Natural Language Equilibrium: Signaling Games

Philip J. Reny*

September 2026

Abstract

We incorporate natural language into signaling games by permitting the sender to make cheap-talk statements that have exogenous literal meaning. We study a convention under which, off path, the receiver interprets any statement by the sender as true according to its literal meaning unless it may be interpreted as a *rational* lie. Sequential equilibria respecting this convention are called *natural language equilibria* (NLE). For a rich language, it is shown that an NLE always exists, and that, generically, every NLE type-action distribution is stable against all perturbations of the sender's strategy in the game obtained from the original by adding n cheap-talk messages for the sender, where n is the number of sender-types. Moreover, while adding language can expand the set of equilibria, the only equilibrium outcomes of the given signaling game that survive the introduction of language are, generically, those that are Kohlberg and Mertens (1986) stable.

Keywords: natural language, language conventions, language games, stable equilibria, sender-stable equilibria, signaling games.

1 Introduction

We seek a theory of equilibrium behavior for noncooperative games in which a formal role is played by a cheap-talk language that is commonly understood by the players. Our focus here is on the class of signaling games largely because its sender-receiver

*Department of Economics, University of Chicago, preny@uchicago.edu. I wish to thank Kyle Bagwell, Dirk Bergemann, Ben Brooks, Eddie Dekel, Songzi Du, Piotr Dworczak, John Geanakopolos, Marina Halac, Emir Kamenica, Frédéric Koessler, Marco LiCalzi, Roger Myerson, Wojciech Olszewski, In-Uck Park, Motty Perry, Andy Postlewaite, Doron Ravid, Joe Root, Joel Sobel, Yair Tauman as well as three referees and the editor for helpful comments, and the Department of Economics at Northwestern University for their hospitality while I was a visiting scholar there from July 2023 to June 2024. Financial support from the National Science Foundation (SES-2049810) is gratefully acknowledged.

structure provides a significant role for language while also allowing for a theory that is relatively simple to describe. Extending these ideas beyond signaling games would be worthwhile.

To be effective, a language must have conventions in place that support informative communication by identifying a wide range of circumstances in which statements can be made and interpreted truthfully according to their literal meanings. Because pre-existing literal meaning can be overridden by equilibrium usage to the contrary, such conventions may be especially important for statements that are made off the equilibrium path. One particularly salient convention, and the one that we will study here, is that under which, *off-path, any statement is interpreted as true unless it may be interpreted as a rational lie,* where rationality is judged relative to the given equilibrium.

To be more precise about this convention, consider a signaling game in which there is a strictly positive prior, p say, over the sender's finite set of private types T, and in which some equilibrium is in effect. Let Q be the set of all probability-vectors over T, let $q = (q(t))_{t \in T}$ be any element of Q, and suppose that the sender deviates by taking an action—which is observed by the receiver—and saying "I have deviated to a strategy (a mapping from my types to randomizations over my actions and statements) that gives my action and this statement positive probability and that, conditional on my action and this statement, induces the Bayes-posterior q over my types."[1] What would be the receiver's posterior belief over the sender's types?

If the receiver were to accept the sender's statement as true according to its literal meaning, the receiver's posterior belief would be q. But the receiver might not accept the sender's statement as true and so his posterior belief, β say, might not be equal to q. In that case, we formalize the language convention by requiring β to be a convex combination of q and some belief γ that gives positive probability only to sender-types for whom the deviation is an equilibrium best reply. We call such beliefs

[1] Note that this statement might be true because the sender's strategy, as a function of his type, might assign the action and statement probability that is proportional to $(q(t)/p(t))_{t \in T}$.

straightforward.

This formalization of the language convention ensures that, after the sender's deviation, the receiver's posterior belief is concentrated on two possibilities, namely, q (the sender's statement is true according to its literal meaning), and γ (the sender's statement is false but, given the receiver's equilibrium response, the sender's statement is rational for every type who makes that statement). So after the sender's deviation, the receiver believes that the sender's statement either is true or is a rational lie.

Under this convention, the receiver's posterior belief can sometimes be a *strict* convex combination of q and γ—i.e., the receiver can sometimes believe that the sender's statement might be true but might also be a rational lie. Then, the receiver will assign positive probability to types t for whom the deviation is best reply (any t with $\gamma(t) > 0$), but he may also simultaneously assign positive probability to types for whom the deviation is strictly unprofitable (this can happen when $q(t) > 0$ and $\gamma(t) = 0$). This may seem puzzling, but it is merely a reflection of the strong influence that literal meaning is assumed to have in our theory. Off-path, literal meaning acts as a focal point, providing a default posterior away from which the receiver's belief may be pulled if the sender's statement can be rationalized as a lie. But even then, the receiver's posterior may continue to give some weight to the statement's literal meaning. Such is the effect of language here. See Examples 5.2 and 5.3, as well as Section 7.5.

To define our solution concept, we add to the original signaling game, henceforth the *base game,* the set of cheap-talk messages Q, resulting in the signaling game $\Gamma(Q)$. Each message $q \in Q$ is interpreted as representing a statement in an underlying natural language such that, if the receiver were to accept that statement as true according to its literal meaning, then the receiver's belief over the sender's types would be q.[2] A *natural language equilibrium (NLE)* is a sequential equilibrium of $\Gamma(Q)$ with straightforward beliefs.

[2]The underlying language may contain many statements that are each represented by q.

To see some of the force of the NLE concept, consider the classic Beer-Quiche game due to Cho and Kreps (1987) shown in Figure 1.[3] There are two sequential equilibria, both pooling. In the "Beer" equilibrium, both sender-types drink beer for breakfast, and in the "Quiche" equilibrium, both sender-types eat quiche for breakfast. In each case, the receiver chooses "Don't Fight" (D) along the equilibrium path. The Quiche equilibrium is eliminated by the Cho-Kreps "intuitive criterion" and by many other refinements. Let us show how natural language equilibrium also eliminates the Quiche equilibrium.

Extend the game by including all of the cheap-talk messages q, namely, all probability vectors over the sender's types, Surly and Wimpy. The two pooling outcomes, Beer and Quiche (with some message q), each followed by Don't Fight, remain the only equilibrium outcomes.

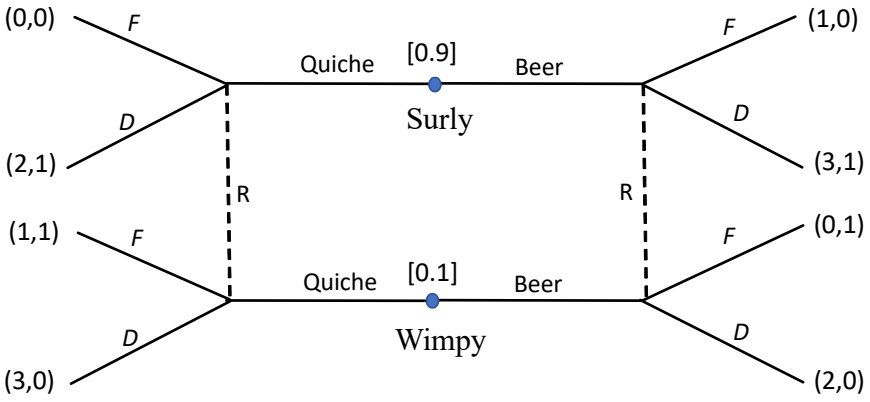


Figure 1. Beer-Quiche

Consider any sequential equilibrium in which, on path, both sender-types choose Quiche and the receiver chooses Don't Fight. Then the surly type's payoff is 2 and the wimpy type's payoff is 3. We will argue that no such sequential equilibrium is an NLE, or, equivalently, that no such sequential equilibrium has straightforward beliefs. For suppose otherwise, and suppose that in such a Quiche equilibrium the surly sender

[3] All game-tree diagrams depict the original signaling game without language—the "base game." Dashed lines indicate R's information sets, the first coordinate in each payoff vector is the payoff to S, and the second coordinate is the payoff to R. The prior is given by the numbers in square brackets.

deviates by choosing Beer and sending the message q such that $q(\text{Surly}) = 1$, which means "I am the surly type." We claim that the receiver will accept this statement as true according to its literal meaning. Otherwise, his posterior belief β would put positive probability on the wimp, implying $\beta(\text{Wimpy}) > 0 = q(\text{Wimpy})$. For such beliefs $\beta \neq q$ to be straightforward, β would have to be a convex combination of q and some γ that puts positive probability only on types for whom the deviation is a best reply. In particular, $\beta(\text{Wimpy}) > 0$ would have to be a convex combination of $q(\text{Wimpy}) = 0$ and $\gamma(\text{Wimpy})$, implying that $\gamma(\text{Wimpy}) > 0$ and that the deviation is a best reply for the wimp. But this is impossible because the wimp gets 3 in equilibrium and can get at most 2 by deviating to Beer no matter how the receiver responds. Hence, the receiver's posterior belief β must put probability one on the surly type as claimed. Intuitively, because the sender's statement "I am the surly type" cannot be rationalized as a lie, the convention is simply to interpret it according to its literal meaning. But this too is impossible because, in that case, the receiver's unique best reply is Don't Fight, which makes the deviation strictly profitable for the surly type, contradicting sequential equilibrium. We conclude that the Quiche equilibrium is not an NLE.[4] Showing that the Beer equilibrium is an NLE is postponed until Section 5.

Our main results are as follows. First, for any finite signaling base-game, an NLE always exists and its type-action distribution—namely, the marginal over the sender's type, the sender's action, and the receiver's action after integrating out the sender's cheap-talk messages—can be induced by an NLE that uses no more than $|T \times Z|$ cheap-talk messages (and no more than $|T|$ for generic base-games) where T is the finite set of sender types, and Z is the finite set of receiver actions. We note that some of these type-action distributions might not arise in any equilibrium of the base game because cheap talk may permit more separation of types than is permitted in any base-game equilibrium. Hence, not surprisingly, the introduction of language can

[4] While the logic of straightforward beliefs may seem similar to that of the intuitive criterion in this illustrative example, we will see presently that the former is typically far more powerful.

expand the set of equilibria.

Second, a type-action distribution ν is an NLE type-action distribution if—and only if for generic base-game utilities—ν is a Nash equilibrium type-action distribution that is stable against all perturbations of the sender's strategy in the signaling game $\Gamma_{|T|}$ that is obtained from the base game by adding $|T|$ cheap-talk messages. Moreover, for generic base-game utilities, there are finitely many NLE type-action distributions.

Finally, for generic base-game utilities, the sequential equilibrium outcomes of the base game that are NLE type-action distributions—these are the equilibrium outcomes of the base game that survive the introduction of language—are precisely the Nash equilibrium outcomes of the base game that are stable in the sense of Kohlberg and Mertens (1986).

We also consider a coarser language that consists only of the $|T|$ statements "my type is t" for each sender-type t. Then, our NLE concept reduces to one that we call *coarse NLE* and it is shown that every coarse NLE is a sequential equilibrium of $\Gamma_{|T|}$ that satisfies the Cho-Kreps (1987) never a weak best reply criterion.[5]

These results for the class of signaling games illustrate the potential benefits of pursuing a broader program to develop a theory of language in games. The benefits include both a theory with sharp predictive power as well as a theory that is based upon intuitive and easily understood conventions for language usage. Language, it seems, may provide a key link connecting strong refinements with simple intuitions.

But incorporating only "off-path" conventions as we do here is insufficient. Indeed, NLE has no refinement power in the sender-receiver games introduced by Crawford and Sobel (1982) and Green and Stokey (2007), and studied by Farrell (1985, 1993). This is because the payoff-irrelevance of the sender's actions there always permits the sender to use strategies with no off-path action-messages.[6] Refining NLE further by developing additional conventions that are effective also in these sender-receiver

[5] Section 7.1 discusses the intermediate language of Farrell (1985, 1993).

[6] However, Farrell's neologism-proofness concept easily applies also to signaling games where NLE is sometimes a strict refinement. See Example 5.3.

games—while still ensuring equilibrium existence—is a high priority.

The remainder of the paper is organized as follows. Section 2 contains notational conventions and other preliminary material, and section 3 defines sequential equilibrium in our setting. Section 4 introduces our solution concept "natural language equilibrium," and provides statements of our main results. Several examples are provided in Section 5. In Section 6 we consider the effect of coarsening the language. Section 7 discusses various points including intermediate languages, the role of counterfactuals, and the introduction of costly messages. Section 8 discusses the literature. Proofs and Lemmas are in Sections 9 and 10.

2 Preliminaries

Throughout, if a set is defined by an upper-case letter, e.g., X, then its elements will be denoted by the corresponding lower-case letter x, x', etc. Then, whenever we write x or x', etc., it is implicit that they are in the set that is their upper-case counterpart, X.

Define the *base game* Γ_0 as follows. There are two players, a Sender (S) and a Receiver (R). Nature chooses S's type t from a non-empty finite set T according to a strictly positive probability measure p. After observing t, S chooses an action a from a non-empty finite set A. Then, R, after observing a but not t, chooses an action z from a non-empty finite set Z. Both players can randomize their action choices. After t, a, and z have been chosen, the players receive their payoffs $u_i(t, a, z)$, $i = S, R$, and the game Γ_0 ends.[7]

Next, let M be a non-empty set of cheap-talk messages m, and suppose that, in addition to taking an action a, the sender S is allowed to send R any message m. The messages in M are payoff irrelevant. So if S's type is t and he chooses the action-message (a, m) after observing t, and R chooses z after observing (a, m), then payoffs

[7] We have assumed that the action sets A and Z are history-independent. This is for simplicity only. All of our results extend to the case in which these sets are history-dependent.

are $u_i(t, a, z)$, $i = S, R$. Let $\Gamma(M)$ denote this cheap-talk extension of the base game Γ_0.

Notice that the base game Γ_0 is equivalent to the game $\Gamma(M)$ whenever M contains exactly one message because this one message must then be sent by all types, rendering it uninformative. We will occasionally make use of this fact.

3 Sequential Equilibria

In this section, we define sequential equilibrium (Kreps and Wilson (1982)) for the game $\Gamma(M)$. These definitions apply also to the base game Γ_0 by setting M equal to a singleton. Since we want to allow the set M to be infinite and we also want the support of the sender's strategy to be well-defined, it will suffice for our purposes to assume henceforth that M is a closed subset of Euclidean space. For any closed subset of Euclidean space, X, let $\Delta(X)$ denote the set of probability measures on the Borel subsets of X.

An *assessment* (σ, ρ, β) consists of a *strategy* $\sigma : T \to \Delta(A \times M)$ for S, a *strategy* $\rho : A \times M \to \Delta(Z)$ for R, and a *system of beliefs* $\beta : A \times M \to \Delta(T)$ for R.[8]

Say that an assessment (σ, ρ, β) is a *sequential equilibrium* of $\Gamma(M)$ iff for every t and z,

(i) $(a, m) \in \arg\max_{a',m'} \sum_{z'} \rho(z'|a', m')u_S(t, a', z')$ for $\sigma(\cdot|t)$ a.e. (a, m),

(ii) $\rho(z|a, m) > 0$ implies $z \in \arg\max_{z'} \sum_{t'} \beta(t'|a, m)u_R(t', a, z')$ for every (a, m), and

(iii) $\sum_{t'} p(t') \int_C \beta(t|a, m)\sigma(a, dm|t') = p(t)\sigma(a, C|t)$, for every a and for every measurable $C \subseteq M$.[9]

Conditions (i) and (ii) are sequential rationality requirements for S and R, respectively, and condition (iii) requires R's beliefs to be Bayes consistent.[10]

[8] When M is uncountably infinite, we require $\rho(z|a, m)$ and $\beta(t|a, m)$ to be measurable functions of m for each t, a, z.

[9] We write $\sigma(a, C|t)$ instead of $\sigma(\{a\} \times C|t)$ and we write $\sigma(a, m|t)$ instead of $\sigma(a, \{m\}|t)$.

[10] Technically, Kreps and Wilson (1982) define sequential equilibrium only for finite games. The definition given here applies whether M is finite or infinite, and coincides with Kreps and Wilson's definition when M is finite because their "consistency" condition on beliefs is equivalent to Bayes'

For any sender-type t, for any action-message (a, m), and for any strategy ρ for R, say that (a, m) is a *best reply for* t *against* ρ iff $(a, m) \in \arg\max_{a',m'} \sum_{z'} \rho(z'|a', m')u_S(t, a', z')$.

For any strategy σ for S, say that an action-message (a, m) is *on-path* (for σ) iff (a, m) is in the support of the probability measure $\sum_{t\in T} p(t)\sigma(\cdot|t)$, and say that a is *on-path* (for σ) iff (a, m) is on-path for some m. If an action or an action-message is not on-path then it is *off-path*.

4 Natural Language Equilibrium

Consider the signaling game $\Gamma(Q)$ that results after adding the set of cheap-talk messages $Q := \Delta(T)$ to the base-game Γ_0. Before defining our solution concept, it will be convenient to provide a second, equivalent, way to formalize the language convention. Up to now, it has been formalized by the requirement that, if after an off-path action-message (a, q), the receiver's posterior belief $\beta(\cdot|a, q)$ is not q, then $\beta(\cdot|a, q)$ must be a convex combination of q and some belief γ that gives positive probability only to sender-types for whom the deviation is a best reply. This can be stated equivalently, but more succinctly, as the requirement that after any off-path (a, q), and for any sender-types t, t', if $\beta(t|a, q)/\beta(t'|a, q) > q(t)/q(t')$ then the deviation must be a best reply for t.[11] Thus, either one of these two equivalent formulations, using convex-combinations of beliefs or ratios of beliefs, can be used as a formal expression of the language convention. Henceforth we will use them interchangeably as convenient.

We can now provide our central definition.

consistency in finite signaling games with or without cheap-talk.

[11] That the former implies the latter follows easily. To see that the latter implies the former, let $b := \beta(\cdot|a, q)$ and let $\lambda := \min_{t\in T:q(t)>0} b(t)/q(t)$ with t^* a minimizer. Then $\lambda \in [0, 1)$ because b and q are distinct and in $\Delta(T)$. Defining $\gamma(t) := (b(t) - \lambda q(t))/(1 - \lambda)$ for each $t \in T$ gives $\gamma \in \Delta(T)$ and $b = \lambda q + (1 - \lambda)\gamma$. Moreover, if $\gamma(t) > 0$ then $b(t) > \lambda q(t) = \frac{\bar{\beta}(t^*)}{q(t^*)} q(t)$ and so $b(t)/b(t^*) > q(t)/q(t^*)$, which implies that the deviation is a best reply for t. This equivalence is noted also in Sobel (2020), Proposition 2.

Definition 4.1 Say that an assessment (σ, ρ, β) is a *natural language equilibrium (NLE)* iff it is a sequential equilibrium of $\Gamma(Q)$ and, for any off-path action-message (a, q) for the sender, and for any pair of sender-types t and t', if $\beta(t|a,q)/\beta(t'|a,q) > q(t)/q(t')$ then (a, q) is a best-reply for t against ρ.[12]

Let us (now formally) call beliefs that satisfy the condition given in Definition 4.1 *straightforward*, and let us call them *straightforward at* (a, q) if the condition holds for the action-message (a, q). Thus, an NLE is a sequential equilibrium of $\Gamma(Q)$ with straightforward beliefs.

In preparation for our results, we introduce some additional terminology. Any terminology for $\Gamma(M)$ applies also to Γ_0 by setting M equal to a singleton.

For a given signaling game with cheap talk $\Gamma(M)$, each strategy profile (σ, ρ) induces a probability distribution over all the endpoints (t, a, m, z) as well as a marginal distribution over the payoff-relevant types and actions (t, a, z). Call the distribution over endpoints the *outcome* of the game and call the distribution over types and actions the *type-action distribution* of the game.[13] If an outcome is the outcome for a Nash equilibrium (σ, ρ), call it a *Nash equilibrium outcome*. If it is the outcome for a sequential equilibrium (σ, ρ, β), call it a *sequential equilibrium outcome*, and so on, and similarly for the type-action distribution. Note that because the endpoints of Γ_0 are the types and actions (t, a, z), the terms "outcome" and "type-action distribution" are equivalent for Γ_0.

For any $\varepsilon \in (0, 1)$, for any $\delta_1, \delta_2 \in (0, \varepsilon)$, and for any strategy profile (σ_0, ρ_0) for $\Gamma(M)$, suppose that when the players choose any strategy profile (σ, ρ) in $\Gamma(M)$, the strategy profile that is implemented is instead $((1-\delta_1)\sigma + \delta_1\sigma_0, (1-\delta_2)\rho + \delta_2\rho_0)$. Call the resulting game an *ε-perturbation of $\Gamma(M)$ toward (σ_0, ρ_0)*. If only the sender's strategy is perturbed, i.e., if $\delta_2 = 0$ and $\delta_1 = \varepsilon$, call the resulting game an

[12] If either one of the two denominators in the strict inequality is zero, the inequality is to be understood as holding if and only if $\beta(t)q(t') > \beta(t')q(t)$. We maintain this convention throughout the paper.

[13] I thank Joel Sobel for suggesting the "type-action distribution" terminology.

ε-perturbation of $\Gamma(M)$ toward σ_0.

Analogous to Kohlberg and Mertens (1986), henceforth KM, when M is finite, say that a Nash equilibrium type-action distribution ν of $\Gamma(M)$ is *KM-stable* iff for any strictly mixed strategy profile, and for any sequence of ε-perturbations of $\Gamma(M)$ toward that strategy profile, there is a corresponding sequence of Nash equilibria of the perturbed games whose type-action distributions converge to ν as ε converges to zero.

Similarly, say that a type-action distribution ν of $\Gamma(M)$ is *sender-stable* iff for any strictly mixed strategy of the sender and for any sequence of ε-perturbations of $\Gamma(M)$ toward that sender-strategy, there is a corresponding sequence of Nash equilibria of the perturbed games whose type-action distributions converge to ν as ε converges to zero.

Sender-stability is less restrictive than stability because it does not require the given type-action distribution to be robust against perturbations of the receiver's strategy. But in terms of restricting the receiver's off-path beliefs, sender-stability and stability are equally powerful since each requires robustness to every perturbation of the sender's strategy.

Define the phrase *for generic base-game utilities* to mean for all base-game utility vectors $u_S(\cdot), u_R(\cdot)$ in a subset of $\mathbb{R}^{T\times A\times Z}$ whose complement is closed and has Lebesgue measure zero.

Our first result states that at least one NLE always exists and that any NLE type-action distribution can be induced by an NLE that has finite support.

Theorem 4.2 *There is at least one NLE. Moreover, every NLE type-action distribution is induced by an NLE that uses no more than $|T \times Z|$ messages in Q.*

Our next result, in part, characterizes the set of NLE type-action distributions. It states that every sender-stable Nash equilibrium type-action distribution of the game obtained by adding $|T|$ cheap-talk messages to the base game is an NLE type-action

distribution and, for generic base-game utilities, conversely. In particular, for generic games, the $|T \times Z|$ messages in Theorem 4.2 can be reduced to $|T|$.

Theorem 4.3 *A type-action distribution ν is an NLE type-action distribution if—and only if for generic base-game utilities—ν is a sender-stable Nash equilibrium type-action distribution of the game obtained from the base game by adding $|T|$ cheap-talk messages for the sender. Moreover, for generic base-game utilities, there are finitely many NLE type-action distributions.*

The genericity results in Theorem 4.3 do not follow from standard genericity results (e.g., Kreps and Wilson (1982), Kohlberg and Mertens (1986)). This is because, after adding cheap-talk messages to the base game, the resulting game is not generic in the usual sense because both S and R are indifferent between any two endpoints that differ only in the sender's cheap-talk message. Therefore, we rely on the genericity results of Park (1997), who studied cheap-talk sender-receiver games. As we note in Section 10, Park's results in fact also cover the more general class of signaling games studied here.

To see the connection between NLE and sender-stability, let M be any finite set of messages and consider a Nash equilibrium type-action distribution ν of $\Gamma(M)$ that is sender-stable. In an ε-perturbation of the game toward a strictly mixed sender-strategy σ_0, any sender-action, a say, that is off-path for ν now receives positive probability. Since ν is sender-stable, there is an equilibrium $(\sigma_\varepsilon, \rho_\varepsilon)$ of the perturbed game whose type-action distribution ν_ε is close to ν. If the sender's strategy σ_ε gives action a probability zero, then the receiver's beliefs after a are the Bayes'-posterior, q_0 say, induced by σ_0.[14] But if the sender's strategy σ_ε puts positive weight on a, then the receiver's beliefs after a (and some message) will be a convex combination of q_0 and the Bayes-posterior, γ_ε say, induced by σ_ε. Moreover, $\gamma_\varepsilon(t) > 0$ only if a is a best reply for t because σ_ε assigns positive weight to a only at types for whom a is a best

[14] Recall that the implemented strategy in the perturbed game is $(1-\varepsilon)\sigma_\varepsilon + \varepsilon\sigma_0$.

reply. Hence, the receiver's beliefs after a are either q_0, or are a convex combination of q_0 and some belief that puts positive weight only on types for whom a is a best reply. Since all of these properties are preserved in the limit as the perturbation vanishes, we obtain straightforward beliefs at (a, q_0) and sequentially rational behavior for the receiver after a that support ν. By varying σ_0 we can similarly construct, for every q, straightforward beliefs at (a, q) and sequentially rational behavior for the receiver after a that support ν.[15] It is then not difficult to construct an NLE whose type-action distribution is ν. This shows that any sender-stable type-action distribution is an NLE type-action distribution. The opposite direction for generic base-game utilities is much harder to prove, but the connections are the same.

We should not expect a clear-cut relationship between the set of sequential equilibria of the base game and the set of natural language equilibria. On the one hand, adding cheap talk to the base game can expand the set of sequential equilibria. On the other hand, restricting R's beliefs to be straightforward refines the set of sequential equilibria.

Nevertheless, it is natural to ask whether any of the sequential equilibrium outcomes of the base game survive the introduction of natural language. Our next result states that at least one sequential equilibrium outcome of the base game must survive. Moreover, generically, those that survive are precisely the KM-stable outcomes, a result that is a direct corollary of a characterization of KM-stable outcomes in generic signaling games due independently to Cho and Kreps (1987) and Banks and Sobel (1987). They show that, for a generic set of base-game utilities, a Nash equilibrium outcome ν of the base-game is KM-stable if and only if after every off-path action a and for every q, there is for R, a belief and sequentially rational behavior, $r \in \Delta(Z)$ say, that support ν and where R's belief is either q or is a convex combination of q and a belief that gives positive probability only to types who would obtain their equilibrium payoff if they deviated to action a and the receiver responded with r. It

[15] Strictly speaking, for q not in the interior of $\Delta(T)$, an additional limit argument is required.

then readily follows that, generically, a Nash equilibrium outcome of the base game is KM-stable iff it is an NLE type-action distribution.[16]

Theorem 4.4 *At least one sequential equilibrium outcome of the base game is an NLE type-action distribution. Moreover, an outcome of the base game is an NLE type-action distribution if—and only if for generic base-game utilities—it is a KM-stable outcome of the base game.*

Remark 4.5 Like sequential equilibria, NLE sometimes employ weakly dominated strategies, though not generically. Even so, there always exists at least one NLE in weakly undominated strategies.

5 Examples

We now consider several examples. Before doing so, it is useful to observe the following.

Remark 5.1 If a sender-action a is on-path in some NLE (σ, ρ, β), then one can always redefine, if necessary, R's beliefs and behavior after (a, q) for every message q—and without changing the equilibrium outcome—so that for every sender-type t, $\beta(t|a, q) > 0$ only if (a, q) is a best reply for t against ρ, thus making R's beliefs straightforward at all of these (a, q).[17] *Consequently, to check that a sequential equilibrium outcome is an NLE outcome, it suffices to check that R's beliefs are straightforward at every (a, q) such that a is off-path.*

[16] Neither Cho and Kreps (1987) nor Banks and Sobel (1987) offer any interpretation of their characterization of KM-stability. Indeed, Cho and Kreps' (1987) conclude their paper by saying, "...if there is an intuitive story to go with the full strength of stability, it is beyond our powers to offer it here."

[17] Bayes-consistency and sequential rationality imply that, for $\sum_{t'} \sigma(\cdot|t')$-a.e. (a, q), $\beta(\cdot|a, q)$ puts positive probability only on types for whom (a, q) is a best reply against ρ. If (a_0, q_0) is one of these, then for every (a_0, q') that is not, we may redefine R's beliefs and behavior after (a_0, q') as being equal to those after (a_0, q_0).

Example 5.2 We return to the Beer-Quiche game of Figure 1, fulfilling here our promise to provide NLE strategies that support the Beer equilibrium. Define S's strategy so that both sender-types choose Beer and send the same message, $\bar{q}$ say.[18]

Define R's beliefs and strategy as follows. After any action-message (Beer, q), R's belief is equal to the prior $p = (0.9, 0.1)$ and he chooses Don't Fight. After any action-message (Quiche, q), if $q(\text{Wimpy}) \geq 1/2$ then R's belief is q and he chooses Fight, while if $q(\text{Wimpy}) < 1/2$ then R's belief is $\beta = (1/2, 1/2)$ and he randomizes uniformly over Fight and Don't Fight.

These strategies form a sequential equilibrium in which the wimp's payoff is 2. To show that they form an NLE, we must show that R's beliefs are straightforward after every off-path action-message. By Remark 5.1 it suffices to check only the off-path messages (Quiche, q) for every q. In fact, since beliefs that coincide with the message q are trivially straightforward, it suffices to check R's beliefs only after messages (Quiche, q) when $q(\text{Wimpy}) < 1/2$. In that case, R's belief is $\beta = (1/2, 1/2)$ and so $\beta(t)/\beta(t') > q(t)/q(t')$ holds only when t is the wimp. Because R's uniform mixture makes the wimp indifferent between the deviation and equilibrium play (both yield the wimp a payoff of 2), R's beliefs are straightforward and we conclude that this Beer equilibrium is an NLE.

In this equilibrium, when S deviates to (Quiche, q) and $q(\text{Wimpy}) < 1/2$, R's posterior belief $\beta = (1/2, 1/2)$ puts positive probability on both sender types, even though that deviation is a best reply only for the wimp! This reflects the focal nature of literal meaning. Off-path, even if S's statement can be rationalized as a lie, its literal meaning, in this case implying that the surly type is more likely than the wimpy type, may still have a residual effect on R's beliefs. This is the import of language here. See also Section 7.5.

[18] Note that $\bar{q}$ need not be the prior. NLE does not pin down which messages are sent on-path. A more complete theory of language should do so, at least to some extent.

Example 5.3 Our second example shows how NLE can eliminate even very refined equilibria. Consider the base game shown in the upper part of Figure 2. The sender S has two equally-likely types, t_1 and t_2, and two actions, l and r. The receiver R has five actions, U', U, C, D, and D'. The lower part of Figure 2 shows R's optimal actions as a function of the probability, denoted by $\Pr(t_1)$, that he assigns to t_1. These are independent of S's action choice. So, for example, after S chooses l (or r), if R believes that there is probability $3/5$ that S's type is t_1, then R has two optimal actions, C and U.[19]

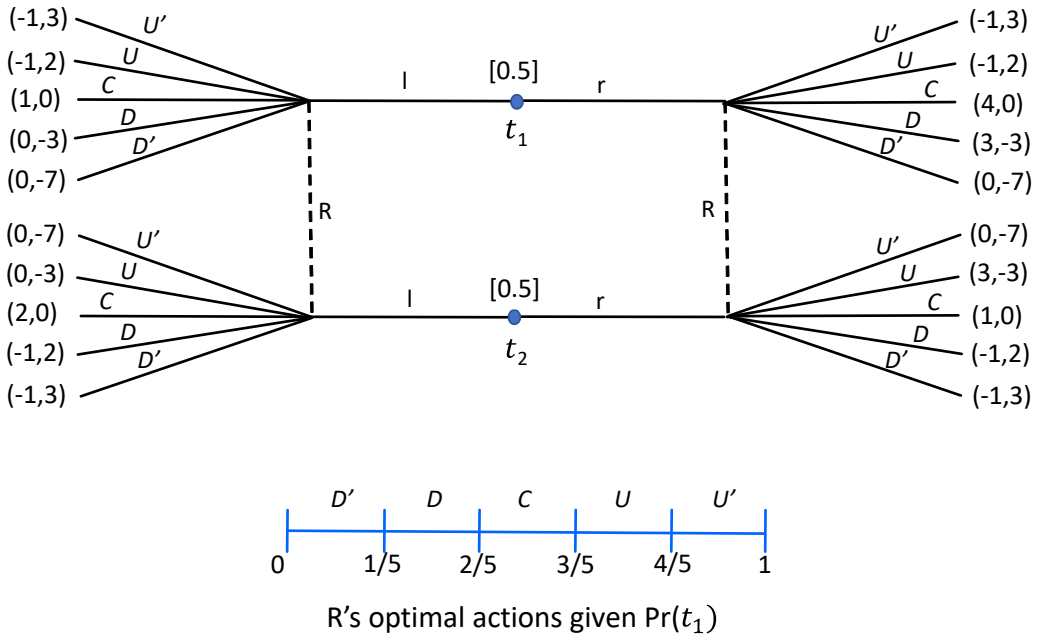


Figure 2.

The base game has two sequential equilibrium outcomes, one in which both sender-types choose l and then R responds with C, giving payoff 1 to t_1 and 2 to t_2, and another in which both sender-types choose r and then R responds with C, giving payoff 4 to t_1 and 1 to t_2.[20] Let us denote these two outcomes by $\nu_{l,C}$ and $\nu_{r,C}$,

[19] The base-game payoffs exhibit coincidences, but all relevant properties hold in an open neighborhood of them.

[20] Both equilibria can be supported by R playing D' if S deviates.

respectively.[21] Both of these sequential equilibrium outcomes survive some of the most demanding equilibrium refinements, including Cho and Kreps' (1987) never a weak best reply criterion.[22] Moreover, both are neologism-proof in the sense of Farrell (1985, 1993).[23] However, only $\nu_{r,C}$ is an NLE type-action distribution and it also happens to be the unique KM-stable outcome of the base game.

In general, the game $\Gamma(Q)$ can have more sequential equilibrium type-action distributions than the base game Γ_0. But, for this example, it can be shown that the only sequential equilibrium type-action distributions of $\Gamma(Q)$ are $\nu_{l,C}$ and $\nu_{r,C}$.

We now show that only $\nu_{r,C}$ is an NLE type-action distribution. To establish this, we first display a sequential equilibrium $(\sigma^*, \rho^*, \beta^*)$ of $\Gamma(Q)$ whose beliefs are straightforward and whose type-action distribution is $\nu_{r,C}$. Here is one of many.

Let p denote the uniform prior over the sender's types, i.e., $p(t_1) = p(t_2) = 1/2$. Define S's strategy σ^* so that both sender-types choose the action-message (r, p) with probability one.

For any action-message (r, q), define R's belief $\beta^*(\cdot|r, q)$ to give both types probability $1/2$, and define R's strategy so that $\rho^*(\cdot|r, q)$ puts probability 1 on C.

For any action-message (l, q), if $q(t_1) \notin (2/5, 3/5)$, then define R's belief $\beta^*(\cdot|l, q)$ to be equal to q and define R's strategy so that $\rho^*(\cdot|l, q)$ is any best reply (given his beliefs) that puts probability zero on C (consult Figure 2 to see that this is possible). If $q(t_1) \in (2/5, 3/5)$, then define $\beta^*(t_1|l, q) = 2/5$, and define $\rho^*(\cdot|l, q)$ so that C and D receive probabilities $2/3$ and $1/3$, respectively. It is easy to verify that $(\sigma^*, \rho^*, \beta^*)$ is a sequential equilibrium. We next show that it is an NLE.

By Remark 5.1, because the only off-path action is l, it suffices to show that the beliefs β^* are straightforward at (l, q) for every message q. Moreover, since beliefs are

[21] More precisely, $\nu_{l,C}$ puts probability one-half each on (t_1, l, C) and (t_2, l, C), and, $\nu_{r,C}$ puts probability one-half each on (t_1, r, C) and (t_2, r, C).

[22] In particular, when both types pool on l, R can respond to the deviation r by mixing 1/3-2/3 on D and D' (resp., U' and U) to make r a best reply for t_1 (resp., t_2).

[23] In each equilibrium both types lose by revealing themselves, and pooling after an action-deviation induces R to play C which makes one type strictly worse off.

trivially straightforward when they are equal to q, we need only check those (l, q) for which $q(t_1) \in (2/5, 3/5)$.

If $q(t_1) \in (2/5, 3/5)$, then $\beta^*(t_2|l, q)/\beta^*(t_1|l, q) = (3/5)/(2/5) > q(t_2)/q(t_1)$. Straightforward beliefs at (l, q) therefore requires (l, q) to be a best reply for t_2, which indeed holds because after (l, q), R puts probability $2/3$ on C and $1/3$ on D, giving t_2 a payoff of 1, which is his equilibrium payoff.[24] Hence $\nu_{r,C}$ is an NLE type-action distribution.

We next show that $\nu_{l,C}$ is not an NLE type-action distribution. Let $(\bar{\sigma}, \bar{\rho}, \bar{\beta})$ be any sequential equilibrium of $\Gamma(Q)$ that induces the type-action distribution $\nu_{l,C}$. Then t_1's payoff is 1 and t_2's payoff is 2. We wish to show that the beliefs $\bar{\beta}$ are not straightforward.

Consider the off-path action r and the message $q = (1/2, 1/2)$. After observing (r, q), R puts probability $\bar{\beta}(t_1|r, q)$ on sender-type t_1. Notice that $\bar{\beta}(t_1|r, q) \neq 1/2$ since, otherwise, sequential rationality would require R to respond with C which would break the equilibrium since t_1 could then profitably deviate to (r, q). So either $\bar{\beta}(t_1|r, q) > 1/2$ or $\bar{\beta}(t_2|r, q) > 1/2$.

Suppose first that $\bar{\beta}(t_1|r, q) > 1/2$. Then, since $q(t_1) = q(t_2) = 1/2$, $\bar{\beta}(t_1|r, q)/\bar{\beta}(t_2|r, q) > q(t_1)/q(t_2)$. Straightforward beliefs at (r, q) then requires (r, q) to be a best-reply for t_1. So (r, q) must yield t_1 his equilibrium payoff 1. However (consult Figure 2), because $\bar{\beta}(t_1|r, q) > 1/2$, this can happen only if $\bar{\beta}(t_1|r, q) = 3/5$ and $\bar{\rho}(\cdot|r, q)$ chooses C and U with probabilities $2/5$ and $3/5$ respectively. But this would break the equilibrium because, in that case, t_2, whose equilibrium payoff is 2, could profitably deviate to (r, q).

Suppose next that $\bar{\beta}(t_1|r, q) < 1/2$. Then, $\bar{\beta}(t_2|r, q)/\bar{\beta}(t_1|r, q) > q(t_2)/q(t_1)$ and straightforward beliefs at (r, q) requires (r, q) to be a best-reply for t_2. So (r, q) must yield t_2 his equilibrium payoff 2. However, this is not possible (consult Figure 2),

[24] Here, R's belief $\beta^*(\cdot|l, q)$ is a strict convex combination of q and γ, where $q(t_1) \in (2/5, 3/5)$ (S's statement is true) and $\gamma(t_1) = 0$ (S's type is t_2 and S's statement is a rational lie). So even though S's statement can be rationalized as a lie, R remains influenced by its literal meaning. See the introduction and also Section 7.5.

because $\bar{\beta}(t_1|r,q) < 1/2$ implies that R will choose some mixture over C, D, and D' which gives t_2 no more than 1. Hence, the beliefs $\bar{\beta}$ are not straightforward and so $\nu_{r,C}$ is the unique NLE type-action distribution.

In our next three examples we use the notation $[t]$ to denote the Dirac measure in $\Delta(T)$ that places probability 1 on type t.

Example 5.4 Our third example illustrates the existence problem that can arise with other approaches and shows how NLE avoids it. The base game shown in Figure 3 is from Farrell (1993), with slightly modified payoffs. The sender has one action, r and so the only equilibrium has both types pooling on r and then R responds with C since both types are equally likely. Type t_2 receives his highest possible payoff of 2. In contrast, type t_1, who also receives a payoff of 2 in equilibrium, could obtain a payoff of 3 if he could reveal his type to R, since R would take action U if he knew S's type was t_1. So it seems credible for type t_1 to say "I am type t_1 and you can be certain that I am being truthful because only I stand to gain by revealing my type to you." If it were understood that R would believe this statement, then t_1 would strictly benefit by making it and this would break the equilibrium.[25]

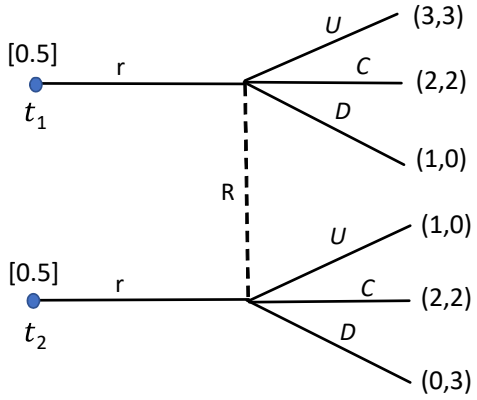


Figure 3.

[25] But some (Joe Stiglitz being perhaps the first) have taken arguments such as this a step further and argued that, because t_2 would then inadvertently reveal his type and receive his lowest possible payoff of zero by *not* making the same statement, t_2 would in fact also make the statement and so t_1 cannot separate himself from t_2 after all. See also Cho and Kreps 1987, p. 203 and footnote 6 there.

Farrell (1985, 1993) calls statements of this kind "credible neologisms." If a credible neologism exists for some equilibrium, then that equilibrium is not "neologism-proof" and Farrell eliminates it from consideration as a possible solution. Since there is a credible neologism for the pooling equilibrium here, no neologism-proof equilibrium exists for this game.[26]

NLE avoids this existence problem by allowing the receiver to be more skeptical. The following is an NLE here. Both sender-types choose (r, p) where p is the prior. After every (r, q), R chooses C and has uniform beliefs.

If, in an attempt to reveal his type, t_1 were to choose the off-path action-message $(r, [t_1])$—recall that $[t_1]$ means "my type is t_1"—then R's response would be to take action C, which makes the deviation a best-reply for *both* types, in particular for t_2. Hence, straightforward beliefs permit R to place positive probability on both types, and indeed R's beliefs do so, placing probability one-half on each type. And given these beliefs, R's choice of C is optimal.

Example 5.5 Our fourth example provides an NLE type-action distribution that is not induced by any strategy profile in the base game because cheap-talk permits additional separation of types. It also shows that the NLE concept sometimes allows too many equilibria.

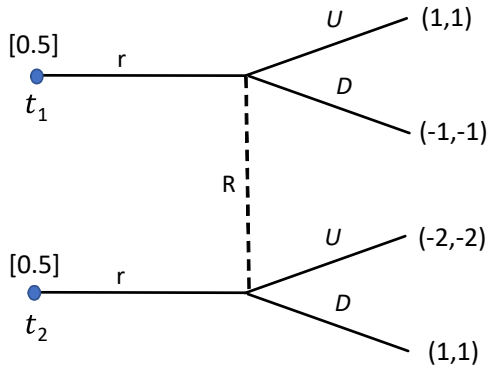


Figure 4.

[26] For critiques of neologism-proofness see Rabin (1990, p.21) and Matthews et. al. (1991).

In the base game shown in Figure 4, the sender has two equally likely types, t_1 and t_2, and has just one action, r. In the unique sequential equilibrium, both types choose r, R's beliefs are uniform, and R chooses D. Payoffs in this "pooling" equilibrium are -1 for t_1, 1 for t_2, and 0 for R.

Adding the cheap-talk messages in Q allows the following separating equilibrium. Type t_1 chooses $(r, [t_1])$ after which R's belief is $[t_1]$ and R takes action U, type t_2 chooses $(r, [t_2])$ after which R's belief is $[t_2]$ and R takes action D. After observing any other (r, q), R's belief is $[t_1]$ and R takes action U. This defines an NLE because after every (r, q), both on-path and off-path, the receiver's belief places probability one on the one type for whom (r, q) is a best reply. Notice that the induced type-action distribution is infeasible in the base game. We note that Farrell's (1985, 1993) neologism-proofness concept uniquely singles out this separating outcome.

But pooling is also an NLE. Indeed, let both sender-types choose (r, p), where p is the uniform prior. After (r, q) for any q, R's belief is uniform and R chooses D. This is an NLE because every (r, q) is a best-reply for *both* types and so all beliefs are straightforward. In particular, after the action-message $(r, [t_1])$ in which the sender's message means "my type is t_1," the receiver can, and does, believe that it is as likely that t_2 sent this message (a rational lie) as it is that t_1 sent it. This exposes a weakness/incompleteness of NLE, which sometimes allows the receiver to interpret a statement as untruthful even when it seems that there would never be an advantage to using that statement untruthfully.

Example 5.6 Our fifth example shows that adding language can make all sender-types worse off. In the base game shown in Figure 5, there is again just one sender-action and so there is a unique sequential equilibrium in which both sender-types take the action r and R responds with C giving both types a payoff of 2.

However, adding the cheap-talk messages in Q permits a sequential equilibrium with separation, e.g., t_1 chooses $(r, [t_1])$ after which R's beliefs are $[t_1]$ and R takes action U, t_2 chooses $(r, [t_2])$ after which R's beliefs are $[t_2]$ and R takes action D, and,

after any other (r, q) R's beliefs are $[t_1]$ and R takes action U. Since after every (a, q), R's beliefs assign positive probability only to types for whom (a, q) is a best reply, this is an NLE. Moreover, both types are now worse off because they each get a payoff of zero.[27] But note that pooling is also an NLE here. In contrast, only the pooling equilibrium is neologism-proof (Farrell (1985,1993)).[28]

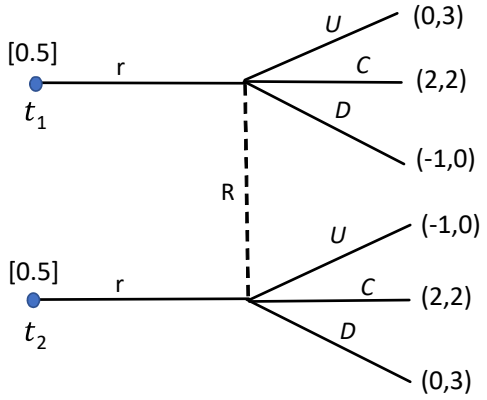


Figure 5.

6 Coarse Natural Language Equilibrium

In this section, we show that even a very coarse language can lead to a powerful equilibrium concept.[29]

Let us coarsen the language by reducing it to the set consisting only of the $|T|$ statements "my type is t," for each $t \in T$. Hence, for this section only, the set of cheap-talk messages is T, and the message $t \in T$ represents the statement "my type is t." Specializing Definition 4.1 to this coarser language leads to the following.

Definition 6.1 Say that (σ, ρ, β) is a *coarse natural language equilibrium* iff it is a sequential equilibrium of $\Gamma(T)$ and, for any off-path action-message (a, t') for the

[27] This once again highlights the incompleteness of our theory. A more complete theory of language in games should exclude this separating equilibrium.

[28] However, for any arbitrarily large number, n say, of equally-likely sender-types, there are examples in which neologism-proofness, by breaking the pooling equilibrium, makes all types but one worse off by n utils, and makes the remaining type better off by just $1/n$ utils.

[29] For settings in which this coarser language may be appropriate see Section 7.2.

sender, and for any $t \neq t'$, if $\beta(t|a,t') > 0$ then (a,t') is a best-reply for t against ρ.

Beliefs that satisfy the condition in Definition 6.1 are called *coarsely straightforward*. Except for the coarser language, the interpretation of coarsely-straightforward beliefs is the same as before, namely, that after any off-path action a and message t', R weighs two possibilities. Either the message "my type is t'" is true or it is a rational lie. This means that R's posterior can put some or all weight on t' and can give positive weight to other types only if the observed action and message is a best reply for those types against R's equilibrium response.

NLE type-action distributions that use more than $|T|$ messages may not be feasible for any strategy profile of $\Gamma(T)$ and hence cannot be coarse NLE type-action distributions. Hence, the existence of a coarse NLE does not follow immediately from Theorem 4.2. Nevertheless, we have the following.

Theorem 6.2 *A coarse NLE exists. In fact, at least one sequential equilibrium outcome of the base game is a coarse NLE type-action distribution.*

We next show that coarse NLE itself is a rather powerful equilibrium concept.

In their influential study of refinements for signaling games, Cho and Kreps (1987) provide a hierarchy of progressively more stringent refinements of sequential equilibrium. The most powerful refinement in this hierarchy, short of stability itself, is a refinement that Cho and Kreps call the "never a weak best reply" criterion which we now define.

For any set of messages M, a sequential equilibrium outcome ν of $\Gamma(M)$ satisfies the (Cho-Kreps) never a weak best reply (NWBR) criterion iff for every action-message (a,m) that is outside the support of ν and is a best reply for some sender-type in some sequential equilibrium of $\Gamma(M)$ with outcome ν, there is a sequential equilibrium (σ,ρ,β) of $\Gamma(M)$ with outcome ν satisfying $\beta(t|a,m) = 0$ for every t such that (a,m) is not a best reply for t in any sequential equilibrium of $\Gamma(M)$ with outcome ν.

Say that a sequential equilibrium of $\Gamma(M)$ satisfies the NWBR criterion iff its outcome satisfies the NWBR criterion.

Cho and Kreps (1987) show that if a sequential equilibrium outcome satisfies their NWBR criterion, then it also satisfies their "intuitive criterion" and two stronger criteria that they call "D1" and "D2," as well as two refinements called "Divinity" and "Universal Divinity" due to Banks and Sobel (1987). Thus NWBR is more powerful than each of these refinements. Our next result states that, in the game $\Gamma(T)$ (which may have more equilibria than the base game), coarse NLE is more powerful still.

Theorem 6.3 *Every sequential equilibrium of* $\Gamma(T)$ *that is a coarse NLE satisfies the Cho-Kreps never a weak best reply criterion, but not conversely (see Example 6.6).*

Remark 6.4 It can also be shown that if an outcome of the base game is a coarse NLE type-action distribution, then it is a sequential equilibrium outcome of the base game that satisfies the NWBR criterion.[30]

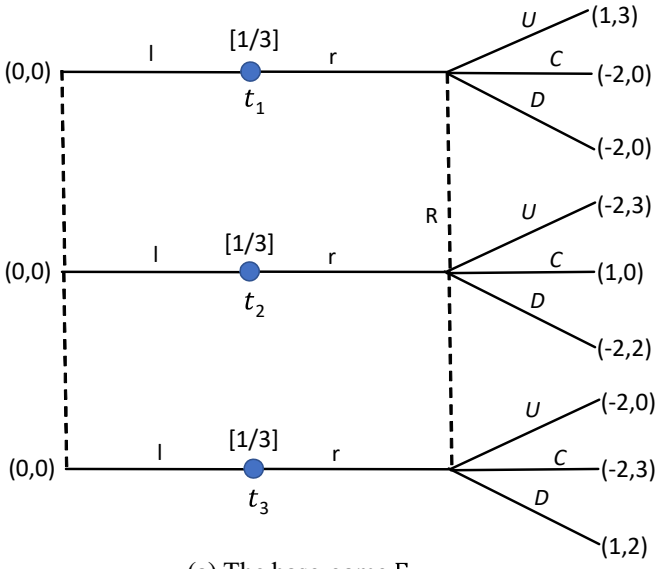
We close this section with two examples.

Example 6.5 Here we provide a coarse NLE type-action distribution that is not an NLE type-action distribution. Recall from Example 5.3 that $\nu_{l,C}$ is not an NLE type-action distribution there. We will show that it is a coarse NLE type-action distribution. Add the set of cheap-talk messages $\{t_1, t_2\}$ to the base game in Example 5.3 and define an assessment as follows. Both sender-types choose (l, t_1) and (l, t_2) with probability one-half each; R chooses C after (l, t_1) and after (l, t_2), R chooses U' after (r, t_1), and R chooses D' after (r, t_2); and R's beliefs put probability $1/2$ on each sender-type after any on-path action-message (l, t), and put probability one on t after any off-path action-message (r, t). It is easy to check that this assessment is a sequential equilibrium. To see that the beliefs are coarsely-straightforward, observe

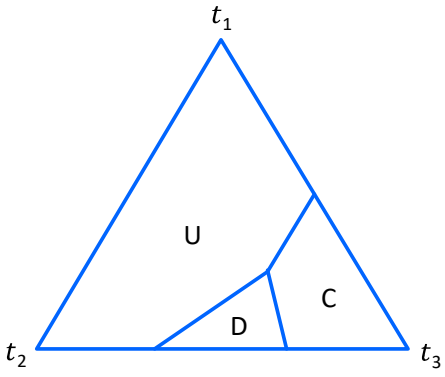
[30] I thank Songzi Du for inquiring about such a result.

that after the only off-path action-messages (r, t') for any t', R's beliefs are concentrated on t' and so are trivially coarsely straightforward because $\beta(t|r, t') > 0$ does not hold for any $t \neq t'$.

Example 6.6 Our final example, due to Cho and Kreps (1987), provides a sequential equilibrium of the base game that satisfies the never a weak best reply criterion but whose outcome is not a coarse NLE type-action distribution. In the base game, the sender has three types, t_1, t_2, and t_3, and two actions l and r; and the receiver has three actions, U, C, and D. If any sender-type chooses l, then R has a strictly dominant action, C say, that gives both players zero. Panel (a) of Figure 6 shows the relevant part of the base-game Γ_0, leaving out the receiver's actions U, C, and D following a choice of l by S because only action C is ever taken and it leads to the payoff vector $(0, 0)$.



(a) The base-game Γ_0



(b) R's best replies after r given beliefs about S's type.

Figure 6.

Panel (b) of Figure 6 shows R's best replies as a function of his beliefs over the sender's types after the sender chooses r. For example U is R's unique best reply when his beliefs are in the interior of the region labelled U, while both U and C are

optimal for R when his beliefs are on the line segment that is between regions U and C. In particular, there is a unique belief for R that makes all three choices U, C, and D optimal for him and it lies at the intersection of the three interior line segments. This belief places probabilities 1/4, 1/4, 1/2 on types t_1, t_2, and t_3, respectively.

Consider any outcome ν of Γ_0 in which every sender-type chooses l. Note that ν is a sequential equilibrium outcome of Γ_0 because R's beliefs after the off-path action r can put probability 1/4 each on t_1 and t_2 in which case mixing uniformly over his actions U, C, and D is optimal for R and makes deviating to r unprofitable for each sender-type. Let (σ, ρ, β) denote this sequential equilibrium.

The outcome ν satisfies the NWBR criterion (see Cho and Kreps (1987)). But ν is not a coarse NLE type-action distribution. Indeed, let (σ, ρ, β) be a sequential equilibrium of $\Gamma(T)$ with type-action distribution ν. It suffices to show that the beliefs β are not coarsely straightforward. Suppose otherwise and consider the off-path action-message (r, t_1). Notice that, after S chooses r, there is no response by R that simultaneously gives any two sender-types an expected payoff of zero and so R's response can make r a best reply for at most one of t_2 and t_3. Hence, because β is coarsely straightforward, either $\beta(t_2|r, t_1) = 0$ or $\beta(t_3|r, t_1) = 0$. Then, Figure 6 panel (b) shows that R's belief is either on the line joining t_1 and t_3 ($\beta(t_2|r, t_1) = 0$) or on the line joining t_1 and t_2 ($\beta(t_3|r, t_1) = 0$). Either way, R's optimal reply is some mixture—possibly degenerate—over U and C. But then (r, t_1) is not a best reply for t_3 and so, by coarsely-straightforward beliefs, we must have $\beta(t_3|r, t_1) = 0$ implying that R's beliefs are on the line joining t_1 and t_2. Given such beliefs, R's unique optimal response is U, making (r, t_1) a profitable deviation for t_1, contradicting sequential equilibrium. We conclude that ν is not a coarse NLE type-action distribution.

The unique coarse NLE type-action distribution here coincides with the unique KM-stable outcome of Γ_0 in which t_1 chooses r and R responds with U, and in which both t_2 and t_3 choose l.

7 Discussion

7.1 Intermediate Languages

We have studied the rich language represented by Q as well as the coarse language represented by T. Farrell (1985, 1993) considers an intermediate language that is represented by the set of nonempty subsets of T where a subset X of T represents a statement that, if interpreted as true according to its literal meaning would induce a posterior equal to the prior p conditioned on X. With this intermediate language, we could define a *semi-coarse* NLE exactly as in Definition 4.1, except that Q would be replaced by the finite set of posteriors obtained by conditioning p on any nonempty subset of T. The set of semi-coarse NLE would, generically, lie between coarse NLE and NLE, and base-game equilibria that are not KM-stable may survive. Unlike both NLE and coarse NLE, semi-coarse NLE would not be invariant to splitting types into copies of themselves. Hence, solutions to equivalent games can be distinct. Splitting types arbitrarily finely would, in the limit, yield NLE with its rich language Q.

7.2 Counterfactuals

According to the language convention here, any off-path statement by the sender either is true or is a rational lie. In particular, the receiver never interprets the sender as having mistakenly stated a falsehood. For statements the sender makes about his strategy—the only statements needed for the language to be sufficiently rich—this can be justified by assuming that the sender knows his strategy. But this assumption may give one pause because it entails counterfactuals. When the sender makes his statement, he knows his type and action, but he may not know how he would have behaved were his type any one of his other possible types. Notice that the issue of counterfactuals does not arise for coarse NLE because there, the only possible statements are of the form "my type is t." Also, if each type t were a different player, then player t does not know any more than R about the randomization over actions

used by any other player t'. So when types are distinct players or when counterfactuals are a concern, coarse NLE may be an attractive concept.

But serious reservations about counterfactuals may be misplaced. Whether the sender knows his strategy or not, we can consider a setting under which he makes claims about his strategy and where the convention for interpreting those claims is as specified here. The proof of this or any other reasonable language convention is, at least to some extent, in the pudding.

7.3 Costly Talk

Our theory can be modified to accommodate statements that are not cheap talk. Let M be the set of messages representing the natural language and allow the sender's payoff $u_S(t, a, m, z)$ to depend on $m \in M$. Assume that each $m \in M$ induces a belief $q_m \in Q$ for R about S's type when R believes that m is true according to its literal meaning and that every belief in Q is induced by some m. The definition of NLE here—and its interpretation—would remain essentially unchanged. Say that (σ, ρ, β) is a *natural language equilibrium* iff it is a sequential equilibrium of $\Gamma(M)$ and, for any off-path action-message (a, m) and for any pair of types t and t', if $\beta(t|a, m)/\beta(t'|a, m) > q_m(t)/q_m(t')$ then (a, m) is a best-reply for t against ρ. Our characterization theorems would of course no longer hold as stated and even establishing existence would require some additional assumptions because there are now potentially infinitely many payoff-relevant action-messages.

A particularly simple such model has messages $M := Q \times [0, \infty)$. The message $m = (q, c)$ has utility cost c and $q_m = q$. So $u_S(t, a, m, z) := u_S(t, a, z) - c$, where $u_S(t, a, z)$ is S's utility in the base game.[31] A model of this kind may resolve the puzzling aspects of Example 5.4 due to Farrell (1993) because now t_1 can, by increasing the cost of his message, separate himself from t_2. This "resolution" of such examples may be as convincing—without sacrificing equilibrium existence—as the never-converging

[31] See Clark (2024) for a similar model.

dynamics discussed in Farrell (1993).

7.4 Stronger Conventions

The "off-path" convention we have studied here can and should be strengthened to ensure that some statements in the language are *always* interpreted as true, *even if they could, in principle, be seen as a rational lie.* Take, for instance, Example 5.5. In this example, there is no conflict between S and R. They both do best when R knows S's type. Hence, for this game, the obvious convention is that each of the statements "my type is t_1" and "my type is t_2" should *always* be interpreted as true. A complete theory of language—which ours is not—should have language conventions in place that, in particular for games of common interest, reduce to the "ideal" convention for effective communication (strategic concerns aside) in which all statements are accepted as true.

7.5 Straightforward Beliefs

In Example 5.2, we remarked that it may seem puzzling that R's (unique) straightforward belief $\beta = (1/2, 1/2)$ following the deviation (Quiche, q) with $q(\text{Wimpy}) < 1/2$ puts positive probability on both sender-types even though the deviation is a best reply only for the wimp. (R has similar beliefs in Example 5.3.) As discussed in the introduction, this reflects the importance that our theory places upon literal meaning, which serves as a focal point that anchors the listener's beliefs. To think about this another way, consider perturbing R's behavior in Example 5.2 after the stated deviation so that R puts slightly more weight on "Fight" than on "Don't Fight." Then R's behavior remains sequentially rational, but now, both sender types strictly lose by deviating. Hence, the receiver's belief $\beta = (1/2, 1/2)$ can now be understood as striking a balance. On the one hand, the receiver, by convention, is drawn toward interpreting the statement represented by q as being true according to its literal meaning and so the receiver is drawn toward believing that the surly type is the more likely since

$q(\text{Wimpy}) < 1/2$. On the other hand, the deviation is only slightly irrational for the wimp and so the receiver may also be drawn toward believing that the wimpy type is the more likely. That β places probability $1/2$ on the surly type no matter how close to rational the deviation may be for the wimp is an expression of the strength of the convention to interpret statements according to their literal meaning, which we consider a hallmark of an effective language. For a generic set of base-game utilities, such a perturbation argument can always be made whenever the receiver's off-path belief assigns positive probability both to types for whom the deviation is a best reply and to types for whom it is strictly unprofitable.[32]

8 Related Literature

Farrell (1985, 1993) initiated the study of language in games and introduced the important concept of a credible neologism. Farrell focussed on cheap-talk sender-receiver games and pointed out that neologism-proof equilibria can fail to exist, a problem that we consider serious. As already remarked, NLE has no refinement power in the cheap-talk games studied by Farrell because, without payoff-relevant actions, the sender can always use a strategy with no off-path action-messages.[33] Additional conventions are needed to refine NLE in these games. Matthews et. al. (1991) and Matthews and Postlewaite (1994), modify Farrell's ideas to resolve various inconsistencies, and Grossman and Perry (1986) define "perfect sequential equilibrium"—a strengthening of Farrell's ideas—for signaling games and sequential bargaining games. None of these authors resolve the existence problem, which, we reiterate, is a serious shortcoming.

Myerson (1989) considers a finite extensive form game setting that includes sig-

[32] That is, for a generic set of base-game utilities, if (σ, ρ, β) is any NLE and a is any off-path action (on-path actions are never a problem by Remark 5.1), then, for any q, we can perturb $\rho(\cdot|a, q)$ arbitrarily slightly so that the assessment remains sequentially rational and so that the deviation (a, q) is strictly unprofitable for every type given positive probability by $\beta(\cdot|a, q)$.

[33] But NLE can strictly refine neologism-proofness in signaling games. See Example 5.3.

naling games as a special case. Myerson's approach to defining credible statements is reminiscent of Farrell's, but with two distinguishing features. First, Myerson allows the players to communicate indirectly through a mediator. Mediated communication yields a set of incentive compatible mechanisms that is convex, which circumvents the non-existence problems that plagued the work of Farrell and others.[34] Our interest here is in direct player to player communication. Second, Myerson considers a convention under which, whenever the mediator, on behalf of another player, recommends an action that is a best reply for the listener, the listener cooperatively carries out that action even if there are many other best replies they could play. Such a "cooperative/bargaining" convention appears to us to go beyond the fundamental connection between a statement's literal meaning and its informational content, being closer in spirit to an equilibrium selection criterion. Our focus here is on literal meaning and the informational content of messages.

To avoid the equilibrium existence problem, Rabin (1990, Section III) and Zapater (1997) modify Farrell's (1993) neologism-proofness concept by using Bernheim (1984) and Pearce's (1984) notion of rationalizability to weaken the definition of credible statements. This can expand the set of solutions well beyond sequential and Nash equilibria.

Mailath et. al. (1993), in response to the "Stiglitz critique,"[35] propose an equilibrium refinement called "undefeated equilibrium" that is neologism-proof and that is not subject to the Stiglitz critique. Undefeated equilibria, which are restricted to pure strategy profiles, do not always exist. However, if the underlying game has ordered action and types sets and satisfies certain conditions that include single-crossing, then undefeated equilibria are shown to exist whenever a pure strategy sequential

[34] In an incentive compatible mechanism for a signaling game, Myerson's (1991) mediator, after receiving the sender's (truthfully) reported type, sometimes stochastically recommends actions for the receiver that are worse for that sender-type than other receiver actions that, if recommended, would be carried out. This can never happen with direct player to player communication and so expands the set of equilibria.

[35] See footnote 25.

equilibrium exists.

Blume and Sobel (1995) define credible messages by considering a particular payoff dominance relation and its associated von Neumann and Morgenstern (1944), henceforth vNM, stable sets of sequential equilibria. They show that there always exists at least one vNM-stable set and that in common interest games any equilibrium in a vNM-stable set must be efficient.[36]

Blume (2026) refines equilibria in cheap-talk sender-receiver games and avoids the equilibrium existence problem by using iterative best replies instead of equilibrium payoffs to determine whether a message is credible. Blume (2026) also considers a number of other interesting issues including determination of the equilibrium meaning of statements in the language and the effects of a language that is not common knowledge.

In the canonical model of Crawford and Sobel (1982), henceforth CS, Gordon et. al. (2023) show that, when the sender is restricted to monotonic strategies, and when the CS regularity condition holds, iterative best reply dynamics and a particular order of iterative elimination of weakly dominated strategies both lead to the most efficient equilibrium.

Sémirat and Forges (2023) also consider an iterative best-reply dynamic for the canonical CS model and show that it always converges to an equilibrium that is undefeated in the sense of Mailath et. al. (1993).

The next several papers perturb the cheap-talk sender-receiver game so that it becomes a signaling game in which the sender has payoff-relevant actions. In these papers, messages are of the form "my type is t" and the authors impose either a cost of lying about one's type or a benefit to telling the truth about one's type and sometimes include additional modifications.

Kartik et. al. (2007) impose a cost of lying about one's type. They also add a small fraction of naive receivers who always believe the sender. A main result is that,

[36] VNM-stable sets for cooperative games are entirely distinct from KM-stable sets for non-cooperative games.

when the sender's type-space is unbounded above, there is a separating equilibrium that features exaggeration—each type overstates his true type.

Kartik (2009) also adds lying costs but assumes a bounded type space. He shows that monotonic strategies together with the bound on types implies partial pooling—hence no separating equilibrium exists. He also shows that a particular class of monotonic equilibria involve exaggeration.

Chen (2011), introduces small fractions of behavioral types for both the sender and receiver similar to Crawford (2003). Honest sender-types truthfully reveal their type and naive receivers believe whatever statement the sender makes. In the limit as behavioral types vanish, and under the Crawford and Sobel (1982), henceforth CS, regularity condition, equilibria involve exaggeration and must satisfy the "no incentive to separate" condition of Chen, Kartik, and Sobel (2008), and so only the most efficient equilibrium survives.

Demichelis and Weibull (2008) suppose that the cost of lying is infinitesimal in that players have lexicographic preferences in which truth is secondary to material payoffs. In generic and symmetric n-player coordination games with a unique Pareto efficient outcome, they show that with such lexicographic preferences for truth, a Nash equilibrium component is evolutionarily stable if and only if it yields the Pareto efficient outcome.

Hart et. al. (2017) study evidence games in which all sender-types prefer the same receiver action. The authors perturb the sender's strategy toward truth-telling, with an additional perturbation toward a sender-preference for truth-telling. With these two perturbations, they show that, in the limit as the perturbations vanish, all sequential equilibrium payoffs for the receiver coincide with his best equilibrium under commitment. That is, commitment has no value for the receiver. Interestingly, the equilibria identified by Hart et. al. are all coarse natural language equilibria (Section 6).

In a learning model, Clark and Fudenberg (2021) study the equilibrium steady

states of a signaling game in which senders can make statements of the form "my type is in this set." If (long-lived) receivers are initially trusting, Clark and Fudenberg identify necessary conditions for a strategy-profile to be a steady state and show that these conditions are satisfied by an equilibrium concept that they call "justified communication equilibrium." These equilibria bear some connection to classic refinements such as Cho and Kreps' (1987) intuitive criterion and D1 criterion. Natural language equilibrium always weakly refines Clark and Fudenberg's (2021) justified communication equilibrium.

Clark (2024) introduces two equilibrium refinements for signaling games, "robust neologism-proofness" and the "flexible speech intuitive criterion." Both of these equilibrium refinements can fail to exist for the same reasons as the classical approaches and so there remains a fundamental difficulty. Even so, Clark (2024) shows that both kinds of equilibria do exist in certain classes of signaling games. In one of these classes, called monotonic games, natural language equilibrium, being a refinement of D1, always weakly refines both of these equilibrium concepts.

9 Proofs

Proof of Theorem 4.2. Existence follows from Theorem 4.4 and the support bound follows from Lemmas 10.1 and 10.2 and the construction in the second part of the proof of Lemma 10.2. ■

Proof of Theorem 4.3. The "if" direction follows from Lemma 10.4 and the "only if" direction follows from Lemma 10.9. The generic finiteness of NLE type-action distributions follows from Lemma 10.1 and Proposition 10.6. ■

Proof of Theorem 4.4.

Let us first prove that if ν is a KM-stable outcome of the base game, then ν is an NLE type-action distribution. Since the base-game Γ_0 is equivalent to the game $\Gamma(\{m_0\})$ obtained by adding a single cheap-talk message m_0, ν is a KM-stable type-

action distribution of $\Gamma(\{m_0\})$. Hence, by a limit argument applied R's pertubed strategy, ν is a sender-stable type-action distribution of $\Gamma(\{m_0\})$. Lemma 10.4 then implies that ν is an NLE type-action distribution.

Next, let us prove that at least one sequential equilibrium outcome of the base-game is an NLE type-action distribution. By Kohlberg and Mertens (1986, paragraph following Remark 1 on p.1027), KM-stable outcomes exist for generic games. Hence, by what we have just shown (and because, as shown by Cho and Kreps (1987), KM-stable outcomes of the base game are sequential equilibrium outcomes), after perturbing base-game payoffs slightly, the perturbed base game has a sequential equilibrium outcome that is an NLE type-action distribution. The desired result follows by observing that the conditions of Lemma 10.2 are preserved under limits.

Lastly, we must show that for a generic set of base-game utilities, every outcome of the base game that is an NLE type-action distribution is a KM-stable outcome of the base game. Using the convex-combination characterization of straightforward beliefs, this follows directly from Banks and Sobel (1987, Theorem 3) and Cho and Kreps (1987, Proposition 4). ■

Proof of Theorem 6.2. By Theorem 4.4 there is a sequential equilibrium outcome ν of the base game that is an NLE type-action distribution. In particular ν is a Nash equilibrium outcome of $\Gamma(\{m_0\})$ for any cheap-talk message m_0. Hence, by Lemma 10.5, ν is a coarse NLE type-action distribution. ■

Proof of Theorem 6.3. Since the second part follows from Example 6.6, we need only prove the first part. Let (σ, ρ, β) be a sequential equilibrium of $\Gamma(T)$ that is a coarse NLE with outcome ν. We may assume that if a is any on-path action for σ, then, for every message $t \in T$, (a, t) is a best reply for every sender-type given positive probability by $\beta(\cdot|a, t)$.[37] We must show that ν satisfies the NWBR criterion.

[37] Indeed, this property holds for every on-path (a, t) by Bayes' consistency. And if (a, t) is off-path for σ, then if necessary we can redefine ρ and β after (a, t) so that they are equal to $\rho(\cdot|a, t')$ and $\beta(\cdot|a, t')$ respectively, where t' is any message such that (a, t') is on-path for σ. The outcome is unchanged and the beliefs remain coarsely straightforward.

So consider any action-message (a_0, t_0) that has probability zero under ν and is a best reply for some type, t_1 say, in some sequential equilibrium of $\Gamma(T)$ with outcome ν. We must show that there is a sequential equilibrium (σ', ρ', β') of $\Gamma(T)$ with outcome ν such that $\beta'(t|a_0, t_0) = 0$ for every t such that (a_0, t_0) is not a best reply for t in any sequential equilibrium of $\Gamma(T)$ with outcome ν.

If a_0 is on-path for σ, then we can take $(\sigma', \rho', \beta') = (\sigma, \rho, \beta)$ since (by the first paragraph) $\beta(\cdot|a_0, t_0)$ gives positive probability only to sender-types for whom (a_0, t_0) is a best reply.

So we may assume that a_0 is off-path for σ. Define $\sigma' := \sigma$, define $\rho' := \rho$ except that $\rho'(\cdot|a_0, t_0) := \rho(\cdot|a_0, t_1)$, and define $\beta' := \beta$ except that $\beta'(\cdot|a_0, t_0) := \beta(\cdot|a_0, t_1)$. (If $t_1 = t_0$, then $(\sigma', \rho', \beta') = (\sigma, \rho, \beta)$.)

Since a_0 has probability zero under σ and under σ', and since the opportunities for deviations for the sender have decreased under (σ', ρ', β') relative to (σ, ρ, β) it is easy to see that (σ', ρ', β') is a sequential equilibrium with the same outcome as (σ, ρ, β), namely ν.

Consider any t such that (a_0, t_0) is not a best reply for t in any sequential equilibrium with outcome ν. It suffices to show that $\beta'(t|a_0, t_0) = 0$. Notice that $t \neq t_1$ by the definitions of t and t_1. If $\beta'(t|a_0, t_0)$ were positive, then $\beta(t|a_0, t_1)$ would also be positive since $\beta'(t|a_0, t_0) := \beta(t|a_0, t_1)$. So, since a_0 is off-path, since $t \neq t_1$, since $\beta(t|a_0, t_1) > 0$, and since β is coarsely-straightforward, (a_0, t_1) is a best reply for t in the sequential equilibrium (σ, ρ, β). Therefore, because messages are cheap talk, (a_0, t_0) is a best reply for t in the sequential equilibrium (σ', ρ', β'), contradicting the definition of t. We conclude that $\beta'(t|a_0, t_0) = 0$, as desired. ■

10 Lemmas

Our first lemma says that at most $|T \times Z|$ messages are needed to support any equilibrium type-action distribution of $\Gamma(M)$. Since related results, all based on

Caratheodory's theorem, are in Hart (1985), Heumann (2020), and Koessler, Laclau, and Tomala (2024), we omit the proof which is available upon request.

Lemma 10.1 *Let M be any measurable space with measurable singleton sets and let $M_{|T\times Z|}$ be any set with $|T\times Z|$ elements. Then any Nash (resp., sequential) equilibrium type-action distribution of $\Gamma(M)$ is a Nash (resp., sequential) equilibrium type-action distribution of $\Gamma(M_{|T\times Z|})$.*

Lemma 10.2 *A type-action distribution ν is an NLE type-action distribution iff ν is a Nash equilibrium type-action distribution of $\Gamma(M)$ for some finite set of messages M, and, for every a satisfying $\sum_{t,z}\nu(t,a,z)=0$ and for every $q\in Q$, there exists $\mu\in\Delta(T)$ and $r\in\Delta(Z)$ such that,*
(a) r is a best reply for R after a when R's beliefs are μ, and
(b) for every t,t', $\pi_S(t)\geq\sum_z r(z)u_S(t,a,z)$ with equality if $\mu(t)/\mu(t')>q(t)/q(t')$, where $\pi_S(t):=\sum_{a,z}\nu(t,a,z)u_S(t,a,z)/p(t)$ is sender-type t's payoff under ν.

Proof. Let us begin with the "only if" statement. Suppose that ν is an NLE type-action distribution. Then there is a sequential equilibrium $(\sigma^*,\rho^*,\beta^*)$ of $\Gamma(Q)$ with straightforward beliefs whose type-action distribution is ν. By Lemma 10.1, there is a message space $M_{|T\times Z|}$ with $|T\times Z|$ messages such that ν is a Nash equilibrium of $\Gamma(M_{|T\times Z|})$, proving the first part of the "only if" statement. To prove the second part, choose any a such that $\sum_{t,z}\nu(t,a,z)=0$ and choose any $q\in Q$. Then, because $(\sigma^*,\rho^*,\beta^*)$ has straightforward beliefs, defining $\mu:=\beta^*(\cdot|a,q)$ and $r:=\rho^*(\cdot|a,q)$ satisfies (a) and (b).

To prove the "if" statement, let $(\hat{\sigma},\hat{\rho})$ be a Nash equilibrium of $\Gamma(M)$ whose type-action distribution is ν. For each $m\in M$ choose a distinct $q_m\in Q$, for each $\hat{\sigma}$-on-path a choose $m_a\in M$ so that (a,m_a) is on path for $\hat{\sigma}$, and for each $\hat{\sigma}$-on-path (a,m) let $\hat{\beta}(\cdot|a,m)$ be the Bayes-posterior on T given (a,m). Define an assessment $(\sigma^*,\rho^*,\beta^*)$ for $\Gamma(Q)$ as follows. For every (a,m) that is on-path for $\hat{\sigma}$, define $\sigma^*(a,q_m|\cdot):=\hat{\sigma}(a,m|\cdot)$, define $\rho^*(\cdot|a,q_m):=\hat{\rho}(\cdot|a,m)$, and define $\beta^*(\cdot|a,q_m):=\hat{\beta}(\cdot|a,m)$. This

defines σ^*, and defines $\rho^*(\cdot|a,q)$ and $\beta^*(\cdot|a,q)$ for every σ^*-on-path (a,q). For any σ^*-off-path (a,q), if $\sum_{t,z}\nu(t,a,z)=0$ then, by hypothesis, we may define $\rho^*(\cdot|a,q)=r$ and $\beta^*(\cdot|a,q))=\mu$ where r and μ satisfy (a) and (b), and if $\sum_{t,z}\nu(t,a,z)>0$ define $\rho^*(\cdot|a,q):=\rho^*(\cdot|a,m_a)$ and $\beta^*(\cdot|a,q)):=\beta^*(\cdot|a,m_a)$. It is easily checked that $(\sigma^*,\rho^*,\beta^*)$ is an NLE with type-action distribution ν (measurability can be obtained from standard results since the correspondence of r's and μ's as q varies has a compact graph). ■

Remark 10.3 *Since the conditions in Lemma 10.2 are preserved under limits, Lemmas 10.1 and 10.2 imply that the correspondence mapping base-game utilities in any compact subset of $\mathbb{R}^{T\times A\times Z}$ to NLE type-action distributions has a compact graph.*

Lemma 10.4 *If M is any non-empty finite set of messages and ν is a sender-stable type-action distribution of $\Gamma(M)$, then ν is an NLE type-action distribution.*

Proof. Since ν is a sender-stable type-action distribution of $\Gamma(M)$, ν is, in particular, a Nash equilibrium type-action distribution of $\Gamma(M)$. For each $t\in T$, let $\pi_S(t)$ denote the payoff to sender-type t from ν, let a_0 be any action such that $\sum_{t,z}\nu(t,a_0,z)=0$ and let q_0 be any element of Q. It suffices to define $r\in\Delta(Z)$ and $\mu\in\Delta(T)$ satisfying (a) and (b) in Lemma 10.2.

Suppose first that $q_0(t)>0$ for every t. Let m_0 be any message in M and let σ_0 be any strictly mixed strategy for S in $\Gamma(M)$ such that $p(t)\sigma_0(a_0,m_0|t)/p(t')\sigma_0(a_0,m_0|t')=q_0(t)/q_0(t')$ for every t,t'.[38] So if S uses σ_0 and R observes (a_0,m_0), then R's Bayes-posterior is q_0.

For each $n=1,2,...$, let G_n be a $\frac{1}{n}$-perturbation of the base-game toward σ_0. Since ν is sender-stable, for each n there is a Nash equilibrium (σ_n,ρ_n) of G_n whose type-action distribution, ν_n say, converges to ν as $n\to\infty$. In particular, because $\sum_{t,z}\nu(t,a_0,z)=0$, we have $\sigma_n(a_0,m_0|t)\to_n 0$ for every t. Let $\tilde{\sigma}_n:=(1-\delta_n)\sigma_n+\delta_n\sigma_0$ be S's perturbed equilibrium strategy in G_n, where $\delta_n\in(0,1/n)$.

[38] Such a σ_0 exists because $\sum_{t,z}\nu(t,a_0,z)=0$ implies that a_0 is not the only action (it is possible that m_0 is the only message).

Since $\tilde{\sigma}_n$ is completely mixed, we may define $\beta_n(\cdot|a_0, m_0)$ to be the Bayes posterior over T given (a_0, m_0), and we may assume that $\lim_n \beta_n(t|a_0, m_0)$ and $\lim_n \rho_n(z|a_0, m_0)$ exist for every t and z. Then we can define $\mu(\cdot) = \lim_n \beta_n(\cdot|a_0, m_0)$ and $r(\cdot) = \lim_n \rho_n(\cdot|a_0, m_0)$. Consequently, because $\rho_n(\cdot|a_0, m_0)$ is a best reply for R after (a_0, m_0) given beliefs $\beta_n(\cdot|a_0, m_0)$ taking limits implies that r and μ satisfy (a) in Lemma 10.2. It remains to show that r and μ satisfy (b) in Lemma 10.2.

Letting $\pi_S^n(t)$ be type t's equilibrium payoff in G_n implies that $\pi_S^n(t) \geq \sum_z \rho_n(z|a_0, m_0)u_S(t, a_0, z) + o(1)$. Since the equilibrium type-action distribution ν_n of G_n converges to ν, $\pi_S^n(t)$ converges to $\pi_S(t)$. Hence, taking limits yields the inequality part of (b). It remains to show that $\pi_S(t) = \sum_z r(z)u_S(t, a, z)$ if $\mu(t)/\mu(t') > q_0(t)/q_0(t')$ for some t, t'.

So consider any t and t' such that $\mu(t)/\mu(t') > q_0(t)/q_0(t')$. This strict inequality implies that $\beta_n(t|a_0, m_0)/\beta_n(t'|a_0, m_0) > q_0(t)/q_0(t')$ for all n greater than some $\bar{n}$. By the definitions of β_n and q_0 this says that,

$$\frac{(1-\delta_n)\sigma_n(a_0, m_0|t) + \delta_n\sigma_0(a_0, m_0|t)}{(1-\delta_n)\sigma_n(a_0, m_0|t') + \delta_n\sigma_0(a_0, m_0|t')} > \frac{\sigma_0(a_0, m_0|t)}{\sigma_0(a_0, m_0|t')} \ (\forall n > \bar{n}).$$

Consequently, $\sigma_n(a_0, m_0|t) > 0$ for all $n > \bar{n}$ which means that, for $n > \bar{n}$, (a_0, m_0) is a best reply for t against ρ_n in G_n, and so $\pi_S^n(t) = \sum_z \rho_n(z|a_0, m_0)u_S(t, a_0, z) + o(1)$. Taking limits establishes (b), as desired. To conclude that (a) and (b) hold also for any $q_0 \in Q$, define (r, μ) to be any limit point of (r_n, μ_n) satisfying (a) and (b) for each $n = 1, 2, ...$, where $q_1, q_2, ...$ is any sequence of strictly positive probability vectors converging to q_0. ■

Lemma 10.5 *If an NLE type-action distribution is a Nash equilibrium type-action distribution of* $\Gamma(M)$ *and* $|M| \leq |T|$, *then it is a coarse NLE type-action distribution.*

Proof. Suppose that $(\tilde{\sigma}, \tilde{\rho}, \tilde{\beta})$ is an NLE, that $(\hat{\sigma}, \hat{\rho})$ is Nash equilibrium of $\Gamma(M)$ where $|M| \leq |T|$, and that ν is their common type-action distribution. We will construct a coarse NLE $(\sigma^*, \rho^*, \beta^*)$ with type-action distribution ν.

Since $|M| \leq |T|$, we may choose, for each $m \in M$ a distinct $t_m \in T$. Define the strategy σ^* for S in $\Gamma(T)$ so that for each a, m, and t, $\sigma^*(a, t_m|t) := \hat{\sigma}(a, m|t)$. Define the strategy ρ^* and the beliefs β^* for R in $\Gamma(T)$ as follows. For every action-message (a, m) that is on-path for $\hat{\sigma}$, define $\rho^*(\cdot|a, t_m) := \hat{\rho}(\cdot|a, m)$ and define $\beta^*(\cdot|a, t_m)$ to be R's Bayes' posterior under $\hat{\sigma}$ conditional on (a, m). This defines $\rho^*(\cdot|a, t)$ and $\beta^*(\cdot|a, t)$ for every (a, t) that is on-path for σ^*. For every (a, t) that is off-path for σ^*, define $\rho^*(\cdot|a, t) := \tilde{\rho}(\cdot|a, \delta_t)$ and define $\beta^*(\cdot|a, t) := \tilde{\beta}(\cdot|a, \delta_t)$.

Since $(\hat{\sigma}, \hat{\rho})$ is a Nash equilibrium of $\Gamma(M)$ and $(\tilde{\sigma}, \tilde{\rho}, \tilde{\beta})$ is an NLE and both have type-action distribution ν, it is easily verified that $(\sigma^*, \rho^*, \beta^*)$ is a coarse NLE with type-action distribution ν. ■

10.1 Systems of Indifferences

Henceforth, $M_{|T\times Z|}$ is a set with $|T \times Z|$ messages, and we denote $(a, m) \in A \times M_{|T\times Z|}$ by am. For any $D \subseteq A \times M_{|T\times Z|}$, and for any $T_{am} \subseteq T$ and $Z_{am} \subseteq Z$ $(\forall am \in D)$ such that $\cup_{am\in D} T_{am} = T$, call $\mathcal{C} := (T_{am}, Z_{am})_{am\in D}$ a *configuration* (following Park (1997)) and consider the following system of linear equations.

$$S(u_R; \mathcal{C}) := \begin{cases} (S.1) \quad \sum_{t\in T_{am}} \sigma(am|t)p(t)u_R(t, a, z) - w(am) = 0, \ \forall z \in Z_{am}, \ \forall am \in D \\ \\ (S.2) \quad \sum_{am\in D: t\in T_{am}} \sigma(am|t) = 1, \ \forall t \in T, \end{cases}$$

with real variables $(\sigma(am|t)_{t\in T_{am}, am\in D})$ and $(w(am))_{am\in D}$, and

$$R(u_S; \mathcal{C}) := \begin{cases} (R.1) \quad \sum_{z\in Z_{am}} \rho(z|am)u_S(t, a, z) - \pi(t) = 0, \ \forall t \in T_{am}, \ \forall am \in D \\ \\ (R.2) \quad \sum_{z\in Z_{am}} \rho(z|am) = 1, \ \forall am \in D, \end{cases}$$

with real variables $(\rho(z|am))_{z\in Z_{am}, am\in D}$ and $(\pi(t))_{t\in T}$. Say that a solution $(\sigma, w), (\rho, \pi)$ to $(S.1), (S.2), (R.1), (R.2)$, which need not be non-negative, is *regular* iff

$\sum_{t \in T_{am}} \sigma(am|t) \neq 0$ holds for every $am \in D$, and, $\rho(\cdot|am) \neq \rho(\cdot|am')$ holds for every distinct $(a,m),(a,m') \in D$. Let $S(u_R;\mathcal{C})$ denote the subsystem $(S.1)$-$(S.2)$, and let $R(u_S;\mathcal{C})$ denote the subsystem $(R.1)$-$(R.2)$.

Because there are finitely many configurations, there is a generic subset $\mathcal{U}_0 \subseteq \mathbb{R}^{T\times A\times Z}$,[39] such that for every configuration $\mathcal{C}$, if the rows (or columns) of the coefficient matrix of $S(u;\mathcal{C})$ are linearly independent for any $u \in \mathcal{U}_0$, then they are linearly independent for every $u \in \mathcal{U}_0$, and similarly for $R(u;\mathcal{C})$.[40]

Given any strategy profile (σ_0,ρ_0) for $\Gamma(M_{|T\times Z|};u_S,u_R)$, define the *indifference system induced by* (σ_0,ρ_0) to be the system $S(u_R;\mathcal{C})$-$R(u_S;\mathcal{C})$ with $\mathcal{C} = (T_{am},Z_{am})_{am\in D}$, where D is the set of σ_0 on-path am, and, for each $am \in D$, $T_{am} := \{t : \sigma_0(am|t) > 0\}$, and $Z_{am} := \{z : \rho_0(z|am) > 0\}$. We then sometimes write $S(u_R;\sigma_0,\rho_0)$ and $R(u_S;\sigma_0,\rho_0)$ instead of $S(u_R;\mathcal{C})$ and $R(u_S;\mathcal{C})$. If (σ_0,ρ_0) is a Nash equilibrium, then the on-path values specified by $((\sigma_0,w_{\sigma_0}),(\rho_0,\pi_{\rho_0}))$ provide a solution to the indifference system induced by (σ_0,ρ_0), where $w_{\sigma_0}(am)$ is the receiver's equilibrium payoff after any on-path am weighted by the probability of am, and where $\pi_{\rho_0}(t)$ is the equilibrium payoff of sender-type t.

We can now state an important genericity result due to Park (1997).

Proposition 10.6 *(Park (1997)). There is a generic set $\mathcal{U}_1 \subseteq \mathbb{R}^{T\times A\times Z}$, such that, if $u_S,u_R \in \mathcal{U}_1$ then $\Gamma(M_{|T\times Z|};u_S,u_R)$ has finitely many Nash equilibrium type-action distributions. Moreover, if for any configuration $\mathcal{C}$, the system $S(u_R;\mathcal{C})$-$R(u_S;\mathcal{C})$ has a regular solution, then the set D in that system satisfies $|D| \leq |T|$, and the coefficient matrices of the subsystems $S(u_R;\mathcal{C})$ and $R(u_S;\mathcal{C})$ are each square and invertible.*[41]

While Park (1997) studies games in which the sender has cheap-talk messages but no payoff-relevant actions, a simple change of variables translates each of the

[39] A set in Euclidean space is *generic* if its complement is closed and has Lebesgue measure zero.

[40] This follows from the fundamental theorem of algebra, Fubini's theorem, and the fact that determinants are multinomial functions of their entries.

[41] Park (1997) considers generic subsets of the product of the two players' utility spaces, but his proofs establish that each player's utility vector can be chosen from the same generic set.

subsystems of indifferences $S(u_R;\mathcal{C})$ and $R(u_S;\mathcal{C})$ to equally-sized systems of indifferences for a surrogate game of the sort studied by Park. Moreover, the latter systems of indifferences have solutions precisely when the former do. For the requisite change of variables, for each $d \in D$, for each $t \in T$, and for each $x \in A \times Z$, define $\hat{\sigma}(d|t) := \sigma(am|t)$ if $d = am$ and $t \in T_{am}$ and define $\hat{\rho}(x|d) := \rho(z|am)$ if $x = (a, z)$, $d = am$, and $z \in Z_{am}$. Then, $(\hat{\sigma}, \hat{\rho})$ determines on-path strategies in the surrogate game in which the sender has only the set of cheap-talk messages $d \in D$, and in which the receiver's action set is $X := A \times Z$, and payoffs are $u_S(t, x)$ and $u_R(t, x)$ when S's type is t and R chooses $x \in X$. Notice that, if $((\sigma, w_\sigma), (\rho, \pi_\rho))$ solves $S(u_R;\mathcal{C})$-$R(u_S;\mathcal{C})$, then, in the surrogate game, each player i's strategy in $(\hat{\sigma}, \hat{\rho})$ makes the other player j indifferent among all pure strategies given positive probability by j's strategy. Since these are precisely the systems of indifferences studied by Park (1997), his results apply also to our more general class of signaling games.

Given $\mathcal{U}_0$ above and $\mathcal{U}_1$ from Proposition 10.6, we henceforth fix the generic set of utility functions $\mathcal{U}^* := \mathcal{U}_0 \cap \mathcal{U}_1$. We note the following useful consequences.

Remark 10.7 If $u_S, u_R \in \mathcal{U}^*$ and (σ, ρ) is any Nash equilibrium of $\Gamma(M_{|T\times Z|}; u_S, u_R)$ such that $\rho(\cdot|am) \neq \rho(\cdot|am')$ for all distinct on-path (a, m) and (a, m'), then any z that is unused by R after any on-path am is a strictly unprofitable deviation for R after am. To see why this is true, note that $((\sigma, w_\sigma), (\rho, \pi_\rho))$ provides a regular solution to $S(u_R;\sigma,\rho)$-$R(u_S;\sigma,\rho)$. By Proposition 10.6 $S(u_R;\sigma,\rho)$ and $R(u_S;\sigma,\rho)$ have square coefficient matrices. Suppose, contrary to our claim, that z is an unused best-reply for R after an on-path am (hence, $\rho(z|am) = 0$ and $z \notin Z_{am}$). If we add z to the set Z_{am}, giving a larger configuration $\mathcal{C}$ say, then, because z is a best-reply for R after am and $\rho(z|am) = 0$, $((\sigma, w_\sigma), (\rho, \pi_\rho))$ provides a regular solution of the resulting system $S(u_R;\mathcal{C})$-$R(u_S;\mathcal{C})$. But the subsystems $S(u_R;\mathcal{C})$ and $R(u_S;\mathcal{C})$ are not square, contradicting Proposition 10.6.

Remark 10.8 If $u_S, u_R \in \mathcal{U}^*$ and (σ, ρ) is a Nash equilibrium of $\Gamma(M_{|T\times Z|}; u_S, u_R)$, then the rows of the, not necessarily square, coefficient matrix of the subsystem

$S(u;\sigma,\rho)$ are linearly independent for all $u \in \mathcal{U}^*$. To see this, we first show that (ρ, π_ρ) uniquely solves $R(u_S;\sigma,\rho)$. If not, some (ρ', π') solves $R(u_S;\sigma,\rho)$ and $\rho'(z_0|a_0m_0) \neq \rho(z_0|a_0m_0) > 0$ for some σ-on-path a_0m_0 and some z_0. Construct $\bar{\sigma}$ from σ by merging ρ-equivalent action-messages (i.e., any σ on-path (a,m) and (a,m') such that $\rho(\cdot|am) = \rho(\cdot|am')$) so that, in particular, $\bar{\sigma}$ uses a_0m_0 whenever any ρ-equivalent a_0m would be chosen by σ. By Proposition 10.6, the (regular) solution provided by $((\bar{\sigma}, w_{\bar{\sigma}}), (\rho, \pi_\rho))$ to $S(u_R;\bar{\sigma},\rho)$-$R(u_S;\bar{\sigma},\rho)$ is unique and so (ρ, π_ρ) provides the unique solution to $R(u_S;\bar{\sigma},\rho)$. But because we can assume that ρ and ρ' have the same sets of equivalent action-messages,[42] (ρ', π') also provides a solution to $R(u_S;\bar{\sigma},\rho)$, contradicting uniqueness because $\rho'(z_0|a_0m_0) \neq \rho(z_0|a_0m_0) > 0$ and a_0m_0 is on-path for $\bar{\sigma}$. Hence, the coefficient matrix of $R(u_S;\sigma,\rho)$ has linearly independent columns, which, since $u_S \in \mathcal{U}^* \subseteq \mathcal{U}_0$, implies that the same holds for $R(u;\sigma,\rho)$ for every $u \in \mathcal{U}_0$ and so for every $u \in \mathcal{U}^*$. The desired result follows because the coefficient-matrix for $S(u;\sigma,\rho)$ is the transpose of that for $R(u;\sigma,\rho)$.

10.2 Sender-Stability

In this section we prove that for generic base-game utilities u_S and u_R, every NLE type-action distribution is a sender-stable Nash equilibrium type-action distribution in the game $\Gamma(M_{|T|};u_S,u_R)$, where $M_{|T|}$ is a set of $|T|$ cheap-talk messages. To do so, we must further refine the generic set $\mathcal{U}^*$.

If (σ,ρ) is a Nash equilibrium of $\Gamma(M_{|T|};u_S,u_R)$ with payoff vector $(\pi(t))_{t\in T}$ for the sender, then, for this fixed π and for any off-path a_0, there is a generic subset of sender utility-vectors $u_S(\cdot,a_0,\cdot) \in \mathbb{R}^{T\times Z}$ such that the following holds. For any non-empty subsets $T_0 \subseteq T$ and $Z_0 \subseteq Z$ such that $|T_0| = |Z_0|$, if

$$\sum_{z\in Z_0} \rho(z) u_S(t,a_0,z) = \pi(t), \forall t \in T_0, \tag{10.1}$$

[42] Otherwise, we could have replaced the two solutions with a pair of distinct solutions on the line joining them, virtually any two of which would have the stated property.

then $\sum_{z\in Z_0}\rho(z)\neq 1$.

Of course, π itself depends on the utility functions u_S and u_R. Nevertheless, (10.1) holds for generic choices of u_S and u_R. This is because, for utilities in $\mathcal{U}^*$ there are finitely Nash equilibrium payoffs π, and the solutions to the indifference systems induced by equilibria do not depend on utility values at off-path actions. So there is a generic subset $\mathcal{V}$ of $\mathcal{U}^*$ such that if $u_S, u_R \in \mathcal{V}$, then (10.1) holds for every Nash equilibrium (σ,ρ) of $\Gamma(M_{|T|};u_S,u_R)$, where a_0 is any off-path sender-action and π is the sender's equilibrium payoff vector. Define $\mathcal{U}^{**}$ to be the intersection of $\mathcal{V}$ with the generic set of utility functions u such that for every $a\in A$, every square submatrix of $[u(t,a,z))_{z\in Z,t\in T},-\mathbf{1}]$ is invertible, where $-\mathbf{1}\in\mathbb{R}^Z$ is a column of -1's.

Lemma 10.9 *If $u_S,u_R\in\mathcal{U}^{**}$, then every NLE type-action distribution is a sender-stable type-action distribution of $\Gamma(M_{|T|};u_S,u_R)$.*

Proof. Let ν be an NLE type-action distribution. By Lemma 10.1, ν is induced by a Nash equilibrium (σ,ρ) of $\Gamma(M;u_S,u_R)$ for some finite set of messages M. Consider the Nash equilibrium $(\bar\sigma,\rho)$ of $\Gamma(M;u_S,u_R)$ that results after merging all action-messages (a,m) and $(a,\tilde m)$ such that $\rho(\cdot|am)=\rho(\cdot|a\tilde m)$. The type-action distribution of $(\bar\sigma,\rho)$ remains equal to ν. Moreover, since $(\bar\sigma,\rho)$ provides a regular solution to its indifference system, Proposition 10.6 implies that $\bar\sigma$ uses at most $|T|$ action-messages and so we may assume that $M=M_{|T|}$. Let $\Gamma_{|T|}:=\Gamma(M_{|T|};u_S,u_R)$ and let A^+ be the set of $\bar\sigma$ on-path actions. Next, construct from $(\bar\sigma,\rho)$, a Nash equilibrium (σ^*,ρ^*) with the same type-action distribution, ν, such that the set of σ^*-on-path action-messages is $A^+\times M_{|T|}$.[43] Let $\pi(t)$ be the payoff of sender-type t in the equilibrium (σ^*,ρ^*) and let $D:=A^+\times M_{|T|}$ denote the set of σ^* on-path action-messages. We will show that ν is a sender-stable type-action distribution of $\Gamma_{|T|}$ by perturbing the equilibrium (σ^*,ρ^*).

Let σ_0 be any strictly mixed strategy for S, and, for any $\varepsilon\in(0,1)$, let G_ε be an

[43] Specifically, for each $a\in A^+$, choose m_a such that am_a is $\bar\sigma$ on-path. Define $(\sigma^*,\rho^*)=(\bar\sigma,\rho)$ except, for each $a\in A^+$ and for each of the n_a messages m such that am is $\bar\sigma$ off-path, define $\sigma^*(am|\cdot)=\sigma^*(am_a|\cdot)=\frac{1}{n_a+1}\bar\sigma(am_a|\cdot)$ and $\rho^*(\cdot|am):=\rho(\cdot|am_a)$.

ε-perturbation of $\Gamma_{|T|}$ toward σ_0. We will show that there is a sequence of sender-strategies σ_ε converging to σ^* as $\varepsilon \to 0$ and there is a receiver-strategy $\hat{\rho}$ such that $(\sigma_\varepsilon, \hat{\rho})$ is a Nash equilibrium of G_ε for all small enough ε (henceforth, for all $\varepsilon \approx 0$) and the type-action distribution of $(\sigma_\varepsilon, \hat{\rho})$ converges to ν as $\varepsilon \to 0$. Consider any $a_0 m_0 \notin D$, we begin by defining σ_ε and $\hat{\rho}$ at $a_0 m_0$.

Define $q_0 \in Q$ by,

$$q_0(t) := \frac{p(t)\sigma_0(a_0 m_0|t)}{\sum_{t'} p(t')\sigma_0(a_0 m_0|t')}, \text{ for every } t \in T. \tag{10.2}$$

Since a_0 is off-path for σ^*, $\sum_{t,z} \nu(t, a_0, z) = 0$. Hence, by Lemma 10.2 and replacing the ratio condition in (b) with its convex-combination equivalent (see footnote 11), there exist $r \in \Delta(Z)$ and $\mu \in \Delta(T)$ such that $\mu = \lambda q_0 + (1-\lambda)\gamma$ for some $\lambda \in [0,1]$ and some $\gamma \in \Delta(T)$, and,

(a) r is a best reply for R after a_0 when R's belief is μ, and

(b) for every t, $\pi(t) \geq \sum_z r(z) u_S(t, a, z)$ with equality if $\mu \neq q_0$ and $\gamma(t) > 0$.

Define $\hat{\rho}(\cdot|a_0 m_0) := r$. Then by (b),

$$\pi(t) \geq \sum_z \hat{\rho}(z|a_0 m_0) u_S(t, a, z), \text{ for all } t \in T.$$

It remains to define $\sigma_\varepsilon(a_0 m_0|t)$ for every $t \in T$.

If $\mu = q_0$ then define $\sigma_\varepsilon(a_0 m_0|t) := 0$ for every $t \in T$. With this definition, no sender-type gives voluntary weight to $a_0 m_0$ and so the only weight that $a_0 m_0$ receives is the involuntary weight from the perturbation σ_0. Thus, R's Bayes posterior after observing $a_0 m_0$ in the perturbed game G_ε will be q_0, and, by (a), $\hat{\rho}(\cdot|a_0 m_0) := r$ is a best reply.

Suppose next that $\mu \neq q_0$.

Claim: $\lambda > 0$.

Proof of Claim: Suppose, by way of contradiction, that $\lambda = 0$. Then $\mu = \gamma$. Let $T_0 := \{t : \gamma(t) > 0\}$ and let $Z_0 := \{z : r(z) > 0\}$. Then,

$$\sum_{z \in Z_0} r(z) u_S(t, a_0, z) = \pi(t), \ \forall t \in T_0 \tag{10.3}$$

$$\sum_{t \in T_0} \gamma(t) u_R(t, a_0, z) = \text{constant}, \ \forall z \in Z_0 \tag{10.4}$$

where (10.3) follows from (b) and (10.4) follows from (a) because given his beliefs $\mu = \gamma$ and the action a_0, R must be indifferent between all actions z assigned positive probability by r.

Suppose that $|Z_0| > |T_0|$. Then there is a subset Z_1 of Z_0 such that $|Z_1| = |T_0| + 1$. But then, letting c denote the constant on the right-hand side of (10.4), we would have $\sum_{t \in T_0} u_R(t, a_0, z)\gamma(t) - c = 0, \forall z \in Z_1$ which, because the square $|Z_1| \times (|T_0| + 1)$ matrix $[u_R(t, a_0, z))_{z \in Z_1, t \in T_0}, -\mathbf{1}]$ is invertible since $u_R \in \mathcal{U}^{**}$, implies that $((\gamma(t))_{t \in T_0}, c)$ is the zero vector contradicting the definition of T_0.

Hence, $|Z_0| \leq |T_0|$. Then we can choose a subset T_1 of T_0 with $|T_1| = |Z_0|$ and so $\sum_{z \in Z_0} r(z) u_S(t, a_0, z) = \pi(t), \forall t \in T_1$ follows from (10.3). Then, $\sum_{z \in Z_0} r(z) \neq 1$ since $u_S \in \mathcal{U}^{**}$, which is a contradiction and completes the proof of the claim.

Since $\lambda \in (0, 1)$, if we define $c_0 := \sum_{t'} p(t')\sigma_0(a_0 m_0 | t')$, then we may choose $\bar{\varepsilon} > 0$ such that $\bar{\varepsilon} < \lambda / (\lambda + (1 - \lambda)c_0)$. Then, for every $\varepsilon \in (0, \bar{\varepsilon})$, there is a unique $\eta_\varepsilon \in (0, 1)$, which depends on $a_0 m_0$, such that $\lambda = \varepsilon c_0 / ((1 - \varepsilon)\eta_\varepsilon + \varepsilon c_0)$. Note that $\eta_\varepsilon \to 0$ as $\varepsilon \to 0$.

So for $\mu \neq q_0$ and for any $\varepsilon \in (0, \bar{\varepsilon})$, we may define $\sigma_\varepsilon(a_0 m_0 | t) := \eta_\varepsilon \gamma(t) / p(t)$ for every $t \in T$. This definition implies that $\sum_{t'} p(t')\sigma_\varepsilon(a_0 m_0 | t') = \eta_\varepsilon$, and that for every $t, \sigma_\varepsilon(a_0 m_0 | t)$ converges to zero with ε.

With this definition of $\sigma_\varepsilon(a_0 m_0 | t)$ when $\mu \neq q_0$, each sender-type t with $\gamma(t) > 0$ places voluntary weight $\eta_\varepsilon \gamma(t) / p(t)$ on $a_0 m_0$. Since, in the perturbed game, the sender's action-messages are generated by the perturbed strategy $(1 - \varepsilon)\sigma_\varepsilon + \varepsilon \sigma_0$,

R's Bayes posterior after observing a_0m_0 is—by our careful definitions of η_ε and of $\sigma_\varepsilon(a_0m_0|\cdot)$—precisely $\lambda q_0 + (1-\lambda)\gamma = \mu$. By (a), r is a best reply for R after a_0 given the beliefs μ. Hence, once again, in the perturbed game R's mixture $\hat{\rho}(\cdot|a_0m_0) := r$ is a best reply after observing a_0m_0.

Hence, for the given $a_0m_0 \notin D$, we have defined $\hat{\rho}(\cdot|a_0m_0)$ and $\sigma_\varepsilon(a_0m_0|t)$ for every t so that,

$$\hat{\rho}(\cdot|a_0m_0) \text{ is a best reply for } R \text{ in } G_\varepsilon \text{ after observing } a_0m_0, \tag{10.5}$$

and

$$\pi(t) \geq \sum_z \hat{\rho}(\cdot|a_0m_0)u_S(t, a_0, z) \text{ for all } t, \text{ with equality if } \sigma_\varepsilon(a_0m_0|t) > 0, \tag{10.6}$$

where, by (b), equality holds in (10.6) when $\sigma_\varepsilon(a_0m_0|t) > 0$ because, in that case, $\mu \neq q_0$ and $\gamma(t) > 0$ (and because $\hat{\rho}(\cdot|a_0m_0) := r$).

Since $a_0m_0 \notin D$ was arbitrary, we may similarly define $\sigma_\varepsilon(am|t)$ and $\hat{\rho}(\cdot|am)$ for every $am \notin D$, and every $t \in T$. For each $t \in T$, let

$$\alpha_\varepsilon(t) := \sum_{a \in A \setminus A^+, m \in M_{|T|}} \sigma_\varepsilon(am|t). \tag{10.7}$$

Observe that $\lim_{\varepsilon \to 0} \alpha_\varepsilon(t) = 0$ because each term in the sum converges to zero. For the remaining action-messages $am \in D$, define $\hat{\rho}(\cdot|am) := \rho^*(\cdot|am)$. We next complete the definition of σ_ε.

Consider the indifference system $S(u_R; \sigma^*, \rho^*)$-$R(u_S; \sigma^*, \rho^*)$ induced by (σ^*, ρ^*), and consider the following perturbation of the subsystem $S(u_R; \sigma^*, \rho^*)$.

$(S.1.\varepsilon)$ $\sum_{t \in T_{am}} \sigma(am|t)p(t)u_R(t, a, z) - w(am) = v_\varepsilon(am, z),\ \forall z \in Z_{am},\ \forall am \in D$

$(S.2.\varepsilon)$ $\sum_{am \in D: t \in T_{am}} \sigma(am|t) = 1 - \alpha_\varepsilon(t),\ \forall t \in T,$

where $v_\varepsilon(am,z) := (-\varepsilon/(1-\varepsilon))\sum_{t\in T}\sigma_0(am|t)p(t)u_R(t,a,z)$, $Z_{am} := \{z : \rho^*(z|am) > 0\}$ and $T_{am} := \{t : \sigma^*(am|t) > 0\}$.

Since the coefficient matrix for $(S.1.\varepsilon)$-$(S.2.\varepsilon)$ is the the same as that for $S(u_R;\sigma^*,\rho^*)$, its rows are linearly independent by Remark 10.8. Hence, because (σ^*, w_{σ^*}) solves $S(u_R;\sigma^*,\rho^*)$, the perturbed system $(S.1.\varepsilon)$-$(S.2.\varepsilon)$ has a solution $(\sigma_\varepsilon(am|t), w_\varepsilon(am))_{t\in T_{am}, am\in D}$ such that $\sigma_\varepsilon(am|t) \to \sigma^*(am|t)$ as $\varepsilon \to 0$ for every $t \in T_{am}$ and for every $am \in D$.

For each $\varepsilon > 0$ and for each $t \in T$, extend $\sigma_\varepsilon(\cdot|t)$ from $\{am \in D : t \in T_{am}\}$ to all of D by defining $\sigma_\varepsilon(am|t) := 0$ if $am \in D$ and $t \notin T_{am}$. Since we have previously defined $\sigma_\varepsilon(am|t)$ for all $am \notin D$, we have now defined σ_ε. Moreover, (10.7) and $(S.2.\varepsilon)$ together imply that $\sum_{am\in A\times M_{|T|}}\sigma_\varepsilon(am|t) = 1$, $\forall t \in T$.

Since $\sigma_\varepsilon \to \sigma^*$, $\sigma_\varepsilon(am|t) > 0$ $(\forall am \in D)$ $(\forall t \in T_{am})$ holds for $\varepsilon \approx 0$. Since we have defined σ_ε to be nonnegative elsewhere, σ_ε is a valid sender-strategy for $\varepsilon \approx 0$. Moreover, since $\hat\rho(\cdot|am) := \rho^*(\cdot|am)$ for $am \in D$, the type-action distribution from $(\sigma_\varepsilon, \hat\rho)$ in G_ε converges to the type-action distribution for (σ^*,ρ^*) namely, ν. So it remains only to show that $(\sigma_\varepsilon,\hat\rho)$ is an equilibrium of G_ε for $\varepsilon \approx 0$.

But first, return to the game $\Gamma_{|T|}$. We claim that, in the equilibrium (σ^*,ρ^*), after R observes any $am \in D$, every $z \notin Z_{am}$ is strictly worse for R than every $z \in Z_{am}$. Indeed, if $\rho^*(z_0|am) = 0$, then by the construction in footnote 43, there is a $\bar\sigma$ on-path (a,m') such that $\rho(z_0|am') = 0$ and R's posterior after am' in the equilibrium $(\bar\sigma,\rho)$ is the same as after am in the equilibrium (σ^*,ρ^*). Therefore, since by Remark 10.7, z_0 is strictly suboptimal for R after am' in the equilibrium $(\bar\sigma,\rho)$, the same is true after am in the equilibrium (σ^*,ρ^*).

Suppose henceforth that $\varepsilon \approx 0$. To show that $(\sigma_\varepsilon,\hat\rho)$ is an equilibrium of G_ε, we first show that $\hat\rho$ is a best reply against σ_ε. For any $am \notin D$, (10.5) says that against σ_ε in G_ε, $\hat\rho(\cdot|am)$ is a best reply for R after observing am. For any $am \in D$, $(S.1.\varepsilon)$—whose right-hand side accounts for the ε-perturbation of S's strategy toward σ_0—says that against σ_ε in G_ε, R is indifferent among all $z \in Z_{am}$ after observing am. Moreover, after observing am, every $z \notin Z_{am}$ is strictly worse for R against

σ_ε than every $z \in Z_{am}$ (by the previous paragraph and because $\sigma_\varepsilon \to \sigma^*$). Hence, every $z \in Z_{am}$ is optimal for R against σ_ε after observing am. Consequently, since $\hat{\rho}(\cdot|am) := \rho^*(\cdot|am)$ puts positive weight only on elements of Z_{am}, we may conclude that $\hat{\rho}$ is a best reply for R against σ_ε.

Finally, we must show that, in G_ε, σ_ε is a best reply for S against $\hat{\rho}$, which, since R's strategy is not perturbed in G_ε, this is equivalent to showing that σ_ε a best reply against $\hat{\rho}$ in $\Gamma_{|T|}$. We first claim that, for every type t, every am given positive probability by $\sigma_\varepsilon(\cdot|t)$ yields t the same payoff (in $\Gamma_{|T|}$), namely $\pi(t)$. Indeed, suppose $\sigma_\varepsilon(am|t) > 0$. If $am \notin D$, then $\sum_z \hat{\rho}(z|am)u_S(t,a,z) = \pi(t)$ by (10.6). If $am \in D$, then $t \in T_{am}$ (otherwise $\sigma_\varepsilon(am|t) = 0$) and so $\sigma^*(am|t) > 0$. Hence, $\sum_z \hat{\rho}(z|am)u_S(t,a,z) := \sum_z \rho^*(z|am)u_S(t,a,z) = \pi(t)$ because $\pi(t)$ is the payoff of sender-type t in the equilibrium (σ^*, ρ^*), proving the claim. It suffices therefore to show that, for each sender-type t, no other action-message yields a payoff above $\pi(t)$.

So consider any a, m, t with $\sigma_\varepsilon(am|t) = 0$. If $am \notin D$, then $\pi(t) \geq \sum_z \hat{\rho}(z|am)u_S(t,a,z)$ by (10.6). If $am \in D$, then $\hat{\rho}(\cdot|am) := \rho^*(\cdot|am)$. Since type t can do no better than to obtain the payoff $\pi(t)$ against ρ^* (because (σ^*, ρ^*) is an equilibrium in which t's payoff is $\pi(t)$), we have $\pi(t) \geq \sum_z \hat{\rho}(z|am)u_S(t,a,z)$ and the proof is complete. ■

References

Banks, Jeffrey. S., and Joel Sobel (1987), "Equilibrium Selection in Signaling Games." *Econometrica*, 55, 647-661.

Bernheim, B. Douglas. (1984), "Rationalizable Strategic Behavior." *Econometrica*, 52, 1007-1028.

Blume, Andreas. (2026), "Meaning in Communication Games." *International Journal of Game Theory*, 55, 1-46. https://doi.org/10.1007/s00182-025-00973-z.

Blume, Andreas, and Joel Sobel: (1995), "Communication-Proof Equilibria in Cheap-Talk Games." *Journal of Economic Theory*, 65, 359-382.

Chen, Ying (2011), "Perturbed Communication Games with Honest Senders and Naive Receivers." *Journal of Economic Theory*, 146, 401-424.

Chen, Ying, Navin Kartik, and Joel Sobel (2008), "Selecting Cheap-Talk Equilibria." *Econometrica*, 76, 117–136.

Cho, In-Koo, David M. Kreps (1987), "Signaling Games and Stable Equilibria." *The Quarterly Journal of Economics*, 102, 179-221.

Clark, Daniel (2024), "Robust Neologism Proofness and the Flexible Speech Intuitive Criterion." working paper, Department of Economics, UCLA.

Clark, Daniel, and Drew Fudenberg (2021), "Justified Communication Equilibrium." *American Economic Review*, 111, 3004–3034.

Crawford, Vincent P., (2003), "Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions." *American Economic Review*, 93, 133-149.

Crawford, V. P., and J. Sobel (1982): "Strategic Information Transmission." *Econometrica*, 50, 1431-1451.

Demichelis, Stefano, and Jörgen. Weibull (2008), "Language, Meaning, and Games: A Model of Communication, Coordination, and Evolution." *American Economic Review*, 98, 1292–1311.

Farrell, Joseph (1985), "Credible Neologisms in Games of Communication." MIT wp 386.

Farrell, Joseph (1993), "Meaning and Credibility in Cheap-Talk Games." *Games and Economic Behavior*, 5, 514-531.

Gordon, Sidartha, Navin Kartik, Melody Pei-yu Lo, Wojciech Olszewski, and Joel Sobel (2023), "Effective Communication in Cheap-Talk Games." working paper, LEDA, Université Paris-Dauphine.

Green, Jerry, and Nancy Stokey (2007), "A Two-Person Game of Information Transmission." *Journal of Economic Theory*, 135, 90-104.

Grossman, Sanford J., and Motty Perry (1986), "Perfect Sequential Equilibrium." *Journal of Economic Theory*, 39, 97-119.

Hart, Sergiu (1985), "Two-Person Repeated Games with Incomplete Information." *Mathematics of Operations Research*, 10, 117-153.

Hart, Sergiu, Ilan Kremer, and Motty Perry (2017), "Evidence Games: Truth and Commitment." *American Economic Review*, 107, 690–713.

Heumann, Tibor (2020), "On the Cardinality of the Message Space in Sender–Receiver Games." *Journal of Mathematical Economics*, 90, 109-118.

Kartik, Navin (2009), "Strategic Communication with Lying Costs." *The Review of Economic Studies*, 76, 1359-1395.

Kartik, Navin, Marco Ottaviani, Francesco Squintani (2007), "Credulity, Lies, and Costly Talk." *Journal of Economic Theory*, 134, 93-116.

Koessler, Frédéric, Marie Laclau, and Tristan Tomala (2024), "A Belief-Based Approach to Signaling." HAL working paper halshs-04455227.

Kohlberg, Elon, and Jean-Francois Mertens (1986), "On the Strategic Stability of Equilibria." *Econometrica*, 54, 1003-1037.

Mailath, George J., Masahiro Okuno-Fujiwara, and Andrew Postlewaite (1993), "Belief-based refinements in Signalling Games." *Journal of Economic Theory*, 60, 241–276.

Matthews, Steven A., Masahiro Okuno-Fujiwara, and Andrew Postlewaite (1991), "Refining Cheap-Talk Equilibria." *Journal of Economic Theory*, 55, 241-273.

Matthews, Steven A., and Andrew Postlewaite (1994), "On Modelling Cheap Talk in Bayesian Games." in *The Economics of Informational Decentralization: Complexity, Efficiency, and Stability: Essays in Honor of Stanley Reiter*, Edited by John O. Ledyard, Kluwer Academic Publishers.

Myerson, Roger B. (1989), "Credible Negotiation Statements and Coherent Plans." *Journal of Economic Theory*, 48, 264-303.

Park, In-Uck (1997), "Generic Finiteness of Equilibrium Outcome Distributions for Sender Receiver Cheap-Talk Games." *Journal of Economic Theory*, 76, 431-448.

Pearce, David G. (1984), "Rationalizable Strategic Behavior and the Problem of Perfection." *Econometrica*, 52, 1029-1050.

Rabin, Matthew (1990), "Communication between Rational Agents." *Journal of Economic Theory*, 51, 144-170.

Sémirat, Stéphan, and Françoise Forges (2023), "Forward-neologism-proof equilibrium and better response dynamics." working paper, Université Paris-Dauphine.

Sobel, Joel (2020), "Lying and Deception in Games." *Journal of Political Economy*, 128, 907-947.

Von Neumann, John and Oscar Morgenstern (1944), *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ.

Zapater, Iñigo (1997), "Credible Proposals in Communication Games." *Journal of Economic Theory*, 72, 173-197.