Lying and Deception in Games

Joel Sobel

University of California, San Diego

This article proposes definitions of lying, deception, and damage in strategic settings. Lying depends on the existence of accepted meanings for messages but does not require a model of how the audience responds to messages. Deception does require a model of how the audience interprets messages but does not directly refer to consequences. Damage requires consideration of the consequences of messages. Lies need not be deceptive. Deception does not require lying. Lying and deception are compatible with equilibrium. I give conditions under which deception must be damaging.

I. Introduction

Communication is an essential part of social interaction. In order to investigate when and why goal-oriented actors communicate effectively, it is useful to clarify basic terms. This paper proposes definitions of lying and deception and derives properties of these definitions in a simple strategic context. My approach is abstract, but the ideas are substantively relevant and topical. Many issues in policy and law involve attempts to limit how people share information. Consumer agencies make and enforce regulations designed to protect agents from deceptive practices. Existing policies lack a consistent framework. I hope that this paper helps to provide

Many colleagues and friends have helped me to develop me ideas on this topic. I thank Andreas Blume, Vincent Crawford, Uri Gneezy, Peicong Hu, Navin Kartik, Kristof Madarász, Daniel Martin, Philip Neary, Janis Pappalardo, Gregory Raiffa, Karl Schlag, Andrew Stivers, Kit Woolard, and loyal participants in an elective course at the University of California, San Diego. Referees and an editor read the paper carefully. They made constructive comments that improved the paper. I am grateful to the National Science Foundation for research funding.

Electronically published February 6, 2020
[*Journal of Political Economy*, 2020, vol. 128, no. 3]

a foundation for future discussions of policies that identify and limit dishonest and deceptive practices.

Loosely, a lie is a statement that the speaker believes is false. This definition requires an accepted interpretation of the meaning of words, but it does not require a model of the speaker's intentions or the expected consequences of the statement. An implication of this definition is that one does not need to know how the audience will interpret a statement in order to evaluate whether the statement is a lie. People do make statements to influence others. I reserve the term "deception" to describe statements—or actions—that induce the audience to have incorrect beliefs. Making these ideas precise requires a formal model, which I introduce in section II. The basic model makes strong assumptions but is general enough for most of the analysis in the paper. In section VIII, I discuss ways to extend the model.

I use the model to develop definitions that parallel a classification introduced by Austin (1975). Austin distinguishes between three different properties of speech acts: "locution" is what the speaker says; "illocution" refers to the interpretation of what she says; and "perlocution" refers to the consequences of the statement.

I define lies purely in terms of locution. Section III uses the model to discuss lying. To conduct the analysis, I add the concept of a common language to a two-player (sender-receiver) model of communication. After I present a definition of lying in section III.A, section III.B briefly reviews the extensive literature that offers definitions of lying and the extent to which my definition conforms with usage in other disciplines. Sections III.C and III.D discuss properties of lying in strategic settings. I identify games in which lying must arise in equilibrium and other situations in which it need not arise. Two central observations are that fully anticipated lies need not interfere with the exchange of information and that the ability to lie may have both positive and negative welfare consequences. These observations are immediate consequences of the definitions. My contribution is to place them in a coherent framework.

Unlike lying, deception does require a theory of mind. The speaker must have a model of how the audience interprets her behavior. I assume that the speaker has beliefs about how the audience will interpret her actions. Deception is therefore an illocutionary act. When the sender contemplates an action, she can figure out how the action influences the receiver. Section IV discusses beliefs. This section begins with a discussion of beliefs and provides a definition of deception. Informally, the sender's behavior is deceptive if (according to her model of the receiver's behavior) it leads the receiver to have inferior beliefs about the state of the world. To make this informal notion formal, I must describe what it means for beliefs to be inferior. In my definition, inducing inferior beliefs means sending a message that leads to beliefs that are farther from the truth than

beliefs that could be induced by another feasible message. Section IV.A gives a precise definition of what it means to be farther from the truth. There are different ways to formalize the notion of inaccurate beliefs and poor decisions. I discuss alternatives in section V.D after I consider the consequences of deception.

I discuss properties of deception in section IV.B. Lies and deception are different. Deception is possible without lying. Lies need not deceive. Deception is possible only if the receiver is open to influence. If the receiver ignores the sender's message, then he cannot be deceived. Deception is not inevitable. The sender can always play in a nondeceptive way. Although deception is not possible in equilibrium in a perfect-information game, it can arise in equilibrium if the sender has information that is not available to the receiver. Even if the receiver has accurate beliefs, he can be deceived (according to my definition) if the sender manipulates her superior information to induce the receiver to have inaccurate beliefs. In particular, mixed-strategy equilibria of zero-sum games in which the receiver observes the action of a privately informed sender are typically deceptive.

Although deception does not directly invoke the preferences of the players, economic analysis requires a discussion of the consequences of behavior. Section V discusses the consequences of behavior from the perspective of the receiver. Hence it discusses what Austin would call perlocutionary acts. I introduce the notion of damage. An action is damaging if the sender has another action available that induces the receiver to make better decisions. In section V.C, I make a connection between deception and damage, providing conditions under which deception must be damaging and damage must be the result of deception.

Section VI also discusses consequences but from the perspective of the sender. I consider bluffs, which are actions or messages that deceive in order to benefit the sender. I discuss the relationship between lies, deception, and bluffs and identify situations in which bluffing arises.

Section VII discusses five models from the literature that illustrate the main ideas. The appendix contains proofs of results (when needed).

"Lying" and "deception" are common English words. The concepts discussed in this paper are important for the study of strategic communication. I hope that the names I attach to these concepts are broadly consistent with common usage.

II. A Basic Model

This section describes a basic framework. There are two agents, an informed sender, S, and an uninformed receiver, R. Nature draws the state of the world, θ, from a set Θ, according to a distribution P. The sender observes θ, sends a message $m \in M$, and takes an action $x \in X$. The receiver observes m (but not x) and makes a decision $y \in Y$. I define $U^i(\theta, m, x, y)$

as the payoff that player i receives given state $\theta \in \Theta$, message $m \in M$, and actions $(x, y) \in X \times Y$. I assume that $U^R(\cdot)$ does not depend directly on m.

Unless I say otherwise, assume that Θ, M, X, and Y are finite. Here $P(\theta)$ gives the probability that the state is θ, while $P(\cdot)$ is positive and satisfies $\Sigma_\theta P(\theta) = 1$. These strong assumptions lead to a restrictive model. Section VIII discusses more general models.

Sometimes I study equilibrium behavior. An equilibrium is a perfect Bayesian equilibrium: a configuration, (m^*, x^*, y^*, μ^*), where (m^*, x^*) is a strategy for the sender, y^* is a strategy for the receiver, and μ^* is a belief function that assigns to each m a probability distribution over Θ, such that the sender's strategy is a best response to the receiver's strategy; the receiver's strategy is a best response to the beliefs μ^*; and μ^* is derived from the sender's strategy using Bayes's rule whenever possible.[1]

The definition of lying requires a comparison between the sender's message m and the true state of the world θ. For each $\Theta_0 \subset \Theta$, there exists a message $m_{\Theta_0} \in M$, and there is a common understanding that m_{Θ_0} means $\theta \in \Theta_0$. Let m_{θ_0} denote the message corresponding to the set $\{\theta_0\}$. To simplify discussions, assume that there is exactly one way to describe each subset.[2] In particular, if $\Theta_0 \neq \Theta_0'$, then $m_{\Theta_0} \neq m_{\Theta_0'}$. When the message m is equal to m_{Θ_0} for some Θ_0, I say that m has an accepted meaning. There may be messages that have no accepted meaning. The purpose of including messages that have accepted meanings is to make it possible to describe lies. I do not require the sender to use a message in the accepted manner or for the receiver to interpret it in the accepted way. It is possible that using messages in a way that violates their accepted interpretation is costly.

Note the distinction between sender messages and actions (m and x). Formally, they are different in two ways. First, the receiver observes m but not x. This means that R's strategy is a function of m and not x. Second, $U^R(\cdot)$ is independent of m.[3] Consequently, S's choice of x can have a direct impact on R's utility but the choice of m does not.[4] One setting in which this formulation applies is a game in which S observes the state of nature θ, chooses m as a function of θ, R hears m, and then S and R play a simultaneous-move game (in which S selects x and R selects y). Section VII contains other examples. I will refer to m as a message. When I

[1] In this definition, m^* is a function from types to messages; x^* is a function from types to sender actions; and y^* is a function from messages to receiver decisions. Throughout the paper, I denote equilibrium strategies and beliefs with an asterisk.

[2] In natural communication, this assumption is not true. There are interesting questions about what happens when there are two ways to communicate the same thing, but I do not consider these issues in this paper.

[3] I permit $U^S(\cdot)$ to depend on m.

[4] The message m will typically have an indirect impact on R's utility because it influences the choice of action, y.

discuss lying, it is important that some messages have an accepted meaning. This interpretation is not necessary in the discussion of deception. I will try to be clear when the analysis requires accepted meanings.

A special case of the model is a cheap-talk game in which X is empty and neither $U^S(\cdot)$ nor $U^R(\cdot)$ depend on m directly.[5] Several times I illustrate results using the special case of the cheap-talk model studied in Crawford and Sobel (1982). In a CS (Crawford and Sobel) cheap-talk game, $\Theta = Y = [0, 1]$; for $i = S, R$, $U^i(\theta, y)$ is continuous, strictly concave in y, and satisfies $U^i_{12} > 0$. A consequence of these assumptions is that $y^i(\theta) \equiv \arg\max U^i(\theta, y)$ is well defined. CS cheap-talk games also impose the property that $y^S(\theta) > y^R(\theta)$ for all θ; and the prior on states has a positive, continuous density. The game has a common language if M consists of all Borel subsets of [0,1] and if the accepted meaning of the message C is "θ is an element of C."

Another special case of the model is a disclosure game (Grossman 1981; Milgrom 1981) in which X is empty, $U^R(\cdot)$ does not depend on m directly, and

$$U^S(\theta, m, y) = \begin{cases} u^S(\theta, y) & \text{if there exists } \Theta_0 \subset \Theta \text{ such that } m = m_{\Theta_0} \text{ and } \theta \in \Theta_0 \\ u^S(\theta, y) - K & \text{otherwise} \end{cases},$$

where K is sufficiently large that the sender of type θ would always select a message of the form m_{Θ_0} for a set Θ_0 that contains m. This specification captures the idea that the sender can withhold information but cannot make false claims.

This formal model is sufficient to describe the basic ideas and formulate results, but it imposes strong assumptions. I discuss extensions in section VIII.

III. Lying

Section III.A defines lying. Section III.B discusses alternative definitions. Section III.C discusses lies in cheap-talk games. Section III.D identifies a situation in which honesty is compatible with equilibrium.

A. *Lying: Definitions*

I now present consistent definitions of lying that illustrate the subtleties underlying the intuitive concept. There are more ambitious and systematic

[5] When discussing cheap-talk games, I suppress the irrelevant arguments (m and x) in the description of $U^i(\cdot)$.

attempts to present definitions of lying. I do not attempt to review this large literature.[6]

Lying arises when the sender says something that she believes to be false. There are several possibilities.

DEFINITION 1 (Lying).

i. The message m is a *lie* given θ if $m = m_{\Theta_0}$ and $\theta \notin \Theta_0$.
ii. The message m is *true* given θ if $m = m_{\Theta_0}$ and $\theta \in \Theta_0$.
iii. The message m is *the whole truth* given θ if $m = m_\theta$.

The definition of lying does not depend on the preferences of either player. Whether a statement is a lie depends on the relationship between the statement, the accepted meaning, and (what is believed to be) the truth. Ultimately, preferences matter. Lies may benefit or hurt either player. It may be intrinsically costly to lie, and this cost may be linked to how the lie influences the behavior of others. This section discusses the definition of lies. Subsequent sections discuss the consequences of lies and the circumstances in which lies will arise.

The use of the message m_{Θ_0} when θ is the state and Θ_0 contains both θ and another state is not as precise as possible. Such a statement could be called a lie of omission, but in this paper I will say it is true but not the whole truth.

The definition describes truth and lies in terms of the message sent. I extend the definition to talk about strategies being honest or dishonest.

DEFINITION 2 (Honest strategies).

i. The strategy $m^*: \Theta \to M$ is honest if $m^*(\theta)$ is true for all θ.
ii. The strategy $m^*(\cdot)$ is completely honest if $m^*(\theta)$ is the whole truth for all θ.
iii. The strategy $m^*(\cdot)$ is dishonest if there exists θ such that $m^*(\theta)$ is a lie.

Every message that has an accepted meaning either is a lie or is true. The language may include messages that do not have an accepted meaning. Messages that have no accepted meaning are neither lies nor true.

According to this definition, impossible statements ("I support you 1,000%") are lies when they have meaning. More formally, what constitutes a lie may depend on the specification of the state space. For example,

[6] Mahon's (2008) article in the *Stanford Encyclopedia of Philosophy* carefully examines the strengths and weaknesses of several possible definitions of lying. I share the reaction of Morris (2009), who writes that the encyclopedia treatment might "make the overwhelmed reader wonder whether lies and lying have any coherent meaning at all." Morris, like me, pursues the more modest goal of trying to clarify the definition of lying without exploring many aspects important to philosophers.

suppose that θ reflects the quality of something; for concreteness, let it describe the score on an examination. It may be known that the score is a nonnegative integer between 0 and 50 so that $P(\theta) = 0$ for $\theta \notin \{0, 1, \ldots, 50\}$. Is it a lie to report a score of 51? It would be if there exists Θ_0 such that $m_{\Theta_0} = 51$ but not otherwise.

By concentrating on the sender's beliefs, the definition of lying ignores the relationship between objective truth and the statement. This limitation may be important when one tries to enforce laws that sanction lying. In those situations, one would want to define truth in terms of an objective standard rather than the sender's beliefs. The distinction arises only when one permits the sender to have statistically inaccurate beliefs.

The literature on strategic communication informally uses lying in a way that is consistent with my definition. These papers often assume $M = \Theta$ and view the commonly accepted meaning of a message as the message itself ($m = m_\theta$). Theoretical papers consider perturbed versions of communication games in which, with positive probability, the sender is a behavioral type who always reports honestly; the receiver is a behavioral type who interprets messages literally (believing that the state is m after receiving the message m); or perturbing preferences include lying costs, which are defined in terms of the difference between the true state and the message.[7] Gneezy (2005) and Fischbacher and Föllmi-Heusi (2008) are examples of experimental papers on communication that associate the message with the state and treat messages as lies if they are not equal to the state.

B. Other Definitions

In this subsection, I briefly discuss alternative definitions of lies and point out empirical support for my definition.

St. Augustine (1887, 469) lists eight types of lies. He focuses on consequences when he discusses the morality of lying and, in particular, tries to determine whether it is ever acceptable to tell a lie.[8] I agree that consequences are important, but it is useful to have a taxonomy that evaluates the veracity of statements without investigating the interpretation of the statements.

Other philosophers follow St. Augustine's tradition and include consequences in their definition of lying. In an article intended for a general audience, Lynch (2016) writes, "to lie is to deliberately say what you

[7] Chen (2011) studies cheap-talk models with behavioral types. Kartik (2009) studies models with costly lies.

[8] Augustine appears to say that it is never right to lie, but he is also flexible: dishonest statements said in jest would not constitute a lie according to his definition but would be lies according to my taxonomy. Under my definition, ironic statements are typically lies. Presumably, Augustine would not view these statements as lies.

believe to be false with the *intention* of deceiving your audience." Frankfurt (2005) appears to use the term in a similar way. Unlike my definition, this definition demands that the sender have beliefs about how the receiver reacts to communication. My use of the term permits lies even without reference to beliefs about others. Consequences are important and require attention. I find it valuable to separate this discussion and discuss the implications of communication on welfare in sections V and VI.

Coleman and Kay (1981) argue that one evaluates whether a statement is a lie by assessing the extent to which it satisfies three criteria:

1. The statement is false.
2. The speaker believes the statement to be false.
3. The intention of the speaker is to deceive.

These three criteria separate statements into eight possible categories (ranging from a true statement known to be true that is uttered with no intention to deceive to a false statement known to be false that is uttered with intention to deceive). Coleman and Kay construct eight stories, one for each category, and ask subjects to rate them.[9] They find that each of the criteria contributes to whether a statement is classified as a lie, but the second criterion—whether the speaker believes the statement to be true—is the most important characteristic, in the sense that the four stories most frequently classified as lies were the ones in which the speaker believes the statement to be false. This suggests that my taxonomy captures an important aspect of lying. It is consistent with the spirit of my definitions to include items from the first category (which would be the same as the second category when the sender's beliefs are correct). Intentions, however, do not play a role in my definition. In order to be a lie in the sense of Frankfurt (2005) and Lynch (2016), a statement must satisfy criteria 2 and 3.

C. Lies in Cheap-Talk Games

This subsection investigates lying in cheap-talk games with a common language. The following properties are simple consequences of facts about equilibria of cheap-talk games.

Remark 1. In a cheap-talk game with a common language, there always exist equilibria involving lies.

When talk is cheap, the existence of a common language does not mean that rational agents will use the language in the conventional way in equilibrium. Remark 1 notes this. Cheap-talk games always have a babbling equilibrium in which all senders send the same message with probability

[9] Subjects reported whether they thought the statement was a lie and how confident they were that others would agree.

one, and the receiver's response does not depend on the message. In equilibrium, if senders use a message that has an accepted meaning of the form "the state is θ_0," then all but one type of sender tells a lie.

An equilibrium is nontrivial if the receiver takes more than one action with positive probability in equilibrium. When an equilibrium is nontrivial, it is possible to relabel messages so that every sender type is lying. To establish this result, I use the following observation:

LEMMA 1. If (m^*, y^*, μ^*) is an equilibrium of a cheap-talk game and $\pi : M \to M$ is a bijection, then $(\tilde{m}^*, \tilde{y}^*, \tilde{\mu}^*)$ is also an equilibrium where $\tilde{m}^*(\theta) = \pi(m^*(\theta))$, $\tilde{y}^*(\pi(m)) \equiv y^*(m)$, and $\tilde{\mu}^*(\cdot \mid n) = \mu^*(\cdot \mid \pi^{-1}(n))$.

Lemma 1 states that the interpretation of messages is arbitrary in an equilibrium of a cheap-talk game. Given an equilibrium, one obtains another equilibrium by relabeling the messages and modifying the action rule so that it respects the relabeling. It is straightforward to verify that $(\tilde{m}^*, \tilde{y}^*, \tilde{\mu}^*)$ is an equilibrium if and only if (m^*, y^*, μ^*) is an equilibrium.

PROPOSITION 1. In a cheap-talk game with a common language, any nontrivial equilibrium type-action distribution can be generated by an equilibrium in which the sender's strategy is dishonest.[10]

If all messages are the whole truth, then the equilibrium must be fully separating (each type θ sends the message m_θ). Hence when separating equilibria do not exist, some agents fail to tell the whole truth.

REMARK 2. In every equilibrium of a CS cheap-talk game, there is a positive probability that the sender does not tell the whole truth.

Remark 2 is not true for all cheap-talk games. In particular, it is not true when the sender and receiver have identical preferences. In CS cheap-talk games, a fully revealing equilibrium does not exist, and hence complete honesty is not possible.

REMARK 3. In a CS cheap-talk game with a common language, every equilibrium type-action distribution can be supported as an equilibrium in which the sender tells the truth.

Remark 3 states that any equilibrium can be interpreted as one in which the sender reports honestly but incompletely. This corresponds to a standard interpretation of the partition equilibria of cheap-talk games. In any equilibrium, one uses lemma 1 to relabel messages so that types in the interval $[\theta_i, \theta_{i+1}]$ report "my type is $[\theta_i, \theta_{i+1}]$." Remark 3 holds for any pure-strategy equilibrium of a cheap-talk game. The result holds for general cheap-talk games provided that the common language is large enough to include descriptions of all probability distributions over types.[11]

[10] Let σ be a mixed strategy for S and α a mixed strategy for R. The type-action distribution generated by a mixed-strategy profile (σ, α) is a probability distribution over (θ, y) pairs where $\alpha(y|m)\sigma(m|\Theta)$ is the density of (θ, y).

[11] When types are not continuously distributed, equilibrium behavior could involve the sender using mixed strategies. In this situation, one needs the entire strategy of the sender to compute the posterior distribution of θ given m.

There is a tension between remark 1 and remark 3 because the former states that lying may be a part of equilibrium, while the latter states that lying need not be a part of equilibrium. The tension arises because cheap-talk games have multiple equilibria. The existence of a common language is not sufficient to guarantee that agents coordinate on the common language when they play a game. Coordination requires three things: that agents have common beliefs about the mapping between language and states (which follows from the existence of a common language); that agents have accurate expectations about the strategy choices of their opponent (which follows from the assumption of equilibrium); and that agents actually use messages in the commonly accepted way (which is consistent with the first two conditions—but not implied by the first two conditions—in cheap-talk games).

The next example is an illustration of this.

EXAMPLE 1 (Lies in a cheap-talk game). Consider a binary cheap-talk game with two messages (A and B), two states (0 and 1), two actions, and an equilibrium in which the sender always sends B when $\theta = 1$ and randomizes between A and B when $\theta = 0$. It is straightforward to identify preferences under which such an equilibrium exists. Suppose that the accepted meaning of A is "the state is 0" and the accepted meaning of B is "the state is 1." In equilibrium, the sender lies (with positive probability) when $\theta = 0$.

Some refinement arguments not only select type-action distributions but make restrictions on the relationship between messages and actions in equilibrium. These restrictions may not be consistent with the accepted meaning of words. For example, Gordon et al. (2020) study a CS cheap-talk game in which M is linearly ordered and in which players are restricted to monotonic strategies.[12] The monotonicity restriction does not eliminate any equilibrium type-action distributions. Nevertheless, Gordon et al. (2020) show that the only equilibrium that survives iterative deletion of weakly dominated strategies uses the highest N^* messages (where N^* is the maximum number of actions induced in any equilibrium). If messages are identified with types in the standard way, agents are systematically dishonest in this outcome: they use messages that exaggerate their type. Identifying messages with types in this way does provide structure for the way that players use and interpret messages in equilibrium. In equilibrium, no one is fooled, but for strategic reasons, no agent reports honestly.

Although it is generally possible for there to be lying in an equilibrium of cheap-talk games with a common language, the receiver always prefers

[12] Monotonicity means that if $\theta' > \theta$, then the sender must send a weakly higher message, and if $m' > m$, then the receiver must take a weakly higher action.

the sender to tell the whole truth and, ex ante (at least for the class of CS cheap-talk games satisfying a monotonicity condition studied in Crawford and Sobel 1982), the sender obtains a higher expected payoff when the receiver is fully informed than in any equilibrium.

D. *Completely Honest Equilibria*

In this subsection, I go beyond cheap-talk games and investigate the possibility of completely honest behavior in equilibrium. On the one hand, if messages are costly, there is no reason to expect agents to be honest. A blunt example is a situation in which there are two states and two messages and it is extremely costly for the sender to use m_θ (the message with accepted meaning θ) when the state is θ. On the other hand, if there is no conflict of interest between the players, there is no strategic reason that rules out the possibility that the sender will report the complete truth in equilibrium.

Given θ, let $y^R(\theta, m, x)$ solve $\max U^R(\theta, y, m, x)$, and let $(\underline{m}(\theta), \underline{x}(\theta))$ solve

$$\max_{m,x} U^S(\theta, m, x, y^R(\theta, m, x)). \tag{1}$$

Hence $(\underline{m}(\theta), \underline{x}(\theta))$ is the set of messages that maximize S's utility, assuming that the messages are revealing and R responds with his optimal action. Assuming that (1) has a unique solution, it is immediate that an equilibrium in which the sender tells the whole truth exists only if $\underline{m}(\theta) \neq \underline{m}(\theta')$ whenever $\theta \neq \theta'$.[13] When $\underline{m}(\cdot)$ is single valued and one-to-one, the game is potentially revealing. It is straightforward to provide conditions on $U^i(\cdot)$ under which a game is potentially revealing.[14]

The next result is an immediate consequence.

REMARK 4. In any potentially revealing game, there exists a specification of language under which there exists an equilibrium in which the sender tells the whole truth with probability one.

Remark 4 gives strong conditions under which honesty is compatible with strategic behavior. In a potentially revealing game with common interests, an honest equilibrium exists if the meaning of $\underline{m}(\theta)$ is "my type is θ."[15] Since $\underline{m}(\theta)$ is determined exclusively from the game while the meaning of the message $\underline{m}(\theta)$ may come from external considerations, it is not

[13] If the solution to (1) is not unique, then the necessary condition is that there exists a selection from the solution correspondence that is one-to-one.

[14] For example, if y, m, and θ are all elements of the unit interval; X is empty; $U^R(\theta, m, y) = -(y - \theta)^2$; $U^S(\theta, m, y) = f(y) + g(y, \theta)$ for differentiable f and twice differentiable g; and $g_{12}(\cdot) > 0$, then the condition holds.

[15] A communication game has common interests if $U^S(\theta, m, x, y) = U^R(\theta, m, y)$ for all x, y, and m; in particular, $U^S(\cdot)$ is independent of x.

necessary for an honest equilibrium to exist even in potentially revealing common-interest games.

Remark 4 requires that the common language be defined in terms of the equilibrium (rather than being specified exogenously). This observation makes the underlying idea behind the existence of truthful equilibria transparent: there exists an equilibrium in which the sender tells the whole truth with probability one if and only if the game has a separating equilibrium. In general, when language is specified independent of the strategic situation, lying may arise in a separating equilibrium. This would happen, as I discuss in section VII, if in equilibrium the sender exaggerates but the receiver accurately takes the possibility for exaggeration into account.

If the language contains a word for each subset of Θ, there is a specification of language in which the sender tells the truth.[16] The specification of language depends on the equilibrium strategies; that is, its conclusion requires that language be defined relative to the equilibrium. If the language is specified independent of the payoffs (which seems natural), then there is no guarantee that there exists an equilibrium in which the sender tells the whole truth with probability one.

Equilibria in which the sender tells the whole truth with probability one need not be efficient (even when S and R have identical preferences). Efficient equilibria need not be honest. In order to create a connection between honesty and efficiency, a reasonable approach would be to assume that lying is costly.

IV. Deception

The definition of lying depends only on the existence of accepted meanings of words. It makes no reference to how S's statements might influence R. To describe these features, I need to investigate how S's behavior influences R's beliefs. This section studies properties of beliefs. Section IV.A provides definitions of central concepts: accuracy and deception. Section IV.B identifies significant properties of deception. Section IV.C discusses deception in cheap-talk games. Section IV.D compares lying and deception.

A. Accuracy and Deception

Assume that each m induces a posterior distribution $\mu(\theta \mid m)$, where $\mu(\cdot \mid m)$ is the posterior belief of the receiver given the message m. In the last two parts of the definition, R forms beliefs by taking into account the

[16] For a mixed-strategy equilibrium, it suffices to have a word for each probability distribution over Θ.

sender's mixed strategy, $\sigma(\cdot)$; $\sigma(m \mid \theta)$ is the probability that the sender with type θ sends the message m. Hence $\sigma(m \mid \theta) \geq 0$ and $\Sigma_{m \in M} \sigma(m \mid \theta) = 1$ for all θ.

DEFINITION 3 (Properties of beliefs).

i. The belief $\mu(\cdot \mid m)$ is *completely inaccurate* given θ if $\mu(\theta \mid m) = 0$.
ii. The belief $\mu(\cdot \mid m)$ is *inaccurate* given θ if $\mu(\theta \mid m) \in [0, 1)$.
iii. The belief $\mu(\cdot \mid m)$ is *accurate* given θ if $\mu(\theta \mid m) = 1$.
iv. The belief $\mu(\cdot \mid m)$ is *rational* given θ and $\sigma(\cdot)$ if

$$\mu(\theta \mid m) = \frac{\sigma(m \mid \theta)P(\theta)}{\sum_{\theta'} \sigma(m \mid \theta')P(\theta')} \tag{2}$$

whenever

$$\sum_{\theta'} \sigma(m \mid \theta')P(\theta') > 0. \tag{3}$$

v. The belief $\mu(\cdot)$ is *rational* given $\sigma(\cdot)$ if equation (2) holds for all m and θ whenever inequality (3) holds.

Beliefs are accurate if they reflect the sender's information and inaccurate otherwise. Completely inaccurate beliefs place zero probability on the true state. Rational beliefs (which are called consistent beliefs in some contexts) are statistically correct given the description of the game (the prior P and the information) and the sender's strategy.

In a strategic setting, the receiver's beliefs are derived from the description of the game (in particular, the prior distribution) and the sender's behavior. In equilibrium, the receiver accurately processes information in the sender's strategy. At this point, I do not want to restrict the receiver to equilibrium behavior. In fact, one can relate properties of beliefs to lying if the receiver believes everything he hears. I maintain the assumption that the game has a language in which for each subset of states $\Theta_0 \subset \Theta$, there exists $m_{\Theta_0} \in M$ that has the meaning "$\theta \in \Theta_0$."

Informally, deception is a deliberate attempt by the sender to induce incorrect beliefs. The notions of inaccurate beliefs are poorly suited for discussing deception because they would lead to calling too many things or too few things deception. On the one hand, unless the receiver is extremely naive or uses a misspecified model, his beliefs will not be completely inaccurate; that is, the sender would rarely have the opportunity to induce completely inaccurate beliefs, and a definition of deception based on this idea would be too narrow. On the other hand, in interesting settings, beliefs are inaccurate. Classifying any message that leads to inaccurate beliefs as deceptive is likely to be too broad. Consequently, I propose another way in which beliefs can be incorrect.

In the following definition, $I(\cdot \mid \theta)$ is the probability distribution that places probability one on θ; that is,

$$I(\theta' \mid \theta) = \begin{cases} 0 & \text{if } \theta' \neq \theta \\ 1 & \text{if } \theta' = \theta \end{cases}.$$

DEFINITION 4 (Deception). Let μ be a probability distribution on Θ.

i. The message m is *deceptive* given θ and μ if there exist a message n such that $\mu(\theta \mid n) > 0$, a number $p \in [0, 1)$, and a distribution ρ satisfying $\rho(\theta) = 0$ such that

$$\mu(\cdot \mid m) = p\mu(\cdot \mid n) + (1 - p)\rho. \tag{4}$$

ii. The message m is *strongly deceptive* given θ and μ if there exist n such that $\mu(\cdot \mid m) \neq \mu(\cdot \mid n)$ and $p \in [0, 1)$ such that

$$\mu(\cdot \mid n) = p\mu(\cdot \mid m) + (1 - p)I(\cdot \mid \theta). \tag{5}$$

The message m is deceptive if there is another message n that leads to different beliefs and if the beliefs induced by m are farther from the truth in the sense that the beliefs induced by m are a mixture of the beliefs given n and some completely inaccurate beliefs. It is immediate that if m is accurate, then it cannot be deceptive, and if m is completely inaccurate, then it must be deceptive, provided that there is some message n that is not completely inaccurate given θ. The restriction that $\mu(\theta \mid n) > 0$ rules out a trivial case. If there were no such message, then it is impossible to convince the receiver that θ is possible, and the receiver will have completely inaccurate beliefs independent of the sender's message. I do not wish to classify any message as deceptive in this case. Notice that if m is deceptive given θ and μ, $\mu(\cdot \mid m) \neq \mu(\cdot \mid n)$ because $p \neq 1$ in (4) and $\mu(\theta \mid n) > 0$.

The definition permits two kinds of deception that do not have adverse consequences for others. The framework allows the possibility of self-deception. In such a situation, I interpret the receiver not as a second player but as a future version of the sender. Why would the sender want to deceive herself? Incentives for self-deception may arise when the present and future versions of the sender have different preferences (as in dual self-models or when there is time inconsistency). Also, there are situations in which S induces inaccurate beliefs in a way that benefits the receiver. The sender's behavior could protect the receiver from painful information (perhaps this happens when a doctor refuses to reveal details of a diagnosis of a terminal illness) or out of paternalism (the sender deceives the receiver in order to prevent the receiver from taking a self-destructive action).

Merely having inaccurate beliefs is not necessarily a sign of deception. Suppose, for example, that $\mu(\cdot \mid m)$ does not depend on m. In this case,

S's message does not influence R's beliefs.[17] Consequently, although R's beliefs may be inaccurate, S is not responsible for the inaccuracy. Hence a necessary condition for deception is that R responds differently to different messages.

Two related aspects of the definition deserve comment. First, deception does not give a privileged position to the prior distribution; that is, one evaluates whether a message is deceptive by comparing the beliefs it induces to beliefs that could have been induced with an alternative message. Second, even when there is a message that corresponds to silence, the definition is agnostic about whether this message is deceptive. I can include these features in the model if I assume that there is a distinguished message $\tilde{m}$ that is interpreted as silence and assume that the receiver does not update given silence ($\mu(\cdot \mid \tilde{m}) = P(\cdot)$). Under these assumptions, definition 4 would say that m is deceptive given θ if $\mu(\cdot|m)$ is farther from θ than the prior and that silence ($\tilde{m}$) is deceptive if the sender could induce beliefs that are closer to the true state than the prior. Of course, in general, $\mu(\cdot \mid \tilde{m}) \neq P(\cdot)$.

The next result provides an alternative characterization of deception.

PROPOSITION 2. The message m is deceptive given θ and μ if and only if there exists a message n such that $\mu(\theta \mid n) > 0$ and $\mu(\cdot \mid m) \neq \mu(\cdot \mid n)$ such that

$$\frac{\mu(\theta \mid m)}{\mu(\theta \mid n)} = \min_{\theta' \in \Theta, \mu(\theta' \mid n) \neq 0} \frac{\mu(\theta' \mid m)}{\mu(\theta' \mid n)}. \tag{6}$$

Proposition 2 is a direct consequence of (4). It states that a message m is deceptive if there is another message n such that the probability of the true state relative to any other state is smaller under beliefs induced by m than by beliefs induced by n; that is,

$$\frac{\mu(\theta \mid m)}{\mu(\theta' \mid m)} < \frac{\mu(\theta \mid n)}{\mu(\theta' \mid n)}.$$

A message m is strongly deceptive if there is an alternative message that leads the receiver to have beliefs that are closer to what the sender believes. In the definition, "closer" refers to a segment connecting S's beliefs to the beliefs induced by m. In general, this definition is quite restrictive. When there are two states, the set of possible beliefs is one-dimensional, and if m and n induce different beliefs, then one of the beliefs will be closer to S's beliefs than the other. When there are more states, however, strongly deceptive beliefs are unusual, as it is unlikely for three beliefs to be colinear. Any strongly deceptive message is deceptive. Conversely, deceptive

[17] It is possible that S's message influences R's action, but if R responds optimally to beliefs, the message will not influence R's utility.

messages are strongly deceptive when there are only two states. In general, there exist deceptive messages that are not strongly deceptive.

EXAMPLE 2 (Strongly deceptive messages). Consider a binary model like example 1 (talk need not be cheap). Assume that the sender uses message B whenever $\theta = 1$ and both messages A and B with positive probability when $\theta = 0$. In this case, the rational receiver has accurate beliefs when he hears A and inaccurate beliefs when he hears B. The message B is strongly deceptive given $\theta = 0$ and equilibrium beliefs.

The definition of deceptive messages leads to a natural definition of deceptive strategies.

DEFINITION 5 (Deceptive strategies). Let μ be a probability distribution on Θ.

i. The strategy $m^* : \Theta \to M$ is deceptive given μ if there exists θ such that $m^*(\theta)$ is deceptive given θ and μ.
ii. The strategy $m^* : \Theta \to M$ is strongly deceptive given μ if there exists θ such that $m^*(\theta)$ is strongly deceptive given θ and μ.

I postpone a discussion of alternative definitions until section V.D.

B. Properties of Deception

This section describes properties of deception.

PROPOSITION 3. If the set of messages is finite, for any μ, the sender will have a strategy that is not deceptive given μ.

Proposition 3 notes that it is always possible for the sender to avoid deception.[18]

REMARK 5. If $\mu(\cdot \mid m)$ is independent of m, then no strategy of the sender is deceptive given μ.

In particular, remark 5 implies that there is no deception in a pooling equilibrium when the receiver responds to off-the-path messages with the same equilibrium action. Remark 5 is a trivial consequence of the definition of deception, but it is an important observation. Proposition 3 and remark 5 guarantee that deception is not inevitable for two reasons. First, it is always possible for the sender to avoid deceiving the receiver. Second, if the receiver ignores the sender, then he cannot be deceived. I view these properties as essential for any notion of deception.

There is another familiar context in which there is no deception in equilibrium.

REMARK 6. If (m^*, x^*, y^*, μ^*) is a separating equilibrium, then m^* is not deceptive given μ^*.

Remark 6 follows because the sender always induces accurate beliefs. Remarks 5 and 6 apply to communication games in general.

[18] This result also holds under mild regularity continuous in continuous models.

Since deception requires the sender's message to (sometimes) influence the receiver's beliefs, cheap-talk games always have a nondeceptive equilibrium: remark 5 implies that there is no deception in a babbling equilibrium.[19]

It is tempting to conjecture that deception is not possible in equilibrium. The truth of the conjecture depends, of course, on the definition of deception. If one equates deception to inducing nonrational beliefs, then deception is inconsistent with equilibrium. My definition is different. Deception, as I have defined it, is consistent with equilibrium. I illustrate this possibility with a simple example, and then, in the next subsection, I discuss deception in cheap-talk games.

EXAMPLE 3 (Deception in a disclosure game). It is deceptive to partially reveal the truth in a disclosure game when full disclosure is feasible. Such an outcome can be the only equilibrium in some situations (e.g., when it is not common knowledge that the sender is informed).

Deception is also possible in equilibrium when revealing the truth is feasible and credible but costly to the sender, for example, in a pooling equilibrium of a labor market signaling game in which the sender has a message that would reveal her to have the highest ability but the highest-type sender prefers to pool.

C. *Deception in Cheap-Talk Games*

In section III.C, I used cheap-talk games to illustrate properties of lying. In this subsection, I discuss deception in cheap-talk games. Recall that m must be linked to accepted meaning to discuss lying, but no such association is necessary for deception.

The first example shows that the conclusion of example 3 holds in cheap-talk games.

EXAMPLE 4 (Deception in cheap-talk games). Consider the binary cheap-talk game of example 1. It is strongly deceptive given $\theta = 0$ to send the message B in an equilibrium in which the sender always sends B when $\theta = 1$ and randomizes between A and B when $\theta = 0$.

Related results are possible in the cheap-talk games studied in Crawford and Sobel (1982) and in the richer environment of Morgan and Stocken (2003). Morgan and Stocken study a cheap-talk model in which the receiver is uncertain about a payoff-relevant state and about the preferences of the sender (some sender types are unbiased and have the same preferences as the receiver, while others have an upward bias). In this way, it is a mixture of a pure-coordination game and a CS game. Equilibrium type-action distributions have the same qualitative features of the equilibria

[19] In a babbling equilibrium, the receiver takes the same action independent of the sender's message.

in CS (in particular, only a finite number of actions are induced). The existence of biased agents makes the messages of the unbiased agents more credible in equilibrium, increasing opportunities to deceive in equilibrium relative to a model with only biased types.

PROPOSITION 4. Any equilibrium type-action distribution in the CS model can be generated by an equilibrium (m^*, y^*, μ^*) such that for each θ, $m^*(\theta)$ is not deceptive given θ and y^*.

Proposition 4 states that deception is not necessary in an equilibrium of a cheap-talk game. It follows because one can generate any equilibrium type-action distribution by beliefs that interpret all messages as on-path messages, and in this situation, no message can provide more accurate information about the true state than the equilibrium message. There are at least two ways in which deception is possible in the CS model. One can imagine an equilibrium in which different beliefs induce the same action. For example, imagine a pooling equilibrium in which all senders send the message m_0; the receiver's beliefs given m_0 equal the prior; the receiver's action given m_0 (best response to prior) is y_0; and given any $m \neq m_0$, R believes that $\theta = \hat{\theta}$, where R's best response to $\hat{\theta}$ is y_0.[20] In this case, it is deceptive for $\theta = \hat{\theta}$ to send the message m_0, but the deception has no consequence because it does not change R's behavior. A more interesting kind of deception is possible. As outlined by Chen, Kartik, and Sobel (2008), there always exists an equilibrium (m^*, y^*) in which $U^S(0, y^*(m^*(0))) > U^S(0, \bar{y}(0))$. In this case, there exists an equilibrium in which R responds to off-path messages with the belief $\theta = 0$ and the action $\bar{y}(0)$. In this case, it is strongly deceptive for $\theta = 0$ to send $m^*(0)$ instead of an off-path message. Typically, one can construct off-path beliefs that pool together all types below a (small) critical type $\hat{\theta}$. In this case, it is deceptive for all types less than $\hat{\theta}$ to follow the equilibrium. This kind of deception—the possibility of supporting an equilibrium with an off-path action that is strictly lower than all of the on-path actions—is possible only for equilibrium outcomes that satisfy the no-incentive-to-separate condition. Chen, Kartik, and Sobel (2008) use this condition as a selection criterion; that is, the argument selects exactly the equilibria that permit deception that influences payoffs.

D. Deception and Lying

There is no reason why lies must be deceptive. The receiver may anticipate a lie and form accurate beliefs after hearing one. If sellers make inflated claims about the quality of a product (or letters of recommendation overstate the abilities of a job candidate) and these claims are fully

[20] If the prior is uniform and $U^R(\theta, y) = -(y - \theta)^2$, then $y_0 = \hat{\theta} = .5$.

anticipated and discounted, then the claims are lies but not deceptive. This form of lying is sometimes called puffery.

There is no reason why honesty must induce accurate beliefs. The receiver may draw inaccurate inferences from an honest statement.[21]

DEFINITION 6. In a communication game, the receiver is *credulous* if $\mu(\cdot \mid m_{\Theta_0})$ is equal to the posterior distribution conditional on $\theta \in \Theta_0$; that is,

$$\mu(\theta \mid m_{\Theta_0}) = \begin{cases} 0 & \text{if } \theta \notin \Theta_0 \\ \dfrac{P(\theta)}{\sum_{\theta' \in \Theta_0} P(\theta')} & \text{if } \theta \in \Theta_0 \end{cases}.$$

If the receiver is credulous, there is a natural connection between lying and inaccurate beliefs.

REMARK 7. Given a communication game,

i. if m is a lie given θ and R is credulous, then $\mu(\cdot \mid m)$ is completely inaccurate given m and θ;
ii. if m is true but not the whole truth given θ and R is credulous, then $\mu(\cdot \mid m)$ is inaccurate given m and θ;
iii. if m is the whole truth given θ and R is credulous, then $\mu(\cdot \mid m)$ is accurate given m and θ.

Remark 7 follows immediately from the definitions.

Remark 7 connects beliefs to lies when the receiver is credulous. Alternatively, we can relate the concepts when the sender is completely truthful and therefore uses the strategy

$$\sigma(m \mid \theta) = \begin{cases} 1 \text{ if } m = m_\theta \\ 0 \text{ if } m \neq m_\theta \end{cases}.$$

REMARK 8. Given a communication game with a common language, if S tells the whole truth and R is rational, then the receiver's beliefs are accurate.

Similarly, the receiver will have inaccurate (but not totally inaccurate) beliefs if the sender tells the truth but not the whole truth.

V. Damage

I discussed lies without introducing the interpretation of messages. Hence I classified lies as locutionary acts. Deception requires making assumptions about how the receiver interprets messages, but my development made no

[21] There are interesting situations in which honest statements may be deceptive. Rogers et al. (2017) call the strategy of using truthful statements to deceive "paltering" and provide natural and laboratory examples.

assumptions about the consequences of deception. Hence deception is an illocutionary act. In this section, I discuss perlocutionary aspects of communication. By studying the consequences of communication, I must take into account how the sender's behavior influences the receiver's utility.

Section V.A defines a damaging message as one that lowers the receiver's payoff. Section V.B discusses the properties of damaging messages. Section V.C identifies connections between damaging and deceptive messages. This subsection connects the idea of inaccurate beliefs introduced in section IV.A to damaging behavior. In particular, it identifies the way in which deceptive (and strongly deceptive) messages are damaging and, conversely, demonstrates a sense in which damaging messages must be deceptive. Section V.D discusses alternative definitions of deception and damage.

A. *Definition of Damage*

The utility the sender believes the receiver will get depends on the sender's behavior, the action the receiver takes, and the state.[22] Let $y^*(m)$ be the receiver's response to the message m. Let $\bar{u}^R(\theta, x, m) = U^R(\theta, x, y^*(m))$ be the receiver's expected utility when S takes action x, S sends the message m, and the true state is θ. Interpret this as the sender's evaluation of the receiver's payoff. Observe that $y^*(m)$ need not maximize $U^R(\theta, x, y)$ even in equilibrium, because when R hears m he may not know what the true state is.

DEFINITION 7 (Damaging behavior).

i. The pair (m, x) is *damaging* given θ and $y(\cdot)$ if there exists a message n such that $\bar{u}^R(\theta, x, y(m)) < \bar{u}^R(\theta, x, y(n))$.
ii. The strategy (m^*, x^*) is a *damaging strategy* given $y(\cdot)$ if there exists a θ such that $(m^*(\theta), x^*(\theta))$ is damaging given θ and $y(\cdot)$.

Definition 7 makes no reference to a common language.

The sender's action choice x does not appear in the definition of lying because x is not evaluated relative to accepted meaning. The sender's action choice does not appear in the definition of deception because the receiver does not observe x and so his beliefs are independent of x.

It is important that m does not enter into the receiver's preferences because in many strategic situations the sender may take actions that harm the receiver, but the damage is not the result of any attempt to deceive. For example, in a dictator game, anytime the dictator takes a positive share of surplus for herself, there is an alternative action that would be strictly better for the other player. One reason for maintaining the distinction

[22] More precisely, "the action the receiver takes" is the action the sender believes that the receiver takes.

between x and m is because there are many natural situations in which the sender's actions directly influence the receiver's utility. I reserve the term damage to describe harm caused to the receiver by the sender's choice of message.

B. *Properties of Damaging Messages*

The following observations are immediate and parallel results about deception (proposition 3 and remarks 5 and 6).

PROPOSITION 5. If the sender's strategy set is finite, then for each θ and $y(\cdot)$, the sender will have a strategy that is not damaging.

REMARK 9. If $y(\cdot)$ is constant, then no sender strategy is damaging given $y(\cdot)$.

Remark 9 asserts that the sender cannot send a damaging message if the receiver's actions are independent of the message received.

REMARK 10. If (m^*, x^*, y^*, μ^*) is a separating equilibrium, then m^* is not damaging given y^*.

Damaging messages are consistent with equilibrium. Take an equilibrium in a CS cheap-talk game in which more than one action is induced and the sender has a uniformly positive bias (with full information, the sender's utility-maximizing action is strictly greater than the receiver's). When the sender is indifferent between sending messages that induce distinct actions, a and $a' > a$, the receiver strictly prefers action a. Hence when the sender's type θ is slightly greater than the type indifferent between a and a', the receiver would strictly prefer the sender to send the message that induces a than type θ's equilibrium message (which induces a'). In this situation, a rational receiver is not deceived on average, but given the receiver's strategy, there are realizations of the state after which the sender causes the receiver to make an inferior decision. If beliefs are rational, R is fully aware that he will be deceived in this way, but short of switching to the babbling equilibrium, there is nothing he can do about it. Hence it is possible for the sender to send damaging messages in informative equilibria of cheap-talk games. Since the receiver ex ante prefers informative equilibria to the babbling equilibrium, this means that allowing damaging messages can be beneficial to the receiver.

C. *Damage and Deception*

This section describes the connections between damaging and deceptive messages. Because, in general, S's action choice can directly harm R, I assume that X is empty and focus on how deceptive behavior—rather than payoff-relevant actions—may have negative consequences for R. First, I relate damage to deception; next, I relate damage to strong deception;

and finally, I talk about the possibility for deception in common-interest games.

For the first result, I assume that the receiver's preferences belong to a restrictive class.

DEFINITION 8. The receiver's preferences are state specific if there is a bijection $v: \Theta \to Y$ and positive numbers $\alpha(\theta)$ for $\theta \in \Theta$ such that

$$U^{R}(\theta, y) = \begin{cases} \alpha(\theta) & \text{if } y = v(\theta) \\ 0 & \text{if } y \neq v(\theta) \end{cases}.$$

If the receiver has state-specific preferences, then he has a unique action $(v(\theta))$ that is a best response when the state is θ. If the receiver takes an action that does not match the best action for the state, he receives payoff zero.

Let $\mathrm{BR}(\mu)$ denote R's best response correspondence; that is, $\mathrm{BR}(\mu) = \arg\max_{y \in Y} \Sigma_{\theta' \in \Theta} U^{R}(\theta', y)\mu(\theta')$.

PROPOSITION 6. Assume that R has state-specific preferences. If m is deceptive given θ and μ, then there exists n such that

$$\max_{y \in \mathrm{BR}(\mu(\cdot \mid n))} U^{R}(\theta, y) \geq \max_{y' \in \mathrm{BR}(\mu(\cdot \mid m))} U^{R}(\theta, y'). \tag{7}$$

If R has at least two actions and m is not deceptive given μ, then there exists a specification of state-specific preferences of R such that

$$\min_{y \in \mathrm{BR}(\mu(\cdot \mid m))} U^{R}(\theta, y) > \max_{y' \in \mathrm{BR}(\mu(\cdot \mid n))} U^{R}(\theta, y') \tag{8}$$

for all $n \neq m$.

Proposition 6 states, on the one hand, that if m is deceptive given θ, then any receiver with state-specific preferences would do at least as well if he heard message n instead of m. On the other hand, if m is not deceptive given θ, then there exists a specification of state-specific preferences for which the message m is not damaging given θ (assuming R responds optimally). The two parts of the proposition are not symmetric when R has multiple best replies. It is possible that R is indifferent between two actions at $\mu(\cdot \mid m)$ and $\mu(\cdot \mid n)$ but that these actions yield different utilities given θ.

The proof of proposition 6 is a straightforward consequence of the characterization result in proposition 2.

Given $\mu(\cdot \mid m) \neq I(\cdot \mid \theta)$, there is a full-dimensional set of beliefs $\mu(\cdot \mid n)$ that satisfy equation (4). This suggests that the definition permits too much deception. Proposition 6 suggests a context under which the definition might be appropriate. If the problem has enough structure so that R's preferences are known to be state specific, then there is a clear connection between deception and damage.

What if R's preferences need not be state specific? The notion of strong deception provides a result that parallels proposition 6 when R's preferences are not restricted. To state the result, I need two preliminary observations.

Suppose that S receives the signal θ and can induce beliefs of the form $pI(\cdot \mid \theta) + (1 - p)\gamma$, where γ is an arbitrary distribution over states. Let $\bar{y}(p)$ be a receiver-optimal response to these beliefs. I claim that R's expected utility is an increasing function of p.

LEMMA 2. $U^{R}(\theta, \bar{y}(p))$ is increasing in p. If $p' > p$ and $\bar{y}(p')$ is not a best response to $p\mu^{S}(\theta) + (1 - p)\gamma$, then $U^{R}(\theta, y(p')) > U^{R}(\theta, \bar{y}(p))$.

LEMMA 3. If R has at least two actions and γ' cannot be written as a convex combination of $\mu^{S}(\theta)$ and γ, then there exists a specification of preferences for the receiver such that $U^{R}(\theta, \bar{y}(\gamma')) > U^{R}(\theta, \bar{y}(\gamma))$.

The next result is a consequence of lemmas 2 and 3. It connects strongly deceptive messages to damaging messages. The result is awkward to state if R's best reply to $\mu(\cdot \mid m)$ is also a best reply to $\mu(\cdot \mid n)$. I view this as a technicality, which I rule out by assumption. For the next results, I call a message *m influentially deceptive* given θ and μ if equation (5) holds and no element of $\mathrm{BR}(\mu(\cdot \mid m))$ is an element of $\mathrm{BR}(\mu(\cdot \mid n))$.

PROPOSITION 7. If m is influentially deceptive given θ and μ, then it is damaging given θ, assuming that R responds optimally to beliefs. If a message is damaging given θ and R's best reply to beliefs for every specification of R's preferences, then it is influentially deceptive.

The next result is essentially a restatement of the first part of proposition 7.

PROPOSITION 8. If the receiver is credulous and all dishonest messages are influentially deceptive, then any lie is damaging.

If R is credulous, then he believes what the sender says. If m is influentially deceptive, then it leads to a payoff for R that is strictly less than what R would have received given an honest message by proposition 7. Hence any lie must be damaging.

I conclude this subsection by observing that the receiver cannot lose because of deception in a game with common interests; that is, if there is a message m that is deceptive (relative to some state and equilibrium beliefs), then R's equilibrium payoff given m is no larger than his payoff given a nondeceptive message. The result does not rule out deception in common-interest games, but it implies that if there is deception, then the deception will not influence payoffs.

PROPOSITION 9. Let (m^*, y^*, μ^*) be an equilibrium of a common-interest communication game; m^* is not damaging given y^*.

D. Discussion

This section discusses alternative definitions of deception. I postponed the discussion until I had introduced damaging messages because some of the alternatives incorporate damage into the definition.

Definition 4 associates deception with inducing beliefs that are farther from the truth and offers two ways to think about what it means for one belief to be farther from the truth than another. But there are many other ways to compare beliefs, so there are many possible ways in which one might talk about deception. The results of section V.C suggest that one can introduce a family of definitions of deception and connect them with damage in different contexts.

Loosely, proposition 6 states that m is deceptive if and only if it is damaging when R's preferences are state specific, and proposition 7 states that m is strongly deceptive if and only if it is damaging for all possible preferences of R. More generally, one can imagine that the receiver's preferences belong in a particular set and then try to identify a "farther from the truth given θ" relationship with beliefs with the property that m induces beliefs farther from the truth than n if and only in cases when R responds optimally to beliefs induced by m, he does worse (in state θ) than when he responds optimally to beliefs induced by n. One set of restrictions seems especially interesting. In many settings, the state space has structure, induced by either a topology or an order. For example, states may have a natural topology, and the receiver's preferences are continuous with respect to the topology (or, even more specifically, R's set of actions Y is equal to the state space θ, and R seeks to minimize the distance between his action and the state). In this setting, I can equate deception with inducing beliefs that are farther from the true state, and the topology on Θ provides guidance about what "farther" means. I believe that there are analogs to propositions 6 and 7 that associate damaging behavior with inducing beliefs that shift weight away from the true state in a given direction. The definitions I use treat all states that are not equal to the true state symmetrically. The alternative could identify nearby states, so that one could say that a distribution that places probability one-half on $\theta - 1$ and $\theta + 1$ is "closer to the truth" than one that places probability one-half on $\theta - 2$ and $\theta + 2$.

I close the section with a brief discussion of three alternative definitions.

My notions of deceptive and damaging messages take as a baseline beliefs and payoffs to the receiver that are available if the sender alters her behavior. An alternative is to take prior beliefs and prior optimal actions as a benchmark. From this perspective, a message would be deceptive if it induces beliefs that are less accurate (in some sense) than the prior and damaging if the message lowers the receiver's payoff relative to the prior optimal action. This alternative definition has two features that I view as shortcomings. First, there are examples in which the sender's message will be classified as deceptive even if it does not influence beliefs (e.g., if the beliefs are independent of the message). Second, there are situations (although they require nonequilibrium behavior) in which the sender cannot avoid sending a deceptive message. There is one economically relevant

situation in which the definition will make alternative classifications: partial disclosure provides information that improves on the prior (so it is not deceptive in the alternate sense), but it is deceptive in my sense when full disclosure is possible. I believe that it is consistent with the literature to call a message persuasive if it leads the receiver to take an action different from the one he would take on the basis of prior information only.

The Federal Trade Commission (FTC; 1983) established guidelines that define deception. It identifies three necessary conditions for deception. First, deception requires doing something that misleads the consumer. Second, it evaluates the impact from the perspective of a consumer who acts reasonably. Third, for a practice to be deceptive, it must have a material impact on a consumer. The FTC notion of misleading information is consistent with my definition of deception. In particular, the policy explicitly states that omitting information may be deceptive. One cannot deceive unless there is an alternative message that provides more useful information. My definition can incorporate the condition that one evaluates impact from the perspective of a reasonable consumer. In my model, what is important is the sender's beliefs about how the receiver interprets messages.[23] Finally, my definition leaves out the third condition. The FTC argues that deception requires studying the effect of a message on what the receiver does. In Austin's terminology, this means it views deception as a perlocutary act. Instead, I classify deception as illocutionary. My notion of damaging statements captures all three of the conditions that the FTC requires for deception. It is sensible for the FTC's concept to include damage, because the purpose of its guidelines are to promote efficient outcomes and protect consumers.

Philosophers also discuss deception. Mahon (2008) provides an overview. He presents several variations, but all involve an action by the sender intentionally causing the receiver to hold a false belief. Consistent with my definition, deception in these treatments involves manipulating beliefs; deception does not directly refer to consequences; and deception is an intentional act.[24] Moreover, these notions separate deception from lying because deception does not require making statements (actions or gestures can be deceptive). My definition associates deception with inducing beliefs that are less accurate than other beliefs that may be induced. In Mahon (2008), false beliefs appear to correspond to what I call inaccurate beliefs. If this is the case, I view the definition as too broad for many

[23] There are (at least) two possible ways in which the receiver can be reasonable: he can be credulous or he can be fully rational. Presumably, mixtures of credulous and rational beliefs may be viewed as reasonable. What constitutes deception will, in general, depend on what is viewed as reasonable behavior.

[24] Fallis (2010) provides a similar treatment. He also introduces the notion of deceptive lies. In my terms, m_θ is a deceptive lie given θ and μ if $\theta' \neq \theta$ and $\mu(\cdot|\theta') \neq I(\cdot|\theta)$.

strategic settings, because it may be that inducing accurate beliefs is not compatible with equilibrium.

VI. Charades and Bluffs

Informally, deception involves purposeful behavior by S that induces R to have inferior beliefs. Damaging messages are purposeful acts that lower R's payoff; that is, when the sender damages the receiver, the receiver loses. Section V discusses perlocutionary aspects of communication from the viewpoint of the receiver. This section studies perlocutionary acts from the perspective of the sender. Although the sender cannot distort her information to influence her (current) beliefs, it is common that she will lie and deceive to influence her own payoffs. Section VI.A contains definitions, and section VI.B describes some properties.

A. *Definitions of Charades and Bluffs*

If R knows the state, then, following the message m, he will take action $y^{R}(\theta, m, x)$ (I allow for the possibility that the receiver's response depends directly on S's action), and the sender's utility is $U^{S}(\theta, m, x, y^{R}(\theta, m, x), m)$.

DEFINITION 9 (Charades). The action (m, x) is a charade given θ and $y^{R}(\cdot)$ if there exists (n, z) such that

$$U^{S}(\theta, m, x, y^{R}(\theta, m, x)) < U^{S}(\theta, n, z, y^{R}(\theta, n, z)).$$

S's behavior is a charade if she does not send the message she would send if R were fully informed. Charades and damaging messages are clearly different in this formulation: the sender's preferences determine what is a charade; the receiver's preferences determine what is damaging. Loosely, a sender engages in a charade if the existence of private information distorts her behavior. Charades need not distort the receiver's beliefs. When a charade does cause the receiver to have incorrect beliefs, it is a bluff.

DEFINITION 10 (Bluffs). The action (m, x) is a bluff given θ and $y^{R}(\cdot)$ if (m, x) is a charade and (m, x) is deceptive.

By analogy with earlier definitions, I say that a strategy (m^{*}, x^{*}) is a charade (bluff) given $y(\cdot)$ if there exists θ such that $(m^{*}(\theta), x^{*}(\theta))$ is a charade (bluff).

Damage, charades, and bluffs refer to payoffs, so they all are perlocutionary properties. The basic model treats the informed sender differently from the uninformed receiver. Consequently, I do not define charades and bluffs, which refer to the sender's payoffs, symmetrically with damage, which refers to the receiver's payoffs. Specifically, there is limited interest in studying actions of senders that damage the sender (presumably, she

would avoid them) and limited interest in studying whether the receiver would act differently with complete information (typically, he would).

B. *Properties*

As with deception and damage, the sender can avoid charades (and hence bluffs) by selecting $(\bar{m}, \bar{x})$ to maximize $U^S(\theta, m, x, y^R(\theta, m, x))$.

It is apparent that charades are not possible in equilibria of perfect-information games, where, by definition, the receiver knows what the sender knows. Charades are also impossible in cheap-talk games.

PROPOSITION 10. If (m^*, y^*, μ^*) is an equilibrium of a cheap-talk game, then for all θ, m^* is not a charade given θ and y^*.

REMARK 11. If (m^*, x^*, y^*, μ^*) is a separating equilibrium, then (m^*, x^*) is not a bluff given any θ and $y^*(\cdot)$.

Remark 11 follows because the receiver has accurate beliefs in a separating equilibrium.

Proposition 10 demonstrates that deception need not be a charade, because deception is possible in cheap-talk games. Conversely, a charade need not be deceptive. Consider the separating equilibrium in a standard Spence signaling game. On the one hand, there is no deception, because the receiver learns the sender's type. On the other hand, the sender's message is a charade. If the receiver knew the sender's type, then there would be no reason for the sender to invest in costly signaling. When education does not add to productivity, it is a charade whenever the sender invests a strictly positive amount. When education does add to productivity, it is a charade to overinvest.

There is a connection between lying and charades. Suppose that the sender believes that the receiver is credulous. The sender expects her message to be taken literally. If she responds optimally to her beliefs, then whenever she chooses to send a dishonest message, she weakly prefers the payoff she receives when dishonest to the payoff received when honest. If this preference is strict, then the lie is a charade.

Poker is perhaps the canonical example of a game in which there is bluffing. In poker, a player bluffs by making large bets with a poor hand.[25] If the bluffer's opponents knew the true quality of her hand, they would be inclined to call the bet, and the bluffer would lose. A bluff is successful precisely because the opponents lack information.

[25] Hörner and Sahuguet (2007) present a model of jump bidding in auctions that can be viewed as a model of poker. This paper describes an overbid from a weak player as a bluff and an underbid from a strong player as a sandbag. They show that both bluffing and sandbagging are possible in equilibrium. Both bluffs and sandbags in the sense of Hörner and Sahuguet are bluffs in my sense. I cannot distinguish the two phenomena because I do not assume that types are ordered.

I make a few more comments about poker to illustrate the definition. First, even rational opponents will associate large bets with good hands and might fold, allowing the bluffer to win. Bluffs are certainly consistent with rational behavior. Second, lies need not be part of the bluff. A bet in poker is an indirect statement about the quality of one's hand, but it is not constrained by natural language. Third, bluffing behavior in poker (or in any zero-sum game) cannot induce accurate beliefs. Bluffs do not arise in two-player zero-sum games in pure-strategy equilibrium, but except in unusual cases, if the sender plays a mixed strategy in equilibrium, she is bluffing.[26]

Charades may arise even if the sender and receiver have common preferences over actions. Suppose, for example, that types, messages, and actions are elements of {0,1} and $U^{S}(\theta, m, y) = -(y - \theta)^2 - .1m$ and $U^{R}(\theta, m, y) = -(y - \theta)^2$. In this case, $m = 1$ is more expensive than $m = 0$ for both types of sender. There are separating equilibria, but one type must send the costly message $m = 1$. If the receiver knew the sender's type, then S would strictly prefer to send $m = 0$. This charade is not a bluff. This game also has a pooling equilibrium in which there is no bluffing. Charades need not arise if one added a round of preplay, cheap-talk communication.

VII. Examples

This section describes five examples and fits them into the paper's framework. The first two examples involve equilibrium models with rational agents; the other examples involve some kinds of behavioral agent.

A. *Costly Lying*

Kartik, Ottaviani, and Squintani (2007) study a model of communication with costly lying. They identify conditions for the existence of separating equilibria in a sender-receiver model in which the sender's message enters directly into the sender's preferences. They offer several applications of their model. In one interpretation, the message space and the state space are identical, and the sender's utility is decreasing in the distance between the message and the true state. This specification identifies the language with the state and provides a tractable way to model costly lying. In this equilibrium, the receiver is fully informed in equilibrium, and the sender's messages are not deceptive. According to my definition, remark 6 implies that there is no deception. Because the equilibrium is separating, there are no bluffs (remark 11). In another interpretation, a fraction of

[26] The strategy will be a bluff if (1) the receiver will change his behavior if he knows the sender's strategy and (2) the receiver's choice influences the sender's payoffs.

receivers are credulous (and believe that the sender honestly reports the true state). In a separating equilibrium, the credulous receivers are deceived and the sender's strategy is deceptive.

Kartik, Ottaviani, and Squintani (2007) "view deception as the act of inducing false beliefs by means of communication, and exploiting them to one's own advantage. Such false beliefs are clearly incompatible with traditional equilibrium analysis" (95, n. 3). The definition of deception that I offer is different in several respects. First, I do not limit deception to communication; actions may be deceptive. Second, for Kartik, Ottaviani, and Squintani, beliefs are false if and only if they are inconsistent with prior information and equilibrium strategies. With this definition, as they say, false beliefs are incompatible with traditional equilibrium analysis. Third, according to my definition, deception need not be disadvantageous to the receiver nor advantageous to the sender. Kartik, Ottaviani, and Squintani appear to view inducing false beliefs as the essence of deception. They treat the property that the sender benefits from deception as an implication of optimizing behavior.

B. Feints

Hendricks and McAfee (2006) present a model of feints. In their setting, an informed sender learns the value of $q \in [0, 1]$. The sender then selects an investment $m \in [0, 1]$. The receiver observes a binary message (in $\{0, 1\}$), which is a stochastic function of m. On the basis of the message that the receiver observes, he selects $y \in [0, 1]$. Let p denote the probability that the receiver hears the message 0 (p depends on m), and let $\hat{q}$ be the expected value of q. When the sender takes the action m and the receiver responds to the message i with y_i, the payoff of the sender is $q(m - py_0 - (1 - p)y_1) + (1 - q)(1 - m - py_1 - (1 - p)y_0)$. The payoff of the receiver is $\hat{q}U(y_1) + (1 - \hat{q})U(y_0)$, where $U(\cdot)$ is increasing and concave. Hendricks and McAfee's interpretation of the model is that the sender and receiver do battle on two fronts. The parameter q determines the relative value of the different fronts. The sender's payoff at each front is the difference in resources directed at the front. The receiver's payoff at a front depends only on the resources he directs to that front.

Under these conditions, the sender would like to apply all of her resources to the more likely front while, at the same time, convincing the receiver to direct his resources to the less likely front. In equilibrium, the sender balances these two incentives. Hendricks and McAfee (2006) demonstrate that there is a mixed-strategy equilibrium. The mixed-strategy equilibrium involves bluffing (because the sender would not follow her strategy if the receiver could directly observe q), damage, and deception (because the choice of m influences the signal the receiver obtains, which

will lead the receiver to take a poor decision with positive probability). I would not classify any message as a lie, because the formulation of the game does not provide an accepted meaning to the signals.

In this example, two features of the general model come into play. First, the receiver does not observe the sender's message perfectly; that is, the binary signal the receiver hears is a stochastic function of the sender's action. I do not permit this in the basic model but discuss the variation in section VIII.A. Second, the receiver's payoff does not depend on the sender's message. This illustrates the importance of the distinction I make between m and x.

C. *Behavioral Model of Lying about Intentions*

The first two examples illustrate how deception arises in models with rational agents. It is perhaps less surprising to observe that deception arises in models with behavioral agents. Nevertheless, it is useful to show how my basic model is general enough to include some behavioral applications.

Crawford (2003) analyzes a behavioral model of cheap-talk about intentions. This study is relevant to my analysis because it demonstrates how deception arises when some agents do not have accurate beliefs about their opponents' behavior. He assumes that there is an underlying 2×2 zero-sum game that is preceded by a round in which one party (the sender) can make a statement. The zero-sum game has a unique equilibrium. In the equilibrium, both players play nondegenerate mixed strategies. The statement is made in a natural language. Either the sender says, "I am going to play up," or she says, "I am going to play down." Following the statement, the sender and receiver play the underlying game. The Nash equilibrium of this game requires the receiver to play the equilibrium (mixed) strategy following either statement, the sender's statement to convey no information, and the sender to play her equilibrium strategy in the underlying game after any message she gives. According to my definitions, the equilibrium involves lying (because the sender's statement does not describe her intentions), no deception (by remark 5, since the receiver's beliefs do not depend on the sender's statement), and no bluffing (because the sender has no private information).

Crawford (2003) analyzes the game assuming that the players respond optimally to beliefs that are not necessarily accurate. He concentrates on a small number of plausible behavioral types: senders who always honestly reveal their intentions; senders who always lie about their intentions; receivers who take the sender's message literally; and receivers who believe that the sender will not do what she says. With these changes, the game fits into the general framework of my model. The credulous receiver type and the truthful sender type create a commonly understood language, and the

sender's message conveys information about her type. For fixed fractions of behavioral types, he characterizes the equilibrium behavior of sophisticated agents who respond optimally to the (nonstrategic) behavior of their behavioral opponents and the (strategic) behavior of their sophisticated opponents. Crawford demonstrates that when the population frequencies of sophisticated agents are low and parameter values are generic, sophisticated agents play pure strategies in equilibrium. Note that, formally, my model considers communication about private information rather than communication about intentions. It still applies to Crawford's analysis because including behavioral types transforms a message that describes intentions in the original game into one about private information. Under this interpretation, for interesting parameter values, the sender will lie and the lies will deceive a fraction of receivers. These receivers will be damaged by the sender. Deception arises in Crawford's model because some agents have incorrect beliefs while other agents exploit these beliefs.

D. A Behavioral Model of Deception

Ettinger and Jehiel (2010) present an alternative model of deception that, like Crawford's model, is based on the possibility that agents hold inaccurate beliefs. They provide a simple example that illustrates their model. The model is analyzed based on a 2×2 zero-sum game Γ, with payoffs given in the table below. The stage game has a unique equilibrium. There is therefore a unique subgame-perfect equilibrium of the twice-repeated game when played by conventional (sophisticated and rational) agents. Deception is not possible.

Ettinger and Jehiel (2010) demonstrate the possibility of deception in an analogy-based sequential equilibrium. In their model, the row player's first-period action can be used to deceive the column player. Specifically, they identify an equilibrium in which the rational row player plays U in the first period and D in the second period if column played L and U otherwise; the coarse row player plays U in both periods; and the column player plays L in the first period and R in the second period if row played U in the first period and L in the second period if the row player begins with D. The outcome of the game is (U, L), followed by (D, R) when the row player is rational and (U, R) when the row player is coarse. In this setting, the row player's move in the first period influences column's beliefs about row's type. It plays the role of the message, m, in my model. The column player observes m. The row player's second-period action plays the role of x in my model. Damage involves second-period payoffs, which do not directly depend on row's first-period choice. The column player views the first-period play of D as evidence that row is rational (because the coarse row player never plays D) and believes that row is coarse with probability

two-thirds following U. This belief is objectively incorrect because row plays U in the first period independent of type. The column player has this belief in an analogy-based equilibrium because column cannot distinguish first- and second-period actions. Hence column thinks that the probability of coarse given U is, by Bayes's rule, $.5/(.5 + .25)$, where the numerator is the probability of both coarse and U (the ex ante probability of coarse, .5, times the probability that coarse plays U, 1), and the denominator is column's view of the probability of observing U (the probability of coarse row and U plus the probability of rational row and U minus the last term is .25, because column, unable to distinguish first- and second-period actions, believes that rational row players play U one-half of the time). For my analysis, the origin of these beliefs is less important than the fact that the rational row player understands them. So the rational row's first-period choice can influence what column expects. If the rational row starts with U, then column will believe that row is more likely to play U than D in the second period; consequently, column will play R. It is (strongly) deceptive and damaging when the rational row begins by playing U. This action leads the column player to obtain a lower payoff than if the first period's action was D. According to my terminology, the rational row is bluffing when she plays U in the first period because she would not follow that strategy if column knew row's intentions. This application shows the value of permitting the sender in my model to select an action as well as a message.

This and the following example illustrate how my model applies to a setting in which beliefs are neither formed as part of a standard equilibrium nor completely credulous.

	L	R
U	$5, -5$	$3, -3$
D	$0, 0$	$7, -7$

E. Projection Bias

Madarász (2016) models another type of bias in beliefs. He assumes that players suffer from projection bias. In his model, a privately informed agent may assume that her opponent has access to her information (even when this is not true). One of the examples in Madarász's paper involves a communication game. The structure of the game is slightly more general than that of the basic games that I study because the receiver has private information. The sender learns the value of a binary state, which is equally likely to be 0 or 1. The receiver has private information about his cost of verifying the sender's message. After learning the state, the sender sends a binary message ($m = k$ has the commonly accepted

meaning "the state is k" for $k = 0$ and 1). The receiver decides first whether to audit the statement; an audit costs $c > 0$ and reveals the state perfectly. Denote by a the audit decision ($a \in \{0, 1\}$; $a = 1$ means audit, and $a = 0$ means do not audit). Second, the receiver takes an action y based on his information. The receiver seeks to minimize the distance between the state and the action ($(y - \theta)^2$) plus auditing costs (if any). The sender's preferences are given by $Bm - Pma(1 - \theta)$ where $P > B > 0$. The first term in the sender's preferences reflects the amount B she earns from making a positive report; the second term reflects the penalty she must pay if she makes the report $m = 1$, she is audited ($a = 1$), and the state is low ($\theta = 0$). In equilibrium (for sensible distributions on auditing costs), the sender reports honestly when $\theta = 1$ and randomizes when $\theta = 0$. The receiver responds to $m = 0$ by never auditing and setting $y = 0$. When $m = 1$, he audits when his cost of auditing is low (and selects $y = \theta$), does not audit when the cost of auditing is high, and sets y equal to the conditional expectation of θ (which is strictly between .5 and 1). According to my definitions, the sender lies with positive probability (because she sometimes reports $m = 1$ when $\theta = 0$); deceives and damages (because when $\theta = 0$ and she reports $m = 1$, the receiver would be better off if the sender had told the truth); and bluffs (because when the sender lies, she both deceives the receiver and makes a statement that she would not have made had the receiver been fully informed). In this situation, deception is consistent with equilibrium. Both players have accurate beliefs, but incomplete information gives the sender an opportunity to mislead R (and take advantage of this). Of course, in equilibrium, R's expectations are rational.

Madarász is interested in what happens when the receiver in this example has projection bias. In this case, the receiver believes with probability ρ that the sender actually knows his cost of auditing (with probability $1 - \rho$, the receiver accurately believes that the sender does not know the auditing cost). The sender is sophisticated (in the sense that she knows R's projection bias); she adjusts the frequency of lies in equilibrium in response to R's bias, but her strategy does not depend on the auditing cost. Equilibrium still involves S reporting honestly when $\theta = 1$ and randomizing when $\theta = 0$. The receiver's strategy depends on the realization of costs. When costs are low, he audits $m = 1$ reports and chooses y optimally. When c is a bit higher, R audits too little (because he projects knowledge onto S and assumes that the lower cost of auditing will deter lies) and selects actions that are higher than optimal. When c is higher still, R audits too much and selects actions that are lower than optimal (because he thinks that S, knowing that auditing is relatively costly, will lie more). Finally, when c is very high, R will not audit and will take actions that are lower than optimal (because R thinks that S will always lie, but S is actually not able to take full advantage of the high cost of auditing).

From the standpoint of the ideas in this paper, when receivers are biased, reporting $m = 1$ when $\theta = 0$ is still a lie, a deception, and a bluff. What changes is the equilibrium frequency of deception and the gains from deceiving.

VIII. Extensions of the Basic Model

I have studied a simple model. The model identifies central concepts but misses important features. This section lists some of the missing elements. The broad lessons of the basic model are the important distinctions between locution, illocution, and perlocution; the notion that dishonesty and deception are compatible with equilibrium; the connection between damage and deception; and the ability of the sender to avoid deception, damage, and bluffs. These features remain in more general models. At the same time, making the model more general allows qualitatively different kinds of dishonest behavior. The variations are almost certainly going to be important when the costs of lying and deception are taken into account.

This section reviews some of the possible variations.

A. *Noise in Messages*

I assume that the receiver hears the message m perfectly. It is possible to extend the model so that the receiver's observation is a (potentially) stochastic function of m. One would need to augment the model with a function $\nu(m'|m)$ that specifies the probability that R observes m' given that S sends the message m. For each m, $\nu(\cdot|m) \geq 0$ and $\Sigma_{m' \in M'} \nu(m' \mid m) = 1$. The difference between the sender's message and the receiver's observation is important in the model of Hendricks and McAfee (2006) that I discussed in section VII.A. Dziuda and Salas (2018) is a recent model that investigates the implications of assuming $m' \neq m$. Another leading example is if sender and receiver speak different dialects or are familiar with different norms of behavior.

The definition of lying does not need modification in this case, but a variation is possible. If the sender knows μ, then she may be able to modify what she says so that what the receiver hears is (approximately) the true state. So it is possible to evaluate lies from the perspective of what the receiver hears rather than what the sender says.

From the sender's perspective, the beliefs she induces are of the form $\mu'(\theta \mid m') = \Sigma_{m'} \mu(\theta \mid m') \nu(m' \mid m)$. One can define accuracy and deception in terms of μ' instead of μ. The analysis of damage, charades, and bluffs carries over without change.

B. *Imperfect Knowledge of State*

I have assumed that the sender knows the state. More generally, one can imagine that the sender obtains imprecise information about the state. A standard way to describe the sender's information is through a joint probability distribution over states θ and signals received by the sender (ω). Viewed in this way, the definition of lying extends trivially, assuming that the accepted meanings correspond to statements about ω; that is, the sender talks about what she observes rather than the state of nature. This extension leaves open a couple of problems of interpretation.

If S receives noisy information, then there are different ways to lie. The sender can make a statement that she believes contains a state that is possible and a state that she believes is impossible. For example, if $\Theta = \{1, 2, 3\}$ and S learns that the state is in {1,2}, then $m_{\{1,3\}}$ is a lie given $\theta = 2$ and is true given $\theta = 1$ but should be evaluated given the sender's information ($\theta \in \{1, 2\}$). Alternatively, the sender can make a statement that excludes states that may be possible. For example, S may learn that $\theta > 1$ and report that $\theta = 3$. I view the first statement as a lie given $\theta = 3$ and the second statement as a lie given $\theta = 2$, but it is also possible to view these statements as lies given what S observes (i.e., the first statement is a lie given $\omega = \{1, 2\}$).

C. *Probabilistic Statements*

I assume that the language associates words to subsets of Θ. It is possible that the language has words for probabilistic statements (e.g., "I believe that θ is equally likely to be 0 or 1").

Suppose there are two states, θ_1 and θ_2, that are equally likely ex ante, and suppose the sender receives an informative but imperfect signal of the true state. If the only statements that have accepted meaning available to the sender are of the form $\theta \in T$ for $T = \{\theta_1\}$, $\{\theta_2\}$, or $\{\theta_1 \cup \theta_2\}$, then it could be the case that both sender and receiver benefit if the sender lies and says, "The state is θ_1," when she believes (only) that θ_1 is highly likely. One can remedy this problem if the sender has access to a language that permits her to make statements about her posterior ("State θ_1 is much more likely than θ_2"). It is difficult to evaluate whether statements like this are honest in one-shot settings.

I am not sure whether to call a probabilistic statement a lie. On the one hand, if we knew that a weather forecaster has expert information that identifies the chance of rain as 95%, then I would want to classify a report that the chance is 5% as a lie. On the other hand, it may be difficult to document that the forecaster has probabilistic information, and if so, an observer cannot know with certainty that the forecaster has lied.

Formally, the issue comes down to what is the right definition of the state space. In order to classify incorrect probabilistic statements as lies, one need only interpret the set of probability distributions over Θ as the state space and have a language that includes messages with accepted meaning for all elements of this larger state space.

Adding noise in messages, imperfect knowledge, and probabilistic statements creates no barriers to formal modeling. I believe that, in practice, the cost of lying and deception depends strongly on these assumptions. Investigating this belief goes beyond this paper.

D. Heterogeneous Beliefs

It is standard to assume that there is a common-knowledge information structure ($P(\cdot)$), but this restriction may be inappropriate in models with boundedly rational agents. It is useful to consider variations of the model in which the sender and receiver have different beliefs about the prior distribution over θ. In such a situation, the concepts that I introduce should be viewed from the perspective of the sender; that is, the model applies to a situation in which different players have different beliefs. The beliefs that appear in definitions will be those of the sender. With this variation especially (but also when there are noisy messages), it is likely that there will be a difference between the intention to deceive and whether, in fact, the sender succeeds in deceiving.

E. More General Games

I studied two-player games with a limited dynamic structure. Adding agents does not influence the definition of lying. It raises the possibility that actions may deceive or damage subsets of opponents. Adding stages creates new context, requiring the need to evaluate honesty of communication not only with respect to the true state but also with respect to the history of play.

Anderson and Smith (2013) study a dynamic model that has features of the static models of Crawford (2003) and Hendricks and McAfee (2006) mentioned in section VII. In this model, an informed agent must trade off short-term benefits from exploiting private information with the long-term benefits of keeping the information hidden. Equilibrium play exhibits generalized versions of deception, damage, and bluffing. Sobel (1985) studies a dynamic model of communication. Communication is cheap-talk, but there is a natural way to add accepting meaning to the messages. The paper characterizes an equilibrium in which the informed sender with interests opposed to those of the receiver deceives the receiver by telling the truth. On the equilibrium path, honest statements make the receiver more

confident that the sender has common interests, enabling the sender to exploit the receiver in the future. Honesty induces inaccurate beliefs because the receiver does not expect honest information in equilibrium.

F. Communication about Intentions

In this paper, the sender's message corresponds to information about the state of the world; that is, S communicates about exogenous information. People also make statements about their intentions. At the end of an interaction, one can interpret statements of the form "I will deliver the product by Friday" or "I am going to the opera." In these situations, messages refer to future actions rather than states of the world. Strictly speaking, my model does not apply to these situations, but it is straightforward to modify definitions so that commonly accepted meanings refer to outcomes rather than states. Without changing the model, it is possible to characterize communication about intentions as deceptive (to the extent that it influences R's beliefs about future actions).

Two of the examples (Crawford 2003 in sec. VII.C and Ettinger and Jehiel 2010 in sec. VII.D) involve communication about intentions. In both cases, the authors add uncertainty about the state of the world into a model so that the receiver learns about intentions by drawing inferences about the state of the world.

G. Incorporating Costs of Lying and Deception

One motivation for providing a systematic treatment of lying and deception is to deal with the informal observation that some people seem to lie less than basic economic models predict.[27] These studies suggest that lying is costly. They also suggest that the nature of the cost may depend on consequence (whether the lie lowers someone else's monetary payoff), context (including whether there is an audience), and expectations (e.g., whether others lie in similar situations). Although I offer a narrow definition of "lie" that focuses only on the relationship between what the sender believes and what she says, these results indicate that the cost of lying depends on many other factors. The good news is that my formulation is flexible enough to include costs that depend on these features. The bad news is that the strict classification of lies as locutionary acts breaks down when one takes costs into account. I have provided a benchmark model that can incorporate lying costs. It is an empirical question to describe

[27] Ariely (2012) is a review of evidence aimed at a popular audience. Gneezy (2005) and Fischbacher and Föllmi-Heusi (2008) are two of many papers that present laboratory experiments demonstrating that many people are unwilling to lie for monetary rewards.

these costs.[28] There is a smaller experimental literature on deception, but again my model provides a way to include costs of deception in a strategic model.

Appendix

Proof of Proposition 1

In a CS (Crawford and Sobel 1982) cheap-talk game, there are only a finite number of actions induced in any equilibrium. The equilibrium partitions the type space into intervals with disjoint interiors. Hence any equilibrium type-action distribution can be generated by an equilibrium that uses only a finite number of messages with positive probability. Let $\phi(\theta)$ be the partition element containing θ. In the equilibrium, if θ', $\theta'' \in \phi(\theta)$, then $y(\theta') = y(\theta'')$. Lemma 1 implies that one can relabel the messages so that $m(\theta)$ has the commonly accepted meaning "my type is θ'" for $\theta' \notin \phi(\theta)$. This establishes proposition 1 for CS cheap-talk games. QED

Proof of Proposition 2

Suppose (6) holds. It follows that $p = \mu(\theta|m)/\mu(\theta|n) \in [0, 1)$. Let ρ be defined by

$$\mu(\theta'|m) = p\mu(\theta'|n) + (1 - p)\rho(\theta').$$

Hence $\rho(\theta) = 0$ and $\Sigma_{\theta' \in \Theta}\rho(\theta') = 1$. Further, if (6) holds, then $\rho(\theta') \geq 0$ for all $\theta' \in \Theta$. Hence n is deceptive.

Conversely, if there exist ρ and p such that (4) holds, then $\mu(\theta|m)/\mu(\theta|n) \leq \mu(\theta'|m)/\mu(\theta'|n)$ for all θ' such that $\mu(\theta'|n) \neq 0$, with strict inequality for at least one θ' if $\mu(\cdot|m) \neq \mu(\cdot|n)$. Hence condition (6) must hold. QED

Proof of Proposition 4

It is known that in any equilibrium type-action distribution in the CS model, there are a finite number of actions induced; denote these actions by $y_1 < y_2 < \cdots < y_N$. There exist $0 = \theta_0 < \theta_1 < \cdots < \theta_{N-1} < \theta_N = 1$, such that types in (θ_{i-1}, θ_i) induce action y_i. Let $m^*(\theta) = m_i$ for $\theta \in [\theta_{i-1}, \theta_i)$ and $m^*(1) = m_N$. Let

$$y^*(m) = \begin{cases} y_i \text{ if } m = m_i \text{ for } i = 1, \ldots, N \\ y_1 \text{ otherwise} \end{cases}$$

and

$$\mu^*(A \mid m) = \begin{cases} \dfrac{P(A \cap [\theta_{i-1}, \theta_i))}{P([\theta_{i-1}, \theta_i))} \text{ if } m = m_i \text{ for } i = 1, \ldots N \\ \dfrac{P(A \cap [\theta_0, \theta_1))}{P([\theta_0, \theta_1))} \text{ otherwise} \end{cases},$$

[28] Abeler, Nosenzo, and Raymond (2019) provide an excellent meta-analysis of the experimental literature on lying costs.

where $P(\cdot)$ is the prior distribution. Here $m^*(\theta)$ is not deceptive given θ and y^* because each sender type can induce only one accurate belief and does so in equilibrium. (If $\theta \in [0, \theta_1)$, then type θ can induce the same accurate belief in many ways.) QED

Proof of Proposition 6

If condition (6) holds, then $\mu(\theta|m)/\mu(\theta'|m) \leq \mu(\theta|n)/\mu(\theta'|n)$ for all θ'. Hence if $y(\theta)$ is a best response for R given $\mu^R(\cdot \mid m)$, then it must be a best response for R given $\mu(\cdot|n)$.

If condition (6) fails, then there exist θ' and $\alpha(\theta)$, $\alpha(\theta') > 0$ such that

$$\frac{\mu(\theta \mid m)}{\mu(\theta' \mid m)} > \frac{\alpha(\theta')}{\alpha(\theta)} > \frac{\mu(\theta \mid n)}{\mu(\theta' \mid n)}.$$

It follows that for significantly small choices of $\alpha(\theta'')$ for $\theta'' \neq \theta, \theta'$, when R's preferences are determined by α, $\mu(\cdot|m)$ induces the action $y(\theta)$ while $\mu(\cdot|n)$ induces the action $y(\theta')$. QED

Proof of Lemma 2

Let $V(p) \equiv pU^R(\theta, \bar{y}(p)) + (1 - p)EU^R(\theta, \bar{y}(p))$, where the expectation is with respect to γ. Let $p' > p$. It follows that

$$V(p) \geq pU^R(\theta, \bar{y}(p')) + (1 - p)EU^R(\theta, \bar{y}(p')) \tag{A1}$$

and so

$$V(p) \geq V(p') - (p' - p)(U^R(\theta, \bar{y}(p')) - EU^R(\theta, \bar{y}(p'))), \tag{A2}$$

and hence

$$U^R(\theta, \bar{y}(p')) - EU^R(\theta, \bar{y}(p')) \geq \frac{V(p') - V(p)}{p' - p}.$$

Similarly,

$$\frac{V(p') - V(p)}{p' - p} \geq U^R(\theta, \bar{y}(p)) - EU^R(\theta, \bar{y}(p)),$$

so that

$$U^R(\theta, \bar{y}(p')) - U^R(\theta, \bar{y}(p)) \geq EU^R(\theta, \bar{y}(p')) - EU^R(\theta, \bar{y}(p)). \tag{A3}$$

The definition of $\bar{y}(p')$ implies that

$$p'U^R(\theta, \bar{y}(p')) + (1 - p')EU^R(\theta, \bar{y}(p')) \geq p'U^R(\theta, \bar{y}(p)) + (1 - p')EU^R(\theta, \bar{y}(p)),$$

so either $U^R(\theta, \bar{y}(p')) \geq U^R(\theta, \bar{y}(p))$ or $EU^R(\theta, \bar{y}(p')) \geq EU^R(\theta, \bar{y}(p))$. It follows that the left-hand side of inequality (A3) is positive, which establishes the first part of the result. If $\bar{y}(p')$ is not a best response to $p\mu^S(\theta) + (1 - p)\gamma$, then the inequalities in (A1) and (A2) are strict, which establishes the second part of the lemma. QED

Proof of Lemma 3

Suppose that the receiver obtains utility 0 in each state if he takes any action $y \neq y^*$. Let $v(\theta)$ be the utility from action y^* in state θ. The lemma follows because, by the separating hyperplane theorem, it is possible to find $v(\cdot)$ such that

$$\sum_{\theta} v(\theta)\gamma(\theta) = \sum_{\theta} v(\theta)\mu^{S}(\theta) > 0 > \sum_{\theta} v(\theta)\gamma'(\theta). \tag{A4}$$

QED

Proof of Proposition 9

In equilibrium, S selects her message m to maximize $U^{S}(\theta, x, y^*(m))$. When $U^{S} = U^{R}$, an equilibrium message must maximize $\bar{u}^{R}(\theta, x, m)$, which means that the message is not damaging. QED

Proof of Proposition 10

In a cheap-talk game, if R learns θ, then S's payoff is $U^{S}(\theta, y^{R}(\theta))$ independent of message. Hence S would not strictly gain by changing her message. QED

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. "Preferences for Truth-Telling." *Econometrica* 87, no. 2 (July): 1115–53.

Anderson, Axel, and Lones Smith. 2013. "Dynamic Deception." *A.E.R.* 103, no. 7:2811–47.

Ariely, Dan. 2012. *The Honest Truth About Dishonesty: How We Lie to Everyone—Especially Ourselves*. New York: Harper Collins.

Augustine. 1887. *De Mendacio*. In *A Select Library of the Nicene and Post-Nicene Fathers of the Christian Church*, vol. 3, edited by Philip Schaff, 457–77. Buffalo, NY: Christian Literature.

Austin, John L. 1975. *How to Do Things with Words*. Cambridge, MA: Harvard Univ. Press.

Chen, Ying. 2011. "Perturbed Communication Games with Honest Senders and Naive Receivers." *J. Econ. Theory* 146, no. 2 (March): 401–24.

Chen, Ying, Navin Kartik, and Joel Sobel. 2008. "Selecting Cheap-Talk Equilibria." *Econometrica* 76, no. 1 (January): 117–36.

Coleman, Linda, and Paul Kay. 1981. "Prototype Semantics: The English Word Lie." *Language* 57, no. 1 (March): 26–44.

Crawford, Vincent P. 2003. "Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions." *A.E.R.* 93, no. 1 (March): 133–49.

Crawford, Vincent P., and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50, no. 6 (November): 1431–51.

Dziuda, Wioletta, and Christian Salas. 2018. "Communication with Detectable Deceit." Technical report, Univ. Chicago, June.

Ettinger, David, and Philippe Jehiel. 2010. "Theory of Deception." *American Econ. J.: Microeconomics* 2, no. 1 (February): 1–20.

Fallis, Don. 2010. "Lying and Deception." *Philosophers' Imprint* 10, no. 11:1–22.

Federal Trade Commission. 1983. "FTC Policy Statement on Deception." Technical report, US Government, October 14.
Fischbacher, Urs, and Franziska Föllmi-Heusi. 2008. "Lies in Disguise: An Experimental Study on Cheating." *J. European Econ. Assoc.* 11:525–47.
Frankfurt, Harry. 2005. *On Bullshit.* Princeton, NJ: Princeton Univ. Press.
Gneezy, Uri. 2005. "Deception: The Role of Consequences." *A.E.R.* 95, no. 1 (March): 384–94.
Gordon, Sidartha, Navin Kartik, Melody Pei-yu Lo, Wojciech Olszewski, and Joel Sobel. 2020. "Effective Communication in Cheap-Talk Games." Technical report, Univ. California, San Diego, January.
Grossman, Sanford. 1981. "The Role of Warranties and Private Disclosure about Product Quality." *J. Law and Econ.* 24:461–83.
Hendricks, Kenneth, and R. Preston McAfee. 2006. "Feints." *J. Econ. and Management Strategy* 15, no. 2:431–56.
Hörner, Johannes, and Nicolas Sahuguet. 2007. "Costly Signalling in Auctions." *Rev. Econ. Studies* 74:173–206.
Kartik, Navin. 2009. "Strategic Communication with Lying Costs." *Rev. Econ. Studies* 76, no. 4:1359–95.
Kartik, Navin, Marco Ottaviani, and Francesco Squintani. 2007. "Credulity, Lies, and Costly Talk." *J. Econ. Theory* 134, no. 1 (May): 93–116.
Lynch, Michael P. 2016. "Fake News and the Internet Shell Game." *New York Times*, November 28.
Madarász, Kristóf. 2016. "Projection Equilibrium: Definition and Applications to Social Investment and Persuasion." Technical report, London School Econ., July.
Mahon, James Edwin. 2008. "The Definition of Lying and Deception." *Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/entries/lying-definition/.
Milgrom, Paul R. 1981. "Good News and Bad News: Representation Theorems and Applications." *Bell J. Econ.* 21:380–91.
Morgan, John, and Phillip C. Stocken. 2003. "An Analysis of Stock Recommendations." *RAND J. Econ.* 34, no. 1 (Spring): 183–203.
Morris, Errol. 2009. "Seven Lies about Lying (Part 2)." *New York Times*, August 6.
Rogers, Todd, Richard Zeckhauser, Francesca Gino, Mike Norton, and Maurice E. Schweitzer. 2017. "Artful Paltering: The Risks and Rewards of Using Truthful Statements to Mislead Others." *J. Personality and Soc. Psychology Interpersonal Relations and Group Processes* 112, no. 3:456–73.
Sobel, Joel. 1985. "A Theory of Credibility." *Rev. Econ. Studies* 52, no. 4 (October): 557–73.