Lying and Deception in Games*

Joel Sobel†

August 2, 2016

Abstract

This article proposes definitions of lying and deception in strategic settings. A critical distinction is that deception requires a model of how the audience responds to messages while lying does not. Lies need not be deceptive. Deception does not require lying. The paper identifies situations in which lying is consistent with equilibrium and when the ability to lie is welfare enhancing. Deception may be consistent with equilibrium behavior.

*Journal of Economic Literature* Classification Numbers: D03; C90; C72. Keywords: lying; deception; behavioral economics.

---

*This draft is preliminary. 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, Kit Woolard, and loyal participants in an elective course at UCSD. I am grateful to the NSF for research funding.

†Department of Economics, University of California, San Diego, La Jolla, CA 92093, U.S.A. E-mail: jsobel@ucsd.edu
1 Introduction

This paper proposed definitions of lying and deception and derives properties of these definitions in a simple strategic context.

Loosely, a lie is a statement that the speaker believes is false. This definition requires an accepted interpretation of the honest 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 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 lead the audience to make incorrect decisions. Making these ideas precise requires a formal model, which I introduce in Section 2.

Section 3 uses the model to define lying. The model adds the concept of a common language to a two-player (Sender-Receiver) model of communication. Section 4 briefly reviews the extensive literature that offers definitions of lying and the extent to which my definition conforms with usage in other disciplines. Sections 5 and 6 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. I point out that lying may be welfare enhancing.

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 behavior. When the Sender contemplates an action, she can figure out how the action influences the Receiver. I discuss beliefs and relate them to honesty in Section 7. Section 8 provides definitions of deception. Informally, the Sender’s behavior is deceptive if it leads the Receiver to take an inferior action. An action is deceptive if the Sender had another action available that induces the Receiver to make “better” decisions. It is sometimes useful to associate the quality of decisions with the quality of beliefs the Receiver has when he makes decisions. I therefore talk about actions that are misleading in the sense that they induce inaccurate beliefs. I explore the connection between beliefs and deception under the assumption that the Receiver responds optimally to his beliefs. If the Sender misleads the Receiver (by inducing “incorrect” beliefs), then she deceives the Receiver. Conversely, an action that is deceptive independent of the preferences of the Receiver must be misleading. Section 8 identifies situations when deception does (and does not) arise in equilibrium.
There are different ways to formalize the notion of inaccurate beliefs and poor decisions. Section 8 describes the implications of several alternatives. Using these formulations it is possible to make logical inferences that are mathematically straightforward, but are potentially useful. First, lies and deception are different. Deception is possible without lying. Lies need not deceive. Second, deception only is possible if the Receiver is open to influence. If the Receiver ignores the Sender’s message, then he cannot be deceived. Third, while deception is not possible in equilibrium in a perfect-information game, it can be an equilibrium phenomenon. In particular, mixed-strategy equilibria of zero-sum games are typically deceptive according to my definition.

Deception involves manipulating the Receiver’s beliefs in a way that is harmful to the Receiver. In Section 9 I consider bluffs, which are actions or messages that are attractive to the Sender only because the Receiver lacks information. I point out that deception need not involve bluffs and bluffs need not be deceptive.

Section 10 discusses four models from the literature that illustrate the main ideas.

Lying and deception are common English words. I hope that the definitions that I provide are broadly consistent with common usage, but it is more important to highlight distinctions between aspects of communication that depend only on the existence of a common language from those that depend on strategic considerations.

2 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, $\theta$ from a set $\Theta$ according to a distribution $\mu$. The Sender observes $\theta$, takes an action $x \in X$, and sends a message $m \in M$. The Receiver observes $m' \in M'$ and makes a decision $y \in Y$. $U^i(\theta, x, y, m)$ is the payoff that Player $i$ receives given state $\theta \in \Theta$, actions $(x, y) \in X \times Y$, and message $m \in M$. I make a formal distinction between $m$ and $x$ because I further assume that $U^R(\cdot)$ is independent of $m$ (neither utility function depends directly on $m'$).

Unless I say otherwise, assume that $\Theta$, $X$, $Y$, $M$, and $M'$ are finite. A function $P(\theta)$ gives the probability that the state is $\theta$. $P(\cdot)$ is positive and satisfies $\sum_\theta P(\theta) = 1$. The Receiver’s observation is a (potentially)
stochastic function of $m$. The function $\nu(m' \mid m)$ is the probability that $R$ observes $m'$ given that $S$ sends the message $m$. For each $m$, $\nu(\cdot \mid m) \geq 0$ and $\sum_{m' \in M'} \nu(m' \mid m) = 1$.

The definition of lying requires a comparison between the Sender’s message $m$ and the true state of the world $\theta$. I assume that the message space $M$ contains ways to describe a rich set of events (subsets of $\Theta$). In particular, for each $\Theta_0 \subset \Theta$, there exists a message $m_{\Theta_0} \in M$ and there is a common understanding that $m_{\Theta_0}$ means “$\theta \in \Theta_0$.” Let $m_{\theta_0}$ denote the message corresponding to the set $\{\theta_0\}$. To simplify discussions, assume that there is exactly one way to describe each subset. In particular, if $\Theta_0 \neq \Theta'_0$, then $m_{\Theta_0} \neq m_{\Theta'_0}$. When the message $m$ is equal to $m_{\Theta_0}$ for some $\Theta_0$, we say that $m$ has an accepted meaning. There may be messages that have no accepted meaning. The purpose of including messages that have conventional meanings is to make it possible to describe lies. I do not require the Sender to use a message in a conventional way or for the Receiver to interpret it in a conventional way. The model is sufficiently general to permit the possibility that using messages in a way that violates their conventional interpretation is costly.

A simple special case of the model is a cheap-talk game in which $X$ is empty; $m = m'$ so that the Receiver observes the Sender’s message without noise; and neither $U^S(\cdot)$ nor $U^R(\cdot)$ depend on $m$ directly.

One feature of the model adds notation, but I rarely exploit the extra generality in this paper. For the most part, I assume that the Receiver observes the Sender’s message so that $m = m'$. I use the possibility that $m \neq m'$ only in the discussion of Hendricks and McAfee [9] in Section 10.

There are several ways to extend the definitions. In some situations it might be useful to study models in which the Sender does not observe $\theta$, but instead learns a noisy signal. Extending the model to allow this possibility is straightforward. While it is standard to assume that there is a common-knowledge information structure ($P(\cdot)$), this restriction makes less sense 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 $\theta$. In such a situation, the concepts that I introduce should be viewed from the perspective of the Sender. That is, the model.

\footnote{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.}
applies to a situation in which different players have different beliefs. The beliefs that appear in definitions will be those of the Sender. Finally, I limit attention to two-player games for simplicity. Extending the definitions and basic insights to games with more players is straightforward.

I assume that the language associates words to subsets of \( \Theta \). It is possible that the language has words for probabilistic statements ("I believe that \( \theta \) is equally likely to be 0 or 1."). I chose to consider messages that are identified with subsets of \( \Theta \) because in the other cases it is difficult to identify lies.

3 Lying: Definitions

In this section I present consistent definitions of lying that illustrate the subtleties underlying the intuitive concept. There are more ambitious and systematic attempts to present definitions of lying. I do not attempt to review this large literature.

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

**Definition 1 (Lying)**

1. The message \( m \) is a lie given \( \theta \) if \( m = m_{\Theta_0} \) and \( \theta \notin \Theta_0 \).

2. The message \( m \) is a lie of omission given \( \theta \) if \( m = m_{\Theta_0} \) and \( \{\theta\} \subset \Theta_0 \).

3. The message \( m \) is true given \( \theta \) if \( m = m_\theta \).

When the Sender sends a message \( m \) that has an accepted meaning (\( m = m_{\Theta_0} \)), she can be inaccurate in two different ways. Either the message is never consistent with (what the Sender believes to be) the state or the message is consistent with states that the Sender believes are impossible in addition to the true state. In the first case, any inference compatible with the natural meaning of the Sender’s statement is not consistent with what \( S \) believes to

\[ ^2 \text{Mahon’s article in the Stanford Encyclopedia of Philosophy} \] carefully examines the strengths and weaknesses of several possible definitions of lying. I share the reaction of Morris 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 meaning of lying without providing a definitive definition.
be true. In the second case, the statement omits information because the statement is consistent with certain states that $S$ thinks are impossible.\footnote{If $S$ receives noisy information a third kind of lie is possible. 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 clearly neither a lie of omission nor the truth. This possibility does not arise when $S$ observes $\theta$ perfectly.}

Formally what the Receiver hears may be an element in a different space ($M'$) than the Sender’s message. While the definition of lying requires only that $S$ believe that there is an accepted connection between what she says and the true state, evaluating the consequences of lies depends on others sharing this interpretation. It is not difficult to imagine situations in which $m' \neq m$. Leading examples include situations in which the Sender and Receiver have different degrees of linguistic skills or speak different dialects. In these situations, the definition describes what constitutes lies in the mind of the Sender.\footnote{The perspective in the paper is that lies are based on the Sender’s interpretation of her message. Taking this perspective to the extreme, lying does not require common knowledge of the conventional interpretation of language. It is sufficient that the Sender has an interpretation of messages. So, for example, the statement “it is raining” is a lie when the speaker believes that it is not raining even if the audience does not understand English.}

The language may include messages that do not have an accepted meaning. If there is no $\Theta_0 \subset \Theta$ such that $m = m_{\Theta_0}$, then $m$ is not a lie. The possibility that $m$ takes on a meaning – $R$’s interpretation of $m$ – makes it possible for $m$ to be misleading. I discuss this possibility in more detail in the Section \textsection.

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. This distinction only arises when one permits the Sender to have inaccurate beliefs.

The definition of lying depends only on the support of the distribution of the state. Suppose that there are two states, $\theta_1$ and $\theta_2$ that are equally likely ex ante and that the Sender receives an informative, but imperfect, signal of the true state. If the only statements that have literal 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 $\theta_1$,” when she believes (only) that $\theta_1$ is highly likely. One can remedy this problem if the Sender has access to a richer language that permits her to make statements about her posterior (“state $\theta_1$ is much more likely than $\theta_2$”). It is difficult to evaluate whether statements like this are honest in one-shot settings.

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 natural language, 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 separates lies from their consequences.

The literature on strategic communication informally uses lying in a way that it consistent with my definition. It is common in these papers to 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; or the Receiver is a behavioral type who interprets messages literally (believing that the state is $m$ after receiving the message $m$); or perturbing preferences to include lying costs, which are defined in terms of the difference between the true state and the message. Fischbacher and Föllmi-Heusi [7] and Gneezy [8] are examples of experimental papers on communication link the message to the state and treat messages as lies if they are not equal to the state.

4 Other Definitions

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

St. Augustine [1, page 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. I agree that consequences are important, but I believe that it is useful to have a taxonomy that

\[5\]


Augustine appears to say that it is never right to lie, but he also gives some flexibility as dishonest statements said in jest would not constitute a lie according to his definition, but would be lies according to my taxonomy.

6

\[6\]
evaluates the veracity of statements without investigating the interpretation of the statements.

Coleman and Kay [3] 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 that is known to be true that is uttered with no intention to deceive to a false statement that is 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. They find that each of the criteria contribute 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 our definitions to include items from the first category (which would be the same as the second category when the Sender’s beliefs are correct). The third category, however, I consider to be deception.

5 Lies in Cheap-Talk Games

This section investigates lying in cheap talk games with a common language. Some of the results impose structure on the game. In a simple cheap-talk game, \( \Theta = [0, 1] \), and \( Y = \mathbb{R} \); 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. Assume in addition that \( y^S(\theta) > y^R(\theta) \) for all \( \theta \); and the prior on states has a positive, continuous density. The game has a common language if \( M \) consists of all subsets (or, at

---

\[7\] Subjects reported whether they thought the statement was a lie and how confident they were that others would agree.

\[8\] I suppress the irrelevant arguments \((m \text{ and } x)\) in the notation.
least, all Borel subsets) of \([0, 1]\) and the conventional meaning of the message \(C\) is “\(\theta\) is an element of \(C\).” Preferences are independent of the messages sent and received.

The following properties are simple consequences of facts about equilibria of cheap-talk games. For the most part, they follow from well known properties of equilibria in simple cheap-talk games.

**Proposition 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. Proposition 1 notes this. Cheap-talk games always have a babbling equilibrium in which all Senders send the same message with probability one and the Receiver’s response does not depend on the message. In equilibrium, if Senders use a message that has a conventional meaning of the form “the state is \(\theta_0\),” all but one type of Sender tells a lie.

An equilibrium is non-trivial if the Receiver takes more than one action with positive probability in equilibrium. That is, non-trivial equilibria are those in which the Sender does not babble. When an equilibrium is non-trivial, it is possible to relabel messages so that every Sender is lying. To establish this result, we use the following simple observation.

**Lemma 1** If \((y^*, m^*)\) is an equilibrium of a cheap talk game, and \(\pi : M \rightarrow M\) is a bijection, then \((\tilde{y}^*, \tilde{m}^*)\) is also an equilibrium where \(\tilde{y}^*(\pi(m)) \equiv y^*(m)\) and \(\tilde{m}^*(\theta) = \pi(m^*(\theta))\).

Lemma 1 states that the meaning 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. Lemma 1 follows from the definition of equilibrium.

**Proposition 2** In a cheap-talk game with a common language, any non-trivial equilibrium type-action distribution\(^9\) can be induced by an equilibrium in which each agent’s message is a lie.

\(^9\)The type-action distribution associated with a mixed-strategy profile \((\alpha, \sigma)\) is a probability distribution over \((\theta, y)\) pairs where \(\alpha(y \mid m)\sigma(m \mid \theta)\) is the density of \((\theta, y)\).
In a simple 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. Let \( \phi(\theta) \) be the partition element containing \( \theta \). Hence any equilibrium type-action distribution can be generated by an equilibrium that uses only a finite number of messages with positive probability. In the equilibrium, if \( \theta', \theta'' \in \phi(\theta) \), 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 \( \theta'' \) for \( \theta' \not\in \phi(\theta) \). This establishes Proposition 2 for simple cheap-talk games.

Notice that if all messages are true, then the equilibrium must be fully separating (each type \( \theta \) sends the message \( m_\theta \)). Hence when separating equilibria no not exist, there must be a positive probability of lying in equilibrium.

Proposition 3 In a simple cheap talk game, there is a positive probability of lying every equilibrium.

Proposition 3 is not true for all cheap-talk games. In particular, it is not true when the Sender and Receiver have identical preferences.

Proposition 4 In a simple cheap-talk game with a common language, every equilibrium type-action distribution can be supported as an equilibrium with only lies of omission.

Proposition 4 states that one can view any equilibrium as one in which agents report 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 \([\theta_i, \theta_{i+1}]\).” Proposition 4 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 rich enough to include descriptions of all probability distributions over types.

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 conventional meaning of words. For example, Kartik and Sobel [12] provide conditions under which 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). In this outcome, agents
are systematically dishonest: They use messages that exaggerate their type. The existence of a common language does provide structure on the way that players use and interpret messages in equilibrium. No one is fooled, but for strategic reasons, no agent reports honestly.

While lying is part of the equilibrium of cheap-talk games with a common language, the ability to lie is damaging to both players in the sense that the Receiver always prefers the Sender to be honest and, ex ante (at least for the class of simple signaling games satisfying a monotonicity condition studied in Crawford and Sobel [5]) the Sender obtains a higher expected payoff when the Receiver is fully informed than in any equilibrium.

6 Honest Equilibria

Honesty is typically incompatible with strategic behavior, but if there is no conflict of interest between the players, there is a possibility that the Sender will report honestly in equilibrium.

Given $\theta$, let $y^R(\theta, m)$ be a solution to $\max U^R(\theta, y, m)$ and let $m(\theta)$ solve:

$$\max U^S(\theta, y^R(\theta, m), m).$$

(1)

Hence $m(\theta)$ is the set of messages that maximize $S$’s utility assuming that the message is revealing and $R$ responds with his optimal action. Assuming that (1) has a unique solution, it is immediate that an honest equilibrium exists only if $m(\theta) \neq m(\theta')$ whenever $\theta \neq \theta'$. When $m(\cdot)$ is single valued and one-to-one, the game is potentially revealing. It is straightforward to provide condition on $U^i(\cdot)$ under which a game is potentially revealing. For example, if $y, m, \theta$ are all elements of the real line, $U^R(\cdot)$ is differentiable, and $U^R(\theta, y, m) = f(y) + g(y, \theta)$ and $g_{12}(\cdot) > 0$, then the condition holds.

It is immediate that

Proposition 5 In any potentially revealing game in which $U^S = U^R$, there exists a specification of language under which an honest equilibrium exists.

Proposition 5 gives strong conditions under which honesty is compatible with strategic behavior. Importantly, the proposition requires that players have identical preferences. In a potentially revealing game with common

10If 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.
interests, an honest equilibrium exists if the meaning of \(m(\theta)\) is “my type is \(\theta\).”

Two aspects of the proposition are worth noting. First, honesty is defined with respect to a common language. The result only states that there exists a language compatible with honest reporting. In general, there will exist equilibria in which the Sender is dishonest. There may be contexts in which this language is inconsistent with equilibrium. Second, the equilibrium constructed is efficient. When there is conflict of interest, there will a temptation to lie for strategic advantage. Common interests are not necessary for an honest equilibrium, but significant conflict rules out honesty. A weaker sufficient condition for the existence of an honest equilibrium in a potentially revealing game is that \(\theta = \theta_0\) solves \(\max U_S(\theta_0, y_R(\theta), m(\theta))\) for all \(\theta_0\). Even this condition is not necessary. It is possible to construct an honest equilibrium if the condition fails, but it involves an inefficient match of actions to messages.

\section{Accurate Beliefs}

The definition of lying depends only on the existence of conventional meanings of words. It makes no reference to how \(S\)'s statements might influence \(R\). To describe these features, I introduce several notions of deception. Before doing so, I need to investigate how \(S\)'s behavior influences \(R\)’s beliefs. Assume that each \(m'\) induces a posterior distribution \(\mu(\theta | m')\). \(\mu(\cdot | m')\) is the posterior beliefs of the Receiver given the message \(m'\). In the last two parts of the definition, \(R\) forms beliefs by taking into account the Sender’s mixed strategy, \(\sigma(\cdot); \sigma(m | \theta)\) is the probability that the Sender with type \(\theta\) sends the message \(m\). Hence, \(\sigma(m | \theta) \geq 0\) and \(\sum_{m \in M} \sigma(m | \theta) = 1\) for all \(\theta\).

\textbf{Definition 2 (Properties of Beliefs)}

1. The belief \(\mu(\cdot | m')\) is \textbf{completely inaccurate} given \(m'\) and \(\theta\) if \(\mu_R(\theta | m') = 0\).

2. The belief \(\mu(\cdot | m')\) is \textbf{inaccurate} given \(m'\) and \(\theta\) if \(\mu_R(\theta | m') \in (0, 1)\).

3. The belief \(\mu(\cdot | m')\) is \textbf{accurate} given \(m'\) and \(\theta\) if \(\mu_R(\theta | m') = 1\).
4. Given the mixed strategy $\sigma(\cdot)$ of $S$, the belief $\mu(\cdot \mid m')$ is rational given $m'$, $\theta$, and $\sigma(\cdot)$ if

$$
\mu^R(\theta \mid m') = \frac{\sum_m \sigma(m \mid \theta) \nu(m' \mid m) P(\theta)}{\sum_m \sum_{\theta'} \sigma(m' \mid \theta') \nu(m' \mid m) P(\theta')}
$$

whenever

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

5. The belief $\mu(\cdot \mid m')$ is rational given $\sigma(\cdot)$ if (2) holds for all $m'$ and $\theta$ whenever (3) holds.

I believe that the standard definition of deception is essentially “an action that induces inaccurate beliefs.” \footnote{The relevant definition from the Oxford English Dictionary is “to cause to believe what is false; to mislead as to a matter of fact, lead into error, impose upon, delude.”} I choose to incorporate goal-oriented behavior in my definition.

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 structure $\nu(\cdot)$) that determines the distribution of messages $m'$ that $R$ hears given that $S$ sends the message $m$.

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. Assume that the game has a common, noise-free language in which $M = M'$ and $m = m'$, so that the Receiver hears the Sender’s message, and for each subset of states $\Theta_0 \subset \Theta$ there exists $m_{\Theta_0} \in M$ that has the conventional meaning “$\theta \in \Theta_0.”$

Definition 3 In a communication game with a common noise-free language, the Receiver is credulous if $\mu^R(\cdot \mid m_{\Theta_0})$ is equal to the posterior distribution conditional on $\theta \in \Theta_0$. That is

$$
\mu^R(\theta \mid m_{\Theta_0}) = \begin{cases} 
0 & \text{if } \theta \notin \Theta_0 \\
\frac{P(\theta)}{\sum_{\theta' \in \Theta_0} P(\theta')} & \text{if } \theta \in \Theta_0.
\end{cases}
$$
If the Receiver is credulous, there is a natural connection between lying and inaccurate beliefs.

**Proposition 6** Given a communication game with a common noise-free language,

1. If $m$ is a lie given $\theta$ and $R$ is credulous, then $\mu^R(\cdot \mid m)$ is completely inaccurate given $m$ and $\theta$.

2. If $m$ is a lie of omission given $\theta$ and $R$ is credulous, then $\mu^R(\cdot \mid m)$ is inaccurate given $m$ and $\theta$.

3. If $m$ is true given $\theta$ and $R$ is credulous, then $\mu^R(\cdot \mid m)$ is accurate given $m$ and $\theta$.

Proposition 6 follows immediately from the definitions.

Proposition 6 connects beliefs to lies when the Receiver is credulous. Alternatively, we can relate the concepts when the Sender is honest and uses the strategy

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

**Proposition 7** Given a communication game with a common language, if $S$ is honest 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 lies of omission.

**8 Deception: Definitions**

Informally, the Sender deceives the Receiver by inducing inaccurate beliefs. More formally, I am interested in how the Sender’s message influences the Receiver’s payoff. To state the definition, I need to describe what utility the Sender believes the Receiver will get following the message $m$. This utility depends on two things, the action the Receiver takes and the state. Let $y(m')$ be the Receiver’s response to the message $m'$. Let $\bar{u}^R(\theta, x, m) = \sum_{m'} U^R(\theta, x, y(m')) \nu(m' \mid m)$ be the Receiver’s expected utility when $S$ takes

\footnote{More precisely, the action the Sender believes that the Receiver takes.}
action \( x \), sends the message \( m \), and the true state is \( \theta \). Interpret this as the Sender’s evaluation of the Receiver’s payoff. (The definition allows the message that \( R \) receives \( m' \) to be a stochastic function of the message sent.) Note 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 4 (Deception)** The message \( m \) is deceptive given \( \theta \) and \( x \) if there exists a message \( n \) such that \( \bar{u}^R(\theta, x, m) < \bar{u}^R(\theta, x, n) \).

The definition states that a message is deceptive if another message is available that would induce \( R \) to take a superior decision, where “superior” means that the action leads to a higher expected utility for \( R \) given what the Sender knows. A message is deceptive if the Sender thinks it will lead the Receiver to make an inferior decision. The Sender could be wrong about this, either because she has incorrect beliefs about the Receiver’s strategy or because she has incorrect beliefs about the state of the world.

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 version of the Sender have different preferences (as in dual self models or when there is time inconsistency).

Deception requires that the Sender’s message can influence the Receiver’s action. It is possible that the Receiver plans to take the same (suboptimal) action independent of what the Sender does. In this case, the Receiver may do badly, but the Sender is not responsible.

Since deception requires the Sender’s message to (sometimes) influence the Receiver’s behavior, simple cheap-talk games always have a non-deceptive equilibrium: Proposition 8 implies that there is no deception in a babbling equilibrium.\[13\]

**Proposition 8** There is no deception if the Receiver’s actions are independent of the message received.

In particular, Proposition 8 implies that there is no deception in a pooling equilibrium when the Receiver responds to off-the-path messages with the

\[13\text{In a babbling equilibrium, the Receiver takes the same action independent of the Sender’s message.}\]
same equilibrium action. There is another familiar context in which there is no deception in equilibrium.

**Proposition 9** Suppose that the Receiver hears the Sender’s message perfectly. There is no deception in a separating equilibrium.

Proposition 9 follows because the Receiver takes the best action given the true state following any message on the equilibrium path in a separating equilibrium. Propositions 8 and 9 apply to communication games in general.

**Proposition 10** Suppose that the Sender and Receiver have identical preferences, there is no deception in equilibrium.

In equilibrium, $S$ selects her message $m$ to maximize:

$$\sum_{m'} U^S(\theta, x, y(m'))\nu(m' | m).$$

When $U^S = U^R$, an equilibrium message must maximize $\bar{u}^R(\theta, x, m)$, which means that the message is not deceptive.

It is tempting to conjecture that deception is not possible in equilibrium. The truth of the conjecture depends, of course, upon the definition of deception. If one equates deception to inducing non-rational beliefs, then deception is inconsistent with equilibrium. My definition is different. Deception, as I have defined it, is consistent with equilibrium. Here is an example. Take an equilibrium in a simple cheap-talk game in which more than one action is induced and the Sender has a uniformly positive bias, that is, 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, the Receiver strictly prefers the lower action. Hence when the Sender’s type is slightly greater than an indifferent type, the Receiver would strictly prefer the Sender to send a non-equilibrium message (the message immediately lower than the equilibrium message). 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 take 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 deception does arise in informative equilibria of cheap-talk games. Since the Receiver
prefers informative equilibria to the babbling equilibrium, this means that allowing deception can be beneficial to the Receiver.

I refer to \( m \) as a message. Unlike lies, which are defined with reference to a common language, the Sender can deceive the Receiver through acts and not words. The Sender’s choice \( m \) need not be “communication.” It is important for the interpretation that \( m \) does not enter directly into the Receiver’s payoff function. In general, the Sender’s efforts to maximize her own payoff may lead her to take actions that lower the Receiver’s payoff. If I applied the definition to cases like this, then the proposer in a dictator game is deceptive if she takes any part of the pie for herself.

It is standard to view the Receiver’s reasoning process in two steps. He first uses the message he receives to update his beliefs and then selects an action to maximize his preferences given these beliefs. Hence it may be useful to think of deception as a property of beliefs.

In the following definition, \( I(\cdot \mid \theta) \) is the probability distribution that places probability one on \( \theta \). That is,

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

**Definition 5 (Misleading Messages)** The message \( m \) is misleading given \( \theta \) if there exists \( n \) and a \( p \in [0, 1) \) such that

\[
\mu^R(\cdot \mid n) = p\mu^R(\cdot \mid m) + (1-p)I(\cdot \mid \theta).
\]

The message \( m \) is truly misleading given \( \theta \) if there exists \( n \) and a \( p \in [0, 1) \) such that

\[
\mu^R(\cdot \mid n) = p\mu^R(\cdot \mid m) + (1-p)I(\cdot \mid \theta) \quad \text{and the Receiver’s best response to } m \text{ is not a best response to } n.
\]

A message \( m \) is misleading 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” means on a segment connecting \( S \)’s beliefs to the beliefs induced by \( m \). Misleading messages induce inaccurate beliefs. In general, this definition is quite restrictive. When there are two states, the set of possible beliefs is one dimension and if \( m \) and \( m' \) induce different beliefs, then one of the beliefs will be closer to \( S \)’s beliefs than the other. When there are more states, however, it is unlikely for three beliefs to be co-linear. So misleading beliefs are unusual. While other notions of closeness are possible, the natural way to measure closeness must make reference to \( R \)’s preferences, which leads back to the definition of deception.

In two-state models, misleading messages arise in familiar situations. Assume \( \Theta = \{0, 1\} \) and consider a game in which the Sender plays a partially
pooling strategy, sending message $m_p$ whenever $\theta = 0$ and sending messages $m_p$ and $m_s$ with positive probability when $\theta = 1$. In this case, the rational Receiver has accurate beliefs when he hears $m_s$ and inaccurate beliefs when he hears $m_p$. The message $m_p$ is therefore deceptive given $\theta = 1$. This example demonstrates that it is possible to mislead a rational Receiver.

Merely having inaccurate beliefs is not necessarily a sign of deception. Suppose for example that $\mu^R(\cdot \mid m)$ does not depend on $m$. In this case, $S$’s message does not influence $R$’s beliefs. Consequently while $R$’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.

There is a formal connection between inaccurate beliefs and deception. Suppose that $S$ receives the signal $\theta$ and can induce beliefs of the form $pI(\cdot \mid \theta) + (1 - p)\gamma$, where $\gamma$ is an arbitrary distribution over states. Let $y(p)$ be a Receiver-optimal response to these beliefs. I claim that $R$’s expected utility is an increasing function of $p$.

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

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

The Appendix contains proofs of these two technical results.

The next result is a consequence of the comment following Lemma 2 and Lemma 3. It connects misleading messages to deception.

**Proposition 11** If $m$ is truly misleading, then it is deceptive. If a message is deceptive independent of the specification of $R$’s preferences, then it is misleading.

Proposition 11 equates misleading messages and deception. It is a weak result because it is difficult for a message to be misleading when there are more than two states.

---

\[14\text{It is possible that } S'\text{’s message influences } R'\text{’s action, but if } R \text{ responds optimally to beliefs, the message will not influence } R'\text{’s utility.}\]
9 Bluffs

Informally, deception involves purposeful behavior by $S$ that misleads $R$ and, because of this, lowers $R$’s payoff. That is, when the Sender deceives the Receiver, the Receiver loses. There are situations in which the Sender will distort her behavior in order to influence the information that the Receiver obtains. I propose to evaluate this kind of behavior in terms of utility losses of the Sender rather than the Receiver.

Definition 6 (Bluff) Given a communication game, the message $m$ is a bluff given $\theta$ if there exists $m' \neq m$ such that

$U^S(\theta, y^R(\theta, m), m) < U^S(\theta, y^R(\theta, m'), m')$.

If $R$ knows the state, then, following the message $m$, he will take action $y^R(\theta, m)$ (I allow for the possibility that the Receiver’s response depends directly on the message) and the Sender’s utility is $U^S(\theta, y^R(\theta, m), m)$. The definition states that $S$ is bluffing if she does not send the message she would send if $R$ were fully informed. A strategic Sender would send a message $m^*$ that maximizes $U^S(\theta, y^R(\theta, m), m)$.

Bluffs and deception are clearly different in this formulation: The Sender’s preferences determine what is a bluff. The Receiver’s preferences determine what is deception.

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. If the bluffer’s opponents knew the true quality of her hand, they would be inclined to “call” the bet. The bluffer would lose. A bluff is successful precisely because the opponents lack information. Even rational opponents will associate large bets with good hands and might “fold,” allowing the bluffer to win. Lies need not be part of the bluff. Bluffing need not be deceptive.

It is apparent that bluffing is not possible in perfect information games, where, by definition, the Receiver knows what the Sender knows. Bluffing is also impossible in cheap-talk games.

Proposition 12 There are no bluffs in cheap-talk games.

Proof. In a cheap-talk game, if $R$ learns $\theta$, then $S$’s payoff is $U^S(\theta, y^R(\theta))$ independent of message. Hence $S$ would not strictly gain by changing her message.
Proposition 12 demonstrates that the Sender can deceive without bluffing, because deception is possible in cheap-talk games. Conversely, it is possible to bluff without deception. Consider the separating equilibrium in a standard Spence signaling game. There is no deception, because the Receiver learns the Sender’s type. On the other hand, the Sender’s message is a bluff (according to my definition). If the Receiver knew the Sender’s type, then there would be no reason for the Sender to invest in costly signaling. Notice that when education does not add to productivity, the Sender is bluffing whenever she invests a strictly positive amount. In general, it is a bluff to over-invest.

If communication is costly, bluffing may arise even if the Sender and Receiver have common preferences. Suppose, for example, that types, messages, and actions are elements of \( \{0, 1\} \) and \( U(y, \theta, m) = -(y - \theta)^2 - .1m \). 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 game also has a pooling equilibrium in which there is no bluffing. Bluffing need not arise if one added a round of pre-play, cheap-talk communication.

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.\(^{15}\)

10 Examples of Deception

This section describes four examples of models of deception in the literature and fits them into the paper’s framework.

Hendricks and McAfee \(^9\) 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 \( \hat{q}U(y_1) + (1 - \hat{q})U(y_0) \).

\(^{15}\)The strategy will be a bluff if (a) the Receiver will change his behavior if he know the realization of the Sender’s strategy and (b) the Receiver’s choice influences the Sender’s payoffs.
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, convince the Receiver to direct his resources to the less likely front. In equilibrium the Sender balances these two incentives. Hendricks and McAfee 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 \( q \)) 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). 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. Second, the Receiver’s payoff does not depend on the Sender’s message.

Crawford [4] 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 \( 2 \times 2 \) 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 non-degenerate 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, for the Sender’s statement to convey no meaning, and for the Sender to play her equilibrium strategy in the underlying game after any message she makes. According to my definitions, the equilibrium involves lying (because the Sender’s statement does not describe her intentions), no deception (because the Receiver’s beliefs do not depend on the Sender’s statement), and no bluffing (because the Sender has no private information).

Crawford analyzes the game assuming that the players best respond to beliefs that are not necessarily accurate. He concentrates on a small number of plausible behavioral types: Senders who always honestly reveal their inten-
tions; 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. For fixed fractions of behavioral types, he characterizes the equilibrium behavior of sophisticated agents, who respond optimally to the (non-strategic) 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. Including behavioral types transforms a message that describes intentions in the original game into one about private information. Nevertheless, to the extent that there is a common language, it is natural to interpret messages as statements about intentions rather than statements about type. Under this interpretation, for interesting parameter values the Sender will lie and the lies will mislead a fraction of Receivers. These Receivers will be deceived by the Sender.

Deception arises in Crawford’s model because some agents have incorrect beliefs and other agents exploit these beliefs. Ettinger and Jehiel [6] present an alternative model of deception based on the same general idea. Ettinger and Jehiel provide a simple example that illustrates their model. They analyze a model based on a $2 \times 2$ zero-sum game $\Gamma$, 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 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 mislead the Column player. Specifically, they identify an equilibrium in which the rational Row player plays $U$ in the first period and plays $D$ in period 2 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 plays $R$ in the second period if Row played $U$ and $R$ otherwise. 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. Deception involves second-period payoffs, which do not directly depend on Row’s first-period choice. The coarse Column player

21
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 $2/3$ following $U$. This belief is objectively incorrect because Row plays $U$ in the first period independent of type. Column 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 $U$ – 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 Rational Row can influence what Column expects by his first-period choice. If 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 deceptive when 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, Rational Row is bluffing when he plays $U$ in the first period because he would not follow that strategy if Column knew Row’s intentions.

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>5, -5</td>
<td>3, -3</td>
</tr>
<tr>
<td>D</td>
<td>0, 0</td>
<td>7, -7</td>
</tr>
</tbody>
</table>

Kartik, Ottaviani, and Squintani [11] study a model of communication with costly lying. They identify conditions for separating equilibria in a model of 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 is a tractable way to model costly lying. In this equilibrium, the Receiver is fully informed in equilibrium and the Sender’s messages are not misleading. According to my definition, there is no deception. In another interpretation, a fraction of Receivers are credulous (and believe that the
Sender honestly reports the true state). In a separating equilibrium, the credulous Receivers are misled and the Sender’s strategy is deceptive.

Kartik, Ottaviani, and Squintani [page 95, Footnote 3]\footnote{11} “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.” Deception in this interpretation is distinct from the notion that a player may choose not to disclose private information in order to exploit the imprecise – but not incorrect – belief induced in a counterpart (e.g., bluffing in poker).” 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 is disadvantageous to the Receiver, but need not be 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.
Appendix

Proof of Lemma 2. Let $V(p) \equiv U^R(\theta, y(p)) + (1 - p)EU^R(\theta, y(p))$, where the expectation is with respect to $\gamma$. Let $p' > p$. It follows that
\[ V(p) \geq U^R(\theta, y(p')) + (1 - p)EU^R(\theta, y(p')) \quad (4) \]
and so
\[ V(p) \geq V(p') - (p' - p) \left( U^R(\theta, y(p')) - EU^R(\theta, y(p')) \right) \quad (5) \]
and hence
\[ U^R(\theta, y(p')) - EU^R(\theta, y(p')) \geq \frac{V(p') - V(p)}{p' - p}. \]
Similarly,
\[ \frac{V(p') - V(p)}{p' - p} \geq U^R(\theta, y(p)) - EU^R(\theta, y(p)) \]
so that
\[ U^R(\theta, y(p')) - U^R(\theta, y(p')) \geq EU^R(\theta, y(p')) - EU^R(\theta, y(p)). \quad (6) \]
The definition of $y(p')$ implies that
\[ p'U^R(\theta, y(p')) + (1 - p')EU^R(\theta, y(p')) \geq p'U^R(\theta, y(p)) + (1 - p')EU^R(\theta, y(p)) \]
so either $U^R(\theta, y(p')) \geq U^R(\theta, y(p))$ or $EU^R(\theta, y(p')) \geq EU^R(\theta, y(p))$. It follows that the left-hand side of inequality (6) is positive, which establishes the first part of the result. If $y(p')$ is not a best response to $p\mu^S(\theta) + (1 - p)\gamma$, then the inequalities in (4) and (5) are strict, which establishes the second part of the lemma.

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 $\theta$. The lemma follows because, by the separating hyperplane theorem, it is possible to find $v(\cdot)$ such that
\[ \sum_\theta v(\theta)\gamma(\theta) = \sum_\theta v(\theta)\mu^S(\theta) > 0 > \sum_\theta v(\theta)\gamma'(\theta). \quad (7) \]
References


