{6}$$

$$\Pi^p_{\tilde{\theta}}(\tilde{\mu} \mid \tilde{b}) = (1-\epsilon)\pi(\tilde{\sigma}, p\hat{\sigma} + (1-p)a^*) + \epsilon\pi(\tilde{\sigma}', \tilde{\sigma}') \tag{7}$$

From (6)−(7), we have:

$$\begin{aligned}&\Pi^p_{\theta}(\tilde{\mu} \mid \tilde{b}) - \Pi^p_{\tilde{\theta}}(\tilde{\mu} \mid \tilde{b})\\ &\quad=(1-\epsilon)\Big[p\left(\pi(a^*, a^*) - \pi(\tilde{\sigma}, \hat{\sigma})\right) + (1-p)\left(\pi(a^*, a^*) - \pi(\tilde{\sigma}, a^*)\right)\Big]\\ &\quad\quad+\epsilon\Big[p\left(\pi(\hat{\sigma}, \tilde{\sigma}) - \pi(\tilde{\sigma}', \tilde{\sigma}')\right) + (1-p)\left(\pi(a^*, \tilde{\sigma}) - \pi(\tilde{\sigma}', \tilde{\sigma}')\right)\Big]\end{aligned} \tag{8}$$

As $\hat{\sigma}$ is a best response strategy against $\tilde{\sigma}$ for type θ^*, $\theta^*(\hat{\sigma}^1, \tilde{\sigma}^2) = \theta^*(a^{1*}, \tilde{\sigma}^2) = 1$ and $\theta^*(\hat{\sigma}^2, \tilde{\sigma}^1) = \theta^*(a^{2*}, \tilde{\sigma}^1) = 1$. Let the set of support actions of $\hat{\sigma}$ be $\hat{A}^1 \times \hat{A}^2$. Then we have $\theta^*(\hat{a}^2, \tilde{\sigma}^1) = \theta^*(\hat{a}^1, \tilde{\sigma}^2) = 1$ for all $\hat{a}^1 \in \hat{A}^1$ and $\hat{a}^2 \in \hat{A}^2$.

Let $\bar{\pi}^* \equiv \pi^1(a^{1*}, a^{2*}) = \pi^2(a^{2*}, a^{1*}) = \pi(a^*, a^*)$.

From $\theta^*(\hat{a}^1, \tilde{\sigma}^2) = 1$ for $\hat{a}^1 \in \hat{A}$, we have one of the following two cases:

(a) $\pi^1(\hat{a}^1, \tilde{\sigma}^2) = \pi^1(a^{1*}, a^2)$ and $\pi^2(\tilde{\sigma}^2, \hat{a}^1) = \pi^2(a^2, a^{1*})$ for $\exists a^2 \in A^2$,
or

(b) $\pi^1(\hat{a}^1, \tilde{\sigma}^2) = \pi^2(a^{2*}, a^1)$ and $\pi^2(\tilde{\sigma}^2, \hat{a}^1) = \pi^1(a^1, a^{2*})$ for $\exists a^1 \in A^1$.

If (a) is satisfied, then we have $\bar{\pi}^* \geq \pi^2(\tilde{\sigma}^2, \hat{a}^1)$ from the second line of Lemma 3 (i). If (b) is satisfied, then we also have $\bar{\pi}^* \geq \pi^2(\tilde{\sigma}^2, \hat{a}^1)$ from the first line of Lemma 3 (i). That is, we have $\bar{\pi}^* \geq \pi^2(\tilde{\sigma}^2, \hat{a}^1)$ for all $\hat{a}^1 \in \hat{A}^1$ in any cases. Because the inequalities hold for all pure strategies composing $\hat{\sigma}^1$, we have $\bar{\pi}^* \geq \pi^2(\tilde{\sigma}^2, \hat{\sigma}^1)$. Similarly, we have $\bar{\pi}^* \geq \pi^1(\tilde{\sigma}^1, \hat{\sigma}^2)$ from $\theta^*(\hat{a}^2, \tilde{\sigma}^1) = 1$. From these two inequalities, we conclude $\bar{\pi}^* \geq \pi(\tilde{\sigma}, \hat{\sigma})$.

On the other hand, we have $\bar{\pi}^* \geq \pi(\tilde{\sigma}, a^*)$ from the definition of a strong NSS. Therefore, we have two cases to be considered.

Case(i) $\bar{\pi}^* > \pi(\tilde{\sigma}, \hat{\sigma})$ or $\bar{\pi}^* > \pi(\tilde{\sigma}, a^*)$

In this case, we have $p\left(\pi(a^*, a^*) - \pi(\tilde{\sigma}, \hat{\sigma})\right) + (1-p)\left(\pi(a^*, a^*) - \pi(\tilde{\sigma}, a^*)\right) > 0$, i.e., the first term of equation 8 is positive. Thus, using a similar discussion to Lemma 2, there exists a $\bar{\epsilon}$ such that for all $\epsilon < \bar{\epsilon}$, $\Pi_\theta^p(\tilde{\mu} \mid \tilde{b}) \geq \Pi_{\tilde{\theta}}^p(\tilde{\mu} \mid \tilde{b})$

Case(ii) $\bar{\pi}^* = \pi(\tilde{\sigma}, \hat{\sigma})$ and $\bar{\pi}^* = \pi(\tilde{\sigma}, a^*)$.

Because a^* is a strong NSS, we have $\pi(a^*, \tilde{\sigma}) \geq \pi(\tilde{\sigma}', \tilde{\sigma}')$ from $\bar{\pi}^* = \pi(\tilde{\sigma}, a^*)$ and the second condition of a strong NSS. We show that $\pi(\hat{\sigma}, \tilde{\sigma}) \geq \pi(\tilde{\sigma}', \tilde{\sigma}')$ from $\bar{\pi}^* = \pi(\tilde{\sigma}, \hat{\sigma})$ and hence Equation (8) is nonnegative in the following.

As we have shown, $\bar{\pi}^* \geq \pi^2(\tilde{\sigma}^2, \hat{a}^1)$ and $\bar{\pi}^* \geq \pi^1(\tilde{\sigma}^1, \hat{\sigma}^2)$, we have $\pi^1(\tilde{\sigma}^1, \hat{\sigma}^2) = \pi^2(\tilde{\sigma}^2, \hat{\sigma}^1) = \bar{\pi}^*$ from $\bar{\pi}^* = \pi(\tilde{\sigma}, \hat{\sigma})$. As discussed, we have two cases, (a) and (b), for $\hat{a}^1 \in \hat{A}^1$ from $\theta^*(\hat{a}^1, \tilde{\sigma}^2) = 1$.
If (a) is satisfied, then we have $\pi^1(\hat{a}^1, \tilde{\sigma}^2) = \pi^1(a^{1*}, a^2) \geq \bar{\pi}^*$ from $\pi^2(\tilde{\sigma}^2, \hat{a}^1) = \pi^2(a^2, a^{1*}) = \bar{\pi}^*$ and Lemma 3 (ii).

If (b) is satisfied, then we also have $\pi^1(\hat{a}^1, \tilde{\sigma}^2) = \pi^2(a^{2*}, a^1) \geq \bar{\pi}^*$ from $\pi^2(\tilde{\sigma}^2, \hat{a}^1) = \pi^1(a^1, a^{2*}) = \bar{\pi}^*$ and Lemma 3 (ii).

That is, in any case, we have $\pi^1(\hat{a}^1, \tilde{\sigma}^2) \geq \bar{\pi}^*$ for all $\hat{a}^1 \in \hat{A}^1$. Thus, $\pi^1(\hat{\sigma}^1, \tilde{\sigma}^2) \geq \bar{\pi}^*$. In the same way, we have $\pi^2(\hat{\sigma}^2, \tilde{\sigma}^1) \geq \bar{\pi}^*$. Therefore, we conclude $\pi(\hat{\sigma}, \tilde{\sigma}) \geq \bar{\pi}^*$. As a^* is a strong NSS, $\bar{\pi}^* \geq \pi(\tilde{\sigma}', \tilde{\sigma}')$. Finally we have $\pi(\hat{\sigma}, \tilde{\sigma}) \geq \pi(\tilde{\sigma}', \tilde{\sigma}')$ if $\pi(\tilde{\sigma}, \hat{\sigma}) = \bar{\pi}^*$.

□

Theorem 6. *Consider a pure strategy $\alpha = (\alpha^1, \alpha^2) \in A^1 \times A^2$. There exists $\bar{p} \in (0,1)$ such that the strategy profile (α, α) cannot be stable in Θ^π for any $p \in [\bar{p}, 1]$ if the following conditions are satisfied for $\ell \in \{1,2\}$:*

(1) $\pi^\ell(\alpha^\ell, \alpha^{-\ell}) < \pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell)$,

(2) *there exists a strategy profile $(\beta^\ell, \beta^{-\ell})$ such that $\pi^\ell(\alpha^\ell, \alpha^{-\ell}) = \pi^{-\ell}(\beta^{-\ell}, \beta^\ell)$, $\pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell) = \pi^\ell(\beta^\ell, \beta^{-\ell})$,*

(3) *If $\Xi^{-\ell}(\beta^\ell) \setminus (\Xi^\ell(\alpha^{-\ell})) \neq \emptyset$, then $\pi^\ell(\alpha^\ell, \alpha^{-\ell}) < \tilde{\pi}^\ell$ for all $(\tilde{\pi}^{-\ell}, \tilde{\pi}^\ell) \in \Xi^{-\ell}(\beta^\ell) \setminus \Xi^\ell(\alpha^{-\ell})$.*

Proof. Suppose that (μ, b) induces a pure strategy profile $(\alpha^\ell, \alpha^{-\ell})$. That is, for all $\theta, \theta' \in C(\mu)$, $b_\theta^\ell(\theta') = b_\theta^\ell(\phi) = \alpha^\ell$ and $b_\theta^{-\ell}(\theta') = b_\theta^{-\ell}(\phi) = \alpha^{-\ell}$. We assume that supporting μ are the payoff-dependent preferences, i.e., $C(\mu) \subset \Theta^\pi$. Then, there exists a function $v_\theta : \mathbb{R} \times \mathbb{R} \to [0,1]$ such that $v_\theta(\pi^r(a^r, a^{-r}), \pi^{-r}(a^{-r}, a^r)) = \theta(a^r, a^{-r})$ for all $a^r \in A^r$ and $a^{-r} \in A^{-r}$ for all $r \in \{1,2\}$. We consider the case where mutants are an indifferent type θ^0 and the post-entry distribution is $\tilde{\mu} = (1-\epsilon)\mu + \epsilon\theta^0$.

Assume to the contrary (μ, b) is stable in Θ^π. Then, it is δ-robust for any $\delta > 0$. Thus, there exists $\bar{\epsilon} > 0$ such that for $\forall \epsilon \in (0, \bar{\epsilon})$, $B_p^\delta(\tilde{\mu} \mid \mu, b) \neq \emptyset$. As the payoff is finite, there exists a function $\gamma(\epsilon, \delta)$ such that for all $\tilde{b} \in B_p^\delta(\tilde{\mu} \mid \mu, b)$, $|\Pi_\theta(\mu \mid b) - \Pi_\theta(\tilde{\mu} \mid \tilde{b})| < \gamma(\epsilon, \delta)$ where $\epsilon \in (0, \bar{\epsilon})$ and $\gamma(\epsilon, \delta) \to 0$ as $\epsilon \to 0$ and $\delta \to 0$. As $\Pi_\theta(\mu \mid b) = (\pi^\ell(\alpha^\ell, \alpha^{-\ell}) + \pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell))/2$. We have:

$$|(\pi^\ell(\alpha^\ell, \alpha^{-\ell}) + \pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell))/2 - \Pi_\theta(\tilde{\mu} \mid \tilde{b})| < \gamma(\delta, \epsilon). \tag{9}$$

We assume that (1)$\pi^\ell(\alpha^\ell, \alpha^{-\ell}) < \pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell)$ (2) $(\beta^\ell, \beta^{-\ell})$ satisfies $\pi^\ell(\alpha^\ell, \alpha^{-\ell}) = \pi^{-\ell}(\beta^{-\ell}, \beta^\ell)$ and $\pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell) = \pi^\ell(\beta^\ell, \beta^{-\ell})$. We have two cases to be considered.

Case (i) $\Xi^{-\ell}(\beta^\ell) \setminus \Xi^\ell(\alpha^{-\ell}) = \emptyset$ (i.e., $\Xi^{-\ell}(\beta^\ell) \subseteq \Xi^\ell(\alpha^{-\ell})$)

For any $\theta \in C(\mu)$, we have

$$(\pi^\ell(\alpha^\ell, \alpha^{-\ell}), \pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell)) \in \underset{(\pi_1,\pi_2)\in\Xi^\ell(\alpha^{-\ell})}{\operatorname{argmax}} v_\theta(\pi_1, \pi_2)$$

Because $(\pi^\ell(\alpha^\ell, \alpha^{-\ell}), \pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell)) = (\pi^{-\ell}(\beta^{-\ell}, \beta^\ell), \pi^\ell(\beta^\ell, \beta^{-\ell}))$ and $(\pi^{-\ell}(\beta^{-\ell}, \beta^\ell), \pi^\ell(\beta^\ell, \beta^{-\ell})) \in \Xi^{-\ell}(\beta^\ell) \subseteq \Xi^\ell(\alpha^{-\ell})$, we have

$$(\pi^{-\ell}(\beta^{-\ell}, \beta^\ell), \pi^\ell(\beta^\ell, \beta^{-\ell})) \in \underset{(\pi_1,\pi_2)\in\Xi^{-\ell}(\beta^\ell)}{\operatorname{argmax}} v_\theta(\pi_1, \pi_2)$$

Consider the case where mutants are an indifferent type θ^0 and always adopt a strategy $(\beta^\ell, \alpha^{-\ell})$. We now assume that incumbents' strategy against mutants is $(\alpha^\ell, \beta^{-\ell})$. In this case, the mutant's payoff is $\Pi_{\theta^0}(\tilde{\mu} \mid \tilde{b}) \geq (1-\epsilon)p\left\{\pi^\ell(\beta^\ell, \beta^{-\ell}) + \pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell)\right\}/2 + (1-\epsilon)(1-p)\underline{\pi} + \epsilon\underline{\pi}$ where $\underline{\pi}$ is the minimum payoff in the material payoff matrix, i.e., $\underline{\pi} = \left(\min_{\{a_1\in A^1, a_2\in A_2\}} \pi^1(a_1, a_2) + \min_{\{a_1\in A^1, a_2\in A_2\}} \pi^2(a_2, a_1)\right)/2$.

On the other hand, the incumbents' payoffs are $\Pi_\theta(\tilde{\mu} \mid \tilde{b}) \leq \left\{\pi^\ell(\alpha^\ell, \alpha^{-\ell}) + \pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell)\right\}/2 + \gamma(\epsilon, \delta)$ As $\left\{\pi^\ell(\beta^\ell, \beta^{-\ell}) + \pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell)\right\}/2 > \left\{\pi^\ell(\alpha^\ell, \alpha^{-\ell}) + \pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell)\right\}/2$, we have $\Pi_{\theta^0}(\tilde{\mu} \mid \tilde{b}) > \Pi_\theta(\tilde{\mu} \mid \tilde{b})$ for large enough p and small ϵ, δ. Thus, there exists $\bar{p} \in (0,1)$ such that for all $p \in [\bar{p}, 1]$, $\forall\bar{\epsilon}$, there exists $\epsilon < \bar{\epsilon}$ and $\delta > 0$ such that $\Pi_{\theta^0}(\tilde{\mu} \mid \tilde{b}) > \Pi_\theta(\tilde{\mu} \mid \tilde{b})$. This implies that (μ, b) is unstable.

Case(ii) $\Xi^{-\ell}(\beta^\ell) \setminus \Xi^\ell(\alpha^{-\ell}) \neq \emptyset$. In this case, $\pi^\ell(\alpha^\ell, \alpha^{-\ell}) < \tilde{\pi}^\ell$ for all $(\tilde{\pi}^{-\ell}, \tilde{\pi}^\ell) \in \Xi^{-\ell}(\beta^\ell) \setminus \Xi^\ell(\alpha^{-\ell})$ from Condition (3) of this theorem.

Case (ii)-(a) $(\pi^{-\ell}(\beta^{-\ell}, \beta^\ell), \pi^\ell(\beta^\ell, \beta^{-\ell})) \in \operatorname{argmax}_{(\pi_1,\pi_2)\in\Xi^{-\ell}(\beta^\ell)} v_\theta(\pi_1, \pi_2)$
In this case, we have the same discussion in Case (i).

Case (ii)-(b) $(\pi^{-\ell}(\beta^{-\ell}, \beta^\ell), \pi^\ell(\beta^\ell, \beta^{-\ell})) \notin \operatorname{argmax}_{(\pi_1,\pi_2)\in\Xi^{-\ell}(\beta^\ell)} v_\theta(\pi_1, \pi_2)$
In this case, for any $(\tilde{\pi}^{-\ell}, \tilde{\pi}^\ell) \in \operatorname{argmax}_{(\pi_1,\pi_2)\in\Xi^{-\ell}(\beta^\ell)} v_\theta(\pi_1, \pi_2)$, we have $\pi^\ell(\alpha^\ell, \alpha^{-\ell}) < \tilde{\pi}^\ell$.
The mutants' material payoff is $\Pi_{\theta^0}(\tilde{b} \mid \tilde{\mu}) \geq (1-\epsilon)p\left\{\tilde{\pi}^\ell + \pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell))\right\}/2 + (1-\epsilon)(1-p)\underline{\pi} + \epsilon\underline{\pi}$.
As $\left\{\tilde{\pi}^\ell + \pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell))\right\}/2 > \left\{\pi^\ell(\alpha^\ell, \alpha^{-\ell}) + \pi^{-\ell}(\alpha^{-\ell}, \alpha^\ell)\right\}/2$, we have $\Pi_{\theta^0}(\tilde{b} \mid \tilde{\mu}) > \Pi_\theta(\tilde{b} \mid \tilde{\mu})$ for large enough p and small ϵ, δ. This implies that (μ, b) is unstable.

As shown, in any case we have mutants that earn larger payoffs than the incumbents do. □

References

Bester, Helmut and Werner Güth, "Is altruism evolutionarily stable?," *Journal of Economic Behavior & Organization*, 1998, *34*, 193–209.

Dekel, Eddie, Jeffrey C. Ely, and Okan Yilankaya, "Evolution of Preferences," *Review of Economic Studies*, 07 2007, *74* (3), 685–704.

Güth, W. and M. Yaari, "Explaining Reciprocal Behavior in a Simple Strategic Game," in U. Witt, ed., *Explaining Process and Change-Approaches to Evolutionary Economics*, University of Michigan Press, 1992, part 3, pp. 23–34.

Güth, Werber, "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives," *International Journal of Game Theory*, 1995, *24*, 323–344.

Herold, Florian and Christoph Kuzmics, "Evolution of Preferences Under Perfect Observability: Almost Anything is Stable," 2008.

Samuelson, Larry, "Introduction to the Evolution of Preferences," *Journal of Economic Theory*, 2001, *97*, 225–230.

Sethi, Rajiv and E. Somanathan, "Preference Evolution and Reciprocity," *Journal of Economic Theory*, 2001, *97*, 273–297.