SAYING AND MEANING, CHEAP TALK AND CREDIBILITY

Robert Stalnaker

In May 2003, the U.S. Treasury Secretary, John Snow, in response to a question, made some remarks that caused the dollar to drop precipitously in value. The *Wall Street Journal* sharply criticized him for "playing with fire," and characterized his remarks as "dumping on his own currency", "bashing the dollar," and "talking the dollar down". What he in fact said was this: "When the dollar is at a lower level it helps exports, and I think exports are getting stronger as a result." This was an uncontroversial factual claim that everyone, whatever his or her views about what US government currency policy is or should be, would agree with. Why did it have such an impact? "What he has done," Alan Blinder said, "is stated one of those obvious truths that the secretary of the Treasury isn't supposed to say. When the secretary of the Treasury says something like that, it gets imbued with deep meaning whether he wants it to or not." Some thought the secretary knew what he was doing, and intended his remarks to have the effect that they had. ("I think he chose his words carefully," said one currency strategist.) Others thought that he was "straying from the script," and "isn't yet fluent in the delicate language of dollar policy." The Secretary, and other officials, followed up his initial remarks by reiterating that the government's policy was to support a strong dollar, but some still saw his initial remark as a *signal* that "the Bush administration *secretly* welcomes the dollar's decline."[1] The explicit policy statements apparently lacked credibility. Perhaps their meaning was not as deep as the meaning of the "secret" signal that currency traders and other observers had read into the initial remarks.

Meaning, whether deep or on the surface, is an elusive and complicated matter. It is difficult to be clear about exactly what is going on in even the most direct, literal acts of communication. My aim in this paper to bring out some of the problems by looking at two very different projects that each try to say something about what it is to mean things. I will first take a look back at Paul Grice's analysis of meaning, and the wider project of which this analysis was the cornerstone. Then I will discuss some attempts by game theorists to give an account of acts of signaling, particularly of acts that are, in a sense to be defined, acts of pure communication. I think that these two projects, which use some similar ideas in very different ways, and face some related problems, throw light on each other.

My discussion will be preliminary, speculative, and indirect, looking only at highly artificial and idealized situations, and reaching only tentative conclusions even about them, so I should begin with the kind of qualifications and expressions of reservation for which Grice was famous. Echoing Grice, "What follows is [only] a sketch of a direction." "We should recognize that at the start we shall be moving fairly large conceptual slabs around a somewhat crudely fashioned board." But I think that if we can get clear about how some basic concepts work in a very simple setting, this will help us to understand the kind of strategic reasoning that is involved in more complex and interesting communicative situations.

My plan is this: I will begin by sketching the Gricean project, as I understand it – its motivation and some of its central ideas. Then I will look at the game theoretic project – a project that focuses on the role of what is called "cheap talk," and on the idea of credibility. I will point to some of the parallels in the two projects, make some suggestions, which are influenced by

[1] Quotations are from the *Wall Street Journal*, May 13, 2003, and from a *Wall Street Journal* editorial.

Grice’s project, about the way the idea of credibility might be characterized, and look at the consequences of these suggestions for some simple games. I will conclude by considering how these ideas might be generalized to slightly more complex and realistic situations, and how they might help to clarify some of the patterns of reasoning involved in real communication. We will see, in the end, if we can get any insight into the question why the secretary of the Treasury didn't just say what he meant, and why others did not take him to be doing so.

The Gricean project was begun in a philosophical environment (Oxford “ordinary language” philosophy of the 1950's) that now seems very distant and alien. Grice was very much a part of this philosophical movement, but was also reacting to it. The ordinary language philosophers shared with the philosophers in the logical empiricist tradition the idea that philosophical problems were essentially problems about language, but they were reacting to that tradition’s emphasis on the artificial languages of logic, and the method of clarification by translation into such formal languages, a procedure that abstracted away from the speaker and context, and from the way natural languages were actually used. The ordinary language philosophers emphasized that speech is a kind of action. To use terminology not current at the time, their focus was on pragmatics rather than just semantics. Meaning was to be understood in terms of the way that a speech act is intended to affect the situation in which it is performed.

Grice’s distinctive project[2] was to provide a philosophical analysis of speaker meaning – to give necessary and sufficient conditions for a claim of the form “S [a speaker] means that *P* by *u* [an utterance].” Speaker meaning (which Grice called “nonnatural meaning,” or “meaning-nn” to contrast it with a sense of “meaning” as a natural sign) was to be the basic semantic concept in terms of which the meanings of statements, sentences and words were to be explained. As with any project of reductive analysis, to clarify the aim, one need to say what is being analyzed in terms of what – what is problematic, and what are the resources of analysis. Grice’s answer to this question is distinctive, and gave his project a character quite different from most philosophical projects of explaining meaning. His project was to explain *semantic* concepts in terms of the beliefs and intentions of the agents who mean things. In contrast, most philosophers, both before and after Grice, who are trying to say what it is for something to *mean* some particular thing are addressing the problem of intentionality – the problem of how words (and thoughts) manage to connect with the world – how they can be *about* something, have propositional content, be true or false. Quine on radical translation, and Davidson on radical interpretation, Michael Dummett on theory of meaning, causal theories of reference, Jerry Fodor on the semantics for the language of thought – all of these projects are attempts to explain both mental and linguistic intentionality in non-intentional terms. The standard strategy for the explanation of intentionality was to begin with language, and then to explain the intentionality of belief, desire and intention as somehow derivative from the intentionality of language. Thinking is “saying in your heart,” (perhaps in the language of thought); the mental act of judgment is the “interiorization” of the act of assertion; believing is being disposed to affirm or assert, where the

[2] Grice’s lectures and papers on meaning and conversation are collected in Grice (1989). See in particular, ch. 14, “Meaning”, originally published in 1957, ch. 5, "Utterer's meaning and intentions"(1969), and the retrospective epilogue.

content of what one is disposed to say is to be explained in terms of the way the language as a whole is used by the speaker's community. Intentionality arises out of the constitutive rules (to use John Searle's term) of an institutional practice of speech.[3]

An important part of the motivation for Grice's project was to reverse the direction of explanation: to return to the idea, more natural from a naive point of view, that speech is to be explained in terms of thought. A speech act is an action that like any rational action should be explained in terms of the purposes for which it is performed, and the agent's beliefs about its consequences. Speech may be an institutionalized social practice, but it is a practice with a function that is intelligible independently of the practice, and we can get clearer about how the practice works by getting clear about what that function is. Grice's idea was that speech is an institution whose function is to provide resources to mean things, and that what it is to mean things needs to be explained independently of the institution whose aim it is to provide the means to do it.

So the project is to explain the distinctive character of *communicative* action, taking for granted the normal resources for the explanation of rational action – beliefs, desires, values and ends, intentions. Step one is the simple idea that a communicative act is an attempt to get someone to believe something, but not every attempt to get someone to believe something is an act of meaning something. I might, for example, try to get the police to believe that the butler did it by putting the murder weapon in the butler's pantry, and to do so would not be an act of meaning anything. The problem is to say what must be true about the way that one intends to induce a belief in order for an act done with that intention to be an act of meaning something.

Step two is to add that the intention to induce a belief must be manifest or transparent (excluding the evidence-planting cases, which can work only if they are not recognized for what they are). It does seem to be a central feature of meaning that it is open – an act is a communicative one only if the intention to communicate is mutually recognized. (Communication can, of course, be devious and deceptive, but a speaker cannot attempt to deceive her interlocutor about what she intends him to understand her to be meaning.) Still, transparency is not enough. Grice used the example of Herod presenting the head of John the Baptist on a charger to Salome to illustrate that more was needed for meaning. Herod's intention to induce the belief that John the Baptist had been beheaded was manifest, but this was not an act of meaning that he had been beheaded. What needed to be added, Grice argued, was that the recognition of the intention must play an essential role in accomplishing the primary intention to induce the belief. In an act of pure communication, the recognition of the intention is what does the work of inducing the belief.
So this was Grice's basic analysis:

[3] See Stalnaker (1984), chs 1 and 2 on the problem of intentionality and the contrast between linguistic and pragmatic strategies for explaining intentionality.

> "We may say that 'A meant_{NN} something by *x*' is roughly equivalent to 'A uttered *x* with the intention of inducing a belief by means of the recognition of this intention.'"[4]

The analysis was later refined and complicated in response to a barrage of counterexamples. Refinement of the Gricean analysis became one of those cottage industries that periodically take hold of the philosophical literature, with evermore complex counterexamples offered, and evermore complex clauses and qualifications added to the analysis in response. We will pass over the details. Our interest is not in vindicating the project of reductive analysis by getting the necessary and sufficient conditions for meaning exactly right, but in what the general ideas of the components of such an analysis might show about the way speakers and addressees reason about communicative acts. In particular, if an act of meaning something is an act of roughly this kind, then we can ask the following two questions about any act of uttering *u*[5] in order to mean that *P*:

(1) Why should uttering *u* be a way for S to get H to recognize her intention to get him to believe that *P*?

(2) Why should getting him to recognize her intention to get him to believe that *P* be a way of getting him to believe that *P*?

Question (1) will be answered in different ways in different situations. It could be that *u* is a natural sign (an act of smiling, frowning, or pointing, for example) that naturally tends to induce a belief, or to make prominent a thought, and as a result has come to be used. Or accidental associations may be noticed, and come to be mutually recognized, and reinforced over time. Obviously, the dominant way of meaning things is by saying them, which is to say by the use of an elaborate conventional system, codified and taught, that associates, in a systematic way, a range of sound patterns with a range of propositions, and Grice thought that this way of meaning things was in some sense central. But it was crucial for his project that meaning be intelligible independently of such institutionalized practices so that one can understand the practice in terms of the function – to mean things – that it is designed to serve, and so that one can better explain why people say what they say, and how sometimes they are able to exploit the rules of a linguistic practice in order to mean things different from what they are saying, or from what the conventional rules imply that they are saying. So while this first question must have an answer, in each particular case, in order for it to be possible for S to mean that *P* by uttering *u*, the question need not be answered in any particular way in order for the act to count as an act of meaning.

[4] Grice (1989), 219.

[5] Grice makes clear that he is using the term "utterance" in an artificially broad way as a label for any act that is a candidate for an act of meaning something.

The second question – why should getting H to believe that S intends him to recognize her intention to get him to believe that *P* be a means of getting him to believe that *P*? – will have a satisfactory answer only if the pattern of priorities and beliefs is (or at least is believed to be) such as to give H reason to think that S would want him to believe that *P* only if *P* were true. (This is one of the things, as we will see, that the game theoretic apparatus can help to sharpen.) If the kind of intention that Grice uses to analyze speaker meaning is really essential to genuine communication, then it will be essential to the possibility of communication that there be a certain pattern of common interest between the participating parties. It will follow from the analysis of meaning that something like Grice's cooperative principle, a principle that plays a central role in his theory of conversational implicature, is essential to the very idea of communication.[6]

Cheap talk signaling games[7]

As many people have noticed, [8] Gricean ideas naturally suggest a game theoretic treatment. The patterns of iterated knowledge and belief that are characteristic of game-theoretic reasoning are prominent in Grice's discussions of speaker meaning, and the patterns of strategic reasoning that Grice discussed in the derivation of conversational implicatures are patterns that game theory is designed to clarify. (Grice's general pattern: one may communicate by saying something that gets the addressee to reason in the following way: what must be true in order that it be rational for S to have said *that*? If the answer is, it must be true that *P*, and if it is transparent that the

[6] This is Grice's cooperative principle:" Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purposes or direction of the talk exchange in which you are engaged." (Grice (1989), 26). About this principle, he says "I would like to be able to show . . . that anyone who cares about the goals that are central to conversation/communication . . . must be expected to have an interest, given suitable circumstances, in participating in talk exchanges that will be profitable only on the assumption that they are conducted in general accordance with the Cooperative Principle and the maxims." (Grice (1989), 30)

[7] I am indebted to work by Robert Farrell and Matthew Rabin on cheap talk and credibility, which got me to appreciate both the complexity of the problems, and some of the constructive ideas that may provide solutions. See their papers in list of references below. The account that is here informally sketched will not (when it is formally spelled out) be exactly the same, in the set of strategies it identifies, as Rabin's notion of credible message rationalizability, but it will be close, and my account was strongly influenced by the examples he used to motivate and raise problems for his own account. I hope in a later more technical paper to focus on some of the examples on which the two accounts differ.

[8] David Lewis was the first, to my knowledge, to connect Gricean ideas to game theory. His analysis of convention, developed in Lewis (1969) drew on work of Thomas Schelling, and discussed the kind of signaling games discussed below. More recent work includes Parikh (2001) and van Rooy (2003), two examples of a growing literature.

speaker intended the addressee to reason in this way, then whatever the literal meaning of what one said, this will count, on a Gricean analysis, as a case of meaning that *P*.) Grice never, to my knowledge, discussed the potential connection between his work and game theory, and some of the developments within game theory that are most relevant to Grice's work (in particular, the explicit modeling of common knowledge and belief, and more generally of the epistemic foundations of game-theoretic reasoning[9]) occurred after his work on meaning and implicature. But game theory provides both some sharp tools for formulating some of Grice's ideas, and some simple idealized models of examples to which those ideas might be applied. And I think Gricean ideas will throw some light on the problems game theorists face when they try to model communicative situations.

I will make some remarks about the general game theoretic setting, and then describe the simple communication games that I will be concerned with. Following this, I will state the problem about meaning that arises in this context, and sketch and refine, informally, a response to the problem.

A game is a sequence of decision problems, usually involving two or more agents, where the outcome depends on the way the actions of different agents interact. To define a game, one specifies the alternative actions available at each point in the playing of the game – which player gets to move, what information that player has about the prior moves of other players, and what the consequences of his or her move are for the subsequent course of the game. Sometimes there are chance moves in the game, in addition to moves by rational players. In such cases, a probability is specified for each chance move, and it is assumed that these probabilities are mutually known, and determine the prior beliefs of all the players about those moves. The definition of a game also specifies each player's motivating values (utilities) for each of the alternative ultimate outcomes of the game. The definition of a game does not specify the beliefs and degrees of belief of the players about the actions of other players. Instead, it is normally assumed that the players will act rationally, and that it is common knowledge that they will act rationally. It is also assumed that the structure of the game is common knowledge among the players.

In the early developments of game theory, there was no formal representation of the idea of common knowledge; it was just a part of the informal commentary used to motivate the notion of Nash equilibrium, and various refinements of it, which were taken to be implicit analyses of an idea of game-theoretic rationality. While it was assumed that rationality required maximizing expected utility when probabilities were given and known, it was not assumed that a player had probabilistic degrees of beliefs about the rational actions of other players, except when it was known or assumed that the other player had chosen a mixed strategy – a strategy that allowed chance to determine his or her choice. In the contrasting Bayesian, or epistemic approach to game theory, developed later, the ideas of common knowledge and common belief were made

[9] see Battigalli and Bonnano (1999) for an excellent survey of the literature in the epistemic approach to game theory. My way of developing an epistemic model theory for games is discussed in Stalnaker (1997). Grice's work had an indirect influence on some of these developments, since it influenced David Lewis's analysis of common knowledge in Lewis (1969), which in turn influenced the development of the epistemic approach to game theory.

formally explicit, and it was assumed that rationality was identified, in all cases, with maximizing expected utility. It was assumed that players have degrees of belief about the behavior of other rational agents, as well as about chance moves. The assumption of common knowledge, or common belief, that all players act rationally may determine those beliefs in some cases, but in other cases, the structure of the game, and the assumption that it is common knowledge that players will make rational choices, given their beliefs, will be compatible with different *models* for the game, where a model for a game provides a full specification of the beliefs and degrees of belief of each of the players about the behavior and beliefs of the others, as well as a specification of what move each player makes, and is disposed to make, at each choice point in the game. Given a model theory for a game, one can give a mathematically precise definition of a solution concept in epistemic terms by specifying a class of models that meet some intuitively plausible epistemic constraints. A strategy or strategy profile satisfies the solution concept if it is realized in some model in the class. So, to take the most basic solution concept, one may define the *rationalizable* strategies of any given game as the set of strategies each of which is realized in some model for the game that satisfies the condition that there is common belief among the players that all players choose rationally. One can then prove that this set of strategies coincides with the set determined by other definitions of rationalizability – for example, rationalizability defined as the set of strategies that survive the iterated elimination of strictly dominated strategies.

The games I will be concerned with in this paper will all be simple sender-receiver games that are designed to model acts whose sole purpose is the communication of information. In these games, one player (the sender, S) has some information (determined by an initial chance move in the game) that is unavailable to the other player (the receiver, R). The chance move (which may model any fact about the state of the world that is determined independently of the choices of the players) determines the sender's *type*, which is simply a label for the state that the information puts the sender in. Only R can act, but the information about S's type that he lacks will normally be relevant in one way or another both to the choice that R would want to make, and to the choice that S would want him to make. All S can do to influence the outcome is to send a signal to R. R can then make his choice depend, in any way he chooses, on which of the alternative signals S sends. In the general case, the signal that S chooses to send might or might not affect the options available to R, and the payoffs to the players, but in a *cheap talk* game, they do not. A cheap talk signal is, by definition, one that has no effect on the subsequent course of the game, except to give R the option of making his choice depend on which signal is sent. That is, the moves available to R, and the consequences of those moves for both S and R are independent of the signal that is sent.

Normally, in game theory, a move in a game is characterized simply by the subsequent options and payoffs that the move makes available – by the subgame that results from the move. In the case of cheap talk, it is true by definition that each of the cheap talk moves available to the sender has exactly the same effect; the subgame that results from one signal is exactly the same as the subgame that results from any other. But the signal will have a point only if it conveys some information, information that is different from the information conveyed by alternative signals. If the theory is to provide any guidance, or any explanation for the choices of players in such games, something must be added to the description of the game that distinguishes the messages in a way that is relevant to the information that they might convey.

Let me illustrate the problem with the following minimal signaling game, example 1, where there is no conflict of interest, and communication should be as easy and unproblematic as it gets:

	a1	***a2***
t1	***10*** ***10***	***0*** ***0***
t2	***0*** ***0***	***10*** ***10***

S is of either type t1 or t2, determined by chance, with equal probability.[10] The columns represent R’s two alternative actions, and the numbers in the cells of the matrix are the payoffs to S and R. Let us suppose that S may send either of two messages, *m1* and *m2*, and that she must send one or the other. So S has four alternative strategies: send *m1* unconditionally, send *m1* if she is of type t1 and *m2* if of type t2, send *m2* if of type t1 and *m1* if of type t2, or send *m2* unconditionally. R also has four alternative strategies for how to respond to the message: he may choose either action unconditionally, or he may make his choice depend on the message in either of the two possible ways. It is clear that if information is to be conveyed, S must choose one of the two conditional strategies, and if the information is to be exploited, R must choose one of his conditional strategies, but nothing about the basic structure of the game favors one of the conditional strategies over the other, for either player. What we need to build in is something about the meaning or content of the messages, and to say how the fact that the messages have the meanings or contents that they have determines or constrains the effect of sending the messages.[11]

[10] In all the examples I will discuss, I assume that the chance move that assigns types to S give equal probability to all types.

[11] The problem was first posed as a problem about equilibria in cheap talk games. On the one hand, it was shown that the addition of a cheap talk move makes possible new equilibrium solutions. But on the other hand, the standard theory provides no basis for favoring one over other symmetrical alternative equilibria. And it was shown that there will always be what was called a “babbling equilibrium” in which the sender chooses her signal at random, and the

The resources available to the game theorist for solving this problem are similar to those available to Grice in his reductive project, which was to explain meaning in terms of a pattern of beliefs and intentions. The game theorist characterizes games and models for games in terms of the beliefs and motivating values of the agents, which in turn determine their intentions and actions, so he or she has the same resources. And there are more specific parallels between Grice's project and the problem of representing meaning in signaling games: it was an important component of Grice's analysis that an action counts as a case of meaning only if the recognition of the intention to induce a belief played an essential part in inducing the belief. The contrast was with the presentation of evidence that is intended to induce a belief by a means independent of facts about the utterer's intentions. The same contrast is implicit in the idea of cheap talk, which contrasts with costly signaling, where something about the sender's beliefs and priorities is demonstrated by an action that has consequences that are independent of the information sent, and that can be seen, on independent grounds, to be irrational unless the proposition the sender intends to communicate is true. (For example, one shows one's wealth by acts of conspicuous consumption that would be prohibitive for one who is not wealthy.)

Grice's analysis suggested that the explanation for an act of meaning divides into two stages, corresponding to the two questions distinguished above that may be asked about why an utterance *u* was able to convey the information that *P*, and it is useful to divide the problem of explaining the meaning of messages in a signaling game in the same way. First, somehow, an action that has no external effect on the situation, and no intrinsic connection with any information (it does not present independent evidence) is able to convey a particular intention of the speaker to induce a belief. Second, the conveying of this intention to induce a certain belief is supposed to succeed in inducing the belief. The central way of explaining the first stage – of answering the first question – was in terms of a conventional device – a language – whose function is to mean things. The central way to mean something is to *say* it. The language provides a mutually recognized systematic correlation between actions that are easy (and cheap) to perform and certain items of information – propositions. So let us suppose that, in our signaling games, S has such a device available to her. In specifying the game, we will specify the conventional meaning of the alternative signals that are available to S. The focus is then on the second stage of the explanation – on the question, under what conditions can such a device be used successfully to mean things – to convey information simply in virtue of the recognition of the sender's manifest desire to send it? This is the question of credibility, which is the central concept in the discussion of cheap talk games.

In general, an epistemic model for a game will contain a state space, or a set of alternative possible worlds that represent the alternative ways that the game might be played, and the alternative belief states that the players might be in. A *proposition* (or in the terminology of the statistician and decision theorist, an event) is represented by a subset of the state space, or a set of possible worlds. So to specify what the available messages *say* we associate with each message a proposition or event. In the general case, a message might express any proposition, but in our simple games, we will restrict possible messages to information about S's type. One

receiver ignores the signal, choosing his response at random. The seminal paper is Crawford and Sobel (1982).

might have a restricted list of available messages, or one might assume that a rich language is available in which anything may be said about S's type (for example, if there are four types, there will be 15 consistent propositions, and so 15 distinguishable messages that are available to be sent. One of them – the tautological proposition – is a message that is equivalent to sending no message at all; four others are determinate propositions that say that S is one particular type; the others convey partial information (for example that S is either of type 2 or type 3, or that S is not of type 2).

The idea of credibility is simple enough, and it is easy to see, intuitively, that in our simple minimal example, once we have endowed our messages with meaning, credible communication will be unproblematic. But as we will see, there are some complications in spelling the definition out in detail. I will characterize the simple idea by giving a rough and unrefined definition of credibility, together with an assumption about the effect of sending a credible message that we can impose as a constraint on the game models we are interested in. The unrefined definition and assumption suffice so long as we don't look beyond the simple and unproblematic cases, and I will illustrate how they work with our minimal example. I will then use some more complex examples to show that the account of credibility need to be refined and qualified, and also to point to some of the complexities of reasoning about communication, and to the possibility that the meaning of a message might diverge from what the message literally says.

First a definition of a preliminary concept, to be used in the definition of credibility:

> A message is *prima facie rational* (pf rational) for player S, of type t if and only if S prefers that R believe the content of the message.

Second, the definition of credibility in terms of pf rationality:

> A message is *credible* if and only if it is pf rational for some types, and only for types for which it is true.

Third, the constraint:

> It is common belief that the content of any credible message that is sent is believed (by R).

This constraint is to be added to the usual constraints that are used to give an epistemic definition of rationalizability: that the structure of the game is common belief, and that it is common belief that both players are rational (that they make choices that maximize their expected utility).[12]

In our simple coordination game (example 1), assume that the message *m1* has the content "S is of type t1" and that message *m2* has the content "S is of type t2". Obviously, *m1* is pf rational

[12] Or, one can add the credibility assumption to any refinement of rationalizability, or to some epistemic conditions that characterize the class of Nash equilibrium strategies.

for t1, but not for t2, and *m2* is pf rational for t2 but not for t1, so both messages are credible. Therefore, by the constraint, it is common belief that R will believe either message, if it is sent, and since it is also common belief that R is rational, it follows that it is true and common belief that R will play the strategy, “a1 if *m1*, a2 if *m2*". S’s best response to this strategy is to send *m1* if she is of type t1 and *m2* if she is of type t2, so our assumptions imply that this is what she will do. Communication, in this simple game, will take place, and will be successful, in any model satisfying our constraints.

But when we move beyond the simple cases, we see that our definitions are not so clear as one might hope, and the required refinements will bring out the holistic and interdependent character of credibility, and will also point to some of the subtleties of strategic reasoning about communication. I will start with a question about how the definition of pf rationality is to be understood: the definition says that for a message to be pf rational, S must prefer that R believe the content of the message, but prefer that to what? It is neither necessary nor sufficient, to capture the intended idea, that S should prefer that R believe the message *rather than to remain in his prior belief state*, since remaining in the prior belief state may not be a feasible option. Consider the following game, example 2:

	a1	**a2**	**a3**	**a4**
t1	5 5	10 10	0 0	0 0
t2	5 5	0 0	6 0	8 1
t3	5 5	0 0	6 6	0 0

If S is type t2, then her first choice is that R get no information at all - to remain in the prior belief state - because that would motivate him to choose a1. But that is not an available option, since it is clear that the message “S is t1" is a credible message that S would be rationally required to send if and only if she were of type t1. So R will infer that S is not t1 if he does not get that message. So sending no message at all would induce the belief that S is either t2 or t3, which (if R didn’t know which of the two it was) would result in action a3, which is a worst outcome for t2. But if t2 is able to reveal her type, R will instead choose a4, which S (if she is of type t2) would prefer to a3. So the message, “S is t2", should be pf rational for t2, since she prefers that R believe that message *to the feasible alternatives to believing it.* Since this message is pf rational *only* for t2, it is credible. Our definitions should ensure that S will reveal her actual type if she is t1 or t2, and that R will believe her and respond appropriately.

The expected effect on R of the feasible alternatives to a given message *m* may depend on whether those alternative messages are credible, which in turn may depend on whether the alternatives to those messages (including *m* itself) are credible. There is a circularity here, but it is not a vicious circularity, since it is not assumed that the players' beliefs can be generated from the definition of the game, and the constraints on credibility. What the circularity implies is that sometimes a message will be credible in one model of a given game, but not in other models of the same game. Example 4, discussed below, will illustrate the phenomenon.

Example 2 showed that sending no message may reveal information, whether the sender wants to reveal it or not. It is also true that sending a credible message may reveal more information than is contained in the explicit content of the message. We have said that a message is credible if it is *sometimes* pf rational,[13] and also pf rational *only* when true; it is not required that it be pf rational, in all cases, when it is true. It might happen that a partial message is pf rational for *some* types for which it is true, and *only* for types for which it is true, but not for *all* types for which it is true. In such a case, if the message is sent, R will believe the message, but will also come to believe more. So, for example, if the disjunctive message, “S is either t1 or t2" is credible, and rational for t1 to send, but not rational for t2 to send, then R would believe the message, if it were sent, but would also come to believe something stronger – that S is of type t1. We need to take account of this possibility in the definition of pf rationality. What S must prefer, for a message to be pf rational, is that R believe the message *in the way he would believe it if the message were received* to all feasible alternatives.

Example 3 illustrates this kind of situation:

[13] This clause was added just so that messages that S *never* would want to be believed do not count as vacuously credible. This does not really matter, since such messages will not be sent by rational players, but it does not seem natural to assume that if, contrary to fact, they were sent, they would be believed.

	a1	a2	a3
t1	5 5	0 0	0 0
t2	5 5	6 0	0 0
t3	5 5	6 6	8 8

Here we assume that there are just two available messages: “S is t1" or “S is not t1". The second message is pf rational for t3, and not for t1 or t2. So it is credible, but will not be sent by t2. The first message is not credible, since if S is of type t2, the message would be false, but she might have a motive to send it, and will definitely have a motive to send it if it is required that one of the two messages be sent. Here we have a case where the meaning of the messages (in Grice’s sense) diverges from what the messages literally say, and (like Grice’s phenomenon of conversational implicature) the divergence is explained in terms of what the messages literally say. Even though the first message literally means that S is t1, it will manifestly express S’s intention to induce the belief that she is either t1 or t2, and will succeed in doing this. It will not credibly communicate its literal content, and so is not strictly speaking credible, but it will credibly convey something weaker. And since it will be mutually recognized that the second message will be sent only by t3, it will induce the stronger belief that it is manifestly intended to induce, that S is t3.

We noted above that credibility is a feature of a model of a game, since it depends on the pattern of S’s beliefs; sometimes a message is determined to be credible, or to be not credible, by the structure of the game, together with the general assumptions that define the relevant class of models. But with some games, a message might be credible in some of the models that conform to the constraints, and not in others. Furthermore, it might happen, with such games, that in some models, R is mistaken or ignorant about whether a message is credible. Credibility, as we have defined it, is a property determined entirely by S’s beliefs and utilities, and while the utilities are assumed to be common knowledge, players’ beliefs are not. If R is mistaken or uncertain about what S believes, he may be mistaken or uncertain about whether her messages are credible. But it is not plausible to assume that credible messages are believed by R in cases where R does not realize that they are credible, so our constraint should not say that the content of a message that is sent and is *actually* credible is believed by R, but rather that the content of a message that is sent, and that is *believed* (by R) to be credible is believed by R. This will not make any difference in the cases where credibility is determined by the structure of the game, but will matter for some potentially ambiguous cases.

Example 4 is an illustration of a situation in which ignorance or error about credibility may arise[14]

	a1	a2	a3
t1	9 0	10 10	0 9
t2	9 0	0 9	10 10

If “S is t1" is credible, and if S believes that R believes that it is credible, then S will definitely send this message, if she is of type t1. But then it will be true, and believed by R to be true, that the alternative message, “S is t2" is *not* pf rational for t1, and this implies that it will also be credible (and believed by R to be credible). Under these assumptions, each message will be sent and believed if and only if it is true; communication will succeed. But the first message might not be credible, since if there is a significant chance that R will believe the first message, but not the second, then S will prefer to send the first message, and to have it believed, even if she is of type t2. In this case, neither message will be credible. Or it might happen that even though the messages are in fact credible, neither message is believed by R to be credible. The credibility of the messages is determined by the pattern of S’s beliefs, and the perceived credibility of the messages is determined by R’s beliefs about the pattern of S’s beliefs; in the case of this game, both are constrained, but not determined, by the structure of the game and the rationality and credibility constraints on the models. S always knows whether a message is credible, since she always knows her own beliefs and utilities, but in cases where R may be mistaken or uncertain about whether a message is credible, S may be unsure whether a credible message will in fact be believed, since she may be unsure whether R realizes that the message is credible. So she may

[14] This example is used by Matthew Rabin (1990) to illustrate the interdependence of the credibility of different messages.

be unsure what effect a given message would have, if sent, and her beliefs about this will effect the *actual* credibility of this message and of others. To take account of S's potential uncertainly about the effect of her messages, we need, in the definition of the pf rationality of a message, to compare S's *expected value* of the hypothesis that the message is sent, and believed, with the *expected value* of sending alternative messages. Here is our final[15] definition:

> A message *m* for S of type t is *prima facie rational* if and only if the expected value, for S, of sending message *m*, and having it believed, is at least as great as the expected value of sending any alternative message.

Credibility is defined as before, and the credibility constraint should be as follows:

> It is common belief that the content of any message that is sent and that is believed by R to be credible is believed by R.

We can then define the class of game models that satisfy this constraint (in the actual world of the model), along with the constraint that there is common belief among the players that both players choose rationally, and the sets of strategies for the players that are played in some model in the class defined.[16]

The simple sender-receiver games are intended to isolate pure communicative acts: to separate them from the complexities of more general strategic contexts. But ultimately, our interest is in the way that communication works in a wider setting, and with the way communicative acts interact with each other and with other kinds of rational decisions. I think this account of credibility can be generalized in a number of ways: first, the definitions apply straightforwardly to cases where the private information available to the sender concerns, not exogenous information determined by chance or nature, but information about other choices that the sender will make, before or after the message is sent. Any game, for example, might be preceded by a cheap talk move in which one player has the opportunity to announce her strategy for the rest of the game. Second, we can consider sequences of communicative moves by different players. There are simple sender-receiver games in which credible communication is not assured, but in which it would be assured if R had the opportunity to send a message to S prior to S's message to R. (Informing her, credibly, that she has the beliefs that are required for credible communication.) Third, in games with more than two players, there may be broadcast messages that must go to many players at once, so that the credibility of the message depends on the effect it will have on players with different interests and different powers. Fourth, in a more general setting, there may be cases where it is uncertain whether or not a sender has certain information; in such cases, credibility requires not just confidence that the sender wants the receiver to know the truth, but also confidence that she knows the truth about the content of the message she is sending.

[15] By "final" I mean final to be offered in this paper. Further refinement will probably be required.

[16] In a future more formal paper, I will spell the theory out more precisely, and explore some of the consequences of this account of credible communication.

In the simple theory, we make no assumptions about the effect of messages that are manifestly not credible, but such messages may have consequences that a more general theory should consider. They do give rise to the question, on the part of the receiver, “why did she say *that*, given it is obvious to both of us that it is not credible?” We considered one very contrived case (example 3) where a literally incredible message managed to convey a meaning. One may hope that future developments in a more general setting will help to explain the role that the content of what is said may play even when it diverges from what is meant.

I am going to conclude with an example that, while it is still a simple sender-receiver game, does gesture toward the kind of phenomenon that is illustrated by our opening story, and at some of the strategic complexities that might arise in a wider context.

	a1	a2	a3	a4	a5
t1	9 -5	0 -5	8 5	3 0	6 0
t2	0 -5	9 -5	3 0	8 5	6 0

Let’s assume that S is actually of type t1. Is there any message that she might like to send? Ideally, S would like to get R to choose a3, yielding a payoff of 5 rather than 0, which is what she would get if she did nothing to change R’s prior 50/50 beliefs. If she could somehow get R to have a degree of belief of about 2/3, rather than 1/2, in the hypothesis that she is of type t1, then he would make this choice. But what might S say to accomplish this? She might try revealing some, but not all, of the evidence that she is of type t1, or she might say something that *could* be taken to be evidence for this, but that might mean something else. She might say something that R already knows to be true, but that might give some support, but only a little, to the conjecture that S said it because she is of type t1. But given the disastrous consequences for S of R fully believing that she is of type t1 (in which case he would choose a1, giving S a payoff of -5), and given that it is common knowledge that S knows whether she is of type t1 or type t2, S would be "playing with fire" if she made such an attempt, since it might get "imbued with deep meaning, whether she wants it to or not."

In a game this simple, with the knowledge and motives of the participants assumed to be transparent, there is nothing that S can do. She had best remain silent and accept her payoff of zero in order to avoid something worse. But in a richer setting where there is perhaps some doubt about what she knows, or about exactly what her motivations are, or about what her messages say and mean, she might try to achieve more, at least if she is “skilled in the delicate language of dollar policy.” She will never be able to succeed by meaning what she wants to convey transparently and openly in the way that Grice’s analysis of meaning was trying to capture, but if she succeeds at all by sending a cheap talk message, then she will do so by exploiting communicative devices that are to be understood in terms of their intended role in this kind of communicative practice.

References

Battigalli, P. and G. Bonanno (1999) “Recent results on belief, knowledge and the epistemic foundations of game theory,” *Research in Economics*, **53**.

Crawford, V. and J. Sobel (1982) “Strategic Information Transmission,” *Econometrica* **50**, 1431-1451.

Grice, P. (1989) *Studies in the Way of Words*. Harvard University Press, Cambridge, MA.

Farrell, J. (1993) "Meaning and credibility in cheap talk games," *Games and Economic Behavior,* **4**, 514-31.

Farrell, J. and M. Rabin (1996) "Cheap talk," *Journal of Economic Perspectives*, **10**, 103-118.

Lewis, D. (1969) *Convention*. Harvard University Press, Cambridge, MA.

Parikh, P. (2001) *The Uses of Language*. CSLI Publications, Stanford, CA.

Rabin, M. (1990) "Communication between rational agents," *Journal of Economic Theory,* **51**, 144-70.

Stalnaker, R. (1984) *Inquiry*. MIT Press, Cambridge, MA.

Stalnaker, R. (1997) "On the Evaluation of Solution Concepts," *Epistemic Logic and the Theory of Games and Decisions*, ed. by M. O. L. Bacharach, L.-A. Gérard-Varet, P. Mongin and H. S. Shin. Kluwer Academic Publisher, 345-64.

van Rooy, R. (2003) “Quality and quantity of information exchange,” *Journal of Logic, Language and Information*, **12**, 423-451.