JOURNAL OF ECONOMIC THEORY 51, 144–170 (1990)

Communication between Rational Agents*

MATTHEW RABIN

Economics Departments, University of California,
Berkeley, California 94720

Received August 30, 1988; revised August 23, 1989

Conventional game-theoretic solution concepts *never* guarantee meaningful communication in cheap-talk games. I define a solution concept which does guarantee communication in some games. I assume full rationality without imposing equilibrium conditions, but add a natural behavioral assumption about how agents use language: agents have a propensity to speak the truth and to believe others speak the truth, but use the game's strategic incentives to check whether such behavior and beliefs are rational. I also define and prove the existence of an equilibrium version of the concept, and present examples where its predictions seem more natural than Farrell's neologism-proof equilibrium. *Journal of Economic Literature* Classification Number: 026. © 1990 Academic Press, Inc.

I. INTRODUCTION

Economists have become used to the idea that agents can take costly actions to signal private information. For instance, Spence [10] shows that workers may undertake costly education to signal their productivity to potential employers, even if the education has no effect on workers' productivity. If the costs of education vary with a worker's type, his willingness to incur these costs may signal his type.

More recently, game theorists have begun to formalize the role of verbal and similar low-cost communication in strategic settings. In many contexts, the potential for low-cost communication is unimportant. For instance, in Spence's example, employers will not believe unverifiable claims by a worker that he is productive; if they did, every worker would claim to be very productive. This captures the intuition that talk is cheap: if making claims is effortless, an agent is likely to make a self-serving claim, not necessarily a truthful one.

* I thank Eddie Dekel, Joe Farrell, Bob Gibbons, Oliver Hart, Edmond Malinvaud, an anonymous referee, and especially Joel Sobel for useful comments. I am very grateful to Drew Fudenberg for extremely helpful comments and conversations on this and related research. Financial assistance from a National Science Foundation Graduate Fellowship and Sloan Foundation Dissertation Fellowship is gratefully acknowledged.

0022-0531/90 $3.00

In situations of pure coordination, on the other hand, such "cheap talk" can be very credible. If two agents agree about what action is optimal contingent on any information, each is likely to believe informational claims by the other. More generally, most strategic situations involve neither pure coordination nor pure conflict, so that the extent of meaningful communication is less obvious. In this paper, I formulate a theory that begins to explore the extent and nature of meaningful communication in strategic situations.

Using standard equilibrium analysis, Crawford and Sobel [3] first demonstrated formally that adding cheap talk to strategic situations can expand the set of possible outcomes. An informed agent might convincingly reveal some of his information so as to induce an uninformed agent to take particular actions. Most strikingly, even if two agents never fully agree about which strategic action would be best, cheap talk still can achieve some degree of coordination.

Unfortunately, standard game theory can *never* guarantee meaningful communication, even in games of pure coordination. When any conventional solution concept is used to analyze cheap talk, there always exist babbling equilibria, in which no communication occurs.[1] In such equilibria, no messages are interpreted by the uninformed agent as being meaningful, so that the informed agent might as well randomize over which messages he sends. If he does so, then the uninformed agent is justified in not believing statements, and he will consequently maintain his prior beliefs no matter what statement he hears. Most strikingly, even when agents fully agree on which actions are appropriate given the private information, there is no guarantee that such private information gets communicated.

Farrell [5] argues that the existence of a rich, common language among agents should be given a stronger role in game theory, so that meaningful communication can be predicted confidently in some situations. A common language consists of (1) a meaningful vocabulary, and (2) a common understanding among agents that it is appropriate to interpret statements according to their literal meaning. Of course, rational uninformed agents will not naively believe all things that an informed agent says if their interests diverge. But it is natural to assume that such agents have a propensity to believe statements if such belief is consistent with rationality. As Farrell

[1] By "conventional solution concepts," I mean any non-cooperative solution concept which attempts to restrict outcomes based on some form of internal consistency or on some notion of strategic stability. This includes Nash equilibrium and its many refinements from Kreps and Wilson's [6] sequential equilibrium (which Crawford and Sobel use) to Kohlberg and Mertens's [7] strategic stability, as well as the non-equilibrium concept of rationalizability (see Bernheim [1] and Pearce [9]). This excludes, for instance, Pareto dominance as a selection criterion among equilibria, and many cooperative solution concepts.

says, "Although honesty may not always be the best policy, it is a focal policy."

I develop below—for the same simple strategic situations examined by Crawford and Sobel and by Farrell—a solution concept, *Credible Message Rationalizability* (*CMR*), which combines Farrell's rich language assumption with the assumption that agents are rational. The solution concept formalizes the notion that honesty is a focal policy, and then provides a test of when such honesty is a reasonable policy for both of the agents. Formally, I show that the theory is consistent with rationality by proving that any permitted behavior by either player is optimal with respect to some beliefs over the permitted behavior of the other player.

Consider the cheap talk situation shown in Example 1. Agent S is informed, and can be of three types, where his type represents the realization of his private information. While Agent S knows his type, an uninformed agent, R, does not, and has beliefs assigning probability $\frac{1}{3}$ to each type. Agent R can take any of three actions, where these actions affect both players. The payoffs for each player are a function of types and actions taken. Before R takes his action, S can send a message in some shared language. What would he say, and what would R believe?

Suppose that S says, "I'm t_1." To choose an appropriate response, player R must figure out which types of player S would say that. If he believed the statement were true, he would play a_1. If he did so, the only type of player S that would ever want to make such a statement would be t_1, who would get his optimal payoff from a_1. The other types would each get their worst outcome possible by claiming to be t_1. So, *if* it is common knowledge that "I'm t_1" will be believed, then it will be true.

The fact that the truth is focal is crucial for inducing useful communication here. Suppose, alternatively, that R thinks that S is just babbling, so that there is no correlation between what S says and his type. Then if he hears "I'm t_1" (or any other statement), R will take action a_3, because this yields the highest expected payoff given his un-updated prior beliefs. This would be "fully rational." It is consistent, for instance, with both sequential

EXAMPLE 1

$p(t_1) = p(t_2) = p(t_3) = \frac{1}{3}$

	S			R		
	t_1	t_2	t_3	t_1	t_2	t_3
a_1	10	0	0	10	0	0
a_2	0	10	10	0	10	0
a_3	0	5	5	0	7	7

equilibrium and rationalizability. But communication does not occur only because neither player expects it to occur, which seems to be unrealistic in this case.

By contrast to the credibility of "I'm t_1," suppose it were common knowledge that R always believed the statement "I'm t_2," and therefore always reacted with his optimal move, a_2. Then both types t_2 *and* t_3 would wish to make the statement, since it yields them both their highest possible payoff. If both types make the statement, then the optimal reaction for R would be to play a_3, not a_2. Naive belief by R that S always tells the truth would lead him to react suboptimally to the statement "I'm t_2." Rationality would thus seem to dictate that R not necessarily believe this statement. CMR formally captures both the credibility of the statement "I'm t_1" and the incredibility of "I'm t_2."

My approach contrasts with most formal game theory in that I do not attempt predictions based solely on notions of internal consistency given the specifications of the game. Rather, I explicitly make additional assumptions about behavior and the beliefs of agents about the use of language. These assumptions are not ad hoc. If agents share a common language, they do not babble; they communicate. Game theory which ignores this tendency unnecessarily loses predictive power.

By incorporating it into a solution concept, I take the view that such a tendency for communication can be usefully integrated into *formal* game theory. Often analysts recognize that some outcomes are more plausible than others, and, game by game, informally select among the outcomes that formal analysis deems possible. Also, more general solution concepts like Pareto-dominant equilibria, as well as Nash equilibrium itself, are sometimes argued to be the natural outcomes in communicationally rich environments. By explicitly formulating a theory of communication, we can better understand whether such selection criteria are indeed natural.

The theory presented is meant to be a non-equilibrium notion, in the spirit of rationalizability as first formulated by Bernheim [1] and Pearce [9]. Equilibrium analysis assumes that, even when there are multiple equilibria, agents know the specific (but possibly mixed) strategy employed by another agent. Rationalizability assumes common knowledge of rationality, but allows that players are perhaps wrong in their conjecture about which reasonable strategy another player might choose. I assume common knowledge of both rationality and a general theory of communication, but likewise do not assume that a player necessarily conjectures correctly about which among many reasonable strategies the other player will employ.

This approach constrasts with that of Farrell [5]. Using the rich language assumption, he defines credible neologisms as deviations from a given sequential equilibrium. From this, he defines a solution concept, Neologism-Proof Equilibrium (NPE), by selecting out those sequential equilibria for which there are no credible neologisms. In addition to NPE's being an equilibrium concept, there are several other differences between CMR and NPE. In Section III, I present some examples comparing the two approaches.[2] To further compare them, I also define an equilibrium version of my theory: *Credible Message Equilibrium* (*CME*). It eliminates certain sequential equilibria by using rich language as an *ex ante* behavioral assumption, instead of as a means of judging the credibility of deviations from an equilibrium.

One goal of this paper is to use a formal model to investigate how communication aids coordination. Indeed, one of the frequent informal arguments for why an equilibrium is likely to obtain in strategic settings is based on cheap talk. It is argued that if agents can cheaply communicate and verbally agree to strategies, then only Nash equilibria will be self-enforcing agreements. As with Farrell [4], this paper can help us better understand the issue of how and whether communication helps equilibrate strategic behavior. As with Farrell, I find only limited support for the hypothesis that communication necessarily yields equilibrium behavior.

CMR puts forth a particular theory of language, but other theories consistent with rationality can be formulated. Indeed, CMR seems to be a weak theory: I feel that all the predicted communication is quite reasonable, but even further communication is likely in many settings. I briefly discuss the possibility of stronger theories of communication in the concluding section. I also briefly discuss the extension of CMR or related concepts to richer, more interesting strategic settings.

II. Credible Message Rationalizability

Consider a *Simple Communication Game*. The information of an informed agent, player S, can be characterized by a finite set of types, $T = \{t_1, t_2, ..., t_N\}$. Player S is assumed to know his type; an uninformed agent, player R, has prior beliefs represented by the probability distribution p over T. Player S sends a message from a finite set of messages, $M = \{m_1, m_2, ..., m_L\}$. The number of messages, L, is assumed to be large, greater than 2^{N+1}. After S sends a message, player R takes an action from a finite set of actions, A. Both players' utility functions, $U^S(a, t)$ and $U^R(a, t)$, depend only on the private information known by S and the action taken by R, and thus do not depend directly on the message sent.

[2] Myerson [8] also uses the rich language assumption. His approach is very similar to that of Farrell: he rules out equilibria based on the credibility of deviating neologisms. While his solution concept does not suffer from non-existence, Myerson can only guarantee existence by assuming the presence of a mediator.

Player S's strategy set, Σ_S, consists of all possible mappings from T to M; that is, he sends a message as a function of his private information. This function can be probabilistic, so that Σ_S includes mixed strategies. Player R's strategy set, Σ_R, consists of mappings from M to A; he takes an action as a function of the message he receives. Player R too can employ mixed strategies.

I will say that a strategy γ by R maps the message m into a set of actions $\tilde{A} \subseteq A$ if it assigns probability 1 that some action $a \in \tilde{A}$ will be employed when the message m is sent. I will define $\sigma(t)$ to be the probability distribution of messages sent by type t if S employs strategy σ. The function $\gamma(m)$ similarly defines the probability distribution over A in response to the message m if R uses strategy γ. Let $V^S(\sigma, \gamma)$, $V^R(\sigma, \gamma)$ denote the expected utilities of S and R respectively when S uses (possibly mixed) strategy σ and R employs (possible mixed) strategy γ. These utilities for both types are taken in terms of expected value over the different types of player S. A strategy σ *strongly dominates* σ' *with respect to a set* Σ_R^* of strategies for player R if for all probability distributions π defined over Σ_R^*, $\sum_\gamma \pi(\gamma)\, V^S(\sigma, \gamma) > \sum_\gamma \pi(\gamma)\, V^S(\sigma', \gamma)$. Strong dominance is defined likewise for player R.

With each subset of types $X \subseteq T$, associate an exclusive set of messages $M(X)$ from M: For all $X_i \neq X_j$, $M(X_i) \cap M(X_j) = \varnothing$. The existence of such messages for all sets of types captures the idea that a meaningful vocabulary exists. Such messages are meant to come from a common language pre-dating the specific strategic situation, but with which the agents can richly describe all relevant strategic issues.

A message $m \in M(X)$ is meant to convey the information "I'm some type in set X."[3] Most likely, the agents will use more natural language, such as "I like fish" or "Money is not important to me." In the setting examined, player S is trying to induce an action by R. Thus, statements suggesting actions rather than direct assertions about the private information, such as "You should invest in my company" or "It will pay you to hire me" are also natural. Given the setting and the incentives, any claim about information implicitly proposes one or more courses of action, and implicit in any proposal for action is an understanding of which types would prefer that action.

What matters is that the agents' understanding of the vocabulary and the strategic situation is clear enough that any preference over actions can be unambiguously conveyed. (Of course, I allow that the informed agent may

[3] If $M(X)$ were a singleton for each X, then there would be only one way to convey the idea, "I'm some type in set X." The main intuition for all of the results would be captured if each $M(X)$ were a singleton. The fact that I assume that there may be many ways to express this idea seems, however, more realistic.

choose not to communicate: he can speak gibberish or can remain silent.) It does not matter for our purposes exactly what language, speaking style, or low-cost mode of communication (e.g., verbal or written) is being employed. In fact, because most of the analysis investigates the incentives for truth-telling, the actual message space is largely suppressed. I focus primarily on which types of S will induce which physical actions, not the actual words used in achieving this.

A *type profile* is a list of exclusive, not necessarily exhaustive, subsets of types of agent S, $\mathscr{X} = \{X_1, X_2, ..., X_D\}$. Let $T_{\mathscr{X}} = \{t \mid \exists X_i \in \mathscr{X} : t \in X_i\}$. $T_{\mathscr{X}}$ is the set of types that are in some subset that is an element of the type profile. Let $M(\mathscr{X}) = \bigcup_{X_j \in \mathscr{X}} M(X_j)$. $M(\mathscr{X})$ is simply the set of self-signaling claims by the subsets of $\mathscr{X}$.

Definitions 1 through 6 construct a type profile that my theory predicts can send a credible message profile. I will motivate the definitions intuitively as developing a set of messages that should always be believed by R (and a set of types that should always send those messages). Yet, to be a theory of communication, I must fully specify what messages all types of S will send, and how R interprets every message in M. Definition 7 explicitly formulates a theory of permissible strategies for each player into which this notion of credible messages can be embedded. Proposition 1 shows that this theory is consistent with rationality, formally capturing the intuition of the earlier definitions.

DEFINITION 1. Let $A^* = \{a^* \in A \mid \exists \pi(T),\ 0 \leqslant \pi(t)\ \forall t$, and $\sum_{t \in T} \pi(t) > 0$, such that $a^* \in \{\operatorname{argmax}_{a \in A} \sum_{t \in T} \pi(t)\, U^R(a, t)\}\}$.

The function $\pi(t)$ represents the beliefs that R can hold about which types he is facing (it is not normalized to add up to 1 over the entire set of types).

A^* is the set of actions for which there are some beliefs by R about who sent a message for which the action is an optimal response. Player S could never hope to induce an action outside this get, so that any type of S will always be fully satisfied if he can induce his best action within this set.

DEFINITION 2. Let $A^*(X_j) = \{a^* \mid a^* \in \operatorname{argmax}_{a \in A} \sum_{t \in X_j} p(t)\, U^R(a, t)\}$.

$A^*(X_j)$ is the set of optimal actions by R if he thought he was facing exactly the types in X_j. (Recall that $p(t)$ are the prior beliefs that R ascribes to type t.) This has clear applications for a theory of credible messages: if R believed a message by S claiming to be any type in X_j were *always* made by types in X_j and never made by other types, then any action in $A^*(X_j)$ would be his rational response.

DEFINITION 3. Let $Y(\mathscr{X}) = T \backslash T_{\mathscr{X}}$.

$Y(\mathscr{X})$ is the set of types that are not in the type profile $\mathscr{X}$.

DEFINITION 4. Let $Y^*(X_j, \mathscr{X})$ be the set of types in $Y(\mathscr{X})$, excluding any type t with the property that either:

(1) $A^*(X_j) = \{a^* \mid a^* \in \text{argmin}_{a \in A^*} U^S(a, t)\}$, or

(2) $\exists X_k \in \mathscr{X} : U^S(a^*, t) < U^S(a, t) \, \forall a^* \in A^*(X_j), \forall a \in A^*(X_k)$.

$Y^*(X_j, \mathscr{X})$ represents all the types of player S that would conceivably want to imitate the message sent by the types X_j, if player R were to believe a self-signaling message by X_j. Types excluded from this set either (1) would always do their worst possible by imitating the set X_j, or (2) could do better by imitating some other set in $\mathscr{X}$.

DEFINITION 5. Let $A^{**}(X_j, \mathscr{X})$ be the set of actions a^* such that $\exists \pi(t)$: $T \Rightarrow [0, 1]$ satisfying

(1) $\pi(t) = p(t) \, \forall t \in X_j$,

(2) $\pi(t) \in [0, p(t)] \, \forall t \in Y^*(X_j, X)$,

(3) $\pi(t) = 0$ else,

such that $a^* \in \text{argmax}_{a \in A^*} \sum_{t \in T} \pi(t) \, U^R(a, t)$.

The set $A^{**}(X_j, \mathscr{X})$ is the set of optimal actions that R could take, if he were sure he was facing all types in X_j, and were not sure which types he was facing in $Y^*(X_j, \mathscr{X})$, and were sure he was not facing any of the other types. (As earlier, $\pi(t)$ represents probabilistic beliefs by R, not normalized to total to 1.)

Using the above definitions, I can propose a set of messages that should be considered credible.

DEFINITION 6. $\mathscr{X}$ is a *Credible Message Profile* if, $\forall X_j \in \mathscr{X}$,

(1) $\forall t \in X_j$, $A^*(X_j) = \{\text{argmax}_{a \in A^*} U^S(a, t)\}$, and

(2) $A^*(X_j) = A^{**}(X_j, \mathscr{X})$.

Intuitively, a message profile is considered credible if (1) when R believes the literal meanings of statements, the types sending the messages obtain their best possible outcomes, so those types will send their messages, and (2) the statements end up being true "enough." The "enough" comes from the fact that some types might also lie to R by pooling with others' credible messages, and R knows this, but the probability of this is small enough that it does not affect R's optimal response.

In what sense, formally, would such a message profile be credible? If the message profile were plugged into a Message Profile Theory (MPT) as defined below, then the resulting theory would be consistent with the rationality of players. An MPT describes one possible form of how the agents might use their common language.

DEFINITION 7. Fix a type profile $\mathscr{X}$. The *Message Profile Theory* (*MPT*) with respect to $\mathscr{X}$, denoted MPT($\mathscr{X}$), is the set of strategy pairs $\{(\sigma, \gamma):(\sigma, \gamma) \in (\Sigma_S^{\mathscr{X}}, \Sigma_R^{\mathscr{X}})\}$, where $(\Sigma_S^{\mathscr{X}}, \Sigma_R^{\mathscr{X}})$ is constructed as follows.

Let $\Sigma_R(0) = \{\gamma \in \Sigma_R \mid \forall X_i \in \mathscr{X}, \forall m \in M(X_i), \gamma$ maps m into $A^*(X_i)$, and $\forall m \in M, \gamma$ maps m into $A^*\}$. Let $\Sigma_S(0) = \{\sigma \in \Sigma_S \mid \forall X_i \in \mathscr{X}, \forall t \in X_i, \sigma$ maps t into $M(X_i)$, and $\forall t \notin X_i \cup Y^*(X_i, \mathscr{X}), \sigma$ maps t into $M \backslash M(X_i)\}$.

Then for all n, let $\Sigma_S(n+1) = \{\sigma \in \Sigma_S(n) \mid \sigma$ is not strongly dominated with respect to $\Sigma_R(n)$ by any $\sigma' \in \Sigma_S(n)\}$, and $\Sigma_R(n+1) = \{\gamma \in \Sigma_R(n) \mid \gamma$ is not strongly dominated with respect to $\Sigma_S(n)$ by any $\gamma' \in \Sigma_R(n)\}$. Choose N so that $\Sigma_S(N+1) = \Sigma_S(N)$ and $\Sigma_R(N+1) = \Sigma_R(N)$. Then $(\Sigma_S^{\mathscr{X}}, \Sigma_R^{\mathscr{X}}) = (\Sigma_S(N), \Sigma_R(N))$.[4]

Intuitively, an MPT is a theory permitting any strategies consistent with there being common knowledge that R will believe the literal meaning of some set of statements and that certain types of S will always make those statements. Note that very few restrictions are placed on Player R's beliefs for messages not in any $M(X_i)$. The iterated strong dominance does, however, guarantee that each player employs minimally rational strategies when sending or receiving messages outside the specified message profile.

Thus, iterated strong dominance guarantees that the players behave rationally *given* that these statements will be believed. Yet, for arbitrary type profiles, it is by no means certain that it is rational for Player R to believe these statements in the first place. In Example 1, for instance, R's believing everything that S says would not be a reasonable theory.

Proposition 1 establishes that, when the type profile used in the MPT can send a credible message profile according to Definition 6, then the resulting theory is consistent with rationality.

PROPOSITION 1. *If* $\mathscr{X}$ *is a credible message profile, then* MPT($\mathscr{X}$) = $(\Sigma_S^{\mathscr{X}}, \Sigma_R^{\mathscr{X}})$ *satisfies*: $\forall \sigma \in \Sigma_S^{\mathscr{X}}$, *$\sigma$ is not strongly dominated with respect to* $\Sigma_R^{\mathscr{X}}$ *by any* $\sigma' \in \Sigma_S$. *Likewise*, $\forall \gamma \in \Sigma_R^{\mathscr{X}}$, *$\gamma$ is not strongly dominated with respect to* $\Sigma_S^{\mathscr{X}}$ *by any* $\gamma' \in \Sigma_R$.

Proof. In Appendix.

The proof of this proposition follows our intuition for what a reasonable theory of credible messages should be. First, as argued above, the iterated

[4] Observe that since the sets of pure strategies are finite, MPT($\mathscr{X}$) is well-defined for all $\mathscr{X}$, because the iteration process of the definition ends in a finite number of steps.

strong dominance in the definition of an MPT guarantees that *given* that it is common knowledge that a certain list of messages would be believed, both players behave reasonably.

It is essential, however, that sending and believing the credible messages be consistent with rationality. This means first that if R will believe the credible messages, then it is optimal for the designated types of S to send them. This is easy to show, as such types get their maximum possible payoff from sending the messages.

More delicate is whether it is optimal for R to believe all credible messages. If there were just one way of conveying each literal meaning, then R would do best by believing the statements. However, since there are multiple ways of saying the same thing, R might think he knows which type of S says a statement which way. If he can differentiate the types fully, then he might be able to take type-specific actions not available to him when some types are indistinguishable from each other.

While this may not violate rationality per se, R is likely to doubt he knows which type says a statement which way. Since all types of S sending a credible message together at least weakly (and sometimes strictly) prefer to pool together, they never have an incentive to differentiate themselves.[5] The proof shows that this intuition holds. The theory does not dictate that all $t \in X_j$ really use identical wording, but only that R does not know their exact behavior.

Thus far, I have shown only that if it were common knowledge that a particular credible message profile were to be believed, it would be sensible both for player S to send these messages and for player R to believe these messages. As a "non-equilibrium" concept, however, it is important that the theory's predictions are natural when the agents enter the situation knowing just the game and the general theory of communication, without assuming that they coordinate on which among many reasonable strategies they each play.

It is desirable, then, that exactly *which* profile of types $\mathscr{X}$ send a credible message profile be common knowledge from the theory itself. This requires a non-arbitrary means of choosing this set, as well as one that gets as much of the natural communication as possible. In particular, the most attractive theory would be that both players believe that all types who have credible messages will send them. To guarantee that this is a coherent theory, it is essential that if profiles of types separately can send credible message profiles, then they can send a joint one. The following lemma establishes that this is the case.

LEMMA 1. *Suppose the type profiles* $\mathscr{X}(1)$ *and* $\mathscr{X}(2)$ *each can send a*

[5] This contrasts with the attempt to rule out babbling equilibria. There, agents often want to be differentiated.

credible message profile. Then $\exists\mathscr{X}$ which can send a credible message profile such that $T_{\mathscr{X}} = T_{\mathscr{X}(1)} \cup T_{\mathscr{X}(2)}$.

Proof. In Appendix.

The basic point of the proof is relatively simple. Credible message profiles are constructed so that no matter which types outside $T_{\mathscr{X}}$ profile R reasonably thinks might send a credible message, his optimal reaction is unchanged. If two credible message profiles are joined into a larger one, then the set of types not included in the profile is smaller than for either of the original profiles, so that R has to worry even less about such types imitating types in the profile. If the possibility of being lied to did not destroy either of the original credible message profiles, then *a fortiori* it will not destroy the joint one.

From Lemma 1, it is immediate that there exists some largest set of types that, appropriately partitioned, can send a credible message profile. It is possible, however, that there may be two different maximal type profiles $\mathscr{X}(1)$ and $\mathscr{X}(2)$ containing the same types, but partitioned differently. Again, if the theory does not assume any unmodelled coordination, it must designate which such partition will send a credible message profile. I now develop such a partition of the maximal set of types upon which the agents can focus.

DEFINITION 8. For any $\mathscr{X}$ with a credible message profile, let $\mathscr{X}(*)$ be the profile of types such that:

(1) $T_{\mathscr{X}} = T_{\mathscr{X}(*)}$, and

(2) $\forall X_i, X_j \in \mathscr{X} : A^*(X_i) = A^*(X_j)$, $\exists X(*) \in \mathscr{X}(*) : X_i \cup X_j \subseteq X(*)$.

$\mathscr{X}(*)$ contains the same set of types as does $\mathscr{X}$, but is partitioned less finely, in that any subsets which induce the same actions are concatenated. Lemma 2 establishes that this repartitioned set of subsets can also send a credible message.

LEMMA 2. *For any $\mathscr{X}$ that can send a credible message profile, $\mathscr{X}(*)$ can send a credible message profile.*

Proof. By definition, $Y(\mathscr{X}(*)) = Y(\mathscr{X})$. Choose $X(*) \in \mathscr{X}(*)$, where $X(*) = \bigcup_{X_i \in K} X_i$, for some set of elements $K \subseteq \mathscr{X}$. By construction of $\mathscr{X}(*)$, $A^*(X_i) = A^*(X_j)$ $\forall X_i, X_j \in K$. Thus

$$\begin{aligned} A^*(X(*)) &= \left\{ a^* \mid a^* \in \operatorname{argmax}_{a \in A^*} \sum_{t \in X(*)} p(t)\, U^R(a, t) \right\} \\ &= \left\{ a^* \mid a^* \in \operatorname{argmax}_{a \in A^*} \sum_{X_j \in K} \sum_{t \in X_j} p(t)\, U^R(a, t) \right\} \\ &= A^*(X_j) \;\; \forall X_j \in K. \end{aligned}$$

This equality holds $\forall X(*) \in \mathscr{X}(*)$, so that $Y^*(X(*), \mathscr{X}(*)) = Y^*(X_j, \mathscr{X})$. This in turn implies $A^{**}(X(*), \mathscr{X}(*)) = A^{**}(X_j, \mathscr{X})$. Thus, since the conditions of Definition 6 hold $\forall X_j \in \mathscr{X}$, they hold $\forall X(*) \in \mathscr{X}(*)$. This means that $\mathscr{X}(*)$ can send a credible message profile. Q.E.D.

The proof is similar in spirit to that of Lemma 1: the additional pooling of types in the new partition can only decrease worries that credible messages will be ruined by some types lying, and imitating types in the profile. The next result establishes that for any set of types that can send a credible message profile, there exists a unique, coarsest partition.

LEMMA 3. *Let $\mathscr{X}_A$ and $\mathscr{X}_B$ be profiles of types that can send credible message profiles such that $T_{\mathscr{X}_A} = T_{\mathscr{X}_B}$.*
Then $\mathscr{X}_A(*) = \mathscr{X}_B(*)$.

Proof. Suppose $\mathscr{X}_A(*) \neq \mathscr{X}_B(*)$. Then, without loss of generality, $\exists t_1, t_2$ such that $\exists X_i \in \mathscr{X}_A(*)$: $t_1, t_2 \in X_i$, but $\exists X_j \neq X_k \in \mathscr{X}_B(*)$: $t_1 \in X_j$, $t_2 \in X_k$. But $t_1, t_2 \in X_i \Rightarrow \{\text{argmax } U^S(a, t_1)\} = \{\text{argmax } U^S(a, t_2)\} \Rightarrow A^*(X_j) = A^*(X_k)$, which contradicts the construction of $\mathscr{X}_B(*)$. Q.E.D.

Lemmas 1, 2, and 3 combine to show that there exists a unique maximal set of types—and a unique coarsest partition of these types—that can send a credible message profile.

DEFINITION 9. Let $\mathscr{X}^{**}$ be the partition of types that can send a credible message profile such that:

(1) $\forall \mathscr{X}: \mathscr{X}$ can send a credible message message profile, $T_{\mathscr{X}} \subseteq T_{\mathscr{X}^{**}}$.

(2) $\forall X_i, X_j \in \mathscr{X}^{**}, A^*(X_i) \neq A^*(X_j)$.

The partition of types $\mathscr{X}^{**}$ is used in the primary definition:

DEFINITION 10. The *Credible Message Rationalizable Strategies* are $(\Sigma_S^{\mathscr{X}}, \Sigma_R^{\mathscr{X}}) = \text{MPT}(\mathscr{X}^{**})$.

The uniqueness of $\mathscr{X}^{**}$ means that CMR is always well-defined, so that there always exists at least one pair of strategies consistent with CMR.

Let us apply this theory of credible communication first to Example 1. In the introduction, I argued that the statement "I'm t_1" should be considered credible. We can check whether it is, by itself, a credible message profile. $A^*(t_1) = a_1$, clearly. That is, R's optimal response to beliefs that S is t_1 is to take action a_1. Type t_1 likes this response: $a_1 = \text{argmax } U^S(a, t_1)$. It only remains to show that no other types would want to make the claim "I'm t_1." But since a_1 is the worst action from the point of view of either of the other types, $Y^*(t_1, t_1) = \varnothing$. Thus $A^{**}(t_1, t_1) = A^*(t_1) = a_1$. Thus, t_1

can send a credible message profile. This means that CMR predicts that t_1 will always induce action a_1.

We can also check whether t_2 can send a credible message. It passes the first rounds of the definition: $A^*(t_2) = a_2 = \text{argmax } U^S(a, t_2)$. But then, whether or not t_1 sends a credible message, $Y^*(t_2, \mathscr{X}) = \{t_3\}$. That is, type t_3 would always want to pool in with t_2. (Because $A^*(t_3) \neq \text{argmax } U^S(a, t_3)$, t_3 will never send a credible message himself, so $t_3 \in Y(\mathscr{X})$ for any $\mathscr{X}$ that can send a credible message profile.) This in turn implies that $A^{**}(t_2, \mathscr{X}) = a_3 \neq A^*(t_2)$. According to our definition, t_2 indeed can not send a credible message.

In Example 1, type t_2 always wishes to distinguish himself from type t_3, but type t_3 always wishes to imitate t_2. Note that t_3 does not want R to believe he can be of either type; he specifically wants to appear as t_2. Thus, t_2 can never credibly separate himself. This is like the Spence example, where t_2 is a productive worker and t_3 is an unproductive worker. The bad worker would always do his utmost to imitate the productive worker, so that there is no scope for credible communication.

The situation is somewhat different in Example 2. There, type t_1 always wishes to distinguish himself from t_2, whereas t_2 prefers that they send the same message. But *if* R believes he knows which type sends which message, type t_2 would rather appear as himself than as t_1. That is, he wants to pool with t_1 if R believes that he is facing both types, so that R will take action a_3. He very much does *not* want to imitate type t_1. This allows t_1 the ability to send a credible message as defined by CMR. Thus, CMR dictates that types will always reveal themselves in certain games, even when some of these types would much prefer that no such revelation takes place.

Consider Example 3. In this case, both types of S will do best by sending the same message, if R reacts optimally given he believes he is facing both types. The statements "I am either type" or "I am not going to tell you what type I am" will be credible as defined by CMR.

EXAMPLE 2

$p(t_1) = p(t_2) = \frac{1}{2}$

	S		R	
	t_1	t_2	t_1	t_2
a_1	1	−2	3	0
a_2	−2	−1	0	3
a_3	0	0	2	2

Example 3

	$p(t_1)=p(t_2)=\frac{1}{2}$			
	S		R	
	t_1	t_2	t_1	t_2
a_1	-1	-2	3	0
a_2	-2	-1	0	3
a_3	0	0	2	2

This example illustrates that CMR can be used to guarantee that communication does *not* matter in situations where standard game theory allows that it might. If there were no opportunity for S to speak, then the unique equilibrium would obviously be the "pooling" equilibrium in which R takes action a_3. When S can talk, however, there is a separating sequential equilibrium in which R, for *all* statements, has very strong beliefs that S is of one type or the other. Given that R will make a strong inference from any statement, S will cooperate by signaling his true type, because he does not want to appear as the other type. This separating equilibrium seems unrealistic, however. It would mean, say, that R infers that the statement "I refuse to tell you what type I am" is a sure signal that S is of type t_1.

The first three examples involved only one credible message, whereas the theory permits non-singleton profiles of credible messages. Indeed, Example 4 illustrates that often the credibility of one message depends crucially on the existence of other credible messages. This follows from the definition of $Y^*(X_j, \mathscr{X})$, the set of types who might want to imitate X_j. This set can be made smaller by the existence of other credible messages which either the types are compelled to send, or with which they would prefer to pool. The profile of messages {"I'm t_1," "I'm t_2"} is credible according to our definition. Both statements will turn out to be true if it is common

Example 4

	$p(t_1)=p(t_2)=\frac{1}{2}$			
	S		R	
	t_1	t_2	t_1	t_2
a_1	0	0	9	9
a_2	10	9	10	0
a_3	9	10	0	10

EXAMPLE 5

$p(t_1) = p(t_2) = p(t_3) = \frac{1}{3}$

	S			R		
	t_1	t_2	t_3	t_1	t_2	t_3
a_1	7	6	0	6	7	0
a_2	6	7	0	7	6	0
a_3	0	0	6	0	0	6
a_4	−1	−1	−1	5	5	5

knowledge that R will believe both of them. Either message alone would not be credible, because the other type might also make the statement, if he were worried that R would react with a_1 to some other message he would send.

Example 5 illustrates that CMR is disturbingly weak in some contexts where communication seems natural.[6] Here, CMR is no more restrictive than rationalizability. It would seem reasonable, however, that types t_1 and t_2 would separate themselves from type t_3. While Player R and all types of Player S strongly prefer that Player S reveal whether or not he is type t_3, Player S would also like to fool R as to which of t_1 or t_2 he is. There is, therefore, no credible message for either t_1 or t_2, because neither can get his best possible outcome. This means, in turn, that type t_3 cannot separate himself, because types t_1 and t_2 might want to pool in with the message "I'm t_3" to avoid their worst possible outcome, a_4. In various ways, the harsh standards of credibility imposed by the definitions above mean that CMR cannot guarantee a natural amount of communication in this example. Refinements of CMR could, presumably, guarantee at least some communication here.

Another sense in which CMR applies high standards to credible messages is that it demands that, for all $t \in X_j$, $A^{**}(X_j, \mathscr{X}) = \{\text{argmax } U^S(a, t)\}$ in order for X_j to send a credible message, rather than just $A^{**}(X_j, \mathscr{X}) \subseteq \{\text{argmax } U^S(a, t)\}$ or $\{\text{argmax } U^S(a, t)\} \subseteq A^{**}(X_j, \mathscr{X})$. That is, a credible message must have the potential to induce *all* payoff-maximizing responses and *only* payoff-maximizing responses for all types sending the message.

In this case, however, the high standards are, I believe, warranted. Consider the credibility of the claim "I'm either t_1 or t_2" in Example 6. If this statement were believed, it would induce the action a_1 by R. This would yield both t_1 and t_2 their maximal utility, and t_3 his worst outcome,

[6] This is very similar to suggested examples by both Joel Sobel and an anonymous referee.

Example 6

$p(t_1) = p(t_2) = p(t_3) = \frac{1}{3}$

	S			R		
	t_1	t_2	t_3	t_1	t_2	t_3
a_0	2	2	2	2	2	2
a_1	10	10	−1	10	10	0
a_2	−1	10	10	0	10	10
a_3	0	0	0	17	0	0
a_4	0	0	0	0	0	17

so that he would never want to make the same statement. Yet CMR does not deem this a credible statement, because t_2 could do as well if he sent a message jointly with t_3.

Suppose CMR did not insist that credible messages induce *all*—but rather just some–of the responses that maximize the utility for the types sending the message. Then "I'm either t_1 or t_2" could be considered a credible statement. By the same logic, the statement "I'm either t_2 or t_3" also would be considered a credible statement. But then R would be believing that t_2 is always sending each of two different messages! If, rather, he tried to conjecture how often t_2 splits between the two messages, he would, for at least one of the messages, respond by playing either a_3 or a_4. This reflects his calculation that either most of the time he hears "I'm either t_1 or t_2," t_1 is speaking, or most of the time he hears "I'm either t_2 or t_3," t_3 is speaking.

Example 7 illustrates the more straightforward effect of the condition that messages are credible only if *any* rational response by R induces a good outçome for S. If both types send the message "I am either type" and R replies rationally and charitably with a_3, then both types do the best

Example 7

$p(t_1) = p(t_2) = \frac{1}{2}$

	S		R	
	t_1	t_2	t_1	t_2
a_1	−1	−2	3	0
a_2	−2	−1	0	3
a_3	0	0	2	2
a_4	−3	−3	2	2

Example 8

	$p(t_1)=p(t_2)=\frac{1}{2}$			
	S		R	
	t_1	t_2	t_1	t_2
a_1	10	6	10	0
a_2	0	8	4	0
a_3	8	0	0	4

they can. If, however, R responds rationally and uncharitably, then both types do the worst they can. CMR does not assume the charitable response. Indeed, in this case, the uncharitable response might represent a natural, spiteful response by R, since the decision by S not to reveal any information has cost R some utility.

All of the examples so far have been such that $Y^*(\cdot,\cdot)=\varnothing$. That is, no types outside $T_{\mathscr{X}}$ have ever desired to send a credible message. This means that credible claims by S are always true. But the definition of CMR considers as credible some claims that are with positive probability false. Consider Example 8, and the message "I'm t_1." $A^*(t_1)=a_1$; R's optimal response to beliefs concentrated on t_1 is move a_1. Yet here, $Y^*(t_1, t_1)=\{t_2\}$. That is, it is not unreasonable for type t_2 to try to pool with t_1. Yet, given the priors, the optimal response to the beliefs that he is facing both types still would be for R to play a_1. Thus, $A^{**}(t_1, t_1)=a_1$. Thus, CMR will predict that type t_1 will always induce a_1. Type t_2 always has the option of doing so, but might try to do better with another message.

CMR thus allows R to be uncertain that he is being told the truth, so long as he is confident enough so that he will play as if he believed the message.[7] The possibility of credible messages being lies with positive probability is not readily handled by relabeling. In Example 8, for instance, we cannot guarantee the truth by considering $\{t_1, t_2\}$ to be the self-signaling set, rather than $\{t_1\}$. The set $\{t_1, t_2\}$ *cannot* send a credible message here, because t_2 might be able to do better by making a different statement. While it may be reasonable for R to interpret the statement "I am both types" as a credible message for type t_1, the formulation of CMR emphasizes that R's thought process focuses in on the certainty of t_1's

[7] This fact may be important for the robustness of the results: In all real situations, there is almost certainly *some* probability of types that, though not getting their maximal payoffs, will want to send the same credible messages sent by higher probability types.

sending the message, rather than on being certain that the statement is true.

Finally, returning to Example 4, we saw that S will communicate his private information fully. This is natural; for both states of nature, the players agree on which action is optimal. Communication in such games seems the minimal requirement of a sensible theory. In fact, the result generalizes to all games of pure coordination, where in all states of nature the two players agree on what would be the optimal action.

DEFINITION 11. A simple communication game is a game of pure coordination if, $\forall t$, $\{\operatorname{argmax}_{a \in A^*} U^S(a, t)\} = \{\operatorname{argmax}_{a \in A^*} U^R(a, t)\}$.

PROPOSITION 2. *In a game of pure coordination,* CMR *predicts that for all t, both agents will get their maximal utility.*

Proof. Consider $\mathscr{X} = \{\{t_i\} \mid t_i \in T\}$, the set of singleton subsets of T. Then $Y^{**}(t_i, \mathscr{X}) = \varnothing$ $\forall t_i$, so that, $\forall t_i$, $A^{**}(t_i, \mathscr{X}) = A^*(t_i) = \{\operatorname{argmax}_{a \in A^*} U^S(a, t_i)\}$. Thus, $\mathscr{X}$ can send a credible message profile, yielding both players their maximal utility in all states of nature. The partition $\mathscr{X}(*)$ which sends a credible message profile in CMR must then have this property. Q.E.D.

III. Credible Communication and Equilibrium Analysis

Credible message rationalizability is a non-equilibrium theory: it does not assume, when there are many strategies permitted by the theory, that players have perfectly coordinated expectations about how the game will be played. I feel that the equilibrium assumption is less appropriate than usual in games where cheap talk is explicitly modeled; one informal justification for equilibrium analysis is exactly that agents will communicate so as to form coordinated expectations. To the extent that communication is the justification for equilibrium analysis, and is modeled but does *not* achieve full equilibrium, it is awkward still to impose further equilibrium conditions.[8]

It is of interest, however, that in many of the examples above, CMR by itself already selects only sequential equilibria as plausible outcomes. Examples 2, 3, and 4 each have two classes of sequential equilibria; informative equilibria where the types induce different actions by R, and pooling equilibria where they induce the same action. (But there are an infinite

[8] Of course, there are other informal justifications for equilibrium analysis which may come into play. For instance, agents could learn over time to coordinate the use of language in a particular context.

number of sequential equilibria of each class, each one using a different combination of messages to induce the physical actions.) In each example, the set of rationalizable outcomes is much larger. In Examples 2 and 4, CMR selects the separating equilibria as the only plausible play, and in Example 3, it selects the pooling equilibrium. Thus, in each of these examples, rationalizable outcomes that are inconsistent with equilibrium are ruled out by my theory of credible communication.

More typical is Example 1, where some coordination is implied, but where, for a range of types and actions, players may employ the non-equilibrium behavior permitted by rationalizability. These four examples thus add to the results of Farrell [4]: credible communication has some of the suspected coordinating effect, but does not appear to be a full justification for equilibrium analysis.

Within the framework of equilibrium analysis, Farrell [5] uses the idea of a rich language by defining credible neologisms as deviations from sequential equilibrium. He then defines a solution concept—Neologism-Proof Equilibrium (NPE)—as the set of sequential equilibria for which no credible neologisms exist. In this section, I define an alternative equilibrium concept also based on the assumption of a rich language, and compare its predictions to NPE. The equilibrium concept, *Credible Message Equilibrium* (*CME*), is a straightforward extension of CMR. It predicts that only sequential equilibria formed by strategies that are consistent with CMR will occur.

As a refinement of Nash equilibrium, CME takes a different approach than is traditional. Most refinements begin with Nash equilibria, and then place certain limitations on behavior by players off the equilibrium path. If such restrictions on behavior will induce deviations by other players, then the equilibrium is eliminated.

CME places *ex ante* restrictions on behavior. The range of permitted strategies is first restricted compared to the strategies permitted by rationalizability. From this smaller set, any combination of strategies which forms a sequential equilibrium is permitted. The restrictions on plausible behavior are thus made without regard to the equilibrium conditions themselves.

Definition 12. A sequential equilibrium is a *Credible Message Equilibrium* if $\forall X_k$: $\exists \mathscr{X}$, $X_k \in \mathscr{X}$, which can send a credible message profile, then $\forall t \in X_k$, some action $a^* \in A^*(X_k)$ is played by R whenever S is of type t.

Of immediate interest is whether such a CME exists for all cheap-talk games. It does.

Proposition 3. *For all simple communication games G, $\exists$ a* CME.

Proof. In Appendix.

The proof constructs an artificial game which forces players to utilize credible messages in a particular way consistent with CMR. This new game has a sequential equilibrium. But given the definition of a credible message profile, I show that for any sequential equilibrium, the strategy employed by each player is optimal even if the restrictions of the artificial game were removed. Therefore, the sequential equilibrium in the artificial game is also a sequential equilibrium in the original game. Since it is consistent with CMR, it is a CME.

NPE defines a credible neologism in terms of a given sequential equilibrium. A neologism consists of a claim by some set of types X such that, if R believed the statement and chose the optimal action accordingly, then exactly those types in X would prefer the outcome to their equilibrium outcome.

How do the predictions of CME compare to those of NPE? Proposition 3 is the first difference. While there always exists a CME, frequently there does not exist an NPE.

Consider Example 9. In this game, all sequential equilibria involve both types of S always inducing action a_3. Yet this equilibrium is not an NPE, because type t_1 can send the credible neologism, "I'm t_1." If believed, only he would prefer it to the equilibrium payoff.

CME does not deem this a credible message, because $Y^*(t_1, t_1) = \{t_2\}$. That is, R believes that t_2 might pool in with the message because it is not the worst that t_2 can do. NPE judges the credibility of a message with respect to a specific counter-factual—a would-be equilibrium—whereas CME judges the credibility with respect to the entire universe of reasonable actions.

The difference is not unrelated to a frequent criticism of deviations-based equilibrium refinements (see, e.g., the "Stiglitz critique" in Cho and Kreps [2]). Such refinements rely on some types of a player deviating from an equilibrium, while other types continue to believe the equilibrium is being

EXAMPLE 9

$$p(t_1) = p(t_2) = \tfrac{1}{2}$$

	S		R	
	t_1	t_2	t_1	t_2
a_1	2	−1	3	0
a_2	−1	−2	0	3
a_3	0	0	2	2

played. If in a pooling equilibrium, it were "known" that t_1 will deviate, then if R does not see a credible neologism, he should infer that he is facing t_2. Realizing this, t_2 will want to follow t_1, and likewise send the neologism. Since CME judges the credibility of statements by standards that do not rely on specific counter-factual play, its validity is independent of this controversy associated with deviations-based refinements.

Even if we are willing to fix an equilibrium payoff and assume that all types are certain that if they play their equilibrium action they will receive their equilibrium payoff, some of the statements considered by NPE to be credible neologisms do not seem particularly compelling.

Consider Example 10. There is a fully pooling sequential equilibrium, where R always plays a_0. This is not an NPE, because there exists any of four credible neologisms which would break it. They would be self-signaling messages sent by any of the sets $\{t_1, t_2\}$, $\{t_2, t_3\}$, $\{t_3, t_4\}$, or $\{t_4, t_1\}$.

Presumably any such credible neologism would be believed by R. If they all were believed, then each type could choose which neologism to send. Their optimal choices are clear. Type t_1 would say "I'm either t_1 or t_2," inducing R to play a_1, which will yield t_1 his best payoff of 10. Similarly, each of the other three types will send their preferred messages. Each type will send a different message. But given these choices, R should *not* respond to the self-signaling sets. For instance, since only t_1 will claim "I am either t_1 or t_2," R's optimal response will be to play a_5, not a_1. This would be a very bad outcome for t_1.

In this example, if S believes that any credible neologism would be believed, and sends messages accordingly, then it would be irrational for R to believe any of the neologisms. If R instead optimally to the neologisms,

EXAMPLE 10

	$p(t_1) = p(t_2) = p(t_3) = p(t_4) = \frac{1}{4}$							
	S				R			
	t_1	t_2	t_3	t_4	t_1	t_2	t_3	t_4
a_0	0	0	0	0	3	3	3	3
a_1	10	5	−1	−1	4	4	0	0
a_2	−1	10	5	−1	0	4	4	0
a_3	−1	−1	10	5	0	0	4	4
a_4	5	−1	−1	10	4	0	0	4
a_5	−1	−1	−1	−1	5	0	0	0
a_6	−1	−1	−1	−1	0	5	0	0
a_7	−1	−1	−1	−1	0	0	5	0
a_8	−1	−1	−1	−1	0	0	0	5

then no type of S will wish to send them. The problem is that the different types of S each have clear preferences among their multiple credible neologisms. CME protects against such a problem: R never is committed to believe that a particular type will send each of several messages, if those messages yield that type different utilities.

CME then permits the fully pooling equilibrium. I believe that it is a very plausible equilibrium. It is, however, Pareto-dominated by another equilibrium. In particular, the equilibrium where $\{t_1, t_2\}$ induces a_1 by R, and $\{t_3, t_4\}$ induces a_3, yields a higher payoff for both R and S in *every* state of nature. It is also Pareto-dominated by the equilibrium where $\{t_1, t_4\}$ induces a_4 and $\{t_2, t_3\}$ induces a_2.

The theory of communication proposed by CMR and CME then suggests that communication among agents need not guarantee that they will coordinate only on undominated play. The problem is that if S were to propose a better equilibrium he might reveal his type more than he would wish, because the different types of S differ on which Pareto-dominant equilibrium they wish to induce.[9]

If CME is usually weaker than NPE in ruling out equilibria, there are cases where it is more restrictive. Consider Example 4 again. This is a game of pure coordination, and CMR and CME predict as the unique outcome that both types reveal their type. NPE, however, does not eliminate the pooling equilibrium.

The problem is that NPE looks for a *single* self-signaling set to send a credible neologism. But for either "I'm t_1" or "I'm t_2," both players would prefer to state this to their pooling-equilibrium payoff. If it were common knowledge that t_1 would deviate with "I'm t_1" *and* t_2 would deviate with "I'm t_2," then both statements would be credible. CME attains extra power by constructing profiles of messages that are credible even when none of the component messages would be credible by themselves.

V. Discussion

Credible Message Rationalizability postulates one particular theory of credible communication. As noted, the theory is based on behavioral assumptions that do not follow from rationality alone. Thus, other reasonable theories based on different behavioral assumptions about the use of language might be plausible.

[9] Caution is in order, however. In this simple setting, I have not allowed R to suggest a method of communication by S; he would likely suggest either of the more informative equilibria. It is not clear whether bilateral communication and bilateral asymmetric information will allow Pareto-inferior equilibria.

I conjecture that stronger theories can be formulated which are consistent with how agents typically use language. While I believe such communication as is predicted by CMR is quite compelling, a message must meet strong standards in order to be deemed credible. In Example 5, for instance, a stronger theory could clearly be developed which would be both consistent with rationality and behaviorally compelling.

More generally, S might reasonably doubt his ability to systematically fool R. Incorporating this more cautious thinking by S into a solution concept might be possible. For instance, we could try to formulate a solution concept which considers claims credible that will yield types their best payoff obtainable without fooling R, rather than their best payoff possible.

Theories of credible communication need not be Message Profile Theories. In MPTs, there are a set of messages which are always believed by R, and are always sent by some types of S. Perhaps there may be reasonable theories where certain messages are always believed, but not necessarily always sent by any types of S. This could be consistent with rationality if it depended on R's reasonable uncertainty about whether certain types will or will not send such messages.

In addition to modifying the theory in this simple setting, further work can be done in applying the ideas to richer settings. For instance, many situations involve bilateral asymmetric information and bilateral communication. New issues also arise when a single agent tries to communicate with many people at once.

Of greater interest than the simple games analyzed above are games where agents both communicate information and take physical actions. In such games, private information can be of two forms. Agents can communicate some exogenous private information, as examined in this paper. Or the private information can be endogenous to the game, with claims about unobservable past physical actions, or promises of future actions (as in [4]). Interesting issues can arise because endogenous private information generated by players' actions will itself depend on their theory of what messages are credible. Finally, some of the equilibrium refinement literature has implicitly or explicitly used the idea that players will communicate their intentions in deviating. Explicit modeling of communication into the physical game could contribute to this literature.

Appendix

Proof of Proposition 1. Choose $\sigma \in \Sigma_S(\mathscr{X})$. Suppose that σ is strongly dominated with respect to $\Sigma_R(\mathscr{X})$ by some $\sigma' \notin \Sigma_S(\mathscr{X})$. Using the notation from the definition of an MPT, either σ' was eliminated in some round of iterated strong dominance, or $\sigma' \notin \Sigma_S(0)$. Suppose that $\sigma' \in \Sigma_S(0)$, and was eliminated in round k of iterated strong dominance. This means that σ' was strongly dominated with respect to $\Sigma_R(k)$ by some strategy $\sigma'' \in \Sigma_S(k)$ that is not eliminated in round k. If σ'' strongly dominates σ' with respect to $\Sigma_R(k)$, then σ'' strongly dominates σ' with respect to $\Sigma_R(\mathscr{X})$. Thus, there exists a strategy in $\Sigma_S(k+1)$ that strongly dominates σ with respect to $\Sigma_R(\mathscr{X})$.

Thus, for all k, if there exists a strategy in $\Sigma_S(k)$ that dominates σ with respect to $\Sigma_R(\mathscr{X})$, then there exists such a strategy in $\Sigma_S(k+1)$. But since there is a last round of iteration in which strategies are eliminated, this means that $\Sigma_S(\mathscr{X})$ contains a strategy which strongly dominates σ with respect to $\Sigma_R(\mathscr{X})$. This contradicts the iterated strong dominance part of the definition: σ would be eliminated if it were strongly dominated by some $\sigma' \in \Sigma_S(\mathscr{X})$.

Thus, $\sigma' \notin \Sigma_S(0)$. Define σ'' as follows: $\forall t$: $\exists X_i$: $t \in X_i$, t sends with probability 1 some $m \in M(X_i)$. $\forall t \notin T_{\mathscr{X}}$, $\forall X_i$: $t \notin Y^*(X_i, \mathscr{X})$, t sends any $m \in M(X_i)$ with probability 0, and sends any other messages with probability greater than or equal to that of σ'. By construction, $\sigma'' \in \Sigma_S(0)$. Further, σ'' weakly dominates σ'. To see this, note that for all $t \in T_{\mathscr{X}}$, σ'' induces the highest payoff possible, and for all $t \notin T_{\mathscr{X}}$, σ'' places at least as high a probability as does σ' on any message which is not certain to induce the set of worst possible outcomes for t.

Therefore, $\forall \sigma' \notin \Sigma_S(0)$, $\exists \sigma'' \in \Sigma_S(0)$ that weakly dominates it with respect to $\Sigma_R(0)$, and therefore with respect to $\Sigma_R(\mathscr{X})$. Therefore, if $\sigma \in \Sigma_R(\mathscr{X})$ is strongly dominated by some strategy not in $\Sigma_S(0)$, it is strongly dominated by some strategy in $\Sigma_S(0)$ as well. This contradicts the above argument. Therefore, no $\sigma \in \Sigma_S(\mathscr{X})$ is strongly dominated with respect to $\Sigma_S(\mathscr{X})$ by any strategy.

The proof that no $\gamma \in \Sigma_R(\mathscr{X})$ is strongly dominated by any other strategy $\gamma' \in \Sigma_R(0)$ is the same, mutatis mutandi, as the above.

For $k \geqslant 0$, define the set $\Sigma_S^*(k) = \{\sigma \in \Sigma_S(k) \mid \text{If } \sigma \text{ maps any type } t \text{ into } m_i \in M(X_j) \text{ with positive probability, then it mixes with equal probability over all } m \in M(X_j)\}$. Define $\Sigma_R^*(k) = \{\gamma \in \Sigma_R(k) : \forall X_i \in \mathscr{X}, \forall m_1, m_2 \in M(X_i), \gamma(m_1) = \gamma(m_2)\}$. Let $(\Sigma_S^*(\mathscr{X}), \Sigma_R^*(\mathscr{X}))$ be these definitions corresponding to $(\Sigma_S(\mathscr{X}), \Sigma_R(\mathscr{X}))$.

Then I claim $\forall \sigma \in \Sigma_S(k)$, $\exists \sigma^* \in \Sigma_S^*(k)$: $\forall t$, $\sigma^*(t)$ places equal probability on each $m \notin M(\mathscr{X})$ as does σ, and $\forall \gamma \in \Sigma_R(k)$, $\exists \gamma^* \in \Sigma_R^*(k)$: $\gamma^*(m) = \gamma(m)$ $\forall m \notin M(\mathscr{X})$. Suppose not. Then $\exists$ some lowest k, $\hat{k}$, for which this is not true. (By construction, it is true for $(\Sigma_S(0), \Sigma_R(0))$.) But if $\sigma^* \notin \Sigma_S(\hat{k})$, then it is strongly dominated by some $\hat{\sigma} \in \Sigma_S(\hat{k}-1)$ w.r.t. $\Sigma_R(\hat{k}-1)$. This means that $\hat{\sigma}$ strongly dominates σ^* w.r.t. $\Sigma_R^*(\hat{k}-1)$, but does not strongly dominate σ w.r.t. $\Sigma_R^*(\hat{k}-1)$. However, $V^S(\sigma, \gamma) = V^S(\sigma^*, \gamma)$ $\forall \gamma \in \Sigma_R^*(\hat{k}-1)$, which means that $\hat{\sigma}$ clearly cannot strongly dominate σ^* without strongly dominating σ. If, conversely, $\gamma^* \notin \Sigma_R(\hat{k})$, then it is strongly dominated by some $\hat{\gamma} \in \Sigma_R(\hat{k}-1)$ w.r.t. $\Sigma_S(\hat{k}-1)$. But this means that $\hat{\gamma}$ strongly dominates γ^* w.r.t. $\Sigma_S^*(\hat{k}-1)$, but does not strongly dominate γ. But again, $V^R(\sigma, \gamma^*) = V^R(\sigma, \gamma)$ $\forall \sigma \in \Sigma_S^*(\hat{k}-1)$, which means that $\hat{\gamma}$ clearly cannot strongly dominate γ^* without also strongly dominating γ.

Now suppose $\gamma' \notin \Sigma_R(0)$ strongly dominates $\gamma \in \Sigma_R(\mathscr{X})$ w.r.t. $\Sigma_S(\mathscr{X})$. Then it strongly dominates γ w.r.t. $\Sigma_S^*(\mathscr{X})$, and, by the claim of the above paragraph, $\Sigma_S^*(\mathscr{X})$ is non-empty. But then define $\gamma'' \in \Sigma_R^*(0) \subseteq \sum_R(0)$: $\gamma''(m) = \gamma(m)$ $\forall m \notin M(\mathscr{X})$. Then γ'' weakly dominates γ' w.r.t. $\Sigma_S^*(0)$, and thus weakly dominates γ' w.r.t. $\Sigma_S^*(\mathscr{X})$. But then γ'' strongly dominates γ w.r.t. $\Sigma_S^*(\mathscr{X})$. But since all strategies in $\Sigma_S(\mathscr{X})$ differ from strategies in $\Sigma_S^*(\mathscr{X})$ only in their distributions of $m \in M(\mathscr{X})$, and since γ'' and γ respond the same way to such messages, this means that γ'' strongly dominates γ w.r.t. $\Sigma_S(\mathscr{X})$. But this contradicts the fact that no $\gamma'' \in \Sigma_R(0)$ strongly dominates $\gamma \in \Sigma_R(\mathscr{X})$. Therefore, there does not exist a $\gamma' \notin \Sigma_R(0)$ which strongly dominates γ w.r.t. $\Sigma_S(\mathscr{X})$. Q.E.D.

Proof of Lemma 1. Let $\mathscr{X}(1) = \{X_1, X_2, ..., X_V\}$ and $\mathscr{X}(2) = \{Z_1, Z_2, ..., Z_W\}$.

Let $F(Z_j) = \{t \in Z_j \mid \forall k, t \notin X_k\}$.

Let $G(X_1) = \bigcup_{Z \in Z^*} F(Z_j)$, where $Z^* = \{Z_j \in \mathscr{X}(2) \mid Z_j \cup X_1 \neq \varnothing\}$.

Let $G(X_{n+1}) = \bigcup_{Z \in Z^*} F(Z_j)$, where $Z^* = \{Z_j \in \mathscr{X}(2) \mid Z_j \cup X_{n+1} \neq \varnothing$, and $\forall k \leqslant n$, $Z_j \cup X_k = \varnothing\}$.

The sets $F(Z_j)$ are each of the sets from $\mathscr{X}(2)$ with all types that are also in $\mathscr{X}(1)$ removed. The sets $G(X_j)$ are the sets of types in $\mathscr{X}(2)$ that are not in $\mathscr{X}(1)$, but who pool with types who are in $\mathscr{X}(1)$.

Relabel $\{Z_j : Z_j = F(Z_j)\}$ as $\{Z_1^*, Z_2^*, ..., Z_r^*\}$. These are the set of subsets in $\mathscr{X}(2)$ that do not overlap with any $X_i \in \mathscr{X}(1)$.

By the construction of these sets, we know that $\mathscr{X} = \{X_1 \cup G(X_1), X_2 \cup G(X_2), ..., X_V \cup G(X_V), Z_1^*, Z_2^*, ..., Z_r^*\}$ contains all types in $\mathscr{X}(1)$ and $\mathscr{X}(2)$, and that these are exclusive subsets. I now show that this set of subsets can send a credible message profile. This will complete the proof by construction.

Since $\mathscr{X}(1)$ and $\mathscr{X}(2)$ can send credible message profiles, $A^*(X_j) = \{a^* \mid a^* \in \text{argmax}_{a \in A^*} U^S(a, t)\ \forall t \in X_j\}$ and $A^*(Z_k) = \{a^* \mid a^* \in \text{argmax}_{a \in A^*} U^S(a, t)\ \forall t \in Z_k\}$. But if X_j and Z_k share an element, then $A^*(X_j) = A^*(Z_k)$. Thus, $A^*(X_j) = \{a^* \mid a^* \in \text{argmax}_{a \in A^*} U^S(a, t)\ \forall t \in X_j \cup F(X_j)\}$.

Clearly $Y^*(X_j \cup G(X_j), \mathscr{X}) \subseteq Y^*(X_j, \mathscr{X}(1))$; the set of types who might want to pool with X_j is smaller in the concatenated message profile, because both exclusionary restrictions in the definition of $Y^*(X_j, \mathscr{X})$ are made more restrictive by concatenation. Further, since $A^*(X_j) = \{a^* \mid a^* \in \text{argmax}_{a \in A^*} U^S(a, t)\ \forall t \in G(X_j)\}$, $G(X_j) \subseteq Y^*(X_j, \mathscr{X}(1))$. Thus $Y^*(X_j \cup G(X_j), \mathscr{X}) \cup G(X_j) \subseteq Y^*(X_j, \mathscr{X}(1))$. By the definition of $A^{**}(\cdot, \cdot)$, this implies that $A^{**}(X_j \cup G(X_j), \mathscr{X}) \subseteq A^{**}(X_j, \mathscr{X}(1))$.

But since $A^*(\cdot) \subseteq A^{**}(\cdot, \cdot)$ always, and $A^{**}(X_j, \mathscr{X}(1)) = A^*(X_j)$ and $A^*(X_j \cup G(X_j)) = A^*(X_j)$, we have $A^{**}(X_j \cup G(X_j), \mathscr{X}) = A^*(X_j \cup G(X_j)) = \{a^* \mid a^* \in \operatorname{argmax}_{a \in A^*} U^S(a, t)\ \forall t \in X_j \cup G(X_j)\}$.

Also, $Y^*(Z_k^*, \mathscr{X}) \subseteq Y^*(Z_k, \mathscr{X}(2))$, so that $A^{**}(Z_j^*, \mathscr{X}) \subseteq A^{**}(Z_k, \mathscr{X}(2))$. Since $Z_k^* = Z_k$, $A^*(Z_k^*) = A^*(Z_k) = A^{**}(Z_k, \mathscr{X}(2))$, we have $A^{**}(Z_k^*, \mathscr{X}) = A^*(Z_k^*) = \{a^* \mid a^* \in \operatorname{argmax}_{a \in A^*} U^S(a, t)\ \forall t \in Z_k\}$.

Thus, $\mathscr{X}$ can send a credible message profile Q.E.D.

Proof of Proposition 3. Consider the simple communication game G and a profile of types $\mathscr{X}$ that can send a credible message profile $M(\mathscr{X})$. Then define a *Profile Game* $G(\mathscr{X}, M(\mathscr{X}))$ as follows:

$\forall t \notin T_{\mathscr{X}}$, player S can send any message $m \in M \backslash M(\mathscr{X})$, or can send the mixed message $m^*(X_k)$, for any X_k: $t \in Y^*(X_k, \mathscr{X})$, where $m^*(X_k)$ consists of mixing with equal probability over all $m \in M(X_k)$. All $t \in X_k$, $\forall X_k \in \mathscr{X}$, must send $m^*(X_k)$. Player R can choose any action $a \in A^*$ in response to $m \in M \backslash M(\mathscr{X})$. He must respond with some $a \in A^*(X_k)$ to any $m \in M(X_k)$.

This is a well-defined game, with finite type and action space, so there must exist some sequential equilibrium. I claim that (1) the associated strategies must be an equilibrium in the original game G, and (2) the associated strategies by each player are permitted by CMR.

Clearly any sequential equilibrium (σ, γ) of $G(\mathscr{X}, M(\mathscr{X})) \in (\sum_S(0), \Sigma_R(0))$, as defined in the definition of an MPT for G, $\mathscr{X}$, and $M(\mathscr{X})$. *If* (σ, γ) is a sequential equilibrium in G, then clearly it will survive iterated strong dominance: each strategy as an optimal response to the belief that the other player will play his equilibrium strategy. Thus, if (σ, γ) is a sequential equilibrium, then $(\sigma, \gamma) \in (\Sigma_S^{\mathscr{X}}, \Sigma_R^{\mathscr{X}})$.

Choose a sequential equilibrium (σ, γ) in $G(\mathscr{X}, M(\mathscr{X}))$. $\forall t \notin \mathscr{X}$, because σ is optimal against γ in $G(\mathscr{X}, M(\mathscr{X}))$, and the only prevented actions would yield these types their lowest possible payoffs, σ is optimal against γ in G. $\forall t \in \mathscr{X}$, each player is getting his maximal payoff of any action $a \in A^*$. Since all a played in γ are contained in A^*, $\forall t \in T$, σ is optimal against γ in the original game G.

By construction of a credible message profile, for any strategy by $S \in \Sigma_S^{\mathscr{X}}$ where any type who sends $m \in M(X_k)$ mixes with equal probability over all $m \in M(X_k)$, it must be optimal for R to respond by playing $a \in A^*(X_k)$. Therefore, since R can choose any $a \in A^*$ in response to $m \notin M(\mathscr{X})$, as in the original game G, γ is an optimal response to σ in G. Q.E.D.

References

1. B. D. Bernheim, Rationalizable strategic behavior, *Econometrica* **52** (1984), 1007–1028.
2. In-Koo Cho and David M. Kreps, Signaling games and stable equilibria, *Quart. J. Econ.* **102** (1987), 179–221.

3. VINCENT P. CRAWFORD AND JOEL SOBEL, Strategic information transmission, *Econometrica* **50** (1982), 1431–1451.
4. JOSEPH FARRELL, Communication, coordination, and Nash equilibrium, *Econ. Lett.* **27** (1988), 209–214.
5. JOSEPH FARRELL, Meaning and credibility in cheap-talk games, *in* "Mathematical Models in Economics" (M. Demster, Ed.), Oxford Univ. Press, Oxford, forthcoming.
6. D. KREPS AND R. WILSON, Sequential equilibrium, *Econometrica* **50** (1982), 863–894.
7. E. KOHLBERG AND J. F. MERTENS, On the strategic stability of equilibria, *Econometrica* **54** (1986), 1003–1038.
8. ROGER B. MYERSON, "Credible Negotiation Statements and Coherent Plans," Discussion Paper No. 691, Center for Mathematical Studies in Economics and Management Science, Northwestern University, August 1986.
9. D. PEARCE, Rationalizable strategic behavior and the problem of perfection, *Econometrica* **52** (1984), 1029–1050.
10. M. SPENCE, Job market signaling, *Quart. J. Econ.* **87** (1973), 355–374.