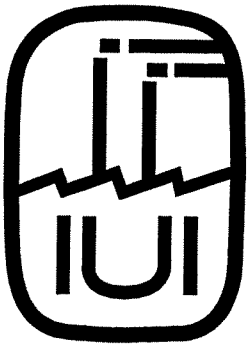

THE INDUSTRIAL INSTITUTE FOR ECONOMIC AND SOCIAL RESEARCH

WORKING PAPER No. 450, 1996

NEUTRALLY STABLE OUTCOMES IN CHEAP TALK GAMES

BY ABHIJIT BANERJEE AND JÖRGEN W. WEIBULL

Neutrally Stable Outcomes in Cheap Talk Games[◇]

ABHIJIT BANERJEE[*]
DEPARTMENT OF ECONOMICS, M.I.T
AND
JÖRGEN W. WEIBULL[†]
DEPARTMENT OF ECONOMICS,
STOCKHOLM SCHOOL OF ECONOMICS

March 5, 1996

ABSTRACT. This paper examines equilibrium and stability in symmetric two-player cheap-talk games. In particular, we characterize the set of neutrally stable outcomes in finite cheap-talk 2×2 coordination games. This set is finite and functionally independent of risk-dominance relations. As the number of messages goes to infinity, this set expands to a countable limit set that has exactly one cluster point, the Pareto efficient Nash equilibrium payoff. In contrast, the set of outcomes that are strategically stable for some finite message set is shown to be dense in the interval spanned by the Nash equilibrium payoffs of the game. We also show that the limit set of neutrally stable outcomes coincides with the set of neutrally stable outcomes for countably infinite message sets. Doc: *cheap5.tex*

◇ We thank Jonas Björnerstedt, David Cooper, Martin Dufwenberg, Glenn Ellison, Drew Fudenberg, Joseph Harrington, Vijay Krishna, Peter Norman and Asher Wolinsky for helpful conversations in connection with earlier drafts of this paper.

* Banerjee thanks the Institute for International Economic Studies, Stockholm University, for its hospitailty during part of this research project.

† The research of Weibull was sponsored by the Industrial Institute for Economic and Social Research, Stockholm, Sweden. He thanks the Institute for Advanced Studies, Vienna, for its hospitailty during part of this research project.

1. Introduction

For some years now game theorists have looked for evolutionary criteria to select among Nash equilibria. One strand of this literature strives to select among strict Nash equilibria. The prototype game is then a 2×2 coordination game with two strict Nash equilibria and one mixed Nash equilibrium. In this spirit Kandori Mailath and Rob [7], Young [24], and Kandori and Rob [8] argue that certain stochastic dynamics - that can be interpreted as processes of social evolution or learning - select the risk dominant equilibrium.[1] Another strand of the literature instead strives to select among the plethora of non-strict Nash equilibria that many games throw out. Researchers then work with static evolutionary criteria such as neutral and evolutionary stability or with such dynamic population models as the replicator dynamics. These latter criteria cannot select among strict equilibria, however. In 2×2 coordination games, the only effect is to reject the mixed Nash equilibrium. However, Wärneryd [21], Robson [16], Matsui [11], Kim and Sobel [9], Bhaskar [3] and Schlag [17], [18] have suggested that if we extend such coordination games to include a pre-play communication stage, then these evolutionary criteria can select among the strict equilibria of the underlying game, usually in favour of the Pareto efficient equilibrium.

The present paper is a contribution to this latter line of research. The setting is standard: there is a symmetric and finite two-player "base game" to be played after a pre-play communication session. Communication takes the form of costlessly and simultaneously sent messages, one from each player. These messages are selected from a finite set of possible messages, and the sent messages are observed without error by both players before they select a strategy in the base game. A pure strategy in this "meta-game" is thus a message to send combined with a "decision rule" that prescribes a pure base-game strategy for every message received from the other player. The main purpose of this study is to obtain a clearer picture of the cutting power of the criterion of neutral stability in cheap talk games, in particular in comparison with the criterion of strategic stability.

Neutral stability is the weakest static evolutionary refinement of the Nash equilibrium concept, and strategic stability is among its most stringent rationalistic refinements. A mixed or pure meta-game strategy is *neutrally stable* (Maynard Smith [12]) if it is a best reply to itself and, moreover, is a weakly better reply to all other best replies than these are to themselves. By a *neutrally stable outcome* we mean a payoff value that arises when some neutrally stable meta-game strategy meets itself. Neutral stability is formally a slight weakening of the evolutionary stability criterion: a strategy is evolutionarily stable if it is a best reply to itself and, moreover, is a

[1] Bergin and Lipman [1] show that this inference is sensitive to the assumptions made concerning the relative magnitude of mutation rates.

strictly better reply to all other best replies than these are to themselves (Maynard Smith and Price [13]). Intuitively, neutral stability allows for the possibility of drift, so that if there is a small shock to the population's behavior, the outcome may change slightly, and hence a series of such shocks may eventually trigger a motion far away from the initial state. However, it may take a very long time before such a motion begins, and in the mean-time the outcome may remain constant, so neutrally stable outcomes may be highly relevant for medium-term predictions (see e.g. Binmore and Samuelson [2] for a similar argument). In contrast, a set of Nash equilibria is *strategically stable* (Kohlberg and Mertens [10]) if it is minimal with respect to the property of being robust against all small trembles in strategies. Strategic stability has been shown to have a number of important implications from a rationalistic viewpoint (see Kohlberg and Mertens [10] or van Damme [5]). A *strategically stable outcome* of a cheap-talk game is here defined as a payoff that arises in some strategically stable set of symmetric meta-game Nash equilibria.

The analysis presented here builds on a straight-forward characterization of symmetric Nash equilibria in symmetric two-player cheap-talk games with arbitrary message sets. We show that the associated set of equilibrium payoffs increases with the number of messages available towards a limit set that is dense in the symmetric convex hull of Nash equilibrium payoffs in the base game. Moreover, for the special case of 2×2 coordination games we show that any payoff value between the worst and best Nash equilibrium payoffs can be approximated by a strategically stable meta-game outcome when the message set is sufficiently large. In this sense, strategic stability - albeit a stringent refinement of the Nash equilibrium concept - has virtually no cutting power in such cheap-talk coordination games.

The picture is quite different for neutral stability. First, the set of neutrally stable outcomes in a symmetric two-player base game need not be monotonically increasing with the number of messages available. Nevertheless, the set of neutrally stable outcomes converges to a limit set as the number of available messages tends to infinity. In the case of a 2×2 coordination game we characterize the set of neutrally stable meta-game outcomes for every finite message set. This set is finite and contains both strict Nash equilibrium payoffs. Indeed, letting the number of messages increase toward infinity, the set of neutrally stable outcomes converges to a countable limit set; if one normalizes the payoffs in the coordination game so that the "good" strict Nash equilibrium payoff is 2 and that of the "bad" strict Nash equilibrium is 1, then this limit set consists of the numbers 1, $1+\frac{1}{2}$, $1+\frac{3}{4}$, ..., $1+\frac{n-1}{n}$, 2. In other words, the set of neutrally stable outcomes contains an infinite number of isolated points between the "bad" and the "good" Nash equilibrium outcomes. Neutral stability in this game therefore gets rid of most of the strategically stable outcomes but at the same time it does admit certain specific convex combinations of the extreme Nash equilibria.

Furthermore, these results concerning the neutrally stable outcomes are independent of risk-dominance properties of the underlying coordination game (payoffs off the diagonal of the payoff matrix play no role). In this sense, neutral stability offers a selection from the set of Nash equilibria which is distinct from those based on Pareto dominance, risk dominance and strategic stability considerations, and which reflects more directly the specific logic of evolution in games.[2]

The results reported above hold when the set of messages is finite. However, in any natural language the set of possible statements is countably infinite. Hence, the above (conventional) assumption that the message set be finite is not self-evident. Of course, one may claim that in any real life interaction there is an upper finite bound on the length of statements that can be made, and hence, since the numbers of signs in any natural language is finite, the set of messages that are effectively available is finite. However, such a finite upper bound may not be common knowledge to all participants, and hence an infinite message set may be more appropriate. It is well known from the repeated games literature that the equilibrium correspondence may be "discontinuous at infinity," i.e., there may be a whole plethora of infinite-horizon outcomes that have no counterpart in the long but finite horizon case (cf. the Folk theorems). An important question for the present study thus is whether also the set of neutrally stable outcomes in cheap talk games is discontinuous "at infinity" in this sense. It turns out that, at least in 2×2 coordination games, this is not the case: The limit set of neutrally stable outcomes for large but finite message sets coincides with the set of neutrally stable outcomes for any countably infinite message set.

The present study can be viewed as an extension of Wärneryd [21] from pure-strategy analysis to mixed-strategy analysis and from finite message sets to (finite and infinite) countable message sets. That paper appears to be the first to point out implications of evolutionary stability properties for social efficiency in cheap-talk coordination games. In particular, Wärneryd showed that no *pure* meta-strategy is evolutionarily stable if the base game is a 2×2 coordination game and there is more than one message. Moreover, he showed that the only outcome compatible with neutral stability in pure meta strategies is the Pareto efficient Nash equilibrium payoff of the coordination game. In contrast, we here allow for mixed strategies, and show that other neutrally and evolutionarily stable outcomes exist. Another related paper is Wärneryd [22], where it is shown that any convex combination of base-game Nash equilibria can be approximated by some meta-game cheap-talk Nash equilibrium if the messages space is sufficiently large.[3] This result follows from the second claim in our

[2]The only cluster point of the set of neutrally stable outcomes is the Pareto dominant Nash equilibrium outcome. As a result one may argue that our results at least weakly favor Pareto dominance over risk dominance.

[3]An observation similar to Wärenryd's is also made in Kim and Sobel [9].

proposition 2. A third related paper is Schlag [17], where it is shown that every finite cheap-talk 2×2 coordination game has exactly one evolutionarily stable strategy, and one outcome that is obtained for a whole set of neutrally stable strategies that together constitute a so-called evolutionarily stable set, i.e., a set of neutrally stable strategies, where each strategy earns a higher payoff against all nearby strategies outside the set than these earn against themselves (Thomas [20]). When played against itself, the unique evolutionarily strategy earns a payoff that lies between that of the "bad" and "good" strict equilibria of the underlying coordination game, and this payoff approaches the "good" payoff as the number of messages increases (see Remark 5 in section 5 below).[4] All strategies in the evolutionarily stable set earn the "good" payoff against each other.

The material is organized as follows. Definitions and preliminaries are given in section 2, symmetric meta-game Nash equilibria are characterized in section 3, and meta-game outcomes are analyzed in section 4. Section 5 characterizes the set of neutrally stable outcomes in 2×2 coordination games. Section 6 extends the results to countably infinite message sets, and section 7 concludes.

2. Definitions and Preliminaries

2.1. Symmetric Two-Player Games. The analysis in the present paper is focused on finite and symmetric two-player games in normal form. Let $S = \{1, 2, ..., n\}$ be the set of pure strategies (the same for both players). Accordingly, a *mixed strategy* is a point σ on the $(n-1)$-dimensional unit simplex $\Delta(S) = \{\sigma \in \mathbb{R}^n_+ : \sum_i \sigma_i = 1\}$ in $\mathbb{R}^n$. The *support* of a mixed strategy $\sigma \in \Delta(S)$ is the subset $C(\sigma) = \{i \in S : \sigma_i > 0\}$ of pure strategies which are assigned positive probabilities. The set of *strategy profiles* will be denoted $\Theta(S) = \Delta(S) \times \Delta(S)$. This is a subset of $\mathbb{R}^{2n}$.

Let a_{ij} be the *payoff* to pure strategy i when played against pure strategy j, and let A be the associated $k \times k$ payoff matrix. Accordingly, the (expected) payoff of a mixed strategy σ when played against a mixed strategy μ is $u(\sigma, \mu) = \sigma \cdot A\mu = \sum_{ij} \sigma_i a_{ij} \mu_j$. The payoff function $u : \mathbb{R}^{2n} \to \mathbb{R}$ so defined is bi-linear, and the payoff to a pure strategy i when played against a mixed strategy μ is $u(e^i, \mu)$, where $e^i \in \Delta(S)$ is the i'th unit vector in $\mathbb{R}^n$. A finite and symmetric 2-player normal-form game will be summarized as a pair $G = (S, u)$.

A *best reply* to a strategy $\mu \in \Delta(S)$ is a strategy $\sigma \in \Delta(S)$ such that $u(\sigma, \mu) \geq u(\sigma', \mu)$ $\forall \sigma' \in \Delta(S)$. For each $\mu \in \Delta(S)$, let $\beta(\mu) \subset \Delta(S)$ be its set of (mixed) best replies. A *Nash equilibrium* is a pair $(\sigma, \mu) \in \Theta(S)$ of mutually best replies; $\sigma \in \beta(\mu)$ and $\mu \in \beta(\sigma)$. A Nash equilibrium (σ, μ) is *strict* if each strategy is the unique best reply to the other. A Nash equilibrium $(\sigma, \mu) \in \Theta(S)$ is *strictly perfect*

[4] An example of this sort was given in Kim and Sobel [9].

if it is robust to all small "trembles" (Okada [14]).[5] A strictly perfect equilibrium is *strategically stable* in the sense of Kohlberg and Mertens [10]. A Nash equilibrium (σ, μ) is *symmetric* if $\sigma = \mu$. By Kakutani's Fixed Point Theorem, every finite and symmetric game $G = (S, u)$ has at least one symmetric Nash equilibrium. Let

$$\Delta^{NE}(S) = \{\sigma \in \Delta(S) : \sigma \in \beta(\sigma)\}. \tag{1}$$

Likewise, let the subset of strict symmetric Nash equilibrium strategies be written $\Delta^{NE+}(S)$, i.e., $\sigma \in \Delta^{NE+}(S)$ if and only if $\beta(\sigma) = \{\sigma\}$.

A strategy σ is *evolutionarily stable* if $\sigma \in \Delta^{NE}(S)$ and, moreover, $u(\sigma, \mu) > u(\mu, \mu)$ for all alternative best replies μ to σ. Likewise, a strategy σ is *neutrally stable* if $\sigma \in \Delta^{NE}(S)$ and $u(\sigma, \mu) \geq u(\mu, \mu)$ for all alternative best replies μ to σ. Let the subset of evolutionarily and neutrally stable strategies be denoted $\Delta^{EES}(S)$ and $\Delta^{NSS}(S)$, respectively. We have

$$\Delta^{NE+}(S) \subset \Delta^{EES}(S) \subset \Delta^{NSS}(S) \subset \Delta^{NE}(S). \tag{2}$$

2.2. Cheap Talk. Costless pre-play communication - "cheap talk" - is modelled in the usual fashion. A finite and symmetric two-player game $G = (S, u)$ is to be played. Before this, each player sends a message to the other player. This is done simultaneously and without cost or possibility of error. Again costlessly and without error they then observe each others messages and both players simultaneously choose a strategy to play in G. We assume that the set M of possible messages is the same for both players, and moreover, that this set is finite. The resulting interaction, including the pre-play communication stage, thus (again) constitutes a finite and symmetric two-player game $\mathcal{G}$ with pure-strategy set H and payoff function v, where both are specified below. In order to distinguish the two games, we will refer to $G = (S, u)$ as the *base game*, and to $\mathcal{G} = (H, v)$ as the *meta-game* associated with G and M.

A pure strategy in $\mathcal{G}$, a *pure meta-strategy*, is a message (to send) and a decision rule specifying what pure strategy in G to play after each possible pair $(m, m') \in M^2$ of sent messages. Without loss of generality one can assume that each player conditions her choice of base–game strategy only on her opponent's message (See e.g. Weibull [23]). Hence, a decision rule can be formally represented as a function $f : M \to S$ that to each message $m' \in M$ received from one's opponent prescribes a

[5] Formally, for any positive perturbation vector $\delta = (\delta_i^1, \delta_i^2)_{i \in S}$ such that $M^k(\delta) = \{\sigma^k \in \Delta(S) : \sigma^k(i) \geq \delta_i^k \text{ for all } i \in S\}$ is non-empty for $k = 1, 2$, let $G(\delta)$ be the two-player game with strategy sets $M^1(\delta)$ and $M^2(\delta)$, and payoff functions $u_1(\sigma^1, \sigma^2) = u(\sigma^1, \sigma^2)$ and $u_2(\sigma^1, \sigma^2) = u(\sigma^2, \sigma^1)$. A strategy profile $(\sigma^1, \sigma^2) \in \Theta(S)$ is *strictly perfect* if for every sequence of perturbations $\delta_t \to 0$ there exists some accompanying sequence of strategy profiles $(\sigma_t^1, \sigma_t^2) \to (\sigma^1, \sigma^2)$ that are Nash equilibria in the corresponding perturbed game $G(\delta_t)$.

pure strategy $i = f(m')$ in G. Let F be the set of all such functions. Formally, a pure meta-strategy thus is a pair $(m, f) \in M \times F$. We write $h = (m, f) \in M \times F = H$.

Since pre-play communication by assumption is costless, the payoff to any pure meta-strategy $h = (m, f) \in H$, when played against some pure meta-strategy $k = (m', g) \in H$, is a_{ij} where $i = f(m')$ and $j = g(m)$: Player 1 receives 2's message m' and thus plays pure strategy $i = f(m')$ in G, while player 2 receives 1's message m and thus plays pure strategy $j = g(m)$ in G. The payoff matrix of the meta-game $\mathcal{G}$ may thus be represented by the $|H| \times |H|$ matrix $\mathcal{A}$ with entries $\alpha_{hk} = a_{ij}$ in each row $h \in H$ and column $k \in H$, where $h = (m, f)$, $k = (m', g)$, $i = f(m)$, and $j = g(m')$. The space of *mixed meta-strategies* is the $(|H| - 1)$-dimensional unit simplex $\Delta(H)$ in $\mathbb{R}^{|H|}$. For any pair of such mixed strategies $p, q \in \Delta(H)$, the *payoff* to meta-strategy p when used against meta-strategy q, is

$$v(p, q) = p \cdot \mathcal{A}\, q = \sum_{h,k \in H} p_h \alpha_{hk} q_k, \tag{3}$$

This defines the meta-game payoff function $v : \Theta(H) \to \mathbb{R}$. The set of (mixed) *best replies* to any meta-strategy $q \in \Delta(H)$ will be denoted $\beta^H(q) \subset \Delta(H)$.

3. Symmetric Meta-Game Nash Equilibria

It turns out to be analytically convenient to group the meta-strategies according to message sent. For any meta-strategy $p \in \Delta(H)$ and message $m \in M$, let $p(m) \in [0, 1]$ denote the probability that message m is sent in p.[6] We say that message m is *used* in p if $p(m) > 0$. Write $M(p) \subset M$ for the subset of messages used in p. For any message m used in p, let $p^m(m') \in \Delta(S)$ be the mixed base-game strategy "played" by message m against any message $m' \in M$. More precisely, given $p \in \Delta(H)$, $m \in M(p)$, and $m' \in M$, let $p_i^m(m')$ be the conditional probability that p assigns to the pure base-game strategy $i \in S$ against message m', given that message m is sent.[7] In particular, for any meta-strategy pair $(p, q) \in \Theta(H)$ in which m is used in p and m' is used in q, the pair $\left(p^m(m'), q^{m'}(m)\right) \in \Theta(S)$ constitutes the base-game strategy profile that messages m and m' play against each other. Using this notation one may decompose the payoff $v(p, q)$ to meta-strategy p against meta-strategy q as follows:

$$v(p, q) = \sum_{m \in M(p)} \sum_{m' \in M(q)} p(m) q(m') u\left[p^m(m'), q^{m'}(m)\right] \tag{4}$$

[6] More precisely, $p(m)$ is the sum of all pure-strategy probabilities p_h where $h = (m, f)$ for some $f \in F$.

[7] Formally: $p_i^m(m') = \sum_{f \in F_{im'}} p_{(m,f)}/p(m)$, where $F_{im'} = \{f \in F : f(m') = i\}$.

It is not difficult to show that a meta-strategy p is in Nash equilibrium with itself, $p \in \Delta^{NE}(H)$, if and only if (i) all used messages play some base-game Nash equilibrium against each other, and (ii) no message earns more than $v(p,p)$.

Proposition 1. *$p \in \Delta^{NE}(H)$ if and only if (i)-(ii) hold.*
(i) $\left(p^m(m'), p^{m'}(m)\right) \in \Theta^{NE}(S) \ \ \forall m, m' \in M(p)$
(ii) $\sum_{m' \in M(p)} p(m') u\left[p^m(m'), p^{m'}(m)\right] \leq v(p,p) \ \ \forall m \in M$

Proof. First let $p \in \Delta(H)$, and suppose (i) does not hold, i.e., $p^{\bar{m}}(\bar{m}') \notin \beta\left[p^{\bar{m}'}(\bar{m})\right]$ for some $\bar{m}, \bar{m}' \in M(p)$. Then some pure strategy $i \in S$ in the support of $p^{\bar{m}}(\bar{m}') \in \Delta(S)$ earns a suboptimal payoff. Let $q \in \Delta(H)$ be like p, except that $q^{\bar{m}}(\bar{m}') \in \beta\left[p^{\bar{m}'}(\bar{m})\right]$. Then

$$u\left[q^m(m'), p^{m'}(m)\right] = u\left[p^m(m'), p^{m'}(m)\right]$$

for all $m \neq \bar{m}$ and all m', as well as for $m = \bar{m}$ and all $m' \neq \bar{m}'$, and

$$u\left[q^{\bar{m}}(\bar{m}'), p^{\bar{m}'}(\bar{m})\right] > u\left[p^{\bar{m}}(\bar{m}'), p^{\bar{m}'}(\bar{m})\right].$$

Since $p(\bar{m}) > 0$ this implies $v(q,p) > v(p,p)$, by (4), so $p \notin \Delta^{NE}(H)$. Hence $p \in \Delta^{NE}(H) \Rightarrow$ (i).

Second, let $p \in \Delta(H)$, and suppose (ii) does not hold, i.e.,

$$\sum_{m' \in M(p)} p(m') u\left[p^m(m'), p^{m'}(m)\right] > v(p,p)$$

for some $m \in M$. Let $q \in \Delta(H)$ be like p, except that $q(m) = 1$ (and thus $q(m') = 0$ for all $m' \neq m$). Then $v(q,p) > v(p,p)$ by (4), so $p \notin \Delta^{NE}(H)$. Hence $p \in \Delta^{NE}(H) \Rightarrow$ (ii).

Third, assume (i) and (ii), and let $q \in \Delta(H)$. By (4), and using first (i), then (ii):

$$\begin{aligned} v(q,p) &= \sum_{m \in M(q)} q(m) \sum_{m' \in M(p)} p(m') u\left[q^m(m'), p^{m'}(m)\right] \leq \\ &\leq \sum_{m \in M(q)} q(m) \sum_{m' \in M(p)} p(m') u\left[p^m(m'), p^{m'}(m)\right] \leq \\ &\leq \sum_{m \in M(q)} q(m) v(p,p) = v(p,p). \end{aligned}$$

Hence $p \in \beta^H(p)$, so (i)-(ii) $\Rightarrow p \in \Delta^{NE}(H)$. **End of proof.**

Remark 1: By decomposition (4), the inequality in (ii) must be an equality for all messages used in a symmetric Nash equilibrium. Hence, $p \in \Delta^{NE}(H)$ implies

$$\sum_{m' \in M(p)} p(m') u\left[p^m(m'), p^{m'}(m)\right] = v(p,p) \quad \forall m \in M(p). \tag{5}$$

Remark 2: The meta-game $\mathcal{G}$, viewed as an extensive-form game, has $|M|^2$ subgames, one for each pair of messages. Since each player moves exactly once in any play of the meta-game, the behavior strategies are the same as the mixed strategies in this game. Moreover, a behavior strategy profile in any finite extensive form game is a Nash equilibrium if and only if it prescribes optimal play at each information set on its path. Conditions (i) and (ii) are equivalent to this requirement: (i) requiring that no deviation pays after the messages have been revealed, and (ii) requiring that no deviation pays before the messages are revealed.

4. Meta-Game Outcomes

4.1. Definitions. Let $V^{NE}(M) \subset \mathbb{R}$ denote the set of symmetric meta-game Nash equilibrium payoff outcomes when the message set is M:

$$V^{NE}(M) = \left\{x \in \mathbb{R} : x = v(p,p) \text{ for some } p \in \Delta^{NE}(H), \text{ for } H = M \times F\right\}. \tag{6}$$

Next, let $C^{NE} \subset \mathbb{R}^2$ denote the convex hull of the set of base-game Nash equilibrium payoff *vectors*.[8] Let $U^{NE} \subset \mathbb{R}$ be the *symmetric base-game payoff values* in this convex hull:

$$U^{NE} = \left\{x \in \mathbb{R} : (x,x) \in C^{NE}\right\}. \tag{7}$$

The set U^{NE} is necessarily convex and compact, hence $U^{NE} = [\underline{x}, \overline{x}]$ for some $\underline{x} \leq \overline{x}$. Moreover, for each of the end-points of this (perhaps degenerate) interval there exists a base-game Nash equilibrium such that the end-point is the average of the two players' payoffs in that equilibrium. More precisely, there exist $(\underline{\sigma}, \underline{\mu}) \in \Theta^{NE}(S)$ such that $\frac{1}{2}[u(\underline{\sigma}, \underline{\mu}) + u(\underline{\mu}, \underline{\sigma})] = \underline{x}$ and $(\overline{\sigma}, \overline{\mu}) \in \Theta^{NE}(S)$ such that $\frac{1}{2}[u(\overline{\sigma}, \overline{\mu}) + u(\overline{\mu}, \overline{\sigma})] = \overline{x}$. Note also that the set U^{NE} always contains the set of symmetric base-game Nash equilibrium payoffs, and that the latter may be a *proper* subset in some games.[9]

A simple example of the latter possibility is the 2×2 "Hawk-Dove" game with payoff matrix

---

[8]The *convex hull* of a set is the smallest convex set containing it. The *set of base-game strategy payoffs* is the set of pairs $(x,y) \in \mathbf{R}^2$ such that $(x,y) = (u(\sigma,\mu), u(\mu,\sigma))$ for some $(\sigma,\mu) \in \Theta^{NE}(S)$.

[9]Formally, the symmetric base-game Nash equilibrium payoffs are the points $x \in \mathbf{R}$ such that $x = u(\sigma,\sigma)$ for some $\sigma \in \Delta^{NE}(S)$.

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{8}$$

Its unique symmetric Nash equilibrium is (σ^*, σ^*), where $\sigma^* = \left(\frac{1}{2}, \frac{1}{2}\right)$. The associated payoff is $\frac{1}{2}$ to each player. However, the game also has two asymmetric (strict) Nash equilibria, namely (e^1, e^2) and (e^2, e^1), both giving payoff 1 to each player. Thus the set of symmetric Nash equilibrium payoffs is the singleton set $\left\{\frac{1}{2}\right\}$, while $U^{NE} = \left[\frac{1}{2}, 1\right]$.

In analogy with the above notation, let $V^{ESS}(M) \subset \mathbb{R}$ and $V^{NSS}(M) \subset \mathbb{R}$ be the sets of evolutionarily and neutrally stable meta-game payoff outcomes, respectively, when the message set is M. By (2):

$$V^{ESS}(M) \subset V^{NSS}(M) \subset V^{NE}(M). \tag{9}$$

4.2. Symmetric Nash Equilibrium Outcomes. We are now in a position to establish some properties of the set $V^{NE}(M)$. First, by proposition 1, this set is a subset of the base-game Nash equilibrium interval $U^{NE} = [\underline{x}, \overline{x}]$. This being true for any finite set M of messages, one may ask how the set $V^{NE}(M)$ depends on the message set M; in particular, if it expands as M expands. It is clear from the definition of the associated meta-game $\mathcal{G}$ that the dependence goes only through the cardinality of M: Any two message sets M and M' with the same number of elements define meta-games that differ only in the labelling of messages. The question may thus be re-phrased as how the set $V^{NE}(M)$ depends on the number $|M|$ of messages. It turns out that the set $V^{NE}(M)$ is non-decreasing in $|M|$, and that it converges towards a limit set W^{NE} that is dense in U^{NE}. In this sense, every point in the interval U^{NE} can be approximated by the outcome in some symmetric meta-game Nash equilibrium. Wärneryd [22] establishes that any payoff in U^{NE} can be approximated by the payoff to some symmetric meta-game Nash equilibrium if the message set is large enough. This follows from (but does not imply) the second claim in the following result.

Proposition 2. *For any base game G and message sets M and M^* with $|M| \leq |M^*|$, $V^{NE}(M) \subset V^{NE}(M^*) \subset U^{NE}$. For any base game G and sequence of messages sets M_k with $|M_k| \to \infty$ as $k \to \infty$, the limit set $W^{NE} = \cap_{k \in \mathbb{N}} \cup_{h \geq k} V^{NE}(M_h)$ exists and is a dense subset of U^{NE}.*

Proof. We prove these two claims in three steps. First we show $V^{NE}(M) \subset U^{NE}$, for any finite set M. Second, we show $V^{NE}(M) \subset V^{NE}(M^*)$ for any pair of sets $M \subset M^*$ where $|M| = k$ and $|M^*| = k+1$. These two steps together establish the first

claim in the proposition. Third, we construct an infinite sequence $V^{NE}(M_k)$, with each $|M_k| = k$, such that the associated limit set W^{NE} is dense in U^{NE}. Existence of the limit follows from the first claim in the proposition, combined with the fact that U^{NE} is compact. It also follows from the first claim that the same limit set W^{NE} obtains from any sequence of messages sets M_k with $|M_k| \to \infty$.

Step 1: By proposition 1 all message pairs $(m, m') \in M^2$ play some base-game Nash equilibrium $(\sigma, \mu) \in \Theta^{NE}(S)$ (symmetric if $m = m'$). By (4) $v(p,p)$ is a convex combination of such base-game payoffs. In case $m \neq m'$, the weight in the convex combination is $p(m)p(m')$ both to the payoff $u(p^m(m'), p^{m'}(m))$ and to the "opposing" payoff $u(p^{m'}(m), p^m(m'))$. Thus, the whole convex combination is symmetric, and thus $V^{NE}(M) \subset U^{NE}$.

Step 2: Assume $M = \{1, ..., k\} \subset M^* = \{1, ..., k, k+1\}$, and $v(p,p) \in V^{NE}(M)$. Let $H^* = M^* \times F$. Without loss of generality assume $k \in M(p)$. We now construct a meta-strategy $q \in \Delta(H^*)$ that mimics p and that treats message $k+1$ just like message k. More precisely, for all $m \leq k$ let $q(m) = p(m)$ (and thus $q(k+1) = 0$). Moreover, for all $m, m' \leq k$, let $q^m(m') = p^m(m')$. For all $m \leq k$, let $q^m(k+1) = p^m(k)$, and for all $m' \leq k$ let $q^{k+1}(m') = p^k(m')$. Let $q^{k+1}(k+1) = p^k(k)$. It follows from this construction that, in meta-strategy q, all used message pairs play base-game Nash equilibria, indeed the same as in p, that every used message earns payoff $v^*(q,q) = v(p,p)$, and no message in M^* earns more. By proposition 1, $q \in \Delta^{NE}(H^*)$.

Step 3: We can construct a meta strategy $p \in \Delta^{NE}(H)$ with payoff $v(p,p)$ arbitrarily close to any given point x in $U^{NE} = [\underline{x}, \overline{x}]$ by letting the messages set M be sufficiently large. Recall that the set $\Delta^{NE}(S)$ is non-empty, take any $\sigma^o \in \Delta^{NE}(S)$, and let $x^o = u(\sigma^o, \sigma^o)$. Any $x \in U^{NE}$ belongs to at least one of the two sub-intervals $[\underline{x}, x^o]$ and $[x^o, \overline{x}]$. Assume $x \in [x^o, \overline{x}]$. To any such point x there exists a $\lambda \in [0,1]$ such that $x = \lambda x^o + (1-\lambda)\overline{x}$. Moreover, for any $\varepsilon > 0$ there exist positive integers t and k such that t is even, $k \geq t+1$, and $\hat{\lambda} = t/k \in [0,1]$ is within distance ε from λ. Now let there be k messages in M and place all messages around a circle. Let $p \in \Delta(H)$ be such that $p(m) = \frac{1}{k}$ for all $m \in M$, each message $m \in M$ plays $\sigma^o \in \Delta(S)$ against itself, $\overline{\sigma} \in \Delta(S)$ against its $t/2$ nearest "clockwise" neighbor messages on the circle, $\overline{\mu} \in \Delta(S)$ against its $t/2$ nearest "counter-clockwise" neighbor messages, and σ^o against all other messages. Then all messages are used in p, all message pairs play base-game Nash equilibria, and all messages earn the same payoff. Thus $p \in \Delta^{NE}(H)$ by proposition 1. Moreover,

$$v(p,p) = \frac{1}{k}[\frac{t}{2}u(\overline{\sigma}, \overline{\mu}) + \frac{t}{2}u(\overline{\mu}, \overline{\sigma}) + (k-t)u(\sigma^o, \sigma^o)] = \hat{\lambda}x^o + (1-\hat{\lambda})\overline{x}.$$

For t and k sufficiently large, $v(p,p)$ is arbitrarily close to x.

The case $x \in [\underline{x}, x^o]$ can be treated in the same way. **End of proof.**

4.3. Neutrally Stable Outcomes. Propositions 2 and the inclusion chain (9) together imply that the set $V^{NSS}(M)$, for any finite set M of messages, is a subset of the base-game Nash equilibrium interval $U^{NE} = [\underline{x}, \overline{x}]$. Like in the case of symmetric Nash equilibrium payoffs, one may ask how the set $V^{NSS}(M)$ depends on the message set M; in particular, if it expands as $|M|$ increases, and if it has a limit, as $|M| \to \infty$, and whether that limit is dense in U^{NE}. While this was shown above to be true for $V^{NE}(M)$, $V^{NSS}(M)$ does not expand monotonically with $|M|$ in all games.

A simple counter-example against monotonicity of $V^{NSS}(M)$ in $|M|$ is the "Hawk-Dove" game with payoff matrix (8) above. It is well-known, and easily verified, that its unique mixed Nash equilibrium strategy σ^* is evolutionarily stable. Hence, $V^{ESS}(M) = V^{NSS}(M) = \left\{\frac{1}{2}\right\}$ when $|M| = 1$. However, one can show that $\frac{1}{2} \notin V^{NSS}(M)$ whenever $|M| > 1$. Take the case of two messages. In order to obtain payoff $\frac{1}{2}$ in such a meta-game, it is necessary, by proposition 1, that all four message pairs play (σ^*, σ^*). But such a meta-strategy p is vulnerable to invasion by the mutant strategy q that sends both messages with equal probability, lets each message play σ^* against itself, one message play pure strategy 1 against the other, and the other message play pure strategy 2 against the first. This meta-strategy is certainly a best reply to q. However, $v(q,q) = \frac{3}{4} > v(p,q) = \frac{1}{2}$. Hence $p \notin \Delta^{NSS}(H)$. It turns out that the reason for this phenomenon of $V^{NSS}(M)$ not being non-decreasing in $|M|$ is that the base-game ESS σ^* is a minimax strategy in the base game - the logic is here similar to that in the Folk theorems.

For any finite and symmetric two-player game let $x_{mm} \in \mathbb{R}$ be its minimax value, i.e.,

$$x_{mm} = \min_{\mu \in \Delta(S)} \max_{\sigma \in \Delta(S)} u(\sigma, \mu). \tag{10}$$

Lemma 1. *For any base game G and message sets M and M^* with $|M| \leq |M^*|$:*
(a) If $x \in V^{NSS}(M)$ and $x > x_{mm}$, then $x \in V^{NSS}(M^*)$
(b) If $x_{mm} \notin V^{NSS}(M)$, then $x_{mm} \notin V^{NSS}(M^*)$

Proof. For (a), assume $x \in V^{NSS}(M)$ and $x > x_{mm}$. Let $p \in \Delta^{NSS}(H)$ have $v(p,p) = x$. It is sufficient to consider the case $M = \{1, ..., k\}$ and $M^* = \{1, ..., k, k+1\}$. Let μ_{mm} be a minimax strategy in G. Thus $u(\sigma, \mu_{mm}) \leq x_{mm}$ for all $\sigma \in \Delta(H)$. Let H^* be the set of pure strategies in the meta-game $\mathcal{G}^*$ associated with message set M^*. Let $q \in \Delta(H^*)$ agree with p on H, have message $k+1$ unused and play μ_{mm} against it. Formally, for all $m \leq k$ let $q(m) = p(m)$ (thus $q(k+1) = 0$). For all $m, m' \leq k$, let $q^m(m') = p^m(m')$. For all $m \leq k$, let $q^m(k+1) = \mu_{mm}$, and for all $m' \in M^*$ let $q^{k+1}(m') = \mu_{mm}$. It follows from this construction that, in meta-strategy q, all used message pairs play the same base-game Nash equilibria as in p, that every used message earns payoff $v^*(q,q) = v(p,p)$, and no message in M^* earns

more. By proposition 1, $q \in \Delta^{NE}(H^*)$. Since $p \in \Delta^{NSS}(H)$: $v(p',p') \leq v(p,p')$ for all $p' \in \beta^H(p)$. Now suppose $q' \in \beta^{H^*}(q)$. Then the support of q' is a subset of H, since message $k+1$ is minimaxed in q. Let $p' \in \Delta(H)$ be the restriction of q' to H. Then $p' \in \beta^H(p)$ and so $v(p',p') \leq v(p,p')$. But $v(q',q') = v(p',p')$ and $v(q,q') = v(p,p')$, which shows that $q \in \Delta^{NSS}(H^*)$.

For (b), assume $x_{mm} \notin V^{NSS}(M)$ and $x_{mm} \in V^{NSS}(M^*)$, for $M = \{1,...,k\}$ and $M^* = \{1,...,k,k+1\}$. Let $p \in \Delta^{NSS}(H^*)$ have $v(p,p) = x_{mm}$. By proposition 1 all messages used in p play some base-game minimax Nash equilibrium against all used messages. Suppose some message is unused in p. Without loss of generality let $m = k+1$ be such. Then the restriction of p to H belongs to $\Delta^{NSS}(H)$, a contradiction. Suppose instead that all messages are used in p. Then all message pairs play some base-game minimax Nash equilibrium. Let $p' \in \Delta(H^*)$ be like p, except that $p'(1) = 1$ (p' only uses message $m = 1$). Then $q \in \beta^{H^*}(p) \Rightarrow q \in \beta^{H^*}(p') \Rightarrow v(q,q) \leq v(p,q) = v(p',q)$, so $p' \in \Delta^{NSS}(H^*)$. But only message $m = 1$ is used in p', and so the restriction of p' to H belongs to $\Delta^{NSS}(H)$, a contradiction. **End of proof.**

Remark 3: This proof does not work for evolutionary stability. For suppose in the proof of (a) above that $p \in \Delta^{ESS}(H)$, and let $q \in \Delta(H^*)$ be defined as in that proof. Suppose $q' \in \beta^{H^*}(q)$, $q' \neq q$. Then the support of q' is a subset of H, since message $k+1$ is minimaxed in q. (But q' may well differ from q at the unused message $m = k+1$.) Thus $v(q',q') = v(q,q')$, which shows that $q \notin \Delta^{ESS}(H^*)$.

Proposition 3. *For any base game G and sequence of messages sets M_k with $|M_k| \to \infty$ as $k \to \infty$, the limit set $W^{NSS} = \cap_{k \in \mathbf{N}} \cup_{h \geq k} V^{NSS}(M_h)$ exists and is a non-empty subsets of U^{NE}.*

Proof. The claim in the proposition follow readily from the above lemma. First note that the sets $V^{NSS}(M_k) \cap (x_{mm}, \overline{x}]$, for $k = 1,2,...$, are increasing in k by Lemma 1(a). Since each such set is a subset of the compact set U^{NE}, the associated limit set $X^{NSS} = \cap_{k \in \mathbf{N}} \left(V^{NSS}(M_k) \cap (x_{mm}, \overline{x}] \right)$ exists, and is a subset of U^{NE}. If, for some k, $x_{mm} \notin V^{NSS}(M_k)$, then $W^{NSS} = X^{NSS}$, by (b). Otherwise, $W^{NSS} = X^{NSS} \cup \{x_{mm}\} \subset U^{NE}$. **End of proof.**

It follows immediately that if $x_{mm} \notin U^{NE}$, which is indeed the case in many games, then the set $V^{NSS}(M)$ is in fact non-decreasing in $|M|$:

Corollary 1. *For any base game G such that $x_{mm} \notin U^{NE}$, and any message sets M and M^* with $|M| \leq |M^*|$: $V^{NSS}(M) \subset V^{NSS}(M^*)$.*

Remark 4: In the Hawk-Dove game (8) with $|M| = 1$ and $|M^*| = 2$ we have $x_{mm} = \frac{1}{2} \in U^{NE}$, $x_{mm} \in V^{NSS}(M)$ and $x_{mm} \notin V^{NSS}(M^*)$.

5. 2 × 2 Coordination Games

We here focus on the special case of symmetric 2 × 2 games with payoff matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{11}$$

for some $a > c$, $d > b$. We will call such games *coordination games*, and it is well-known that their set of evolutionarily stable strategies consists of the two pure strategies, and, moreover, that their unique mixed Nash equilibrium strategy is not neutrally stable: $\Delta^{NSS}(S) = \Delta^{ESS}(S) = \{e^1, e^2\}$.

Consider any game G with payoff matrix as in (11), where $a < d$, i.e., a is the "bad" and d the "good" strict Nash equilibrium payoff . Let this be the base game in a cheap-talk game $\mathcal{G}$ with finite message set M. Observe that, irrespective of each player always receives the same payoff as the other player, both in the base game and in the meta game.

By proposition 2 the set $V^{NE}(M)$ of symmetric meta-game Nash equilibrium outcomes, for any finite message set M, is a subset of the set $U^{NE} = [\underline{x}, \overline{x}]$. In a coordination game with payoff matrix (11), $\underline{x} = \frac{ad}{a+d}$ is the mixed-strategy base-game Nash equilibrium payoff and $\overline{x}$ is the "good" strict Nash equilibrium payoff d.[10] By (2), also all neutrally and evolutionarily stable meta-game payoff outcomes, respectively, belong to the interval $U^{NE} = [\underline{x}, \overline{x}]$. A point $x \in U^{NE}$ will be called a *neutrally (evolutionarily) stable cheap-talk outcome* for the coordination game with payoff matrix (11) if there exists some finite message set M and associated meta-game strategy p that is neutrally (evolutionarily) stable in the meta-game, and that has $v(p,p) = x$. We will give a complete characterization of these sets: it turns out that they coincide, and that this set is infinite but countable, consisting of points between a and d, including these, and having d as its unique cluster point. Hence, all but one of the neutrally (and likewise evolutionarily) stable outcomes are isolated points, the unique exception being the Pareto efficient outcome.

This result is in stark contrast with the corresponding result for the set of *strategically stable cheap-talk outcomes*. These are defined as the points $x \in U^{NE}$ for which there exists some finite message set M and associated meta-game strategy p such that (p,p) is a strategically stable Nash equilibrium in the meta-game with payoff $v(p,p) = x$. It turns out that this is a dense subset of U^{NE}.

As a first step towards our characterization of the set of neutrally stable meta-game outcomes, we show that neutral (and hence also evolutionary) stability in the meta game requires that all present messages play pure strategies against each other.

[10] Note that the "bad" strict Nash equilibrium payoff a is a point in the interior of the interval U^{NE}.

Consequently, neutrally stable meta-game outcomes belong to the sub-interval $[a,d] \subset U^{NE}$.

We will say that a message $m \in M(p)$ is *nice* to a message $m' \in M$ in $p \in \Delta(H)$ if m plays the "good" strict Nash equilibrium against m': $p^m(m') = e^2$. We establish that if some used message is nice to itself, then every used messages is nice to all used messages. Consequently, the payoff is then maximal.

Lemma 2. *Suppose* $p \in \Delta^{NSS}(H)$.
(i) If $m, m' \in M(p)$, *then* $p^m(m') = p^{m'}(m) \in \{e^1, e^2\}$,
(ii) If $p^m(m) = e^2$ *for some* $m \in M(p)$, *then* $v(p,p) = d$.

Proof. (i) By proposition 1 it suffices to show that m and m' do not play the mixed base-game Nash equilibrium with each other. Suppose they would. Then let $q \in \Delta(H)$ be like p, except for $q^m(m') = q^{m'}(m) = e^2$. Then $q \in \beta^H(p)$, and $v(q,q) = d > v(p,p)$, so $p \notin \Delta^{NSS}(H)$.

(ii) Suppose $m \in M(p)$, $p^m(m) = e^2$ and $v(p,p) < d$. Let $q \in \Delta(H)$ be such that $q(m) = 1$ and $q^m(m'') = p^m(m'')$ for all $m'' \in M$. Then $q \in \beta^H(p)$, and $v(q,q) = d$. However, $v(p,q) = p(m)d$, where $p(m) < 1$ since $v(p,p) < d$. Thus $v(q,q) > v(p,q)$, and hence $p \notin \Delta^{NSS}(H)$. **End of proof.**

For any meta strategy p and message m, let $N(m,p) \subset M$ be the subset of messages that are nice to m in p:

$$N(m,p) = \{m' \in M : m' \text{ nice to } m \text{ in } p\}. \tag{12}$$

We call a subset $M' \subset M(p)$ *polite* in $p \in \Delta(H)$ if every message in M' plays the "good" strict Nash equilibrium strategy e^2 against all *other* messages in M' and the "bad" strict Nash equilibrium strategy e^1 against itself. A meta-strategy $p \in \Delta(H)$ is said to be in *politeness class* n if some non-empty subset of messages $M' \subset M$ with $|M'| = n$ is polite in p, and no larger subset of M is polite in p. The next result establishes a lower bound on the neutrally stable meta-game outcomes in terms of politeness classes. The higher politeness class, the higher is this lower bound.

Lemma 3. *Suppose* $p \in \Delta^{NSS}(H)$ *is of politeness class* n. *Then* $v(p,p) \geq \frac{1}{n}a + \left(1 - \frac{1}{n}\right)d$.

Proof. Let $\emptyset \neq M' \subset M$ be polite in $p \in \Delta^{NE}(H)$, with $|M'| = n$. Let $q \in \Delta(H)$ be such that $q(m) = \frac{1}{n}$ for all $m \in M'$, and $q^m(m') = p^m(m')$ for all

$m, m' \in M'$. Then

$$\begin{aligned} v(q,p) &= \frac{1}{n}\sum_{m\in M'}\sum_{m''\in M(p)} p(m'')u\left[p^m(m''), p^{m''}(m)\right] \\ &= \frac{1}{n}\sum_{m\in M'} v(p,p) = v(p,p), \end{aligned}$$

so $q \in \beta^H(p)$. Moreover, $v(q,q) = \frac{1}{n}a + \left(1-\frac{1}{n}\right)d$, and $v(p,q) = v(p,p)$. Hence $p \notin \Delta^{NSS}(H)$ if $v(p,p) < \frac{1}{n}a + \left(1-\frac{1}{n}\right)d$. To see that $v(p,q) = v(p,p)$, first note that

$$\begin{aligned} v(p,q) &= \sum_{m\in M(p)} p(m)\sum_{m'\in M'}\frac{1}{n}u\left[p^m(m'), p^{m'}(m)\right] = \\ &= \sum_{m\in M'} p(m)\sum_{m'\in M'}\frac{1}{n}u\left[p^m(m'), p^{m'}(m)\right] \\ &\quad + \sum_{m\notin M'} p(m)\sum_{m'\in M'}\frac{1}{n}u\left[p^m(m'), p^{m'}(m)\right] \\ &= \sum_{m\in M'} p(m)\frac{1}{n}\left[a+(n-1)d\right] + \sum_{m\notin M'} p(m)\sum_{m'\in M'}\frac{1}{n}u\left[p^{m'}(m), p^m(m')\right]. \end{aligned}$$

In the last equality we have used the fact that q mimics p on $M' \subset M(p)$ (for the first term) and the fact that p there lets all message pairs play symmetric base-game (for the second term). Reversing the order of summation in the second term, and using Remark 1, we obtain

$$\begin{aligned} v(p,q) &= \sum_{m\in M'} p(m)v(q,q) + \frac{1}{n}\sum_{m'\in M'}\sum_{m\notin M'} p(m)u\left[p^{m'}(m), p^m(m')\right] = \\ &= \sum_{m\in M'} p(m)v(q,q) + \frac{1}{n}\sum_{m'\in M'}\left(v(p,p) - \sum_{m\in M'} p(m)u\left[p^{m'}(m), p^m(m')\right]\right) = \\ &= v(p,p) + \sum_{m\in M'} p(m)v(q,q) - \frac{1}{n}\sum_{m'\in M'}\left(p(m')a + \left[1-p(m')\right]d\right) = \\ &= v(p,p) + v(q,q) - v(q,q) = v(p,p). \end{aligned}$$

End of proof.

For any non-empty subset $M' \subset M$ of messages and meta-strategy $p \in \Delta(H)$, let $\Pr(M' \mid p)$ be the probability that a message from M' is sent in p.

Lemma 4. *For any $\emptyset \neq M' \subset M$ and $p \in \Delta(H)$:*

$$\Pr\left[\cap_{m\in M'} N(m,p) \mid p\right] \geq \sum_{m\in M'} \Pr\left[N(m,p) \mid p\right] - |M'| + 1. \tag{13}$$

Proof. For any probability measure μ on a set X with $k \geq 1$ μ-measurable subsets B_i: $\mu[\sim \cap_i B_i] \leq \sum_i \mu(\sim B_i)$. Equivalently,

$$\mu[\cap_i B_i] \geq 1 - \sum_i \mu(\sim B_i) = 1 - k + \sum_i \mu(B_i).$$

End of proof.

Lemma 5. *Suppose $v(p,p) < d$ and $p \in \Delta^{NE}(H)$ is of politeness class n. Then $v(p,p) \leq \frac{1}{n}a + \left(1 - \frac{1}{n}\right)d$.*

Proof. Let $M' \subset M$ be polite in p, with $|M'| = n$. Since no $M'' \subset M$ with $|M''| > n$ is polite in p, no $m'' \notin M'$ is nice to all $m' \in M'$. Since $v(p,p) < d$, no $m' \in M'$ is nice to itself, by Lemma 2. Hence $\cap_{m'\in M'} N(m',p) = \emptyset$. Moreover, by Proposition 1 each $m \in M(p)$ earns payoff $a + \Pr[N(m,p)](d-a) = v(p,p)$. Since $M' \subset M(p)$ this equation holds for all $m' \in M'$. An application of lemma 4 to the set M' gives

$$0 \geq n\frac{v(p,p) - a}{d - a} - n + 1,$$

which is equivalent to the claimed inequality. **End of proof.**

Lemma 6.

$$V^{NSS}(M) \subset \left\{a, \frac{a+d}{2}, \frac{a+2d}{3}, ..., \frac{a+(|M|-1)d}{|M|}, d\right\} \tag{14}$$

Proof. Every $p \in \Delta^{NSS}(H)$ is either of politeness class n for some integer $n \in [1, |M|]$ or else $v(p,p) = d$. Lemmas 3 and 5 give (14). **End of proof.**

The following proposition establishes that the inclusion in lemma 6 in fact is an equality. This result thus characterizes the sets of neutrally stable outcomes in all finite cheap-talk extensions of 2×2 coordination games.

Proposition 4.

$$V^{NSS}(M) = \left\{a, \frac{a+d}{2}, \frac{a+2d}{3}, ..., \frac{a+(|M|-1)d}{|M|}, d\right\} \tag{15}$$

Proof. Let $n = |M|$. By lemmas 1 and 6 it is sufficient to show $\frac{a+(n-1)d}{n} \in V^{NSS}(M)$. We will in fact establish $\frac{a+(n-1)d}{n} \in V^{ESS}(M)$. For this purpose, let $p(m) = \frac{1}{n}$ and $p^m(m) = e^1$ for each $m \in M$, and $p^m(m') = p^{m'}(m) = e^2$ for all $m, m' \in M$ with $m' \neq m$. Then $v(p,p) = \frac{1}{n}a + (1 - \frac{1}{n})d$. To see that $p \in \Delta^{ESS}(H)$, first note that $q \in \beta^H(p) \Rightarrow q^m(m') = p^{m'}(m) = p^m(m')$ for all $m \in M(q)$ and $m' \in M(p) = M$. Since q and p let all message pairs play symmetric and pure base-game strategy profiles against each other, the off-diagonal elements b and c in the payoff matrix A are never used, and so we may assume without loss of generality that $b = c$. Thus $\mathcal{G}$ is doubly symmetric, and consequently for any $q \in \beta^H(p)$ we have $v(p,q) = v(q,p) = v(p,p)$. It thus suffices to show that $v(q,q) < v(p,p)$ for all $q \in \beta^H(p)$, $q \neq p$. By (4),

$$\begin{aligned} v(q,q) &= \sum_{m \in M(q)} q(m) \left[a + (d-a) \sum_{m' \in M(q) \setminus m} q(m') \right] \\ &= a + (d-a) \sum_{m \in M(q)} q(m) \left[1 - q(m) \right] = d - (d-a) \sum_{m \in M(q)} q^2(m). \end{aligned}$$

Thus $v(q,q)$ is maximal when $\sum_{m \in M(q)} q^2(m)$ is minimal. This sum is minimal precisely when $M(q)$ is maximal and all $q(m)$ are equally large, i.e., when $M(q) = M$ and $q(m) = \frac{1}{n} = p(m)$ for all $m \in M$.[11] In sum, $q \in \beta^H(p) \Rightarrow v(q,q) \leq v(p,p)$ with equality only when $q = p$. Hence $p \in \Delta^{ESS}(H)$. **End of proof.**

The set Ω^{NSS} of *neutrally stable cheap-talk outcome* for the coordination game with payoff matrix (11) is defined (as indicated above) as the set of points $x \in U^{NE}$ for which there exists some finite message set M and associated meta-game strategy p that is neutrally stable in the meta-game and that has payoff $v(p,p) = x$. Formally, $\Omega^{NSS} = \cup_{M \text{ finite}} V^{NSS}(M)$. Likewise, let $\Omega^{ESS} = \cup_{M \text{ finite}} V^{ESS}(M)$. It follows from the above proposition, together with the observation in its proof that $\frac{a+(|M|-1)d}{|M|} \in V^{ESS}(M)$, that the sets of neutrally and evolutionary cheap-talk outcomes coincide:

$$\Omega^{NSS} = \Omega^{ESS} = \left\{ \frac{a+(n-1)d}{n} : \text{ for some } n \in \mathbb{N} \right\} \cup \{d\}.$$

Remark 5: The finding above that $\frac{a+(|M|-1)d}{|M|} \in V^{ESS}(M)$ is consistent with Schlag's [17] result that this is the payoff to the unique evolutionarily stable strategy.

[11] First fix $M(q) = M'$. The program to minimize the sum $\sum_{m \in M'} q^2(m)$, subject to the constraint that all $q(m)$, for $m \in M'$, are non-negative and sum to one, has the unique solution $q(m) = \frac{1}{k}$ for all $m \in M'$, where $k = |M'|$. Geometrically, this is equivalent to finding the point in the unit simplex in $\mathbb{R}^k$ that is closest to the origin. The minimum value, for $M(q) = M'$ fixed, is thus $\frac{1}{k}$. Hence, k should be chosen as large as possible, i.e., $M' = M$.

He also shows that the Pareto efficient outcome d is obtained in an evolutionarily stable set (Thomas [20]). Clearly the difference between his and our results derive from the difference between neutral and evolutionary stability. (For more on this, see Section 7.)

Remark 6: The fact that the set of neutrally stable outcomes in 2×2 coordination games always includes the Pareto efficient point does not mean that this holds for all games. In fact, there are games in which a unique Pareto dominant (but non-strict) Nash equilibrium is unstable. An example is given by the following payoff matrix:

$$A = \begin{pmatrix} 1 & 2-\alpha & 0 & -\gamma \\ 0 & 1 & 2-\alpha & -\gamma \\ 2-\alpha & 0 & 1 & -\gamma \\ -\gamma & -\gamma & -\gamma & \beta \end{pmatrix} \tag{16}$$

where $\alpha \in (0,1)$, $\beta \in (0, 1-\frac{\alpha}{3})$ and $\gamma \geq 0$. The three first rows and columns together constitute a generalized "Rock-Scissors-Paper" game which has a unique Nash equilibrium, and in this equilibrium both players randomize uniformly over the three strategies and each player obtains the payoff $1-\frac{\alpha}{3}$. It is well-known that this equilibrium is unstable in the replicator dynamics (see e.g. Hofbauer and Sigmund [6] or Weibull [23]). For non-negative values of γ, this "Rock-Scissors-Paper equilibrium" remains a Nash equilibrium in the full game. However, the full game has two more Nash equilibria, each of which is symmetric. One is the strict equilibrium in which both players use only strategy 4, resulting in payoff β to both players - by hypothesis a lower payoff than in the "Rock-Scissors-Paper equilibrium." The third Nash equilibrium is completely mixed and its payoff can be made arbitrarily low by choosing γ sufficiently large. However, the unique Pareto-dominant Nash equilibrium, giving payoff $1-\frac{\alpha}{3}$ to each player, is not Lyapunov stable in a cheap-talk extension this game, for any finite message set. For if $p \in \Delta^{NE}(H)$ has $v(p,p) = 1-\frac{\alpha}{3}$, then all used messages earn the same payoff and all active message pairs play the "Rock-Papers-Scissors equilibrium." When such a message meets itself, the situation is exactly the same as in the absence of communication, and so the associated sub-population state is dynamically unstable in the replicator dynamics. It follows that p is not neutrally stable, since neutral stability implies Lyapunov stability in the replicator dynamics (Thomas [20], Bomze and Weibull [4]).

The set $\Omega^{SS}$of *strategically stable cheap-talk outcomes* is defined as the set of points $x \in U^{NE}$ for which there exists some finite message set M and associated strategically stable set of meta-game strategy pairs $(p,p) \in \Theta^{NE}(H)$ with payoff $v(p,p) = x$. The following result establishes that any payoff x in the interval U^{NE} can be approximated

by some strategically stable Nash cheap-talk outcome. In fact, it is shown that x can be approximated by the payoff to a strategically stable *singleton* set:

Proposition 5. *Ω^{SS} is dense in U^{NE}.*

Proof. For every $x \in [\underline{x}, \overline{x}]$ and $\varepsilon > 0$ there exist positive integers k and n, with k even and $n \geq k$, such that $y = \lambda \underline{x} + (1-\lambda)\overline{x}$, for $\lambda = k/n$, is within distance ε from x. Let $|M| = n$, and let $p \in \Delta(H)$ have $p(m) = 1/n$ for all $m \in M$. Order all messages in a ring, and let each message play the base-game Nash equilibrium strategy σ^* to its $k/2$ nearest neighbors on each side, and let it play e^2 to all other messages, and to itself. Then all messages play base-game Nash equilibria with each other, and all messages earn the same payoff

$$v(p,p) = \left[k\underline{x} + (n-k)\overline{x}\right]/n = \lambda \underline{x} + (1-\lambda)\overline{x}.$$

It follows from proposition 1 that $p \in \Delta^{NE}(H)$.

To see that $(p,p) \in \Theta(H)$ is strictly perfect, let $\delta = (\delta_h^1, \delta_h^2)_{h \in H}$ be such that $P^k(\delta) = \left\{p \in \Delta(H) : p_h \geq \delta_h^k \text{ for all } h \in H\right\}$ is non-empty for $k = 1, 2$, and let $\mathcal{G}(\delta)$ be the associated (possibly asymmetric) two-player perturbed meta-game with strategy sets $P^1(\delta)$ and $P^2(\delta)$. For δ sufficiently small this game has a Nash equilibrium (p', p') arbitrarily close to (p,p). Let $p'(m) = p(m) = 1/n$ for all $m \in M$ and let each message play the base-game Nash equilibrium strategy σ^* to its $k/2$ nearest neighbors on each side, and let it place maximal probability on the decision rule that assigns the pure base-game strategy e^2 to all other messages, and to itself. Since $e^2 \in \Delta^{NE+}(S)$, p' is a best reply to itself in the perturbed meta-game $\mathcal{G}(\delta)$, granted the vector $\delta > 0$ is sufficiently small. **End of proof.**

Remark 7: The argument of the above proof can be used, *mutatis mutandis*, to establish that the meta-strategy p in the proof of proposition 5 is strictly perfect.[12]

6. Infinite message sets

In any natural language the set of possible statements is infinite and countable. Hence, the above assumption that the message set M be finite is not as innocent as it may look. It is well known from the repeated games literature that the equilibrium correspondence may be discontinuous (more precisely lack lower hemi-continuity) "at infinity," i.e., as one moves from a finite but arbitrarily distant time horizon to an infinite time horizon. In that context, the limit of finite horizon equilibrium

[12]This observation may be compared with van Damme's [5] general result that if a mixed strategy σ in a finite and symmetric two-player game is evolutionarily stable, then (σ, σ) is a proper equilibrium. The somewhat stronger conclusion drawn here is due to the special structure of coordination games.

outcomes always constitute equilibrium outcomes also in the infinite horizon case (the equilibrium correspondence is upper hemi-continuous) but there may be a whole plethora of infinite horizon outcomes that have no counterpart in the finite but distant horizon case. An important question for the present analysis thus is whether the relevant solution correspondences for cheap talk games are continuous "at infinity," i.e., as one moves from a finite but arbitrarily large message set to an infinite message set.

For the purpose of investigating this question, we now assume $M = \mathbb{N}$, and re-examine all results established above for finite message sets. The first question that arises is how to define payoffs and solution concepts when M, and hence also the pure-strategy set H of the meta game $\mathcal{G}$, is infinite. Since the base-game G is finite and thus has bounded payoffs, all methods easily generalize. First, payoffs may still be defined as in equation (11) since the set of numbers α_{hk}, for $h, k \in H$, is bounded. Consequently, the definitions of Nash equilibrium, evolutionary and neutral stability etc. may be extended to an infinitely countable message set. (Existence of Nash equilibria is no longer guaranteed, however.) The decomposition formula (4) still holds, and the proof of proposition 1 applies.

We focus on neutrally stable outcomes in the special case of 2×2 coordination games. Inspection of the proofs of lemmas 2 through 5 reveals that these are valid for any countable set M, positive integer n, and finite subset $M' \subset M$. The counterpart to Proposition 4 is

Proposition 6. $V^{NSS}(\mathbb{N}) = \Omega^{NSS}$.

Proof. We show (a) $V^{NSS}(\mathbb{N}) \subset \Omega^{NSS}$, (b) $d \in V^{NSS}(\mathbb{N})$ and (c) $\frac{a+(n-1)d}{n} \in V^{NSS}(\mathbb{N})$ for all $n \in \mathbb{N}$.

(a) In view of the fact that lemmas 2-5 can be generalized as claimed above, it is sufficient to show that if $p \in V^{NSS}(\mathbb{N})$ is not of politeness class n, for any $n \in \mathbb{N}$, then $v(p,p) = d$. Thus suppose $p \in V^{NSS}(\mathbb{N})$ is not of politeness class n for any positive integer n. Then either (a1) there exists no used message that plays e^1 against itself, or (a2) there exist an infinite set $M' \subset M(p)$ of used messages that play e^2 against each other and e^1 against themselves.

In case (a1) all used messages play e^2 against themselves, by lemma 2. If there is only one used message, then $v(p,p) = d$. If there is more than one used message and $v(p,p) < d$, then some pair (m, m') of used messages, $m \neq m'$, play e^1 against each other. But then $p \notin V^{NSS}(\mathbb{N})$ since an alternative best reply to p then is the meta strategy $q \in \Delta(H)$ that lets all message pairs play like in p, but uses only, say, message m. Clearly $v(q,q) = d > v(p,q) = v(q,p) = v(p,p)$.

In case (a2), suppose $v(p,p) < d$. Then $v(p,p) < \frac{a+(n-1)d}{n}$ for some $n \in \mathbb{N}$. But then $p \notin V^{NSS}(\mathbb{N})$, since there exist alternative best replies to p that earn more

against themselves than p earns against them. For instance, let $q \in \Delta(H)$ have all message pairs play against each other like they do in p, but let q use only, say, $n+1$ of the infinitely many messages in $M(p)$, with equal probability for all. Formally, let $M(q) \subset M(p)$, $|M(q)| = n+1$ and $q(m) = \frac{1}{n+1}$ for all $m \in M(q)$. Clearly $v(q,q) = \frac{a+nd}{n+1} > \frac{a+(n-1)d}{n} > v(p,p) = v(q,p) = v(p,q)$.

(b) It is easily verified that if $p \in \Delta(H)$ lets all message pairs play (e^2, e^2), then $v(p,p) = d$ and $p \in \Delta^{NSS}(H)$.

(c) Let $n \in \mathbb{N}$, $M' = \{1, ..., n\}$, and let all pairs of messages from M' play as in the proof of proposition 4, while the remaining (unused) messages behave just like message $m = n$, and are treated exactly like that message by all messages. Let all messages be used, and let the first $n-1$ messages be used with probabilities, $p(m) = \frac{1}{n}$. This will turn out to define a neutrally stable strategy with payoff $\frac{a+(n-1)d}{n}$. Formally, let $p \in \Delta(H)$ be defined as follows: $p(m) = \frac{1}{n}$ for $m < n$, and $p(m) > 0$ for all $m \in \mathbb{N}$. (Hence $\sum_{m \geq n} p(m) = \frac{1}{n}$). Let each of the $n-1$ first messages play e^2 against all other messages, and e^1 against itself. Let each message $m \geq n$ play e^2 against all of the $n-1$ first messages, and otherwise e^1. Then

$$\begin{aligned} v(p,p) &= \left(1-\frac{1}{n}\right)\left[\frac{1}{n}a + \left(1-\frac{1}{n}\right)d\right] + \frac{1}{n}\left[\frac{1}{n}a + \left(1-\frac{1}{n}\right)d\right] \\ &= \frac{1}{n}a + \left(1-\frac{1}{n}\right)d. \end{aligned}$$

All messages are used in p, all message pairs play (pure strategy) base-game Nash equilibria, and it is easily verified that every message earns $v(p,p)$. Hence, $p \in \Delta^{NE}(H)$ by proposition 1.

In order to show that $p \in \Delta^{NSS}(H)$, first note that $q \in \beta^H(p)$ implies that $q^m(m') = p^{m'}(m) = p^m(m')$ for all $m \in M(q)$ and $m' \in \mathbb{N}$. Since q and p let all "active" message pairs play symmetric and pure base-game strategy profiles against each other, the off-diagonal elements b and c in the payoff matrix A are never used, and so we may assume without loss of generality that $b = c$. Thus $\mathcal{G}$ is doubly symmetric, and consequently for any $q \in \beta^H(p)$ we have $v(p,q) = v(q,p) = v(p,p)$. It thus suffices to show that $v(q,q) < v(p,p)$ for all $q \in \beta^H(p)$. Assume $q \in \beta^H(p)$, and let $Q = \sum_{m \geq n} q(m)$. By (4),

$$\begin{aligned} v(q,q) &= \sum_{m<n} q(m)\left(aq(m) + d\left[1-q(m)\right]\right) + \sum_{m \geq n} q(m)\left(aQ + d\left[1-Q\right]\right) \\ &= (a-d)\sum_{m<n} q^2(m) + (1-Q)d + aQ^2 + dQ\left(1-Q\right). \end{aligned}$$

We now investigate the maximum value of $v(q,q)$ for $q \in \beta^H(p)$. Proceed just as in the proof of proposition 4: for any $q \in \beta^H(p)$, with accompanying sum Q, the sum

of squares, $\sum_{m<n} q^2(m)$, is minimized (subject to the constraint that these $q(m)$'s add up to $1-Q$) if and only if $q(m) = \frac{1-Q}{n-1}$ for all $m < n$. Given this, we have

$$v(q,q) = d - (d-a)\left[Q^2 + \frac{1}{n-1}(1-Q)^2\right].$$

This is a parabola in Q with unique maximum at $Q = \frac{1}{n}$. Hence, $q(m) = \frac{1}{n}$ for all $m \leq n$, and thus $v(q,q)$ is maximized for $q = p$. Consequently, $p \in \Delta^{NSS}(H)$. **End of proof.**

It is not difficult to show that the situation is radically different for evolutionarily stable outcomes: there simply does not exist any evolutionarily stable strategy when the message set is infinite:

Proposition 7. $V^{ESS}(\mathbb{N}) = \emptyset$.

Proof. Suppose $p \in \Delta^{ESS}(H)$. If p does not have full support, then alternative best replies to p exist that differ only at unused messages, and such alternative best replies earn just as much against themselves as p earns against them. Hence, it is necessary that p have full support. But then all message pairs must play (e^2, e^2), and there are lots of alternative best replies to p that earn just as much against themselves as p earns against them. For instance, let q use only one message and have all message pairs play (e^2, e^2). **End of proof.**

7. Concluding comments

An alternative approach to formally study stability with respect to evolutionary forces is to set up an explicitly dynamic model of some evolutionary selection process and then look for outcomes that are stable in that dynamics. One well-studied evolutionary dynamics is the so-called replicator dynamics (Taylor and Jonker [19]). One then imagines a large population of pure-strategists who are randomly matched to play the game in question, here a cheap-talk game. A mixed strategy represents a population state, with probabilities interpreted as population shares of pure strategists. The payoff $v(p,p)$ of a meta-strategy p when playing against itself then is the average payoff in population state p.

It has been shown that evolutionary stability implies asymptotic stability (Taylor and Jonker [19]), and that neutral stability implies Lyapunov stability (Thomas [20], Bomze and Weibull [4]), in the replicator dynamics. Hence, the above analysis of finite cheap talk 2×2 coordination games implies that each payoff in the finite set $V^{NSS}(M)$ is the average payoff in some Lyapunov stable population state in the replicator dynamics, as applied to a cheap-talk coordination game with message set M. Hence, if the population state happens to be such a state, then no small shock

can bring it to move far away. Indeed, the payoff may remain unchanged under a wide range of small and moderate shocks. In the very long run one should expect that the population state, if subject to an infinite sequence of small random shocks, should end up in some asymptotically stable set of population states. Evolutionarily stable sets, studied in the context of 2×2 coordination games by Schlag ([17], [18]), indeed have this property. However, for many economics applications the "medium term" may be more relevant for predictive purposes than the "very long run" (Binmore and Samuelson [2]). For such applications our results may serve as a guide to what is going to happen.

References

[1] Bergin, J. and B. L. Lipman (1995), "Evolution with state-dependent mutations", mimeo. Economics Department, Queen's University.

[2] Binmore, K. and L. Samuelson (1994), "Evolutionary drift", *European Economic Review* 38, 859-867.

[3] Bhaskar V. (1991), "Noisy comminucation and the evolution of cooperation", mimeo. Delhi School of Economics.

[4] Bomze I. and J. Weibull (1995), "Does neutral stability imply Lyapunov stability?", *Games and Economic Behavior* 11, 173-192.

[5] van Damme E. (1987), *Stability and Perfection of Nash Equilibria*, Springer Verlag, Berlin.

[6] Hofbauer J. and K. Sigmund (1988), *The Theory of Evolution and Dynamical Systems*. Cambridge, Cambridge University Press.

[7] Kandori, M., G. Mailath, and R. Rob (1993), "Learning, mutation, and long-run equilibria in games", *Econometrica* 61, 29-56.

[8] Kandori M. and R. Rob (1995), "Evolution of equilibria in the long run: A general theory and applications", *Journal of Economic Theory* 65:2, 383-414.

[9] Kim Y.-G. and J. Sobel (1991), "An evolutionary approach to pre-play communication", mimeo, University of Iowa and University of California at San Diego.

[10] Kohlberg E. and J.-F. Mertens (1986), "On the strategic stability of equilibria", *Econometrica* 54, 1003-1037.

[11] Matsui A. (1992), "Cheap talk and cooperation in society", *Journal of Economic Theory* 54, 245-258.

[12] Maynard Smith, J., *Evolution and the Theory of Games,* Oxford Universtiy Press, 1982.

[13] Maynard Smith, J. and G.R. Price (1973), "The logic of animal conflict", *Nature* 246, 15-18.

[14] Okada A. (1981), "On stability of perfect equilibrium points", *International Journal of Game Theory* 10, 67-73.

[15] Ritzberger K. and J. Weibull (1995), "Evolutionary selection in normal-form games", *Econometrica* 63, 1371-1399.

[16] Robson A.J. (1990), "Efficiency in evolutionary games: Darwin, Nash and the secret handshake, *Journal of Theoretical Biology* 144, 379-396.

[17] Schlag K. (1993), "Cheap talk and evolutionary dynamics", Bonn University Economics Disc. Paper B-242.

[18] Schlag K. (1994), "When does evolution lead to efficiency in communication games", Bonn University, Economics Department, Disc. paper B-299.

[19] Taylor P. and L. Jonker (1978), "Evolutionary stable strategies and game dynamics", *Mathematical Biosciences* 40, 145-156.

[20] Thomas B. (1985), "On evolutionarily stable sets", *Journal of Mathematical Biology* 22, 105-115.

[21] Wärneryd K. (1991), "Evolutionary stability in unanimity games with cheap talk", *Economics Letters* 36, 375-378.

[22] Wärneryd K. (1992), "Communication, correlation, and symmetry in bargaining", *Economics Letters* 39, 295-300.

[23] Weibull J., *Evolutionary Game Theory,* MIT Press, Cambridge (USA), 1995.

[24] Young, P. (1993), "Evolution of conventions", *Econometrica* 61, 57-84.