Games and Economic Behavior **32,** 1–24 (2000)
doi:10.1006/game.1999.0756, available online at http://www.idealibrary.com on IDEAL®

Neutrally Stable Outcomes in Cheap-Talk Coordination Games[1]

Abhijit Banerjee

*Department of Economics, Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139*

and

Jörgen W. Weibull

*Department of Economics, Stockholm School of Economics,
and the Research Institute of Industrial Economics*

Received January 20, 1997

This paper examines equilibrium and stability in symmetric two-player cheap-talk games and specifically characterizes the set of neutrally stable outcomes in cheap-talk 2×2 coordination games. With a finite message set, this set is finite. As the number of messages goes to infinity, the set expands toward a countable limit. The Pareto efficient Nash equilibrium payoff is its unique cluster point. By contrast, the corresponding limit set of strategically stable outcomes is dense in the interval spanned by the Nash equilibrium payoffs of the underlying game. Journal of Economic Literature Classification Number: C70. © 2000 Academic Press

Key Words: coordination games; cheap-talk games; evolutionary stability; neutral stability.

1. INTRODUCTION

There has been a substantial literature on evolutionary aspects of pre-play communication in games, in particular on the possibility that communication and evolution together lead to socially efficient equilibria. Robson

[1] The authors are grateful for helpful comments from two anonymous referees, an associate editor, as well as from Jonas Björnerstedt, V. Bhaskar, David Cooper, Martin Dufwenberg, Glenn Ellison, Drew Fudenberg, Joseph Harrington, Ebbe Hendon, Vijay Krishna, Peter Norman, and Asher Wolinsky. Banerjee thanks the National Science Foundation and the Sloan Foundation for support of this research.

0899-8256/00 $35.00
Copyright © 2000 by Academic Press
All rights of reproduction in any form reserved.

(1990), Wärneryd (1991, 1992, 1993), Sobel (1993), Schlag (1993, 1994), Blume, et al. (1993), Kim and Sobel (1995), and Bhaskar (1998) show how evolutionary criteria, in a variety of settings, have a tendency to select against socially inefficient equilibrium outcomes. Evolutionary solution concepts which have been used in this context are evolutionary stability and neutral stability. A strategy is *evolutionarily stable* (Maynard Smith and Price, 1973) if it is a best reply to itself and a better reply to all other best replies than these are to themselves. A strategy is *neutrally stable* (Maynard Smith, 1982) if it is a best reply to itself and a weakly better reply to all best replies than these are to themselves.

The present paper examines the weaker of these two concepts, neutral stability, and compares its cutting power with that of stringent noncooperative refinements. 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. 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 and a "decision rule" that prescribes a pure base-game strategy for every message pair. Much of the analysis is focused on the resulting payoff outcomes, rather than on the strategies that generate these. An outcome will be called neutrally stable if it is the payoff that results when a neutrally stable strategy meets itself.

We begin by characterizing the set of symmetric Nash equilibria in finite and symmetric two-player cheap-talk games. It is necessary and sufficient that each pair of used messages play a Nash equilibrium of the base game against each other, and that no message, if hypothetically sent, earns more than a used message. This implies, in particular, that adding cheap talk does not take us out of the convex hull of the base-game Nash equilibrium payoffs, a point made in Wärneryd (1992). We go on to study how the set of neutrally stable outcomes changes as the set of messages becomes larger. We show that any neutrally stable outcome which exceeds the minmax payoff of the base game remains neutrally stable when the message set is enlarged. In other words, the set of neutrally stable and strictly individually rational outcomes is nondecreasing in the number of messages available to the players.

We next turn to the special case of 2 × 2 coordination games, which is the setting in which cheap talk has been most closely studied. We first consider the effect of refining the Nash equilibrium concept by way of "trembles" in strategies, and show that any payoff value between the worst and best base-game Nash equilibrium payoffs can be approximated by the outcome of a strictly perfect Nash equilibrium in the cheap-talk game, granted that the message set is sufficiently large.[2] A Nash equilibrium is *strictly perfect* (Okada, 1981) if it is robust to all "trembles" in strategies (Okada, 1981). Viewed as a singleton set, such an equilibrium is a strategically stable set in the sense of Kohlberg and Mertens (1986). Hence, even these stringent refinements have effectively no cutting power in this class of games.

The picture is quite different for neutral stability. In the case of 2 × 2 coordination games, we exactly characterize the set of neutrally stable cheap-talk outcomes: the set of neutrally stable outcomes is a certain finite set that contains *both* strict Nash equilibrium payoffs, for any finite message set. As the number of available messages increases toward infinity, the set of neutrally stable outcomes converges to a countable limit set. If the payoffs in the underlying coordination game are such that the "good" strict Nash equilibrium payoff is 2, and that of the "bad" strict Nash equilibrium is 1, this limit set consists of the number 2 and all numbers $2 - \frac{1}{n}$ for positive integers n. The limit set of neutrally stable points thus contains the "bad" and the "good" strict Nash equilibrium outcomes, and an infinite set of isolated points between these extremes, with the "good" Nash equilibrium outcome as the unique cluster point of the set. This result is independent of the payoffs off the diagonal of the base-game payoff matrix. Neutral stability thus offers a selection from the set of Nash equilibria which is distinct from equilibrium selection criteria based on Pareto dominance, risk dominance, perfection, and strategic stability.[3]

The strategies that support the neutrally stable outcomes between the "good" and the "bad" strict Nash equilibrium outcomes have a particular structure. They let a subset of messages form a "group." All messages in the group are sent with equal probability, and messages outside the group are not sent. All pairs of distinct messages from such a "group" play the "good" strict Nash equilibrium, and every message in the group plays the "bad" equilibrium when matched against itself. To anticipate the vocabulary we use later in the paper, the messages in such a group are "polite" to each other. Messages outside the group, however, are "punished" by play of some unfavorable base-game strategy, either the pure strategy associated with the "bad" strict Nash equilibrium, or the minmax strategy. This type of cheap-talk strategy is immune against a small "invasion" of any "mutant" cheap-talk strategy: it is clearly not worthwhile for a mutant to send

[2] Wärneryd (1992) shows that any symmetric payoff outcome in the convex hull of the Nash equilibrium payoffs of the underlying game (the set U^{NE} below) can be approximated by a symmetric Nash equilibrium outcome in a cheap-talk game with a sufficiently large message set.

[3] By contrast, Kandori et al. (1993), Young (1993), and Kandori and Rob (1995) show that certain stochastic dynamics, that can be interpreted as processes of social evolution, select the risk-dominant equilibrium in 2 × 2 coordination games.

a message that does not belong to the group of messages used by the incumbent strategy, since this only results in the "punishment" payoff in the base game. Thus mutants, if they are to be successful, should send messages that belong to the group. However, the best a mutant can do when meeting a message in the group is to mimic the incumbents. So the most a mutant can earn in the post-entry population is the payoff earned by the incumbent strategy, and thus the incumbent strategy is invasion-proof in the sense of neutral stability. It is, however, not evolutionarily stable, unless the group comprises *all* messages. For if there exists an unused message, then there are alternative best replies to the incumbent strategy that do just as well against themselves as the incumbent does against them.

This result is formally established in the paper. Moreover, we establish that this is a complete characterization, in the sense that no other cheap-talk strategy is neutrally stable. These results generalize a result by Schlag (1993, Theorem 5.1). He establishes that the cheap-talk strategy that engages *all* messages in a "polite group" as described above (and hence uses no base-game punishment) constitutes the only evolutionarily stable strategy.

The results reported above hold for any finite set of messages, while in any natural language the set of messages is countably infinite. It is not clear that the two cases should have the same qualitative properties. It is well known that the set of equilibrium outcomes in a repeated game may "explode" as one moves from a finite but arbitrarily distant time horizon to an infinite time horizon. Indeed, there may be a whole plethora of infinite-horizon outcomes that have no counterpart in the finite-horizon case—the Folk theorems establish just this. An important question thus is whether also the set of neutrally stable outcomes in cheap-talk games may "explode at infinity," i.e., as one moves from finite but arbitrarily large message sets to an infinite message set. We show that this does not happen in 2×2 coordination games: The set of neutrally stable outcomes for countably infinite message sets *coincides* with the limit set for finite message sets.

The model developed here admits alternative, more biological or sociological interpretations. Instead of thinking of individuals who send messages, one could think of individuals who are endowed with physical traits or attributes—such as feathers or clothes—along with the ability to distinguish between these.[4] Randomization across messages, as in a mixed strategy, can then be interpreted as a statistical population distribution of these traits.[5]

[4] An earlier version of this paper, Banerjee and Weibull (1993), was explicitly written in this vein.

[5] If all individuals play pure "cheap-talk" strategies, then each neutrally stable outcome corresponds to a dynamically (Lyapunov) stable population state in the replicator dynamics; see Thomas (1985) or Bomze and Weibull (1995).

The implication of our results, if we were to interpret the model in this way, is that variety in observable traits allows for conditioned behaviors, which expands the set of possible evolutionary outcomes, but only in a very specific way. The neutrally stable strategies in the associated “cheap signalling” 2×2 coordination games represent “group behaviors” of exactly three possible types. The first type yields to the highest payoff, the “good” strict Nash equilibrium outcome. In this case, all traits present in the population play the “good” strict Nash equilibrium with each other, and “punish” all traits that are absent from the incumbent population, for example by playing the “bad” strict Nash equilibrium strategy against them. The second type of group behavior results in the “bad” strict Nash equilibrium. This behavior is neutrally stable if this payoff exceeds the minmax payoff in the base game. Here all individuals play the “bad” strict Nash equilibrium with each other, and punish all traits that are absent from the incumbent population by playing the minmax base-game strategy against them. The third type of group behavior that is compatible with neutral stability is based on “group politeness” between the traits present in the population: individuals with differing traits within the group play the “good” strict Nash equilibrium when they meet, while individuals with the same trait play the “bad” strict Nash equilibrium. If an individual appears with a trait that is absent from the incumbent population, then this individual is again “punished” by the incumbents. It also follows from our analysis that the more traits are present in the population (the more messages are sent in equilibrium), the higher is the average payoff (fitness) in the population. In this sense, the analysis suggests that diversity in traits or attributes in a population leads to high population fitness—in interactions that can be modelled as 2×2 coordination games.

To the best of our knowledge, we are the first to characterize the set of neutrally stable outcomes in finite symmetric cheap-talk 2×2 coordination games. While cheap-talk games in evolutionary settings have been widely studied, the emphasis has been on finding settings where the long-run outcome is Pareto efficient. As was first observed by Robson (1990), one way to get this is to assume that an unused message will become available at some point, a message which can be used as a “secret handshake” between mutants—to allow them to play the efficient equilibrium when they meet and therefore to do better than the rest of the population. Wärneryd (1991) analyzes cheap talk in 2×2 coordination games in this vein. He restricts the analysis to pure strategies, with the implication that at least one message is unused in equilibrium (granted at least two messages are available) and that “mutants” cannot be sufficiently severely “punished” (minmaxing in the base game requires randomization). Kim and Sobel (1995) make a similar argument for long-run efficiency, but in their model it is the stochastic drift in the population distribution of strategies which eventually leads to the existence of an unused message.

Instead of studying this kind of process, which inevitably converges to the efficient outcome, we focus on neutral stability. Under neutral stability, the presence of unused messages does not per se undermine the equilibrium, because of the possibility of punishing anyone who sends a new message.[6] Punishing absent messages does not cost anything in equilibrium. However, there are also no benefits to punishing individuals who send such messages. Therefore, there is nothing to stop the population from drifting in the direction of those who do not punish new messages but are all in other respects identical to the incumbents. Eventually, when there are enough incumbents who do not punish such "entrants," new messages may enter. Neutral stability ignores this possibility as being something that can happen in the very long run but is unlikely in the medium run. Evolutionary stability, by contrast, takes the long-run consequences of such drift very seriously, and therefore rejects equilibria with unsent messages.

The role of drift in eliminating unused messages also explains the difference between our results and those of Schlag (1993). He analyzes evolutionary stable outcomes in a setting very similar to ours. He does not make assumptions that guarantee him unsent messages. Therefore, the long-run outcomes in his model are not necessarily fully efficient. He does, however, obtain almost full efficiency when the number of messages is large. These results rest entirely on the fact that evolutionary stability rules out outcomes with unused messages.

Finally, Bhaskar (1998) analyzes neutrally stable outcomes in a setting which is similar to ours, except in that (a) he allows individuals to condition their strategies on their assigned role in the game, and (b) he assumes that communication is noisy in the sense that all messages have some risk of being misinterpreted. The latter assumption rules out unused messages, while the former is shown to rule out mixed-strategy equilibria of the base game, leaving only the efficient outcome. We see this as an alternative and interesting setting but do not feel that it makes our results, for the more standard setting, any less useful.

The material is organized as follows. Notation and definitions are introduced in Section 2. Section 3 provides some preliminaries concerning symmetric two-player cheap-talk games, including a characterization of symmetric Nash equilibria, and a monotonicity result for neutrally stable outcomes, in such games. Section 4 contains our main result for 2×2 coordination games. Section 5 extends the result in Section 4 from finite to countably infinite message sets. Section 6 concludes.

[6] Indeed, many of the outcomes we study here do have unsent messages.

2. DEFINITIONS

2.1. *Symmetric Two-Player Games*

This study is focused on finite and symmetric two-player games in normal form. Let $S = \{1, 2, \ldots, n\}$ be the set of pure strategies, the same for both player positions. Accordingly, a *mixed strategy* is a point σ on the $(n-1)$-dimensional unit simplex $\Delta(S) = \{\sigma \in \mathbb{R}^n_+ \colon \sum_i \sigma_i = 1\}$ in $\mathbb{R}^n$. The *support* of a mixed strategy $\sigma \in \Delta(S)$ is the subset $C(\sigma) = \{i \in S\colon \sigma_i > 0\}$ of pure strategies which are assigned positive probabilities. The set of *strategy profiles* will be denoted $\Theta(S) = \Delta(S) \times \Delta(S)$. Let a_{ij} be the *payoff* to pure strategy i when played against pure strategy j, and let A be the associated $n \times n$ 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 $u\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ 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 ith unit vector in $\mathbb{R}^n$. A finite and symmetric two-player normal-form game will be summarized as a pair $G = (S, u)$.

For each $\mu \in \Delta(S)$, let $\beta(\mu) \subset \Delta(S)$ be its set of (mixed) best replies, and let $\Theta^{\mathrm{NE}}(S)$ denote the set of Nash equilibria. A Nash equilibrium is *strict* if each strategy is the unique best reply to the other. A Nash equilibrium is *strictly perfect* if it is robust to all small "trembles" in strategies (Okada, 1981).[7] A strictly perfect equilibrium, viewed as a singleton set, is strategically stable in the sense of Kohlberg and Mertens (1986). A Nash equilibrium (σ, μ) is *symmetric* if $\sigma = \mu$. By Kakutani's Fixed Point Theorem, every finite and symmetric game has at least one symmetric Nash equilibrium. Let

$$\Delta^{\mathrm{NE}}(S) = \big\{\sigma \in \Delta(S)\colon\ \sigma \in \beta(\sigma)\big\}.^{8} \tag{1}$$

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

A strategy σ is *evolutionarily stable* if $\sigma \in \Delta^{\mathrm{NE}}(S)$ and, moreover, $u(\sigma, \mu) > u(\mu, \mu)$ for all alternative best replies μ to σ. Likewise, a strategy σ is *neutrally stable* if $\sigma \in \Delta^{\mathrm{NE}}(S)$ and $u(\sigma, \mu) \geq u(\mu, \mu)$ for all

[7] Formally, for any positive perturbation vector $\delta = (\delta_i^1, \delta_i^2)_{i \in S}$ such that $M^k(\delta) = \big\{\sigma^k \in \Delta(S)\colon\ \sigma^k(i) \geq \delta_i^k$ for all $i \in S\big\}$ is nonempty 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 games $G(\delta_t)$.

[8] Though the set of Nash equilibria depends not only on the strategy set, but also on payoffs, we suppress payoffs in our notation, since the payoffs will be fixed throughout our analysis, while the strategy set will vary.

best replies μ to σ. Let the subset of evolutionarily and neutrally stable strategies be denoted $\Delta^{\mathrm{ESS}}(S)$ and $\Delta^{\mathrm{NSS}}(S)$, respectively. We have[9]

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

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 error. Again costlessly and without error, the two players then observe each other’s messages, and they simultaneously choose a strategy in G. The set M of possible messages is taken to be the same for both players, and, in the following two sections, this set is finite. The resulting interaction, including the pre-play communication stage, thus constitutes a finite and symmetric two-player game $\mathscr{G}$. Its pure-strategy set H and payoff function v will both be specified below. We will refer to $G = (S, u)$ as the *base game*, M as the *message set*, and call $\mathscr{G} = (H, v)$ the *meta-game* associated with G and M.

A pure strategy in $\mathscr{G}$, a *pure meta-strategy*, is a message to send and a decision rule specifying what pure strategy in G to play after each pair $(m, m') \in M^2$ of sent messages. Such a decision rule can be formally represented as a function f: $M^2 \to S$ that to each message pair $(m, m') \in M$, where m is the own message and m' the opponent’s message, assigns a pure strategy $i = f(m, m')$ in G. Let F be the set of all such functions. A pure meta-strategy thus is a pair $h = (m, f) \in M \times F = H$.

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 $i = f(m, m')$ and $j = g(m', m)$. The payoff matrix of the meta-game $\mathscr{G}$ may thus be represented by the $|H| \times |H|$ matrix $\mathscr{A}$ with entries $\alpha_{hk} = a_{ij}$ in each row $h \in H$ and column $k \in H$, where $h = (m, f)$, $k = (m', g)$, $i = f(m, m')$, and $j = g(m', m)$. The set of *mixed meta-strategies* is the $(|H| - 1)$-dimensional unit simplex $\Delta(H)$ in $\mathbb{R}^{|H|}$. For any pair of mixed strategies $p, q \in \Delta(H)$, the *payoff* to meta-strategy p, when used against meta-strategy q, is

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

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

[9] Here and elsewhere in the paper, we use the sign $\subset$ to denote *weak* inclusion.

3. PRELIMINARIES

3.1. *Characterization of Symmetric Cheap-Talk 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.[10] 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 $m' \in M$, $p_i^m(m')$ is the conditional probability that p assigns to the pure base-game strategy $i \in S$ against message m', given that it sent message m.[11] 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 $(p^m(m'), q^{m'}(m)) \in \Theta(S)$ 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 when played against meta-strategy q as follows:

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

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

LEMMA 1. *$p \in \Delta^{\mathrm{NE}}(H)$ if and only if* (*i*)-(*ii*) *hold.*

(i) $\big(p^m(m'), p^{m'}(m)\big) \in \Theta^{\mathrm{NE}}(S)$, $\forall m, m' \in M(p)$.

(ii) $\sum_{m' \in M(p)} p(m') u\big[p^m(m'), p^{m'}(m)\big] \le v(p,p)$, $\forall m \in M$.

Proof. First, let $p \in \Delta(H)$, and suppose (i) does not hold, i.e., $p^{\bar{m}}(\bar{m}') \notin \beta\big[p^{\bar{m}'}(\bar{m})\big]$ 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 $q^{\bar{m}}(\bar{m}') \in \beta\big[p^{\bar{m}'}(\bar{m})\big]$. Then

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

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

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

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

[11] Formally, $p_i^m(m') = \sum_{f \in A(i,m,m')} p_{(m,f)} / p(m)$, where $A(i, m, m') = \{f \in F\colon\ f(m, m') = i\}$.

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

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

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

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 Eq. (4), so $p \notin \Delta^{\mathrm{NE}}(H)$. Hence, $p \in \Delta^{\mathrm{NE}}(H) \Rightarrow$ (ii).

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

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

Hence, $p \in \beta^H(p)$, so (i)–(ii) $\Rightarrow p \in \Delta^{\mathrm{NE}}(H)$. ■

REMARK 1. By decomposition (4), the inequality in (ii) is an equality for all messages that are used in a symmetric Nash equilibrium: if $p \in \Delta^{\mathrm{NE}}(H)$ and $m \in M(p)$, then

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

3.2. *A Relevant Set of Cheap-Talk Nash Equilibrium Outcomes*

A base-game equilibrium payoff vector is a pair $(x, y) \in \mathbb{R}^2$ such that $(x, y) = (u(\sigma, \mu), u(\mu, \sigma))$ for some $(\sigma, \mu) \in \Theta^{\mathrm{NE}}(S)$. Let $P^{\mathrm{NE}}(S) \subset \mathbb{R}^2$ denote the convex hull of this set, and let

$$U^{\mathrm{NE}} = \Big\{x \in \mathbb{R}\colon\ (x, x) \in P^{\mathrm{NE}}(S)\Big\}. \tag{6}$$

This set turns out to be relevant for the subsequent analysis. It is nonempty and convex by definition, and compact since in a finite game the set of Nash equilibria is compact and payoff functions are continuous. Hence, $U^{\mathrm{NE}} = \big[\underline{x}, \overline{x}\big]$ for some $\underline{x} \leq \overline{x}$. The set U^{NE} clearly contains all payoffs that a player can receive in symmetric base-game Nash equilibrium.[12] In some

[12] Formally, the set of payoffs that a player can receive in symmetric base-game Nash equilibrium is $\{x \in \mathbb{R}\colon\ x = u(\sigma, \sigma) \text{ for some } \sigma \in \Delta^{\mathrm{NE}}(S)\}$.

games it also contains other payoffs. An example of this possibility is the 2×2 game with payoff matrix

$$A = \begin{pmatrix} 0 & a \\ a & 0 \end{pmatrix} \tag{7}$$

for some $a > 0$. Its unique symmetric Nash equilibrium is (σ^*, σ^*), where $\sigma^* = (1/2, 1/2)$. Hence, the only payoff that a player can receive in symmetric base-game Nash equilibrium is $a/2$. However, the game also has two asymmetric Nash equilibria, namely, (e^1, e^2) and (e^2, e^1), both giving payoff a to each player. Thus, U^{NE} is the whole interval $[a/2, a]$.

3.3. *Evolutionarily and Neutrally Stable Cheap-Talk Outcomes*

Let $V^{\mathrm{NE}}(M)$ denote the set of payoff outcomes in symmetric meta-game Nash equilibria when the message set is M,

$$V^{\mathrm{NE}}(M) = \left\{x \in \mathbb{R}\colon\ x = v(p, p) \text{ for some } p \in \Delta^{\mathrm{NE}}(H)\right\}. \tag{8}$$

Likewise, let $V^{\mathrm{NSS}}(M)$ be the set of neutrally stable meta-game payoff outcomes when the message set is M,

$$V^{\mathrm{NSS}}(M) = \left\{x \in \mathbb{R}\colon\ x = v(p, p) \text{ for some } p \in \Delta^{\mathrm{NSS}}(H)\right\}, \tag{9}$$

and likewise for $V^{\mathrm{ESS}}(M)$. By Eq. (2) and Lemma 1,

$$V^{\mathrm{ESS}}(M) \subset V^{\mathrm{NSS}}(M) \subset V^{\mathrm{NE}}(M) \subset U^{\mathrm{NE}}. \tag{10}$$

It is easily shown that there exists no evolutionarily stable strategy when the message set contains more than one message.

PROPOSITION 1. *$V^{\mathrm{ESS}}(M) = \emptyset$ if $|M| > 1$.*

Proof. Suppose $p \in \Delta^{\mathrm{NE}}(H)$ and $p_h > 0$ for $h = (m, f)$. Let $g \in F$ agree with f whenever the own message is m, but differ from f when the own message is some other message m^o. Formally, $g(m, m') = f(m, m')$ for all $m' \in M$, but $g(m^o, m') \neq f(m^o, m')$ for some $m^o \neq m$ and $m' \in M$. Let $q \in \Delta(H)$ differ from p only with respect to the two pure strategies $h = (m, f)$ and $k = (m, g)$, in such a way that $q_h = 0$ and $q_k = p_h + p_k$. Then $q \neq p$, and $v(q, p) = v(p, p) = v(p, q)$, so $p \notin \Delta^{\mathrm{ESS}}(H)$. ■

REMARK 2. It was mentioned in the Introduction that Schlag (1993) identified a certain cheap-talk strategy as being evolutionarily stable. The above proposition may appear to be at variance with that result. The reason for this difference is that our analysis and that of Schlag take place in different normal-form representations of the cheap-talk game. He analyzes the "reduced" normal form that arises when players can condition their base-game strategy choice only on their opponent's message, and not, as here, on both messages. Therefore, the normal-form game in his analysis contains fewer pure strategies. The "spurious copies" of pure meta-strategies that differ only when the own message differs from the "intended" own message have been taken away—and his holds for that normal form. Moreover, his main results concern evolutionarily stable sets (Thomas, 1985), and strategy sets that are asymptotically stable in the replicator dynamics (Taylor and Jonker, 1978), and those results are insensitive to the presence of "spurious" duplicates of pure strategies. The present study is focused on the outcomes associated with the solution concepts of Nash equilibrium, neutral stability, strictly perfect equilibrium, and strategic stability. Also for such analyses, it is immaterial whether one uses the full or the reduced normal form. The reason is that the decision rule f in a pure strategy $h = (m, f)$ is then only applied to message pairs $(\tilde{m}, m')$ where the own message $\tilde{m}$ is m. Without loss of generality, one may thus let f respond in the same way to any pair $(\tilde{m}, m')$, where $\tilde{m} \in M$, as it responds to (m, m'). This reduction of the set of pure strategies is not appropriate, however, in analyses of (point-wise) evolutionary stability, since that solution concept is sensitive to multiplicity of best replies, even if these result in the same outcome.[13]

We are interested in how the set $V^{\text{NSS}}(M)$ varies with the message set M. In particular, one may ask if this set increases as $|M|$ increases. It turns out that this is not always the case.

An example of this possibility is the game with payoff matrix (7). It is well known (since this is equivalent to a "Hawk-Dove" game), and easily verified, that its unique mixed Nash equilibrium strategy σ^* is evolutionarily stable. Hence, $V^{\text{ESS}}(M) = V^{\text{NSS}}(M) = \{a/2\}$ when $|M| = 1$. However, $a/2 \notin V^{\text{NSS}}(M)$ whenever $|M| > 1$. To see this, consider the case of two messages. In order to obtain payoff $a/2$ in such a meta-game, it is necessary, by Lemma 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, $v(q, q) = 3a/4 > v(p, q) = a/2$. Hence, $p \notin \Delta^{\text{NSS}}(H)$.

The reason why, in this example, $V^{\text{NSS}}(M)$ is not nondecreasing in $|M|$ is that the base-game strategy σ^* happens to be a minmax strategy. For any

[13] We thank an anonymous referee for urging us to clarify these distinctions.

finite and symmetric two-player game, let $x^o \in \mathbb{R}$ be its minmax value, i.e.,

$$x^o = \min_{\mu \in \Delta(S)} \max_{\sigma \in \Delta(S)} u(\sigma, \mu). \tag{11}$$

LEMMA 2. *For any base game G and message sets M and M^+ with $|M| \leq |M^+|$:*

(a) *If $x \in V^{\text{NSS}}(M)$ and $x > x^o$, then $x \in V^{\text{NSS}}(M^+)$.*

(b) *If $x^o \notin V^{\text{NSS}}(M)$, then $x^o \notin V^{\text{NSS}}(M^+)$.*

Proof. For (a), assume $x \in V^{\text{NSS}}(M)$ and $x > x^o$. Let $p \in \Delta^{\text{NSS}}(H)$ have $v(p, p) = x$. It is sufficient to consider the case $M = \{1, \dots, k\}$ and $M^+ = \{1, \dots, k, k+1\}$. Let σ^o be a minmax strategy in G. Thus, $u(\sigma, \sigma^o) \leq x^o$ for all $\sigma \in \Delta(H)$. Let H^+ be the set of pure strategies in the meta-game $\mathscr{G}^+$ associated with message set M^+. Let $q \in \Delta(H^+)$ agree with p on H, have message $k+1$ unused, and play σ^o 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 $q^m(k+1) = \sigma^o$, and for all $m' \in M^+$, let $q^{k+1}(m') = \sigma^o$. 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 Lemma 1, $q \in \Delta^{\text{NE}}(H^+)$. Since $p \in \Delta^{\text{NSS}}(H)$: $v(p', p') \leq v(p, p')$ 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 minmaxed 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 $q \in \Delta^{\text{NSS}}(H^+)$.

For (b), assume $x^o \notin V^{\text{NSS}}(M)$ and $x^o \in V^{\text{NSS}}(M^+)$, where $M = \{1, \dots, k\}$ and $M^+ = M \cup \{k+1\}$. Let $p \in \Delta^{\text{NSS}}(H^+)$ have $v(p, p) = x^o$. By Lemma 1, all messages used in p play some base-game minmax 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 $\Delta^{\text{NSS}}(H)$, a contradiction. Suppose instead that all messages are used in p. Then all message pairs play some base-game minmax Nash equilibrium. Let $p' \in \Delta(H^+)$ be like p, except that $p'(1) = 1$ (p' only uses message $m = 1$). Then $q \in \beta^{H^+}(p) \Rightarrow q \in \beta^{H^+}(p')$, and thus $v(q, q) \leq v(p, q) = v(p', q)$, so $p' \in \Delta^{\text{NSS}}(H^+)$. But the restriction of p' to H belongs to $\Delta^{\text{NSS}}(H)$, a contradiction. ■

It follows immediately that if $x^o \notin U^{\text{NE}}$, which is indeed the case in many games, then the set $V^{\text{NSS}}(M)$ is in fact nondecreasing in $|M|$: For any base game G such that $x^o \notin U^{\text{NE}}$, and any message sets M and M^+ with $|M| \leq |M^+|$, $V^{\text{NSS}}(M) \subset V^{\text{NSS}}(M^+)$.

4. CHEAP TALK IN 2×2 COORDINATION GAMES

We here focus on the special case where the base game is a symmetric 2×2 game with payoff matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{12}$$

for some $a > c$, $d > b$. We will call such games *coordination games*. It is well known that their evolutionarily stable strategies are the two pure strategies, and that their unique mixed Nash equilibrium strategy is not even neutrally stable: $\Delta^{\mathrm{NSS}}(S) = \Delta^{\mathrm{ESS}}(S) = \{e^1, e^2\}$. The payoff to each player in the unique mixed-strategy Nash equilibrium, which is symmetric, is $e = (ad - bc)/(a - c + d - b)$.

Consider any game G with payoff matrix as in Eq. (12), 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 $\mathscr{G}$ with finite message set M. Here $\overline{x}$ is the "good" strict Nash equilibrium payoff d, and $\underline{x}$ is the minmax (and also Nash equilibrium) payoff $\min\{a, e\}$. It is easily verified that $\underline{x} = e$ if and only if $a \geq b$. By Eq. (10), all neutrally and evolutionarily stable meta-game payoff outcomes, along with all symmetric meta-game Nash equilibrium outcomes, belong to the interval $U^{\mathrm{NE}} = [\underline{x}, \overline{x}]$.

4.1. *Strictly Perfect Cheap-Talk Outcomes*

In analogy with the notation for neutrally and evolutionarily stable outcomes, let $V^{\mathrm{SP}}(M)$ be the set of strictly perfect meta-game payoff values when the message set is M. Formally, $x \in V^{\mathrm{SP}}(M)$ if and only if there exists a strictly perfect meta-game Nash equilibrium (p, p) with payoff $v(p, p) = x$. Evidently, the set $V^{\mathrm{SP}}(M)$ is a subset of $U^{\mathrm{NE}} = [\underline{x}, \overline{x}]$, for any message set M.

We proceed to show that any payoff in the interval U^{NE} can be approximated by a strictly perfect payoff when the number $|M|$ of messages is large. To state this formally, let $\limsup_{|M| \to \infty} V^{\mathrm{SP}}(M)$ be the smallest set containing all sets $V^{\mathrm{SP}}(M)$ for $|M|$ large, i.e.,

$$\limsup_{|M| \to \infty} V^{\mathrm{SP}}(M) = \bigcap_{n \in \mathbb{N}} \bigcup_{|M| \geq n} V^{\mathrm{NSS}}(M), \tag{13}$$

where $\mathbb{N}$ denotes the set of positive integers.

Proposition 2. *$\limsup_{|M| \to \infty} V^{\mathrm{SP}}(M)$ is dense in U^{NE}.*

Proof. Let $n \in \mathbb{N}$, $x \in U^{\mathrm{NE}}$, and $\varepsilon > 0$. There exists a positive integer $n' \geq n$ and a nonnegative even integer $k \leq n'$, 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 $\underline{\sigma}$ that results in payoff $\underline{x}$ to its $k/2$ nearest neighbors on each side, and let it play e^2 to all other messages, and to itself. More exactly, let the strategy $\underline{\sigma}$ be the unique mixed Nash equilibrium base-game strategy σ^* if $a \geq b$; otherwise, let it be the "bad" pure base-game strategy e^1. Then all messages play base-game Nash equilibria with each other, and all messages earn the same payoff

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

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

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 $P^i(\delta) = \{p \in \Delta(H)\colon\ p_h \geq \delta_h^i \text{ for all } h \in H\}$ is nonempty for $i = 1, 2$. Let $\mathscr{G}(\delta)$ 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 $m \in M$. If $\underline{\sigma} = \sigma^*$, let each message play σ^* to its $k/2$ nearest neighbors on each side, and let it place maximal probability on the pure base-game strategy e^2 against all other messages and against itself. If $\underline{\sigma} = e^1$, let each message place maximal probability on the base-game strategy e^1 against its $k/2$ nearest neighbors on each side, and let it place maximal probability on the pure base-game strategy e^2 against all other messages and against itself. Since $e^1, e^2 \in \Delta^{\mathrm{NE}+}(S)$, and σ^* is completely mixed in the base game, p' is a best reply to itself in the perturbed meta-game $\mathscr{G}(\delta)$, granted the vector $\delta > 0$ is sufficiently small.

The same construction as for strategy p above works for any multiple of the pair (n', k). Hence, the payoff value, $\lambda\underline{x} + (1-\lambda)\overline{x}$, approximating the given payoff $x \in U^{\mathrm{NE}}$, belongs to $V^{\mathrm{SP}}(M)$ for $|M| = n', 2n', 3n', \ldots$, and so on. Thus, $\lambda\underline{x} + (1-\lambda)\overline{x} \in \limsup_{|M|\to\infty} V^{\mathrm{SP}}(M)$. ■

In sum: strict perfection, one of the most stringent noncooperative refinements of the Nash equilibrium concept, has virtually no cutting power on the outcomes in the studied class of games. Since a strictly perfect equilibrium, viewed as a singleton set, is strategically stable in the sense of Kohlberg and Mertens (1986), strategic stability has virtually no cutting power either.

4.2. *Neutrally Stable Outcomes*

We now turn to the main result of this study, a characterization of the set of neutrally stable meta-game outcomes. As a first step towards this goal, we show that neutral stability in the meta-game requires that all used messages play pure strategies against each other. We will say that a message $m \in M(p)$ is *nice* in $p \in \Delta(H)$ to a message $m' \in M$ if m plays the "good" strict Nash equilibrium against m', i.e., if $p^m(m') = e^2$. We establish that if some used message is nice to itself, then every used message is nice to all used messages. Consequently, the payoff is then maximal.

LEMMA 3. *Suppose* $p \in \Delta^{\mathrm{NSS}}(H)$.

(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 Lemma 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 $p \notin \Delta^{\mathrm{NSS}}(H)$.

(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 $p \notin \Delta^{\mathrm{NSS}}(H)$. ■

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:

$$N(m, p) = \{m' \in M\colon\ m' \text{ nice to } m \text{ in } p\}. \tag{14}$$

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 nonempty subset of messages $M' \subset M(p)$ with $|M'| = n$ is polite in p, and no larger subset of $M(p)$ 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 is the politeness class, the higher is this lower bound.

LEMMA 4. *Suppose* $p \in \Delta^{\mathrm{NSS}}(H)$ *is of politeness class n. Then* $v(p, p) \geq \frac{1}{n}a + (1 - \frac{1}{n})d$.

Proof. Let $\emptyset \neq M' \subset M(p)$ be polite in $p \in \Delta^{\mathrm{NE}}(H)$, 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, $p \notin \Delta^{\text{NSS}}(H)$ 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

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

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 strategy profiles (for the second term). Reversing the order of summation in the second term, and using Remark 1, we obtain

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

■

For any nonempty 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 5. *For any $\emptyset \neq M' \subset M$ and $p \in \Delta(H)$,*

$$\Pr\left[\bigcap_{m\in M'} N(m,p) \mid p\right] \geq 1 - \left|M'\right| + \sum_{m\in M'} \Pr[N(m,p) \mid p]. \tag{15}$$

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

LEMMA 6. *Suppose $v(p,p) < d$ and $p \in \Delta^{\mathrm{NE}}(H)$ is of politeness class n. Then $v(p,p) \leq \frac{1}{n}a + \left(1 - \frac{1}{n}\right)d$.*

Proof. Let $M' \subset M(p)$ be polite in p, with $\left|M'\right| = n$. Since no $M'' \subset M(p)$ with $\left|M''\right| > n$ 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 3. Hence, $\bigcap_{m'\in M'} N(m',p) = \emptyset$. Moreover, by Lemma 1, each $m \in M(p)$ earns payoff $a + \Pr[N(m,p)|p](d-a) = v(p,p)$. Since $M' \subset M(p)$, this equation holds for all $m' \in M'$. An application of Lemma 5 to the set M' gives $0 \geq 1 - n + n[v(p,p) - a]/(d-a)$, which is equivalent to the claimed inequality. ■

LEMMA 7.

$$V^{\mathrm{NSS}}(M) \subset \left\{a, \frac{a+d}{2}, \frac{a+2d}{3}, \ldots, \frac{a+(|M|-1)d}{|M|}, d\right\}. \tag{16}$$

Proof. Every $p \in \Delta^{\mathrm{NSS}}(H)$ is either of politeness class n for some positive integer $n \leq |M|$ or else $v(p,p) = d$. Lemmas 4 and 6 give Eq. (16). ■

The following proposition establishes that the inclusion in Lemma 7 in fact is an equality, thus characterizing the set of neutrally stable outcomes in every finite cheap-talk extension of every 2 × 2 coordination game.

PROPOSITION 3.

$$V^{\mathrm{NSS}}(M) = \left\{a, \frac{a+d}{2}, \frac{a+2d}{3}, \ldots, \frac{a+(|M|-1)d}{|M|}, d\right\}. \tag{17}$$

Proof. Let $n = |M|$. The minmax payoff is $\underline{x} = \min\{a, e\} < \frac{a+d}{2}$. To see that $a, d \in V^{\mathrm{NSS}}(M)$, it suffices to note that if all messages are used and all messages play e^1 (e^2) against all messages, then the payoff a (d) results, and the associated meta-strategy p is neutrally stable. Hence, by Lemmas 2 and 7, it is sufficient to show $\frac{a+(n-1)d}{n} \in V^{\mathrm{NSS}}(M)$. 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 $p \in \Delta^{\mathrm{NSS}}(H)$, 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. 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 $v(q, q) \leq v(p, p)$ for all $q \in \beta^H(p)$. By Eq. (4),

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

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 $m \in M$.[14] In sum, $q \in \beta^H(p) \Rightarrow v(q, q) \leq v(p, p)$. Hence, $p \in \Delta^{\text{NSS}}(H)$. ■

REMARK 3. The argument used in the proof of this proposition can be used, *mutatis mutandis*, to establish that the meta-strategy pair (p, p) forms a strictly perfect equilibrium.[15]

REMARK 4. By the same argument as in the proof of Proposition 3, one can establish that $[a + (n - 1)d]/n$ is an evolutionarily stable outcome in the "reduced" normal-form game where the choice of base-game strategy is conditioned on the opponent's message only. As mentioned in the Introduction, this finding conforms with Schlag's (1993) result that this is the only outcome that can arise from an evolutionarily stable strategy in the reduced game.[16]

[14] 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 nonnegative and sum to 1, 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., $M' = M$.

[15] This observation may be compared with van Damme's (1987) 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.

[16] He also shows that the Pareto efficient outcome d is obtained in an evolutionarily stable set. An *evolutionarily stable set* (Thomas, 1985) is a closed set of symmetric Nash equilibrium strategies such that strategies in the set earn at least the same payoff against all nearby best replies as these earn against themselves, with a strict inequality if the best reply is outside the set.

REMARK 5. Suppose that, although the game is symmetric, all individuals can distinguish the two player positions in the game and are allotted one of these positions in each matching. Then individuals can condition their cheap-talk strategy on the position that they happen to be assigned in an interaction. It is well known that in such a setting, evolutionary stability is equivalent to strict Nash equilibrium (Selten, 1980). Hence, in any cheap-talk 2 × 2 coordination game with more than one message, evolutionarily stable outcomes cease to exist. It is easily verified that the set of neutrally stable outcomes then is reduced to the set $\{a, d\}$.[17] Other outcomes are vulnerable to invasion of mutants who send one message in one position of the game and another message in the other. Such mutants will earn the same payoff as the incumbents when meeting these, and earn the "good" base-game equilibrium payoff when meeting each other.

An immediate consequence of Proposition 3 is that the sequence of sets $V^{\mathrm{NSS}}(M)$, for $|M| = 1, 2, \ldots$, is growing. Each time a new message is added to a finite message set M, all neutrally stable outcomes remain neutrally stable in the new cheap-talk game, and one more neutrally stable outcome is added, viz., the payoff $\frac{a+|M|d}{|M|+1}$. Thus, the limit set $\lim_{|M|\to\infty} V^{\mathrm{NSS}}(M)$ is well-defined, and we have established[18]

$$\lim_{|M|\to\infty} V^{\mathrm{NSS}}(M) = \left\{x \in \mathbb{R}\colon\ x = \frac{1}{n}a + \left(1 - \frac{1}{n}\right)d \text{ for some } n \in \mathbb{N}\right\} \cup \{d\}.$$

5. INFINITE MESSAGE SETS IN 2 × 2 COORDINATION GAMES

In any natural language, the set of possible statements is uncountably infinite. Hence, the above assumption that the message set M be finite is not innocuous. It is well known from the repeated-games literature that the equilibrium correspondence may be discontinuous (lack lower hemicontinuity) "at infinity." More exactly, the limit set of finite-horizon equilibrium outcomes, as the number of time periods goes to infinity, is always a subset of the set of equilibrium outcomes when the number of time periods

[17] We thank an anonymous referee for pointing this out to us.

[18] The formal definition of this limit set is

$$\begin{aligned}\lim_{|M|\to\infty} V^{\mathrm{NSS}}(M) &= \bigcap_{n\in\mathbb{N}} \bigcup_{|M|\geq n} V^{\mathrm{NSS}}(M)\\ &= \bigcup_{n\in\mathbb{N}} \bigcap_{|M|\geq n} V^{\mathrm{NSS}}(M),\end{aligned}$$

granted the two sets coincide, which they do since the sequence $\{V^{\mathrm{NSS}}(M)\colon\ |M| \in \mathbb{N}\}$ is increasing.

is infinite, but there may also be lots of infinite-horizon outcomes that cannot be approximated in the finite-horizon case. One may hence ask whether the same is true in the present context: Do there exist neutrally stable outcomes in the case of an infinite message set that cannot be approximated by using a finite, but arbitrarily large, message set?

For the purpose of investigating this question, we now assume $M = \mathbb{N}$, and reexamine all results established above for finite message sets. We then need 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 Eq. (3) since the set of numbers α_{hk}, for $h, k \in H$, is bounded by $\pm \max_{i,\, j \in S} |a_{ij}|$. Consequently, the definitions of Nash equilibrium, evolutionary and neutral stability, etc., may be accordingly extended.[19] Likewise, the decomposition formula (4) still holds, and the proof of Lemma 1 applies to any countable set M. Inspection of the proofs of Lemmas 3 through 6 reveals that these are valid for any countable set M, positive integer n, and finite subset $M' \subset M$. This fact can be used to establish that the set of neutrally stable outcomes is "continuous at infinity."

PROPOSITION 4. $V^{\text{NSS}}(\mathbb{N}) = \lim_{|M| \to \infty} V^{\text{NSS}}(M)$.

Proof. It remains to show (a) $V^{\text{NSS}}(\mathbb{N}) \subset \lim_{|M| \to \infty} V^{\text{NSS}}(M)$, (b) $a, d \in V^{\text{NSS}}(\mathbb{N})$, and (c) $[a + (k-1)d]/k \in V^{\text{NSS}}(\mathbb{N})$ for all $k \in \mathbb{N}$, $k > 1$.

(a) In view of the fact that Lemmas 3–6 can be generalized as claimed above, it is sufficient to show that if $p \in V^{\text{NSS}}(\mathbb{N})$ is not of politeness class n, for any $n \in \mathbb{N}$, then $v(p, p) = d$. If $p \in V^{\text{NSS}}(\mathbb{N})$ is not of politeness class n for any positive integer n, then either (a1) no used message 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 3. 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 $p \notin V^{\text{NSS}}(\mathbb{N})$ 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 $p \notin V^{\text{NSS}}(\mathbb{N})$, since there exist alternative best replies to p, that earn more against themselves than p earns against them. For instance,

[19] However, neutral stability no longer guarantees a uniform "invasion barrier"; see Bomze and Pötscher (1989) for alternative evolutionary stability criteria for infinite games.

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) To see that $a, d \in V^{\mathrm{NSS}}(M)$, it suffices to note that if all messages are used ($M(p) = \mathbb{N}$), and all messages play e^1 (e^2) against all messages, then the payoff a (d) results, and the associated meta-strategy p is neutrally stable.

(c) Let $k \in \mathbb{N}$, for $k > 1$, and let $p \in \Delta(H)$ be defined as follows. First, $p(m) = \frac{1}{k}$ for all messages $m \le k$. Second, each of the first k messages play e^1 against itself, e^2 against all other messages $m \le k$, and e^1 against all messages $m > k$. Third, each message $m > k$ plays e^1 against all messages. Then $v(p,p) = \frac{1}{k}a + (1 - \frac{1}{k})d$, a number that exceeds $a = v(e^1, e^1)$. All message pairs play base-game Nash equilibria in p, and no message earns more than $v(p,p)$. Hence, $p \in \Delta^{\mathrm{NE}}(H)$ by Lemma 1.[20]

In order to show that $p \in \Delta^{\mathrm{NSS}}(H)$, first note that $q \in \beta^H(p)$ implies $M(q) \subset M(p)$ and $q^m(m') = p^{m'}(m) = p^m(m')$ for all $m \in M(q)$ and $m' \le k$. Since q and p have all message pairs play symmetric pure base-game strategy profiles, the off-diagonal elements b and c in the payoff matrix A are never used. As in the proof of Proposition 3, it thus suffices to show that $v(q,q) < v(p,p)$ for all $q \in \beta^H(p)$. The same argument as used there applies here too, implying that $v(q,q)$ is maximized when $q(m) = \frac{1}{k}$ for all $m \le k$. Hence, $v(q,q) \le v(p,p)$ for all $q \in \beta^H(p)$.[21] ■

It is not difficult to show that, just as in the case of finite message sets, there does not exist any evolutionarily stable strategy when the message set is infinite. Unlike in the case of finite message sets, though, this last claim is valid even when players condition their base-game strategy on their opponent's message only. To see this, first note that evolutionary stability requires that all messages be used, since otherwise deviations at unused messages result in alternative best replies that do just as well against themselves as the incumbent strategy does against them. This is just as in the finite case. In particular, no meta-strategy is of finite politeness class. Since all messages necessarily earn the same payoff in equilibrium (Remark 1), and not all messages can be used with the same probability in the infinite case, no equilibrium meta-strategy can be of infinite politeness class either.

[20] We thank an anonymous referee for pointing out an error in our earlier proof of claim (c).

[21] However, unlike in the case of a finite message set, $v(q,q) = v(p,p)$ does not imply $q = p$ in the reduced normal form. For here q necessarily has unused messages and can thus be altered at these without any payoff consequences.

The only remaining possibility is that all messages play the "good" base-game equilibrium against all messages. However, this does not constitute an ESS even in the reduced normal form, since the probabilities with which messages are sent can be altered without payoff consequences.

6. 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 process and then look for outcomes that are stable in that dynamics. One well-studied evolutionary dynamics is the replicator dynamics (Taylor and Jonker, 1978). 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, in the replicator dynamics, evolutionary stability implies (local) asymptotic stability (Taylor and Jonker, 1978), and that neutral stability implies Lyapunov stability (Thomas, 1985; Bomze and Weibull, 1995). Hence, the above analysis of finite cheap-talk 2 × 2 coordination games implies that each payoff in the finite set $V^{\mathrm{NSS}}(M)$ 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. Consequently, 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, however, one should expect that the population state, if subject to an infinite sequence of small random shocks, will end up in some asymptotically stable set of population states. However, for many economics applications, the "medium term" may be more relevant for predictive purposes than the "very long run"; see Binmore and Samuelson (1994, 1997) who argue for this view. A challenge for future research is to identify the "relative stability," or "relative size of basins of attraction," of the different Nash equilibrium components that correspond to each of the neutrally stable outcomes, a challenge that may require a fair amount of numerical computer simulations.

REFERENCES

Banerjee, A., and Weibull, J. (1993). "Evolutionary Selection with Discriminating Players," Economics Department Working Paper 1616, Harvard University.

Bhaskar, V. (1998). “Noisy Communication and the Evolution of Cooperation,” *J. Econom. Theory* **82**, 110–131.

Binmore, K., and Samuelson, L. (1994). “Drift,” *European Econom. Rev.* **38**, 859–867.

Binmore, K., and Samuelson, L. (1997). “Muddling Through: Noisy Equilibrium Selection,” *J. Econom. Theory* **74**, 235–265.

Blume, A., Kim, Y.-G., and Sobel, J. (1993). “Evolutionary Stability in Games of Communication,” *Games Econom. Behav.* **5**, 547–575.

Bomze, I., and Pötscher, B. (1989). *Game Theoretical Foundations of Evolutionary Stability*. Berlin: Springer-Verlag.

Bomze, I., and Weibull, J. (1995). “Does Neutral Stability Imply Lyapunov Stability?”, *Games Econom. Behav.* **11**, 173–192.

Kandori, M., Mailath, G., and Rob, R. (1993). “Learning, Mutation, and Long-Run Equilibria in Games,” *Econometrica* **61**, 29–56.

Kandori, M., and Rob, R. (1995). “Evolution of Equilibria in the Long Run: A General Theory and Applications,” *J. Econom. Theory* **65**, 383–414.

Kim, Y.-G., and Sobel, J. (1995). “An Evolutionary Approach to Pre-Play Communication,” *Econometrica* **63**, 1181–1193.

Kohlberg, E., and Mertens, J.-F. (1986). “On the Strategic Stability of Equilibria,” *Econometrica* **54**, 1003–1037.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. London: Oxford Universtiy Press.

Maynard Smith, J., and Price, G.R. (1973). “The Logic of Animal Conflict,” *Nature* **246**, 15–18.

Okada, A. (1981). “On Stability of Perfect Equilibrium Points,” *Int. J. Game Theory* **10**, 67–73.

Robson, A. J. (1990). “Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake,” *J. Theoret. Biol.* **144**, 379–396.

Schlag, K. (1993). “Cheap Talk and Evolutionary Dynamics,” Bonn University Economics Department Disc. Paper B-242, Bonn University.

Schlag, K. (1994). “When Does Evolution Lead to Efficiency in Communication Games,” Economics Department Disc. Paper B-299, Bonn University.

Selten, R. (1980). “A Note on Evolutionarily Stable Strategies in Asymmetric Animal Conflicts,” *J. Theoret. Biol.* **84**, 93–101.

Sobel, J. (1993). “Evolutionary Stability and Efficiency,” *Econom. Lett.* **42**, 301–312.

Taylor, P., and Jonker, L. (1978). “Evolutionary Stable Strategies and Game Dynamics,” *Math. Biosci.* **40**, 145–156.

Thomas, B. (1985). “On Evolutionarily Stable Sets,” *J. Math. Biol.* **22**, 105–115.

van Damme, E. (1987). *Stability and Perfection of Nash Equilibria*. Berlin: Springer-Verlag.

Wärneryd, K. (1991). “Evolutionary Stability in Unanimity Games with Cheap Talk,” *Econom. Lett.* **36**, 375–378.

Wärneryd, K. (1992). “Communication, Correlation, and Symmetry in Bargaining,” *Econom. Lett.* **39**, 295–300.

Wärneryd, K. (1993). “Cheap Talk, Coordination, and Evolutionary Stability,” *Games Econom. Behav.* **5**, 532–546.

Weibull, J. (1995). *Evolutionary Game Theory*. Cambridge, MA: MIT Press.

Young, P. (1993). “Evolution of Conventions,” *Econometrica* **61**, 57–84.