Equilibrium Refinement in Signaling Games as Truth Conditions of Counterfactuals*

Christina Pawlowitsch†

June 21, 2018

*I would like to thank Hari Govindan and Joel Sobel for comments and discussion.

†LEMMA–Laboratoire d'Économie Mathématique et de Microéconomie Appliquée, Université Paris II, 4 Rue Blaise Desgoffe, 75006 Paris, France. E-mail: christina.pawlowitsch@u-paris2.fr

Abstract

Equilibrium refinement based on restrictions on beliefs "off the equilibrium path" can be related to Lewis's (1973) account of counterfactuals. In signaling games with two states of the world, two signals, and two actions in response to signals, "forward induction" (Govindan and Wilson 2009), which for this class of games coincides with "divinity" (Banks and Sobel 1987), is equivalent to Lewis's accessibility condition relying on the similarity between the actual world and other possible worlds. The formal results are illustrated in a game-theoretic model of communicative implicatures driven by politeness.

1 Introduction

In a game in *extensive form*—a game given by a tree—what can be an equilibrium typically depends on what players would do at a point in the game "off the equilibrium path," that is, a point in the game that can in principle be reached but that is not reached in the equilibrium under study. In classical, rationality-oriented game theory, what a player does at some point in the game has to have a foundation in what he or she believes about the possible states of the world and the strategy choices of the other players. Game theorists have pointed out a number of examples in which players' beliefs at points off the equilibrium path do not seem plausible—due to inferences that a player would draw at that point if it were reached—and they have proposed *refinements* of Nash equilibrium that operate on the principle of *restricting players' beliefs off the equilibrium path*.

A point in the game *off the equilibrium path* can be understood as a *counterfactual*, an event that could have happened but that did not happen in the sequence of events that makes up the *actual world*—if the actual world is taken to be the equilibrium under study. In this paper, I try to connect the game-theoretic research program of restricting beliefs off the equilibrium path to Lewis's (1973) account of counterfactuals.

Equilibrium refinement—placing extra conditions on Nash equilibrium beyond the requirement that one player's strategy has to be a best response to the other players' strategies—is an approach taken by game theorists in response to the observation that many games have multiple Nash equilibria. A class of games in which refinements based on restrictions on beliefs off the equilibrium path are productive are *signaling games*; in particular, signaling games in which the cost-benefit of using a signal is a function of the state of nature.

In signaling games, what is a belief "off the equilibrium path" is easily accessible to intuition: it is what someone would think in case that a signal were received that is in principle part of the game but that is never used in the equilibrium under study. Some of the most prominent belief-based refinements, like the "intuitive criterion" (Cho and Kreps 1987) and a criterion known as "divinity" (Banks and Sobel 1987), have been specifically developed for this framework.

This is also the framework that I adopt here to study the relation between belief-based refinements of Nash equilibrium and Lewis's theory of counterfactuals. More specifically, I will develop this connection for the simple class of games with two possible states of the world, two signals, and two possible actions for the receiver. Games of this structure, while being sufficiently simple to keep the problem formally tractable, are sufficiently rich to yield nontrivial results and interesting applications. I will show, for this class of games, a correspondence between "forward induction" (Govindan and Wilson 2009), which here coincides with "divinity" (Banks and Sobel 1987), and Lewis's criterion of accessibility of other possible worlds from the actual world that he uses to

evaluate truth conditions of counterfactuals. I conclude with an application of the ideas developed to communicative implicatures driven by politeness (Grice 1975, Brown and Levinson 1987).

2 Signaling games

Consider a simple signaling game with two states of nature, "high" and "low"; two signals, a marked signal $S^\star$ and an unmarked signal S; and two possible reactions to signals, "accept" (a) and "do not accept" ($\bar{a}$). Such a game in represented in figure 1a. The story that goes with a game of this form is the following: The first player, for example, a candidate for a job, knows the state of nature; he knows whether he is of the "high" or "low" productivity type. The second player, which in the example will be the employer, has no direct access to that information, but she has a probability assessment—a *belief*[1]—over the possible states of nature, represented by the probability p that she attaches to the first player being of the high type, and she can observe which signal the first player has emitted; for example, if the candidate has a certain degree $S^\star$ or not S.

In a game tree, different states of nature are represented by possible moves of nature. If a player at some point in the game has to take an action but has not observed some move of nature or a personal player before him, then this will be represented by placing all nodes that this player cannot distinguish in the same *information set*. In figure 1a, this in indicated by ovals: when the second player has to make her choice, she can see which signal the first player has emitted, but she still does not know the state of nature: the two nodes after $S^\star$, and respectively S, are therefore in the same information set.

A game form—a tree—alone does not define a game. To define a game in extensive form, one also needs to specify the payoffs that players have at each end node of the tree. An interpretation of the payoffs indicated in figure 1a is the following: Player 1 and 2 are related by a basic strategic conflict that can be called an *identification problem*: if the first player is of the high type, they have coinciding interests (if the second player takes a, they both get a base payoff of 1, and if the second player takes $\bar{a}$, they both get a base payoff of 0), but if the first player is of the low type, they have perfectly opposed interests (if the second player takes a, the first player gets a base payoff of 1 and the second player of 0, and vice versa if the second player takes $\bar{a}$). On top of these base payoffs, player 1, in his different types, incurs different costs for using different signals: the marked signal $S^\star$ bears a cost of production of c_1 to the first player's high type and of c_2 to the first player's low type, while the unmarked signal S can be produced with no cost by both types. Endowed with this payoff structure, the game in figure 1a can be seen as a discrete variant of Spence's (1973) model of signaling in the job market.

[1] I use the term *belief* for a player's probability assessment over possible states of nature, and the term *conjecture* for a player's probability assessment over strategy choices of some other player.

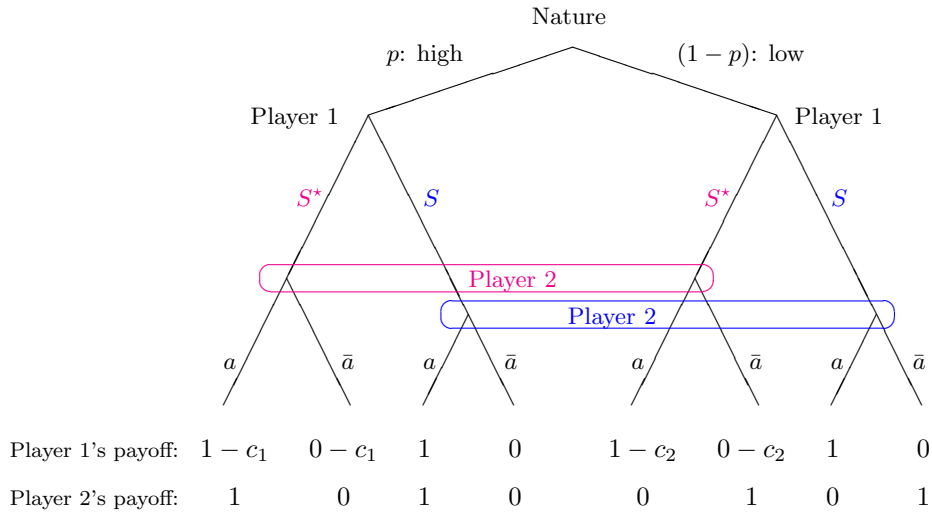


Figure 1a. A simple signaling game with differential costs of producing the signal: $0 < c_1 < c_2 < 1$.

	aa	$a\bar{a}$	$\bar{a}a$	$\bar{a}\bar{a}$
$S^\star S^\star$	$1 - pc_1 - (1-p)c_2,\ p$	$1 - pc_1 - (1-p)c_2,\ p$	$-pc_1 - (1-p)c_2,\ 1-p$	$-pc_1 - (1-p)c_2,\ 1-p$
$S^\star S$	$1 - pc_1,\ p$	$p(1-c_1),\ 1$	$-pc_1 + (1-p),\ 0$	$-pc_1,\ 1-p$
$SS^\star$	$1 - (1-p)c_2,\ p$	$(1-p)(1-c_2),\ 0$	$p - (1-p)c_2,\ 1$	$-(1-p)c_2,\ 1-p$
SS	$1,\ p$	$0,\ 1-p$	$1,\ p$	$0,\ 1-p$

Figure 1b. The normal form of the game in figure 1a. Nash equilibria for the case $0 < c_1 < c_2 < 1$:

- If $p < 1/2$, there is: (PS1) a *partially separating equilibrium* in which player 1 takes a mixed strategy between $S^\star S^\star$ and $S^\star S$ with a probability of $p/(1-p)$ on the first, and player 2 mixes between $a\bar{a}$ and $\bar{a}\bar{a}$ with a probability of $y = c_2$ on the first, and (P1) an equilibrium component in which player 1 takes SS and player 2 mixes between $a\bar{a}$ and $\bar{a}\bar{a}$ with a probability of $y \in [0, c_1]$ on the first.
- If $p > 1/2$, there is: (PS2) a *partially separating equilibrium* in which 1 mixes between SS and $S^\star S$ with a probability of $(1-p)/p$ on the first, and player 2 mixes between aa and $a\bar{a}$ with a probability of $1 - c_1$ on the first, (P2) an equilibrium component in which player 1 takes $S^\star S^\star$ and player 2 mixes between aa and $a\bar{a}$ with a probability of $y' \in [0, 1 - c_2]$ on the first, and (P3) an equilibrium component in which player 1 takes SS and player 2 any mix between aa and $\bar{a}a$.

Which outcomes of such a game can be regular patterns of interaction? To answer that question classical game theory relies on equilibrium analysis. A *Nash equilibrium* (Nash 1950, 1951) is a combination of strategies, specifying one strategy for each player, such that no player has an incentive to deviate; or to say it differently, such that each player's strategy is a *best response* to the other players' strategies.[2] *What are strategies for the two players in this game?* A strategy

[2]It is worth noting the "static" nature of this solution concept: every player chooses a best response to the other players' strategies *as if he or she had been told what the strategies of the other players were*. However, the concept of Nash equilibrium remains silent about how players get to their, ex-post correct, anticipation of the other players' strategies.

for the first player is a plan of action whether to produce the marked signal $S^\star$ or the unmarked signal S as a function of his type. For the second player, a strategy is a plan of action, a or $\bar{a}$, conditional on which signal she has observed. Each player then has four possible *pure strategies*:

Pure strategies for player 1:	Pure strategies for player 2:
$(S^\star, S^\star)$: If high, then $S^\star$; if low, then $S^\star$	(a, a): If $S^\star$, then a; if S, then a
$(S^\star, S)$: If high, then $S^\star$; if low, then S	$(a, \bar{a})$: If $S^\star$, then a; if S, then $\bar{a}$
$(S, S^\star)$: If high, then S; if low, then $S^\star$	$(\bar{a}, a)$: If $S^\star$, then $\bar{a}$; if S, then a
(S, S): If high, then S; if low, then S	$(\bar{a}, \bar{a})$: If $S^\star$, then $\bar{a}$; if S, then $\bar{a}$

Players' strategies can also be *mixed*, that is, in terms of a probability distribution over their respective set of pure strategies. To determine when one strategy is a best response to another, one needs to make assumptions about how the players in the game deal with the uncertainty that they face.

The standard approach to solve games with uncertainty is by *Bayesian Nash equilibrium* (Harsanyi 1967), an extension of the general notion of Nash equilibrium to games of incomplete information under the assumption that players maximize expected payoffs given the probabilities of the states of nature, which are assumed to be common knowledge among the players of the game. Under this assumption, one can calculate the representation of the game in *normal form*, which gives for every possible combination of complete contingent strategies the expected payoff for each player, given the commonly known probability distribution on the states of nature. Then, for any fixed value of p, the Bayesian Nash equilibria of the game can be calculated as the usual Nash equilibria in this normal form. Figure 1b shows the game of figure 1a in normal form and indicates its Nash equilibria.

In a classical game-theoretic perspective, where the idea is that players are rational, an equilibrium should be supported by how players' beliefs over the states of nature evolve in the course of play in response to what they observe the other players doing. For this purpose, players' strategies are modeled in terms of *behavior strategies*, that is, probability distributions over available actions for each information set, or node, where the respective player has to move. For example, a behavior strategy for the first player is: "If you are of the high type, use $S^\star$ with 60% and use S with 40%; if you are of the low type, use S for sure." A behavior strategy for the second player is, for example: "If you observe the costly signal $S^\star$, take a for sure; if you do not observe it, take a with a probability of 50% and $\bar{a}$ with 50%." Every mixed strategy in the normal form is equivalent to some behavior strategy in the sense that both give the same probability distribution over the end nodes of the tree for any fixed strategy profile of the other players (Kuhn 1953). The behavior strategy for the first player given above, obviously, is induced by a mixed strategy of $S^\star S$ and SS with a probability of 60% on the first and 40% on the second; that for the second player by a

mixed strategy of aa and $a\bar{a}$ with a probability of 50% on each of them.

Sequential Bayesian Nash equilibrium (Kreps and Wilson 1982) requires that players' choices of actions at the nodes or informations sets where they are called to move have to be supported by a *belief*—a probability distribution—over the states of nature such that players' beliefs are *consistent with Bayes' rule along the path being played* and players' choices are best responses to their beliefs and the other players' choices in the remainder of the game.[3] For the games discussed here, a *sequential Bayesian Nash equilibrium* is a profile of behavior strategies together with a vector of beliefs specifying for each information set a probability distribution over the two possible states of nature such that

- one player's behavior strategy is a best response to the other player's behavior strategy, and
- the second player's behavior strategy is *consistent with Bayes' rule along the path being played*, that is, her decision to take a or $\bar{a}$, when she comes to move, is a best response to her belief about the type of the first player that results from her prior belief by a Bayesian update, given the probabilities with which the first player uses $S^\star$ or respectively S as a function of his type as prescribed by the first player's strategy.

One recovers in this definition the typical "static" nature of Nash equilibrium: the second player has to update her prior belief *as if she knew the first player's strategy* (the probabilities with which the first player takes $S^\star$ and S depending on his type), and the first player has to make his choice based on his expectations about how the second player reacts to which signal. However, how the players come to know these probabilities is not part of the solution concept. Table 1 shows the sequential Bayesian Nash equilibria for the game in figure 1a.

Sequential Bayesian Nash equilibrium, it is important to notice, requires that players' choices are best responses to their beliefs and the other players' choices in the remainder of the game not only at nodes or information sets that are reached but also at nodes or information sets that are not reached in that equilibrium—nodes or information sets "off the equilibrium path." In this quality it translates the idea of *backward induction* (Selten 1975). For signaling games of the simple extensive form shown in figure 1a, this requirement does not add anything over and above the general notion of Bayesian Nash equilibrium. Without additional restrictions on beliefs (other than that they be compatible with Bayes' rule along the path being played), the sequential Bayesian Nash equilibria of the extensive form in figure 1a and the Nash equilibria of the normal form in figure 1b coincide.

[3]Kreps and Wilson (1982), in their definition of *sequential equilibrium* require a further condition concerning the *consistency* of player's beliefs off the equilibrium path, namely that they can be derived from Bayes' rule after a small perturbation of the behavioral strategies, which for signaling games is however trivially fulfilled (see the section on equilibrium refinements).

Table 1: Sequential Bayesian Nash equilibria of the game in figure 1, $0 < c_1 < c_2 < 1$

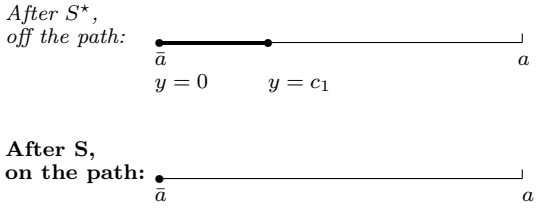
- If $p < \frac{1}{2}$:

(PS1) There is a *partially separating equilibrium* in which player 1's high type uses the costly signal $S^\star$ and player 1's low type uses it with a certain probability x (and does not use it with the complementary probability $1 - x$), namely, such that if player 2 observes the costly signal $S^\star$, her updated belief $p_{S^\star}$ will make her indifferent between a and $\bar{a}$, which will be the case if:

$$p_{S^\star} = \frac{p}{p + (1-p) \cdot x} = \frac{1}{2} \qquad \Longleftrightarrow \qquad x = \frac{p}{1-p}.$$

Player 2, if she does not observe the costly signal (S), chooses $\bar{a}$ (which will be a best response to her updated belief $p_S = 0$), and if she observes the costly signal $S^\star$, will choose a with probability $y = c_2$, which is the probability that will make player 1's *low type* indifferent between using the costly signal and not using it, while ensuring that using the costly signal is a best response for player 1 if he is of the high type. In this equilibrium, the absence of the costly signal (the unmarked S), fully reveals the low type. The costly signal $S^\star$ does not perfectly reveal the high type, but still will push up player 2 belief that player 1 is of high type, namely to $p_{S^\star} = 1/2$. In an equilibrium of this form, the lower the prior on the high type p, the lower has to be the probability x with which the low type sends the costly signal. Obviously this has to be so, because the lower p, the "stronger" the costly signal has to be in its discriminating between the two types.

(P1) Furthermore, there is an equilibrium outcome with *perfect pooling* in which both types of player 1 *never use the costly signal* (use the unmarked S). In such an equilibrium, player 2, when she does not observe the costly signal, will have the same belief as her prior belief, $p_S = p < 1/2$, and will therefore choose $\bar{a}$. If player 2 observes the costly signal $S^\star$—which will be a situation off the equilibrium path—she will either believe that player 1 is of the high type with a probability of less than $1/2$ and will choose $\bar{a}$, or will believe that player 1 is of the high type with a probability of $1/2$ (at which she will be indifferent between a and $\bar{a}$) and will choose a with a probability $y \in [0, c_1]$, which will be low enough so that player 1's high type, and a fortiori player 1's low type, has no incentive to deviate from his equilibrium strategy of using S.



- If $p > \frac{1}{2}$:

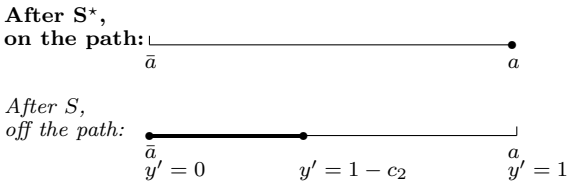
(PS2) There is a *partially separating equilibrium* in which player 1's low type *does not* use the costly signal, and player 1's high type *does not* use it with a certain probability $1 - x$ (and will send it with the complementary probability x), namely, such that if player 2 sees that the costly signal *has not been sent*, her updated belief p_S will make her indifferent between a and $\bar{a}$, which will be the case if:

$$p_S = \frac{p \cdot (1-x)}{p \cdot (1-x) + (1-p)} \quad \Leftrightarrow \quad 1 - x = \frac{1-p}{p}.$$

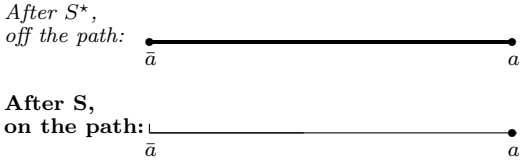
Player 2, if she sees the costly signal $S^\star$, chooses a (which will be a best response to her updated belief $p_{S^\star} = 1$), and if she does not see it, chooses a with probability $1-c_1$, which is the probability that will make player 1's *high* type indifferent between sending and not sending the costly signal, while ensuring that not sending the costly signal is a best response for player 1's *low* type. Under this equilibrium, the costly signal fully reveals the high type. Not sending the costly signal does not fully reveal the low type, but still will have the effect of bringing down player 2's belief that player 1 is of the high type, namely, to $p_S = 1/2$.

Furthermore, there are two equilibrium outcomes with *perfect pooling*:

(P2) One in which *both types of player 1 use the costly signal* $S^\star$, and therefore, player 2, when she observes the costly signal, will have the same belief as her prior belief, $p_{S^\star} = p > 1/2$, and will choose a. If the costly signal has not been sent, which will be "off the equilibrium path," player 2 either believes that player 1 is of the high type with a probability of less than $1/2$ and therefore will choose $\bar{a}$, or believes that 1 is of the high type with a probability of $1/2$ and chooses a with a probability $y' \in [0, 1-c_2]$, which will be low enough to prevent player 1's low type, and a fortiori player 1's high type, from deviating from his equilibrium strategy of using $S^\star$.



(P3) And one in which player 1 *never uses the costly signal* no matter what his type. In such an equilibrium, player 2, when she sees that the costly signal has not been sent, will have the same belief as her prior belief, $p_S = p > 1/2$, and hence will choose a, and in case that the costly signal has been sent, which will be "off the equilibrium path," can have any belief and best respond to it.



Signaling can be socially beneficial or a curse—depending on the prior p:

For any fixed value of p, the payoff of player 2 is the same in all equilibria that exist under this value of p: if $p < 1/2$, then player 2 has a payoff of $1 - p$; if $p > 1/2$, then p; and if $p = 1/2$, then $1/2$. The equilibrium payoffs vary only for player 1. Comparing the payoffs for player 1 then allows one to compare the equilibria in terms of their welfare properties.

- If $p < 1/2$, signaling is socially beneficial: player 1's low type gets the same payoff in the two equilibrium outcomes, PS1 and P1, namely 0; while the high type gets $c_2 - c_1 > 0$ in PS1, and 0 in the "no-signaling" equilibrium P1. In this situation, then, the possibility to signal is improving, in the sense of Pareto, over the situation in which no signaling takes place: nobody is made worse off, but at least someone (the high type of player 1) is made better off.
- If $p > 1/2$, the availability of a signaling mechanism creates a social dilemma: in PS2 both of player 1's types get $1 - c_1$; in P2 the high type gets $1 - c_1$, and the low type $1 - c_2$; and in P3 they both get 1. Both types of player 1 lose then in PS2 and P2 relative to the "no-signaling" outcome P3, in which here, due to the high prior, player 2 takes a. In PS2, this loss comes from the fact that the signaling that takes place makes it at least partially possible for player 2 to discriminate between the two types, and this is not beneficial for any of them: the low type cannot fully free-ride on the high prior, and the high type has to use the costly signal at least sometimes. In P2, the payoff loss relative to P3 is even more pronounced—and dramatic: In both P2 and P3, because both types use the same signal, observation of the respective signal does not carry any information: the updated belief after observation of the commonly used signal is the same as the prior belief. But in one case the commonly used signal costs something, and in the other case it costs nothing. The difference between P2 and P3 hinges on the (expected) reaction to signals of the second player: in P2, player 2 takes a only if she has observed the costly signal; while in P3, she takes a without having observed the costly signal. In P2 signaling can be said to be truly wasteful: the signal carries no information, but given expectations about player 2's reaction to signals (that have formed or jump into place for whatever reason), everybody has to use the costly signal.

A warning is in place though: this is a particularity of *this* extensive form. There are games in extensive form that have the same reduced normal form as that in figure 1b in which not all Nash equilibria of that normal form are sequential Bayesian Nash equilibria.[4] Such an extensive-form game is shown in figure 1c. Observations of this sort play a role in some theories of equilibrium refinement. I will come back to the issue of alternative extensive-form representations in the discussion of refinements.

[4]The reduced normal form of a game is the normal form of this game from which all strategies whose payoff vectors (rows or columns) are linear combinations of those of other strategies have been removed.

When uncertainty about the state of nature is involved—no matter if the game is looked at in some extensive form or its normal form—which equilibria the game has typically depends on the prior probability distribution over the states of nature. For the game in figure 1a, and its normal-form representation in 1b, the equilibrium structure falls into two generic cases according to whether p is below or above $1/2$, which is the probability at which the second player is indifferent between taking a and $\bar{a}$ in any of her information sets: If $p < 1/2$, there is a partially separating equilibrium, in which the high type uses the costly signal $S^\star$ for sure and the low type with some probability, and an equilibrium outcome in which both types use the unmarked S; if $p > 1/2$, there is a partially separating equilibrium, in which the high type uses the costly signal $S^\star$ with some probability and low type uses always the unmarked S, and there are two equilibrium outcomes in which both types use the same signal: an equilibrium outcome with pooling in $S^\star$ and an equilibrium outcome with pooling in S.[5]

It is worthwhile to notice the difference between *equilibrium* and *equilibrium outcome*. An *equilibrium* is given by a profile of strategies that satisfy equilibrium conditions (that one strategy is a best response to the other); an *equilibrium outcome* is a probability distribution over the end nodes of the tree induced by an equilibrium strategy profile. Every equilibrium maps to one outcome; but an equilibrium outcome might be the image of several equilibria. This comes from the fact that a *strategy* specifies actions at all nodes or information sets where a player potentially comes to move but not all nodes or information sets of a player are necessarily reached under a given profile of strategies of the other players. At a node or information set "off the equilibrium path," there might be multiple choices of the player acting there that sustain a given outcome as an equilibrium outcome. In signaling games, this concerns information sets after a signal that is never used in the equilibrium. Take, for example, the equilibrium outcome P1 of the game in figure 1a under the prior $p < 1/2$, in which the first player uses the unmarked S in both of his types, and the second player, in response to S, takes $\bar{a}$. This observed way of behaving is sustained as an equilibrium outcome if player 2 in the hypothetical (or, if one wishes "counterfactual") case that she observed $S^\star$ would take a with a probability $y \in [0, c_1]$. Each specific value of $y \in [0, c_1]$ designates another equilibrium (geometrically, a point in the set of all mixed-strategy profiles); but all these equilibria map to the same outcome, namely that both types use S and the second player in response to S takes $\bar{a}$. Given the prior p, this induces a probability distribution over the end nodes of the tree. Geometrically, in the space of all mixed-strategy profiles, an equilibrium outcome in which both types use the same signal is a connected component of equilibria.

[5] In the knife-edge case $p = 1/2$, from which I abstract here, the two generic cases melt into each other: there will be two continua of equilibrium outcomes, one with pooling in $S^\star$, which absorbs what is the partially separating equilibrium PS1 in case that $p < 1/2$; and one with pooling in S, which absorbs what is the partially separating equilibrium PS2 in case that $p > 1/2$.

2.1 Beliefs off the equilibrium path

Sequential Bayesian Nash equilibrium has one fateful property: while it spells out the role of beliefs *along the equilibrium path*, it does not, at least not for signaling games of this simple form, impose any restrictions on beliefs in situations *off the equilibrium path*—in the hypothetical situation that player 2 observed a signal that is part of the game but that is actually never used in the equilibrium under study.[6]

But are all beliefs plausible? Take again the equilibrium outcome P1 as an example. To think that player 2 in case that she received the costly signal $S^\star$ would take a with a probability in the interval $[0, c_1]$, is to think that player 2 after $S^\star$ would attribute to the high type a probability of not more than $1/2$, and hence to the low type a probability of at least $1/2$, because otherwise (given that she responds optimally to her beliefs) she would have to take a for sure in response to $S^\star$. However, one could have the following *intuition*:

> Given that the high type has a lower cost of producing the costly signal $S^\star$ relative to the low type, and this is commonly known between the players of the game, wouldn't it be more plausible to believe that the costly signal $S^\star$ came from the high type?

If this intuition is correct and commonly shared among the players of the game, the equilibrium outcome P1 cannot be maintained as a plausible prediction of the model. Investigating this sort of hypotheses is what the theory of equilibrium refinements based on the plausibility of beliefs off the equilibrium path is concerned with.

2.2 A digression: Existence of perfectly separating equilibria

The equilibrium structure of the game in figure 1, besides the fact that it falls into two generic cases as a function of p, has the conspicuous property that for the specification $0 < c_1 < c_2 < 1$ no fully separating equilibrium in which the high type uses $S^\star$ and the low type S exists. The question when, under which parameter specifications, perfectly separating equilibria exist has been intensively discussed in the theory of costly signaling in economics and biology.[7] For the game in figure 1, a fully separating equilibrium in which the high type uses $S^\star$ and the low type S, and hence both signals perfectly reveal the type who uses it, exists only if $c_2 \geq 1$. This condition, which mimics a condition for continuous signaling games known as the *single-crossing property*, is rather strong. It says that for the low type the costly signal is so costly that it consumes the positive payoff to be had if player 2 chooses action a. In this case, no matter what the prior p, there will always be a separating equilibrium, in which the high type uses the costly signal and

[6] Bayes' rule simply is not defined at a point off the equilibrium path.

[7] See, for example, the surveys by Kreps and Sobel (1994) and Sobel (2009) for the discussion in economics, and the studies by Számadó (2011) and Zollman et al. (2013) for the discussion in biology.

the low type does not use it and player 2 takes action a if the costly signal has been sent and $\bar{a}$ if the costly signal has not been sent.

The existence of such a fully separating equilibrium does, however, not do away with the multiplicity of equilibria. If $0 < c_1 < c_2 = 1$, in addition to the fully separating equilibrium, there will still be partially separating equilibria and pooling equilibria similarly as in the generic case, and if $c_2 > 1$, while partially separating equilibria will vanish, there will still be the pooling equilibria, in which both types use the same signal. The questions of equilibrium refinement and the plausibility of beliefs off the equilibrium path hence persist. I concentrate on the case $0 < c_1 < c_2 < 1$, because partially separating equilibria are better suited to show Bayesian updating of beliefs (in a fully separating equilibrium, the Bayesian update becomes trivial, because observation of each signal makes the updated belief "jump" to sure knowledge about the state of nature) and also more interesting for applications; notably in the study of language—which is what I have in mind with this investigation.

3 Equilibrium refinement by restrictions on beliefs off the equilibrium path

A minimal condition on beliefs off the equilibrium path that should be required is that beliefs be *consistent* in the sense that they can be deduced from Bayes' rule after a small perturbation of the behavior strategies (Kreps and Wilson 1982). It is straightforward to show that in signaling games, beliefs off the equilibrium path are always consistent and hence the equilibrium always *sequential*.[8]

What triggered research in equilibrium refinements in the 1980s, was in fact the realization that in many games, and this concerned also signaling games, equilibria that are supported by consistent beliefs still seemed unplausible, for some other reason, similarly to what I have called above the *intuition* that destabilizes the equilibrium outcome in which both types never use the costly signal under the prior $p < 1/2$. A refinement known as "divinity" casts this intuition into a formal concept.

[8] Let $(p, 1-p)$ be the initial prior for $(\text{high}, \text{low})$. Suppose that in equilibrium a specific signal is never sent. Let $(p^*, 1-p^*)$ be player 2's belief off the equilibrium path when he receives that signal. Suppose that player 1 perturbs his behavior strategies as follows: the high type sends the signal which in the original equilibrium is never sent with probability $\varepsilon(1-p)p^*$, where ε is very small, and the low type sends this signal with probability $\varepsilon p(1-p^*)$. By Bayes' rule, the updated belief is $(p^*, 1-p^*)$.

3.1 "Divinity"

"Divinity" (Banks and Sobel 1987), applied to the games considered here, requires that player 2's belief after a signal off the equilibrium path does not attribute any positive probability to type t if for the other type, say t', the set of mixed-strategy best responses of player 2 to the off-the-equilibrium-path signal that would make him *strictly better off* if he deviates and uses the off-the-equilibrium-path signal compared to the payoff that he gets in the equilibrium outcome under study is non-empty and a superset of the set of mixed-strategy best responses responses of player 2 to the off-the-equilibrium-path signal that would make type t *indifferent or better off* if he uses the off-the-equilibrium-path signal compared to the payoff that he gets in the equilibrium under study (a set that is possibly empty).[9]

In the game in figure 1a, for the prior $p < 1/2$, "divinity" discards the equilibrium outcome P1, in which both types use the unmarked S and player 2 takes $\bar{a}$. For the prior $p > 1/2$, "divinity" does not exclude any of the three equilibrium outcomes that exist in that case, which means in particular that it cannot discriminate between the maximally inefficient equilibrium outcome P2, in which both types have to use the costly signal $S^\star$ in order to make player 2 take a, and P3, in which both types use the unmarked S and player 2 takes a. Table 2 shows how the arguments works out in detail.

"Divinity," one can say, translates into a formal criterion the *intuition* evoked above with respect to the "no-signaling" equilibrium outcome P1 that when the signal "costs" more for type t than for type t', it shall be more plausible that type t' deviates from the convention of not using the signal to using it than type t. Still and all, while this formal criterion can be rigorously verified in any game, it does not explain where the intuition that it expresses comes from. It is after all not so clear on which grounds this intuition is built for it is not demanded by that formal criterion that type t' *always* (for any possible reaction of player 2 to the signal under question) would win, nor that type t never could win from using the off-the-equilibrium-path signal.

3.2 Other belief-based refinements

"Divinity" can be interpreted as a test of *strategic stability* of the equilibrium under study that fits into the following pattern:

[9] Banks and Sobel define different variants of "divinity" along two dimensions, namely, (1) whether the "domination" of a type-signal pair has to be by a single other type t' (Banks and Sobel's criterion D1) or whether it can be by a combination of other types (D2); and (2) whether a type disfavored by such a comparison has to be discarded completely from the support of equilibrium beliefs ("universal divinity") or, less strictly, whether the belief on such a type just cannot be augmented relative to the probability that player 1 is of some other type (what they call simply "divinity"). For the simple class of games with two states of the word, two signals, and two actions of the second player, these different versions of "divinity" all coincide to the criterion given above.

(1) Suppose the equilibrium outcome under study is the established signaling convention.

(2) What is the second player to infer about the first player if she observes a deviation from that convention—that is, a signal "off the equilibrium path"? Which beliefs about the type of the first player are plausible under this observation?

(3) Is the equilibrium outcome under study compatible with what the second player has to infer? If yes, the equilibrium outcome is "stable" under the applied rule.

This protocol can be understood as a thought experiment conducted by the players in the game. Given that it is common knowledge among the players of the game that they are both rational, it can be argued that only an equilibrium outcome that is stable under this thought experiment should be considered a plausible outcome of the game. This protocol, which is implicit in many belief-based refinements, has been posited explicitly by Cho and Kreps (1987) as a framework to investigate and relate different refinement concepts that rely on restrictions on beliefs off the equilibrium path. Different refinements, in this scheme, differ at step 2 where they posit different rules to restrict beliefs off the equilibrium path. Clearly, this is where the potential trouble with this approach is: Where does the rule to restrict beliefs off the equilibrium path come from? Game theorists have given different answers to that question and defended them of different grounds. It is, one can say, the game theorists' debate of counterfactual conditionals.

Another prominent refinement for signaling games that fits into the above protocol has come to be known under the name of the "intuitive criterion" (Cho and Kreps 1987). The "intuitive criterion" imposes stricter conditions to exclude a type from the support of the belief after a signal off the equilibrium path. Applied to the games considered here, it says that player 2's belief after a signal off the equilibrium path should exclude a type t only if this type gets a *strictly higher* payoff in the equilibrium under study than the maximum payoff that he could get from sending the off-the-equilibrium-path signal (for any possible reaction of player 2 to that signal), provided that this is not true for all types. An equilibrium outcome that does not satisfy the "intuitive criterion," will also not be "divine" (Banks and Sobel 1987, Cho and Kreps 1987). Obviously, for if a type t is excluded after an off-the-equilibrium-path signal by the "intuitive criterion," the set of responses of player 2 to this off-the-equilibrium-path signal that would make this type indifferent or strictly better off if he sends the off-the-equilibrium path signal is empty and because that is required not to be true for both types, this set will be trivially contained in the set of responses of player 2 to the off-the-equilibrium-path signal that would make the other type t' better off sending the off-the-equilibrium path signal.[10] However, an outcome that it not "divine" might still satisfy

[10]If it is the case that both types get a *strictly higher* payoff in the equilibrium under study than the maximum payoff that they could get from sending the off-the-equilibrium-path signal, then, after Cho and Kreps, the equilibrium passes the test if every belief after the off-the-equilibrium-path signal supports the equilibrium under

the "intuitive criterion": "divinity" is stronger as a selection tool than the "intuitive criterion." The game in figure 1 is a case in point: the "intuitive criterion" is not strong enough to discard the no-signaling equilibrium outcome P1, which exists for $p < 1/2$. What I have called above the *intuition* on which grounds $P1$ would be discarded is actually not captured by the criterion officially known under the name of the "intuitive criterion."

There are, it should be noted for completeness, refinements based on plausibility assumptions concerning beliefs off the equilibrium path that cannot be so easily construed as tests of strategic stability in the sense of the protocol outlined above. One of these is McLennan's (1985) concept of *justifiable beliefs*, which is by many considered the origin of belief-based refinements (see, for example, the discussion in Cho and Kreps 1987). McLennan's justifiable beliefs differs from the intuitive criterion in that the benchmark in the payoff comparisons for a type is not the payoff of this type in *the equilibrium outcome under study* but the *minimal payoff* of that type *in any equilibrium outcome of the game*. For the game in figure 1a, this criterion coincides with the intuitive criterion. Another concept that is defined on the set of equilibrium outcomes is Mailath et al.'s (1993) *undefeated equilibrium*, which requires that beliefs off the equilibrium path be compatible with the beliefs at *some other equilibrium*.

I concentrate on strategic stability in the sense of the protocol indicated above and the intuition that feeds "divinity," because this is the concept that will provide a link to Lewis's account of counterfactuals. The rest of this section can be understood as a quest for a foundation of this concept—or better the *intuition* on which it thrives—from a game-theoretic point of view.

3.3 Foundations: strategic stability and "forward induction"

The general idea of *strategic stability* in the sense of robustness of an equilibrium against deviations of a player that induce another player to draw certain conclusions that break the equilibrium under study goes back to Kohlberg and Mertens (1986).

Kohlberg and Mertens's investigation is inspired not by signaling games but *outside-option games*—games in which one player has the choice of either entering into a game with the other player or "opting out" by taking an action that determines the payoff of both players and thereby ends the game. In such games, Kohlberg and Mertens point out Nash equilibria that have player 1 taking the outside option that are perfectly compatible with backward induction but still seem strategically unplausible in that they do not withstand what they call the "forward-induction" logic, namely that player 1 by not taking the outside option could have forced the second player into some inference about what he would do in the future—of which "strategic type" he is, if

consideration (Cho and Kreps 1987, p. 196), which in the simple games considered here will be trivially fulfilled. In this case, then there is no type for whom there is a non-empty set of responses of player 2 that will make him better off, and therefore the equilibrium outcome will also satisfy "divinity."

one wishes—that would break the outside option as an equilibrium outcome of the game. The typical example is a coordination game with two equilibria in pure strategies (in which players have different payoffs) preceded by an outside option for player 1 that secures player 1 a payoff that is between the payoff that he gets in the two Nash equilibria of the coordination subgame. The forward-induction logic claims that the outside-option equilibrium outcome is not strategically stable because if the second player sees that the first player deviates form such an equilibrium, that is, does not take the outside option and moves into the coordination game, the second player shall exclude that in the ensuing coordination game the first player will take the strategy leading to the equilibrium within the subgame in which the first player gets the lower payoff. "Forward induction," one might say in the language of philosophers of language, is some form of "implicature." Interestingly, Kohlberg and Mertens explain the forward-induction logic as a silent speech act. It is, they say, as if player 1 by not taking the outside option were telling player 2:

> "Look, I had the opportunity to get 2 for sure [his payoff from the outside option], and nevertheless I decided to play in the subgame, and my move is already made. And we both know that you can no longer talk to me, because we are *in* the game, and my move is made. So think now well, and make your decision" (Kohlberg and Mertens 1987, p. 1013).

Kohlberg and Mertens notice that in such games, the outside-option equilibrium outcome will also be discarded by iterative deletion of strategies that are *weakly dominated* in the normal form of the game. And they make another astonishing observation: the outside-option equilibrium outcome might fail to satisfy backward-induction in an extensive form that has the same reduced normal form as the original extensive-form game. Kohlberg and Mertens conclude that "a good concept of 'strategically stable equilibrium' should satisfy both the backwards induction rationality of the extensive form and the iterated dominance rationality of the normal form, and at the same time be independent of irrelevant details in the description of the game," by which they refer to differences in the extensive form that do not show in the reduced normal form (Kohlberg and Mertens 1986, p. 1004).

The challenge emanating from Kohlberg and Mertens's program is that it is not defined in a fully axiomatic fashion but heuristically, starting from certain phenomena that shall be avoided by "a good concept of 'strategically stable' equilibrium," like not being robust against the "forward-induction logic." Understanding what the "forward-induction logic" is in more general games and how is shall be defined in general proved to be challenging questions that stimulated much research and debate in game theory. Interestingly, and here is where Banks and Sobel as well as Cho and Kreps come in, arguments that can be advanced to discard certain equilibrium outcomes in signaling games seem to have a "forward-induction flavor," namely in the sense that an equi-

librium outcome is not stable in the light of inferences that a player would draw from a deviation of another player from the equilibrium under study, which here takes the form of using a signal that was supposed to be never used in the equilibrium under study. One of the fundamental questions that emerges out of the Kohlberg–Mertens program is whether there is a unifying definition of "forward induction" as a refinement of sequential Nash equilibrium that accommodates both forward-induction reasoning as described in outside-option games and restrictions on beliefs off the equilibrium path in signaling games.

One of the approaches pursued in this endeavor, building on the observation by Kohlberg and Mertens that equilibria that do not satisfy "forward-induction" might disappear as backward-induction equilibria in alternative extensive forms, is to derive "forward induction" from "backward induction plus invariance"—the axiomatic requirement that an equilibrium should satisfy backward induction in any extensive form that has the same reduced normal form. In games with uncertainty, backward induction is captured by the notion of sequential Bayesian Nash equilibrium. Hence, the axiomatic requirement that only equilibria that are sequential in any extensive form that has the same reduced normal form should be maintained (see, for example, Hillas and Kohlberg 2002).

Govindan and Wilson (2009) propose a criterion of "forward induction" as a refinement of sequential Bayesian Nash equilibrium that is implied by "sequentiality plus invariance." In other words, if an equilibrium outcome does not satisfy the forward-induction criterion defined by Govindan and Wilson, it can be discarded on the grounds that it is not "sequential" in every possible extensive form that has the same reduced normal form. Govindan and Wilson's forward-induction criterion is a generalization of a concept that has been known for some time, which consists in deleting strategies from the normal form that are never an alternative best response to any equilibrium of the game, which has come to be known under the term "never a weak best response" (Banks and Sobel 1987, Cho and Kreps 1987).

For the simple class of games considered here, an equilibrium outcome in which both types use the same signal satisfies the forward-induction criterion defined by Govindan and Wilson if this outcome can be supported by a belief of player 2 in response to the off-the-equilibrium-path signal that is restricted to types for whom it is true that there exists an equilibrium with the outcome under study in which this type is *indifferent* between sending the equilibrium signal and sending the off-the-equilibrium-path signal—provided that such a type exists; if, in the terminology of Govindan and Wilson, player 2's belief after the off-the-equilibrium-path signal has support in "relevant strategies"—provided that the information set of player 2 after the off-the-equilibrium-path signal is "relevant."

Both Banks and Sobel (1987) and Cho and Kreps (1987) show that "never a weak best response" implies the respective concept that they defend, "divinity" and respectively the "intuitive criterion." The reason that they do not embrace it seems to be that they do not see a foundation for it in

terms of some thought process, some intuition or story, that would go with it. Govindan and Wilson (2009) show that the general criterion of "forward induction" that they give, for signaling games, implies "never a weak best response," "divinity" and the "intuitive criterion." For the simple class of games studied here, one can say a bit more: "divinity" and "forward induction" ("never a weak best response") coincide.

Proposition 1. *For signaling games with two types, two signals, and two actions of the second player, "forward induction" as defined by Govindan and Wilson (2009), which here takes the form of "never a weak best response," and "divinity" (Banks and Sobel 1987) coincide.*

Proof. I first replicate the argument already shown by the authors mentioned above that an equilibrium outcome that does not satisfy "divinity" will also not satisfy "forward induction" ("never a weak best response"). Of course, if there is a type t for whom the set of mixed-strategy responses of player 2 to the off-the-equilibrium-path signal that would make this type indifferent or better off compared to the payoff that this type gets in the equilibrium under study (a subset of $[0,1]$) is contained in the set of mixed-strategy responses of player 2 to the off-the-equilibrium-path signal that would make the other type t' strictly better off as compared to the payoff that t' gets in the equilibrium outcome under study (another subset of $[0,1]$), then, by continuity of the payoff functions, there exists a response of player 2 to the off-the-equilibrium path signal (that is, a probability with which he takes a) at which type t' is equally well off between using the off-the-equilibrium-path signal and the equilibrium signal, while type t is strictly better off using the equilibrium signal.

To see the converse, why an outcome that does not satisfy "forward induction," will not satisfy "divinity," note that if for a type t there is no equilibrium with the outcome under study, that is, no mixed-strategy response in $[0,1]$ of player 2 to the off-the-equilibrium-path signal such that type t is indifferent between using the equilibrium signal and the off-the-equilibrium-path signal, but a type t' for whom this is true (in Govindan and Wilson's terminology, the information set of player 2 after observation of the signal that is off the equilibrium path is "relevant"), then of course, by continuity of the payoff functions with respect to player 2's reaction to the off-the-equilibrium-path-signal, the set of responses of player 2 to the off-the-equilibrium-path signal that would make type t' strictly better off if he sends the off-the-equilibrium-path signal compared to the payoff that he gets in the equilibrium outcome under study (a subset of $[0,1]$) is a superset of the set of responses of player 2 to the off-the-equilibrium-path signal that would make type t when sending the off-the-equilibrium-path signal strictly better or equally well off compared to the payoff that he gets in the equilibrium outcome under study (a subset of $[0,1]$, possibly empty). If there is no type who would be indifferent between sending the equilibrium signal and sending the off-the-equilibrium-path signal for any possible reaction that player 2 could have to the off-the-

equilibrium path signal (the information set off-the-equilibrium-path is "not relevant," in the terms of Govindan and Wilson), the equilibrium outcome trivially passes the test of forward induction and divinity. □

Table 2. "Divinity" and "forward induction" for the game in figure 1a

Generic case: $0 < c_1 < c_2 < 1$

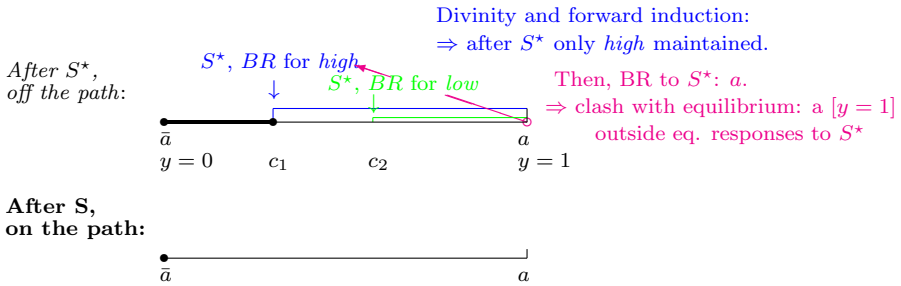
- $p < \frac{1}{2}$:

  (P1) In this outcome, both types use the unmarked signal S, and player 2 takes $\bar{a}$. The starting point for any of the refinements considered here is the payoff of player 1's types in the equilibrium outcome under consideration, which here is 0 for both types.

  Divinity. Player 1's *high type* will be *strictly better off* if he deviates from his equilibrium strategy S and sends the costly signal $S^\star$ if player 2 in response to $S^\star$ chooses action a with a probability in the interval $(c_1, 1]$; player 1's *low type* will be *indifferent or strictly better off* if he deviates from his equilibrium strategy S and sends the costly signal $S^\star$ if player 2 in response to $S^\star$ chooses action a with a probability in the interval $[c_2, 1]$. Since $c_1 < c_2$, evidently, $(c_1, 1]$ contains $[c_2, 1]$, and as a consequence, according to "divinity," any belief that puts some positive probability on the low type after observation of $S^\star$ has to be discarded.

  Forward induction. There exists an equilibrium with this outcome, namely the one in which player 2 if she were to see the costly signal $S^\star$ would take a with probability c_1, in which player 1's *high type* is indifferent between his equilibrium strategy S and $S^\star$. According to forward induction, then, any belief after observation of the off-the-equilibrium-path signal $S^\star$ has to be restricted to player 1's high type and accordingly put zero probability on player 1's low type.

  The consequence for divinity and forward induction is the same: In any equilibrium with the outcome under study, player 2 after observation of $S^\star$ must take $\bar{a}$ with a probability of at least $1 - c_1$, that is, put a belief of at least $1/2$ on player 1's low type, which is however the type that has to be excluded after observation of $S^\star$. The equilibrium outcome therefore does not satisfy these two refinements.



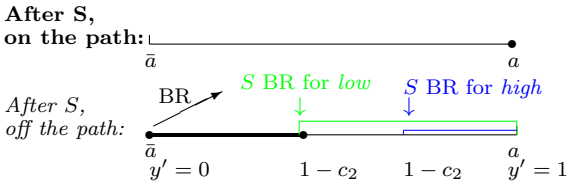
- $p > \frac{1}{2}$:

  (P2) In this outcome, both types use the costly signal $S^\star$ and player 2 takes a. Both of player 1's types if they were to send the off-the-equilibrium-path signal S, depending on player 2's response to S, would get either 0 or 1, while in the equilibrium outcome under study they get $1 - c_1$ and $1 - c_2$, respectively.

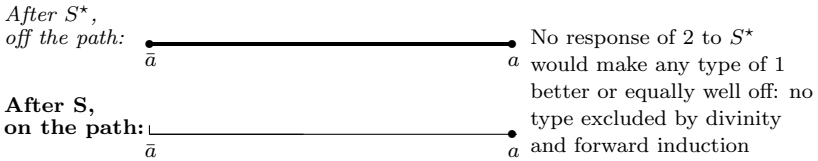
  Divinity. Player 1's *low type* will be strictly better off if he deviates from his equilibrium strategy and uses S, if player 2 in response to S takes a with a probability in the interval $(1 - c_2, 1]$; while player 1's *high type* will be indifferent or strictly better off if he uses S if player 2 in response to S takes a with a probability in the interval $[1 - c_1, 1]$. Obviously, $(1 - c_2, 1]$ contains $[1 - c_1, 1]$.

  Forward induction. There exists an equilibrium with that outcome in which player 1's *low type* is indifferent between $S^\star$ and the off-the-equilibrium signal S, namely the one in which player 2 in response to S would take a with probability $1 - c_2$.

  According to divinity and forward induction, then, any belief that puts some positive probability on the high type after S has to be discarded—which is perfectly in line with the beliefs of player 2 that support the equilibrium outcome under study, and so the equilibrium outcome passes both "divinity" and forward induction.



  (P3) The equilibrium outcome in which both types of player 1 use S, and player 2 takes a also survives both "divinity" and forward induction. Surely, because *any* belief off the equilibrium path supports the equilibrium under study, and therefore no equilibrium can be sorted out by any refinement based on restrictions on beliefs off the equilibrium path. It is still insightful to remark that both types of player 1 get a *strictly higher* payoff, namely 1, in the equilibrium outcome under study than the maximum payoff that they could get by deviating from their equilibrium strategy S and sending the costly signal $S^\star$. Therefore, for none of player 1's types, there is a response of player 2 to the off-the-equilibrium-path signal $S^\star$ that would make him indifferent or better off between sending the equilibrium signal S and $S^\star$, which is to say that none of the types can be discarded after $S^\star$, neither on the basis of "divinity" nor forward induction.



For the game in figure 1a, it follows from this equivalence that "forward induction" and "divinity" will select the same equilibrium outcomes. In table 2, as a matter of illustration, I still show the argument in full for both of these criteria.

If one leaves the simple class of signaling games with two states of nature, two signals, and two actions, "divinity" no longer coincides with "forward induction" ("never a weak best response"), which in general is a stricter condition and hence stronger as a selection criterion. To break the equivalence between "divinity" and "forward induction" ("never a weak best response") it suffices, for example, that there be three actions in response to signals. An example showing this is given by Banks and Sobel (1987), which in a variant is also discussed by Cho and Kreps (1987).

3.4 Strategic stability as stability under different representations

Govindan and Wilson's forward-induction criterion is as such a static notion—as is "divinity" and the "intuitive criterion." The forward-induction logic, which is after all a movement that captures some form of inference, comes off from the idea that the stability test encapsulated by it is a thought experiment that both players run through. It is not just that player 2 asks herself what she should think given that she has received the signal that was supposed to be never used in the equilibrium under study, but player 1, when deciding to deviate, anticipates that this is the thought experiment that player 2 is going to do if she receives the signal that he, player 1, was supposed not to use; and player 2 knows that this is what player 1 thinks she is going to think, etc. It is the anticipation of that thought experiment and the common knowing that player 1 has pushed player 2 into that thought experiment that makes the "forward induction." In this perspective, the intuitive criterion, one can hold, also captures some forward-induction movement—only that player 2 when she sees the deviation is assumed to draw less restrictive conclusions from that deviation, which leads back to the crucial question: *What is the "right" restriction on beliefs off the equilibrium path? (What is the right way to draw inferences in case of counterfactual events?)*

Govindan and Wilson's criterion of forward induction stands out from other criteria in that it is founded not on some "intuition" or "story" in a specific example but on some decision-theoretic axiom—the requirement that the equilibrium under study be a sequential Bayesian Nash equilibrium *in any extensive form that has the same reduced normal form.* For signaling games one can say that Govindan and Wilson's contribution provides a foundation for "never a weak best response" in terms of "invariance plus sequentiality"; for signaling games with two types, two signals, and two actions, in which "never a weak best response" coincides with "divinity," one hence has a foundation for the *intuition* that underlies "divinity" in terms of "invariance plus sequentiality."

The idea that the solution of a game should be invariant under any extensive form that has the same reduced normal form is certainly radical but not unfounded. A game tree is a *representation* of the situation of interaction in which individuals find themselves with other individuals. But

this representation is itself an abstraction, and if such a representation is not unique, which is the case, it makes sense to ask if there is a solution that is stable under any possible representation.[11]

One might, of course, ask whether, why and how (through which mechanisms that might be part of a larger game) one representation might become fixed. But before doing that it is of interest to know whether there is a solution that is robust under any representation of the same basic strategic situation as captured by the reduced normal form. Asking that question might give some insight into the first: If an equilibrium does not satisfy the requirement of sequentiality in every extensive form that has the same reduced normal form, then there exists at least one extensive form in which this equilibrium is not a sequential Bayesian Nash equilibrium. Such an extensive form might be a good candidate for the representation of the game that becomes fixed. Strategic stability—and this is one of the deep insights of Kohlberg and Mertens—is transformed into a question of stability under different representations.

Figure 1c shows an alternative extensive form of the game in figure 1a (one that has the same reduced normal form) in which, for the prior $p < 1/2$, the no-signaling equilibrium outcome (P1) is not sequential, and therefore only the partially separating equilibrium (PS1) survives as a sequential Bayesian Nash equilibrium. This alternative extensive form consists in "splitting off" player 1's pure strategy SS: player 1, before even nature makes here move, first has to make a principled decision whether to play according to SS (right) or not (left), and only after that decides which of the strategies that have their support in $S^\star S^\star$, $S^\star S$, $SS^\star$ to use in case that he moves right. After each of the four resulting nodes, the extensive form of the game in figure 1a ensues but with player 1's actions prescribed by his previous choice: that is, nature chooses player 1's type, which, given the choice of pure strategy already taken by player 1, prescribes using $S^\star$ or S with the specified probabilities, and then player 2, who can only observe whether $S^\star$ or S has been taken, has to decide whether to take a or $\bar{a}$.

This alternative extensive form obliges player 1 in case that he uses SS to say what he would do if he were not using SS, which in turn obliges player 2 to have a hypothesis about that. And this is all it takes to "destabilize" the no-signaling equilibrium outcome in case that $p < 1/2$: from the node onwards at which player 2 has to choose between $S^\star S^\star$, $S^\star S$ and $SS^\star$, the only profile of strategies that guarantees that in the continuation what one player does is a best response to what the other does is such that player 1 plays according to the mixture between $S^\star S^\star$ and $S^\star S$ that will make player 2, when she receives $S^\star$ indifferent between a and $\bar{a}$; and player 2, when she receives $S^\star$, mixes between a and $\bar{a}$ such that player 1 is indifferent between $S^\star S^\star$ and $S^\star S$, and when she receives S, takes $\bar{a}$ (play as in the partially separating equilibrium PS1). If player 2 reacts to signals in that way, player 1, at his first node is better off rejecting SS and moving to

[11] For a discussion of extensive-form transformations that leave the normal form unchanged see Kohlberg and Mertens (1986) and Elmes and Reny (1994).

his second node, that is, playing effectively according to PS1, and consequentially, the equilibrium outcome P1, in which player 1 chooses SS and player 2 takes $\bar{a}$ in response to S, cannot be a sequential.

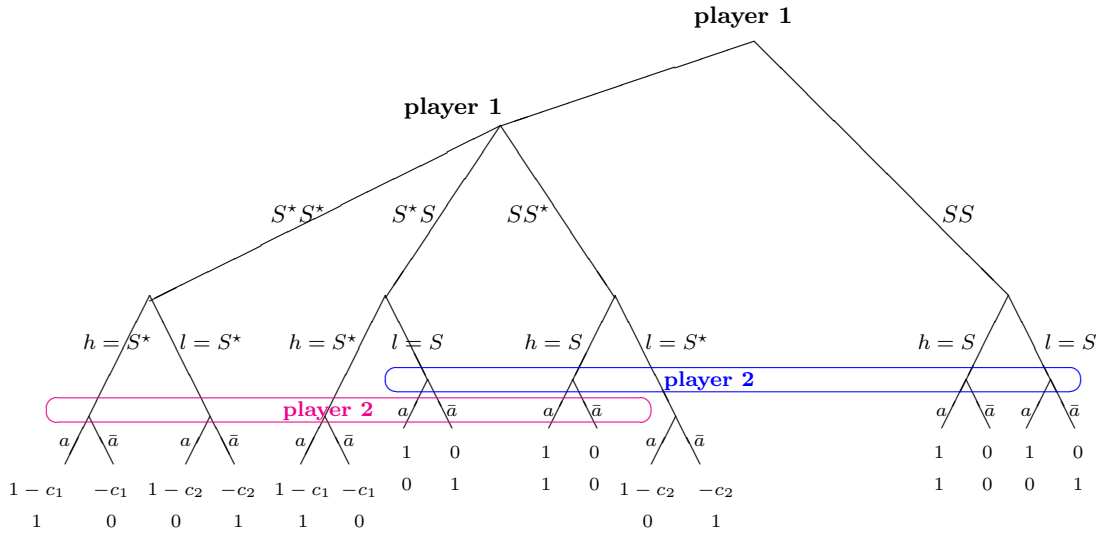


Figure 1c. An game in extensive form with the same reduced normal form as that in figure 1a.

This extensive-form game has the flavor of an outside-option game. If player 1 takes it for granted that in response to S, player 2 choses $\bar{a}$, then taking SS in fact becomes an outside option that yields him 0. In this alternative extensive form, when player 2 observes the costly signal $S^\star$, she knows, and it will be commonly known, that player 1 did not make the choice to take SS. The very fact that player 2 receives $S^\star$ becomes like an implicit speech of player 1 by which he tells player 2:

> "Look, I could have played according to SS. I did not. You know this from the very fact that you have received $S^\star$. Now, we both know that we are both sequentially rational. So you understand that what you see me doing, if you think that what I am doing can be rationalized as a best response to something that you are doing that can be rationalized as a best response to something that I am doing, etc., can only come from me playing according to the mixture between $S^\star S^\star$ and $S^\star S$ that is such that you, if you update your beliefs in a Bayesian rational way, when you receive $S^\star$, are going to be indifferent between a and $\bar{a}$."

Now, since player 2 upon receiving $S^\star$ will be indifferent between a and $\bar{a}$, this inference does not "force" her into any particular reaction.[12] But because she is indifferent, there is also nothing

[12] Other than in outside-option games in the style of Kohlberg and Mertens, where player 2 when he sees that player 1 did not take the outside option is "forced" into taking a particular pure strategy in the ensuing subgame.

standing in the way that she takes a and $\bar{a}$ with the probabilities required in equilibrium. The splitting-off of SS has, I will argue in the closing section, a natural interpretation when using this machinery for modeling communicative implicatures.

4 Lewis's account of counterfactuals as equilibrium refinement

In game theory, David Lewis is mostly known for the notion of *common knowledge* that he introduces in *Conventions* (1969). This work, to my knowledge, is however never mentioned in the debate on strategic stability. And yet, Lewis's interpretation of Nash equilibrium as a convention, which he derives from his understanding of Nash equilibria as self-enforcing agreements, silently pervades this approach to refine the equilibrium concept.

Cho and Kreps (1987), in response to the critique that the intuitive criterion accords a crucial role to the particular equilibrium outcome under study, but when it works, works to discredit that equilibrium outcome, say:

> An equilibrium is meant to be a candidate for a mode of self-enforcing behavior that is common knowledge among the players. (Most justifications for Nash equilibria come down to something like this [...]) In testing a particular equilibrium (or equilibrium outcome), one holds to the hypothesis that the equilibrium (outcome) is common knowledge among the players, and one looks for "contradictions" (Cho and Kreps 1987, p. 203).

Restrictions on beliefs off-the-equilibrium-path, in this perspective, are additional, necessary requirements that an equilibrium has to satisfy to qualify as self-enforcing—as a convention. An equilibrium in order to be self-enforcing has to be stable against the type of thought experiment captured by strategic stability tests. Belief-based refinements of the strategic-stability type, on this view, are a development in the spirit of Lewis's interpretation of Nash equilibria.

More interestingly still, a link can be developed between Lewis's account of *counterfactuals* (Lewis 1973, 1978) and the equilibrium refinement program based on the plausibility of beliefs off the equilibrium path. A counterfactual, put crudely, is an event that can happen in the universe of all possible worlds but that does not happen in the actual word. If one replaces "actual world" by "equilibrium under study," a connection seems to be within reach.

4.1 Possible worlds and game trees

The representation of a situation of strategic interaction by a game tree fits very well Lewis's *realism* about possible worlds. In *Counterfactuals* (1973) Lewis declares:

> "I emphatically do not identify possible worlds in any way with respectable linguistic entities. I take them to be respectable entities in their own right. When I profess realism about possible worlds, I mean to be taken literally. Possible worlds are what they are, and not some other thing. If asked what sort of thing they are, I cannot give the kind of reply my questioner probably expects: that is, a proposal to reduce possible worlds to something else. I can only ask him to admit that he knows what sort of thing our actual world is, and then explain that other worlds are more things of *that* sort, differing not in kind but only in what goes on at them. Our actual world is only one among others. We call it alone actual not because it differs in kind from all the rest but because it is the world we inhabit. The inhabitants of other worlds may truly call their own worlds actual, if they mean by 'actual' what we do; for the meaning we give to 'actual' is such that it refers at any world i to that world i itself. 'Actual' is indexical, like 'I' or 'here', or 'now': it depends for its reference on the circumstances of utterance, to wit the world where the utterance is located" (Lewis 1973, p. 86).

A game tree can be understood as a representation of all possible worlds. This representation not only provides an inventory of all possible worlds, but also shows exactly where, by which move (either a move of nature or the move of a personal player), one possible world starts to differ from another possible world. Every profile of strategies together with a realization of the possible moves of nature and the random draws that define mixed strategies specifies a unique path through the tree leading to one of its end nodes.

While a game tree keeps track of everything that possibly can happen, Bayesian Nash equilibrium is based on an abstraction: it treats the possible moves of nature as a constant—not by fixing one, but by working with the probability distribution over all possible moves of nature—and it is interested only in the variations in players' choices.[13] As far as moves of nature go, possible worlds are differentiated only on the level of the probability distribution over these moves. In the class of games studied here, this probability distribution is summarized by the single parameter p, the prior probability that nature's move makes player 1 a high type. Any value of p in the interval $[0, 1]$ stands for another set of possible worlds. Similarly, when a player uses a mixed strategy: equilibrium is defined in terms of the probability distributions induced by these strategic random variables. As a consequence, an equilibrium outcome does not necessarily specify one end node of the tree, but more generally a probability distribution over end nodes.

In adapting Lewis's possible-worlds framework to games, I make the following postulates:

[13]This is not specific to Bayesian Nash equilibrium, which shares this with virtually all other solution concepts for games involving uncertainty about moves of nature. Of course this is so, because game theory's subject is how individual decisions shape aggregate outcomes keeping the conditions of the environment in which this interaction happens constant.

- A *possible world* is a combination of a prior p over possible moves of nature together with a strategy profile specifying a behavior strategy for the entire game for every player and a profile of players' beliefs specifying for each information set a belief of the player acting there.

- And, the *actual world* is the given prior p together with the *equilibrium under consideration*.

Any such possible world defines a probability distribution over the end nodes of the tree, but more than this, also specifies under which conditions (which beliefs and conjectures about actions, at nodes reached and not reached) these outcomes are obtained.

4.2 Accessibility conditions and the similarity of possible worlds

Lewis (1973) proposes to evaluate truth conditions of counterfactuals—propositions of the sort "if it were that ϕ, then it would be true that ψ"—based on *accessibility relations* determined by similarity between the actual world and other possible worlds. A counterfactual conditional, Lewis says, is true at a world i (the actual world) if and only if ψ holds at certain ϕ worlds. But, so Lewis, not all ϕ worlds matter. Some worlds are "too far away" from our actual world (Lewis 1973, p. 8). Which are those that matter?

"Respects of similarity and difference trade off," Lewis says. "If we try too hard for exact similarity to the actual world in one respect, we will get excessive differences in some other respect." Lewis explains this in the context of his leading example: *If kangaroos had no tails, they would topple over*. "Are we to suppose," Lewis develops, "that kangaroos have no tail but that their tracks in the sand are as they actually are? Then we shall have to suppose that these tracks are produced in a way quite different from the actual way. Are we to suppose that kangaroos have no tails but that their genetic makeup is as it actually is? Then we shall have to suppose that genes control growth in a way quite different from the actual way." And Lewis concludes:

> "It therefore seems as if counterfactuals are strict conditionals corresponding to an accessibility assignment determined by similarity of worlds—overall similarity, with respects of difference balanced off somehow against respects of similarity. Let S_i, for each world i, be the set of all worlds that are similar to at least a certain fixed degree to the world i. Then the corresponding strict conditional is true at i if and only if the material conditional of its antecedent and consequent is true throughout S_i; that is, if and only if the consequent holds at all antecedent-worlds similar to at least that degree to i" (Lewis 1973, p. 9–10).

Lewis (1973) models accessibility more precisely by a system of *spheres of accessibility* $\mathscr{S}_i$ from the actual world i, which are, as he explains, to be understood as a nested sequence of sets S_i

containing the actual world i, all subsets of the set of all possible worlds. I will in the following, to keep exposition simple, transcribe Lewis's formal notation of counterfactual conditions into ordinary language by using "If–then" phrases, and I will use the term "set" instead of "sphere." The formula that Lewis proposes within such a system to evaluate truth of counterfactual conditionals then is:

> "If it were that ϕ, then it would be true that ψ" if and only if either
>
> (1) no ϕ-world belongs to any set S_i in $\mathscr{S}_i$, or
>
> (2) some set S_i in $\mathscr{S}_i$ does contain at least one ϕ-world, and "if ϕ then ψ" holds at every world in S_i (Lewis, 1973, p. 16).

I will refer to this as Lewis's (1973) *accessibility condition*. If this condition holds with (1), then the counterfactual conditional is trivially, or as Lewis says "vacuously," true.

In his article on fiction (1978), Lewis presents this condition in the following form:

> "If it were that ϕ, then it would be true that ψ" is non-vacuously true iff some possible world where both ϕ and ψ are true differs less from our actual world, on balance, than does any world where ϕ is true but ψ is not true" (Lewis 1978).

In that form I will refer to it as Lewis's (1978) *similarity condition*.

4.3 Adapting Lewis's accessibility condition for games

In adapting Lewis's accessibility condition for games, I make the following postulates:

- In any possible world, players are Bayesian rational, that is, they respond optimally to their beliefs and conjectures about other players' strategies, and this is common knowledge among the players of the game.
- A departure from the actual world, a counterfactual ϕ is an action by a personal player that he or she never takes in the equilibrium under study, under no possible move of nature.

This is how I "trade off respects of similarity and difference."

In signaling games, the departure of interest, the antecedent ϕ of the counterfactual conditional, is a signal off the equilibrium path, which I denote by S'. The consequence ψ then stands for the belief p' of player 2 in response to S', which should be expressed in the form that p' belongs to a certain subset P' of $[0, 1]$. What remains to be determined is what the set S_i stands for, how accessibility is defined, and respectively how "differs less" is to be evaluated.

The assumption that in every possible world players are Bayesian rational payoff maximizers becomes effective here. Since in all possible worlds players are Bayesian rational, which means that

once they have a belief and conjecture about the strategy choice of the other player, they act on it (in a quasi automatic way), what are possible worlds, and hence the similarity between possible worlds, has to be defined in terms of beliefs and conjectures about the other player's strategies that are the basis of action. More specifically, one should base a measure of the similarity between the actual world and another possible worlds in which the departure from the actual world happens on beliefs and conjectures at nodes that lead to the node where the move that brings about the deviation from the actual world is taken. And, respectively, a possible world in which a departure from the actual world happens should be considered accessible from the actual world if beliefs and conjectures up to the point where the departure happens make the departure possible. In signaling games, this concerns a single variable: the conjecture that player 1 has about player 2's strategy choice in response to the counterfactual, off-the-equilibrium-path-signal S'. In the games studied here, since player 2 has only two actions, such a conjecture will be captured by a single variable, namely $y \in [0,1]$ representing the probability with which player 2 takes action a in response to the off-the-equilibrium-path signal S'. Let y be the conjecture that player 1 has about player 2's response to the off-the-equilibrium-path signal S' in the actual world (the equilibrium under study) and $y' \in [0,1]$ the conjecture in some other possible world. Any subset Y'_y of $[0,1]$ that contains the actual y is a set of other possible worlds accessible from the actual world.

Definition 1. *For signaling games with two types, two signals and two actions, Lewis's accessibility condition is:*

> *"If it were that S', then it would be true that $p' \in P' \subset [0,1]$" if and only if either*
>
> *(1) no world in which S' belongs to any $Y'_y \subset [0,1]$, or*
>
> *(2) some Y'_y does contain at least one world in which S', and the strict conditional "If S', then $p' \in P' \subset [0,1]$" holds at every world in Y'_y.*

How does this work out practically?

(1) Is there a possible world accessible from the actual world in which S' can be true at all? (If not, any counterfactual "If it were that S', then it would be true that $p' \in P' \subset [0,1]$" will be vacuously true.)

Yes, if there is at least one type of player 1 who at the actual conjecture $y \in [0,1]$ that player 1 has about player 2's choice of action in response to S' is indifferent between using the equilibrium signal S and the off-the-equilibrium-path signal S'.

In order to make sure that the counterfactual can be non-vacuously true, let there be at least one type from whom that is true, say t', and fix as a candidate consequence of the antecedent S' the belief p' that puts full weight on type t' (the set P' contains as its single element the belief p' that puts full weight to type t').

(2) The belief p' that puts full weight on type t' will be true, according to Lewis's accessibility criterion, if there is a set of possible worlds $Y'_y \subset [0,1]$ containing the actual world y (a set of conjectures about player 2's response to S'), for which S' is possibly true and that satisfies the condition that "If S' then p'." Is there such a set?

Yes, if there is the set of conjectures about player 2's choice in response to S' (some interval in $[0,1]$) such that at this conjecture type t' is indifferent or better off using S' while the other type, type t, is still strictly better off sticking to S. (Of course, because if type t were indifferent or better off using S', then the belief p' that puts full weight to t' would no longer be a strict consequence of S'.) There are two possible cases under which this can be true: Either (i) there is a probability $y'_t > y$ with which player 2 takes a after S' such that type t is indifferent between S and S'. Then $[y, y'_t)$ is the set Y'_y in which the strict conditional "If S', then p'" holds (of course, because at y'_t, type t could use S', and then the belief p' that puts full weight on type t' is not necessarily a consequence of the antecedent S'). Or (ii) type t is always strictly better off sticking to S, then $[y,1]$ is the set Y'_y.

But this, one realizes, is nothing but the condition that constrains beliefs in response to the off-the-equilibrium-path signal S' to type t' according to "forward-induction," which in the simple model examined implies discarding this type also by "divinity." It is straightforward to show that the converse holds also, namely that if a type t is discarded by "divinity" and hence "forward-induction," then one can find a set Y'_y that makes a belief p' that excludes t' non-vacuously true. The case that the condition holds with (1), and is hence non-vacuously true, is the case that S' can never possibly be observed because for no type there is a possible reaction of player 2 to S' that would make any of the types at least equally well off if they deviated to S'—the case that the information set after S' is, as say Govindan and Wilson, "not relevant." If point (2) holds with (i), type t will be excluded by "forward induction" respectively "divinity" but not by the "intuitive criterion"; if (2) holds with (ii), then type t is also excluded by the "intuitive criterion."

For signaling games with two types, two signals and two actions, due to the fact that other possible worlds can be captured by a single variable y' that takes values in $[0,1]$, it is possible to express in a mathematically precise way what it means that some possible world y' "differs less" from the actual world y than some other possible world: it can be evaluated by the distance $\mid y - y' \mid$. This distance captures the "certain fixed degree" in which another possible world y' differs from the actual world y, all the rest (that players are Bayesian rational and how player 2 reacts to the signal that is on the equilibrium path) being kept as it is in the actual world. This allows one to derive from Lewis's (1978) similarity condition a criterion of proximity in the sense of a well-defined mathematical distance.

Definition 2 (Lewisian proximity criterion)**.** *For signaling games with two types, two signals and*

two actions, let $y'_{P'}$ stand for a conjecture of player 1 about player 2's strategy choice after S' that makes a belief $p' \in P'$ after observation of S' a consequence of player 1's payoff maximizing and Bayes' rule, and $y'_{\neg P'}$, if it exists, for a conjecture about player 2's strategy choice after S' that makes $p' \in P'$ after observation of S' incompatible with player 1's payoff maximizing and Bayes' rule, then Lewis's (1978) similarity condition can be written in the form of the following proximity criterion*:*

> *"If it were that S', then it would be true that $p' \in P'$" is non-vacuously true iff*
>
> $$| y - y'_{P'} | < | y - y'_{\neg P'} | \quad \text{for any } y'_{P'} \text{ and } y'_{\neg P'}$$
>
> *if $y'_{\neg P'}$ exists.*

Obviously, this condition holds for any $y'_{P'} \in [y, y'_t)$ and any $y'_{\neg P'} \in (y'_t, 1]$, if y'_t exist (case i); if y'_t does not exist (case ii), there is no $y'_{\neg P'}$.

Definition 3. *For signaling games with two types, two signals and two actions, a sequential Bayesian Nash equilibrium outcome is called*

- counterfactual proof according to Lewis's accessibility condition *if there exists a sequential Bayesian Nash equilibrium with that outcome that is supported by beliefs off the equilibrium path that are true according to Lewis's accessibility condition for games (definition 1),*
- counterfactual proof in the sense of the Lewisian proximity criterion *if there exists a sequential Bayesian Nash equilibrium with that outcome that is supported by beliefs off the equilibrium path that are true according to the Lewisian proximity criterion (definition 2).*

Proposition 2. *For signaling games with two types, two signals, and two actions of the second player: (1)* counterfactual proofness according to Lewis's accessibility condition *is equivalent to "forward induction" (Govindan and Wilson 2009); (2)* counterfactual proofness in the sense of the Lewisian proximity criterion *is equivalent to "divinity" (Banks and Sobel 1987); and the two concepts coincide.*

The argument given above shows necessity of point (1). The proof below shows sufficiency of point (2). The missing arguments are easily obtained from the equivalence of "forward induction" and "divinity."

Proof. Suppose that according to "divinity" type t has to be excluded from the support of the belief after the off-the-equilibrium-path-signal S' and that hence only the belief p' that puts full weight to type t' can be maintained (the set P' contains a single element: the probability p' that puts full weight to type t). Then for type t' there will be an equilibrium with the outcome under study, and that is to say an equilibrium conjecture y regarding player 2's reaction to S', at which

type t' is indifferent between S and S', while type t strictly prefers to stick to S. This equilibrium conjecture y will be the smallest conjecture about player 2's reaction to S' that makes the belief p' after S' a consequence of player 1's payoff maximizing and Bayes' rule (of course, because for type t taking S' is payoff maximizing while it is not for type t). There will also be a largest conjecture about player 2's reaction to S' that makes p' after S' a consequence of player 1's payoff maximizing and Bayes' rule, which will be either (i) the probability y'_t with which player 2 has to take a after S' to make type t indifferent between S and S', if such a probability exists (of course, because for any probability beyond that value, both types will be strictly better off using S' and so the p' that puts full weight on t is not compatible anymore with player 1's payoff maximization and Bayes' rule), or (ii) equal to 1, namely if type t is always strictly better off sticking to his equilibrium strategy S no matter how player 2 reacts to S' (which is the case if type t has to be discarded after S' not only by divinity but already by the intuitive criterion). If (i), for any $y'_{P'} \in [y, y'_t)$, the p' after S' that puts full weight to type t is a consequence of player 1's payoff maximizing and Bayes' rule, and for any $y'_{\neg P'} \in (y'_t, 1]$, p' after S' is not compatible with player 1's payoff maximizing and Bayes' rule. Given the equilibrium y,

$$\mid y - y'_{P'} \mid < \mid y - y'_{\neg P'} \mid \quad \text{for any } y'_{P'} \in [y, y'_t) \text{ and } y'_{\neg P'} \in (y'_t, 1],$$

and hence the p' that puts full weight on t' after S' is non-vacuously true. If (ii), then player 1's payoff maximizing and Bayes' rule make the belief p' after S' a necessity, and hence the p' after S' is non-vacuously true. □

The equilibrium outcome P1 in the game in figure 1 (which exists under the prior $p < 1/2$) provides a good illustration of the result (table 3). It also should be noted that the equilibrium outcome $P3$ (which exists under the prior $p > 1/2$) is a case where any counterfactual conditional is trivially true (the costly signal $S^\star$ can never be observed in any possible world, because in the actual world, by using the costless S, both types get a strictly higher payoff then they could get from using $S^\star$, for any possible reaction that player 2 could have to $S^\star$).

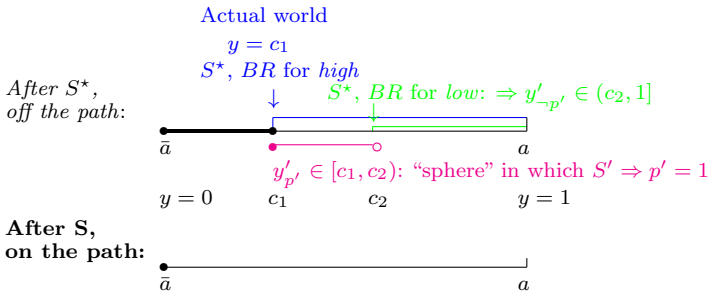
The interpretation of this result is twofold: A possible conclusion is that for the class of games considered, the intuition that feeds "divinity" can also be expressed as Lewis's similarity condition expressed as a criterion of proximity in the mathematical sense. In this sense, philosophy gives support to what game theorists have proposed. On the other hand, one can also draw the conclusion that the foundation of the forward-induction criterion proposed by Govindan and Wilson in terms of "invariance and sequentiality" also provides a foundation to Lewis's idea of "accessibility determined by similarity"—at least when it is translated into a game setting of the simple form considered here. In that sense, game theory (for the setting of games) offers a foundation for a pattern of reasoning that philosophers have proposed.

Table 3. Counterfactual proofness for the game in figure 1a

The equilibrium outcome P1, in which both types use the unmarked signal S, is supported by those strategy choices of player 2 in which in response to the off-the-equilibrium-path signal $S^\star$, player 2 takes a with a probability $y \in [0, c_1]$. We know that by divinity, and respectively, forward induction, the low type has to be excluded after $S^\star$, that is, full belief has to be put to the high type. Therefore, the belief p' that shall be shown to be "non-vacuously true" in case that the off-the-equilibrium path signal $S^\star$ is observed is $p' = 1$. Which conjectures about player 2's strategy choices guarantee that the belief $p' = 1$ after $S^\star$ is a consequence of Bayes' rule and player 1's payoff maximizing behavior? It is given by any $y'_{p'} \in [c_1, c_2)$. Surely, because for any probability in this interval, the high type will be equally well or strictly better off using $S^\star$, while the low type will be strictly better off sticking to the equilibrium S. On the other hand, for any strategy choice of player 2 that puts a probability strictly higher than c_2 on a after $S^\star$, it cannot be that $p' = 1$, because then the low type would also be better off using $S^\star$. Hence, $y'_{\neg p'} \in (c_2, 1]$. Now, for any equilibrium y, any $y'_{p'} \in [c_1, c_2)$ and $y'_{\neg p'} \in (c_2, 1]$,

$$| \, y - y'_{p'} \, | < | \, y - y'_{\neg p'} \, |,$$

obviously, because $c_1 < c_2$. That is: a possible world in which observation of $S^\star$ reveals the high type ($p = 1$) differs less from the actual world (the equilibrium under consideration) than does any possible world where observation of $S^\star$ does not reveal the high type.



The coincidence of "forward induction" and "divinity," as pointed out before, does not go beyond the class of 2-states–2-signals–2-actions games. My conjecture is that beyond that class, Lewis's accessibility criterion can also no longer be captured by "proximity" as expressed by the mathematical distance between the possible reactions to off-the-equilibrium-path signals, so that in general one has the formula

"forward induction" = "accessibility," and

"divinity" = "similarity based on proximity."

5 Applications in the study of language

When are these results relevant? For which kind of games with two states, two signals, and two actions do they have any selection force?

5.1 Arbitrary conventional meaning

A class of games in which none of the criteria considered here is ever potentially effective (never sorts out any equilibria) are signaling games in which signals are of no costs, or more generally, in which the costs or benefits of using a signal do not depend on the type of the first player (the state of nature). These are signaling games—and that is an unfortunate interrelation—as Lewis has introduced them in *Convention* (1969) to support the argument that a conventional signaling system, in which signals have *arbitrary* meaning, can be sustained in a decentralized way because it is supported by Nash-equilibrium conditions.

Lewis's signaling games differ from the game in figure 1 also in that the sender and the receiver have coinciding interests throughout—in every state of nature. To transform the game in figure 1 into a signaling game in the style of Lewis (1969) one has to interchange the payoffs of the low type after a with those after $\bar{a}$, so that in case of the "low state" player 1 also prefers player 2 to take $\bar{a}$; and one has to set $c_2 = c = c_1$. In such a game, "high" and "low" have no sensible interpretation. Call the two states rather "state 1 and "state 2." Such a game will have two perfectly separating equilibria: (1) a perfectly separating equilibrium in which player 1, if state 1 occurs, takes $S^\star$, and if state 2 occurs, takes S, and player 2 in reaction to $S^\star$ takes a and in reaction to S takes $\bar{a}$; and (2) a perfectly separating equilibrium in which player 1, if state 1 occurs, takes S, and if state 2 occurs, takes $S^\star$, and player 2 in reaction to $S^\star$ takes $\bar{a}$ and in reaction to S takes a. Depending on the prior p and the cost c, there might also be equilibria in perfectly mixed strategies; and there are always pooling equilibria in which in both states player 1 uses the same signal. Depending on the prior and the cost c, it might be the case that one of the perfectly separating equilibria is "more efficient" than the other (gives both players a higher payoff). However, no matter what the specific parameters of such a game, all equilibria will always pass any test based on the plausibility of beliefs off the equilibirum path. Surely, in a perfectly separating equilibrium (even when it is inefficient) there is no signal off the equilibrium path, and as a consequence, no restriction on beliefs off the equilibrium path ever possibly applies in a non-vacuous way. In an equilibrium in which player 1 uses the same signal in both states and there is hence a signal "off the equilibrium path," none of the refinements discussed here has the force to exclude any of the states after an off-the-equilibrium-path signal, because there is no intrinsic relation between the states and any of the signals. Assume there is an equilibrium outcome in which player 1 uses S in both states. True, if $c > 0$, the off-the-equilibrium path signal $S^\star$ costs something, but because this cost is the

same for player 1 no matter what the state, the response of player 2 to $S^\star$ that would make player 1 equally well off from using $S^\star$ (compared to what he gets from using S in the equilibrium under study) is the same no matter what the state. Any counterfactual conditional will be as valid as any other. Counterfactuals play no role in these games for the very reason that what a signal means in equilibrium is arbitrary.

What one needs in general for any belief-based refinement to be potentially effective are different incentives of different sender types to use the off-the-equilibrium-path-signal. This is typically the case when different sender types face different costs for producing the signal (as in the game in figure 1), but can also be satisfied if using the signal procures different positive payoff increments to different types,[14] or through a combination of the two. There has to be some intrinsic relation between the nature of signals and the payoffs of different sender types. The domain of applications of belief-based refinements is not where it is about arbitrary meaning of signals but where signals acquire their meaning *pragmatically*. In the following, I outline two possible applications in the study of language.

5.2 Sociolinguistics

The game in figure 1, which, as mentioned, can be interpreted as a model of signaling in the job market in the sense of Spence (1973) with the costly signal $S^\star$ standing for a certain educational credential, has a direct application related to language: language competences often come to function as such educational signals. This can concern foreign language competences, but also competences in one's own language, for example, to speak and write grammatically correct, to express oneself in certain ways, to be eloquent, witty, or to speak with a certain accent. And this, of course, goes beyond the job market. To be able to speak in certain *styles*, to have or not to have a certain accent, to speak or not to speak in a certain dialect, or the sheer ability to shift between different such styles, might come to function as a signal for certain social qualities, for example, the ability to recognize and to abide to certain rules, the ability to adapt to different social environments, etc.

The game in figure 1 might be a suitable model to investigate certain aspects of problems in sociolinguistics related to accents, dialects, code-switching and style shifting *if it can be argued that different types incur differential costs from producing the costly signal*, or in any case, *differential gross costs from using the costly signal*, which of course has to be evaluated on linguistic grounds. It might, for example, give some insight into the phenomenon that two different styles, an "official style" and a "dialect," coexist. The coexistence of two styles might appear as inefficient, at first

[14] That is the case in Cho and Kreps's (1987) leading example, the so-called *beer-quiche game*, a 2-states–2-signals–2-actions game, in which different types do not pay of cost for emitting different signals (having a beer or a quiche for breakfast) but get some positive payoff increment from emitting these two different signals.

sight, a waste of social resources, but it might fulfill a social function, namely providing a signaling mechanism. If everybody is forced to be fluent in two different styles and to be able to switch between them in differently marked situations, different individuals might do that differently well, and this can be exploited as a signaling mechanism.

A signal, it is important to keep in mind when debating applications, might operate on conditions found in the environment that have evolved for some other reason than the function of providing a signaling mechanism; maybe just by chance. In the study of animal signals this has been explicitly pointed out, notably with respect to the interplay between intra and interspecies signaling. What functions as a signal to enhance cooperation between members of the same species (which might have evolved for that very reason) can easily come to function as a costly signal that gives away information to a predator (see, for example, Dawkins and Krebs 1978). Similarly when it comes to human language: saying that a signaling mechanism might thrive on the coexistence of different accents or dialects is not to say that this is why different accents or dialects have evolved or that that would be their main function.

Nevertheless, even if it can be sustained that different styles, or the ability to switch between different styles, functions as a signal, the assumption that different individuals face different costs for doing this and that these differences correlate with some other social quality (which is what defines the different "types") is very restrictive. This is notably true when the signal is not something in the say "physical quality of speech" (like an accent or a dialect) but when the signal is something in the semantic quality of what is said. To model this requires a game slightly different from that in figure 1.

5.3 An application in linguistic pragmatics: politeness implicatures

A number of problems in linguistics pragmatics can be phrased in terms of a marked form $S^\star$ and an unmarked form S, with the marked form $S^\star$ having a differential cost over the unmarked form S. The cost of a marked form $S^\star$ will often be explained in terms of *complexity* or *social obligations* that emanate from the speech act (or both); and this cost can be carried either by the sender or the receiver (or both of them). Communicative implicatures based on politeness are a typical example (Brown and Levinson 1987). More polite forms are typically longer, more elaborate, more complex, and often they will also impose some social cost on the speaker.

The formulation of the problem as presented by Brown and Levinson, with the assumption of rational individuals whose motives are expressed as their caring about their *face*, which is inspired by the rational individual as it appears in economics, lends itself to a formalization in terms of games in the game theoretic sense.

The basic underlying situation of strategic interaction between the sender and the receiver, at least in some of the situations that might be relevant for politeness, might correspond to a game

with the same base payoffs as that of the game in figure 1 (the payoffs before the costs of the signal are deducted) with the following interpretation: in one state of nature, the two players have coinciding interests: player 1 cares about the face of the other player as much as he cares about his own—a situation that corresponds to the "high" type in figure 1; in the other state of nature, the players have opposed interests: player 1 does not care about the face of the second player—a situation that corresponds to the "low" type. And, of course, player 2 wants to take the right decision, which depends on which type player 1 is: if player 1 is of the high type, she wants to take a; if player 1 is of the low type, she wants to take $\bar{a}$. The action a might stand for many things: pass player 1 the salt, lend him money, show him the way—to echo some of the typical examples used in discussions of politeness. More generally, these transactions (passing the salt, lending money, etc.) can be framed as a change in the relationship type, and of course, so the assumption of the model then, if this change is desirable for the second player depends on the type of the first player.

Polite forms can be construed as costly in a double sense: first, in that polite forms usually are more complex than not polite forms: "Can you please pass the salt" versus "Pass the salt"; and second in that polite forms often bear a social cost to the sender, for example, in that he takes on a social or moral debt (he exposes himself to a "face threat"): "I'll never be able to pay you back if you .."; or by debasing his own face: "I must be absolutely stupid but I simply can't understand this map."

At least two of the basic features of the game in figure 1 then are met by models of politeness: (1) there is an underlying identification problem between the sender and the receiver of a speech act (the signal), and (2) different speech acts have different costs.

The game in figure 1 has one feature, though, that is difficult to justify in applications to politeness, or pragmatics in general, namely that different types incur *differential costs in producing the marked form* $S^\star$ or more generally *differential gross costs from using the marked form* $S^\star$. If costs are grounded in complexity or the social cost that they bear on the sender, then the payoff deduction that results from using the marked form should be the same no matter of which type player 1 is. "I must be absolutely stupid but I simply can't understand this map" is longer than "Show me the way," and threatens the face of the person who utters it, no matter if it is said by someone who cares about the face of the addressee or not.[15]

[15]Van Rooy (2003) has first pointed to this problem in a slightly different development. The solution that he proposes, with reference to Zahavi's (1975) *handicap principle*, is that the cost of being polite is easier to bear by a sender with a higher background payoff. This is however—and this is why van Rooy's argument is not fully closed in a game-theoretic sense—not the mechanism that one can see at work in the formal model presented by van Rooy. In the game presented by van Rooy, the high type incurs a cost of 1/2 when he uses the polite request and the low type of 2/3, and these costs are deducted from a uniform background payoff for the two types. In van Rooy's game, if player 2 takes the "good" action, both the high and the low type have a base payoff of 1, and if player 2 takes

5.4 A game that models the "handicap principle"

What is needed to account for politeness is a game with two signals, a marked form $S^\star$ and an unmarked form S, and a cost-benefit structure that explicitly models the *handicap principle* (Zahavi 1975, Grafen 1990): that is, a uniform "gross" cost for producing or using the marked form $S^\star$ that is deducted from differential background payoffs for the two types. Figure 2 shows such a game. This game differs from the game in figure 1 in two aspects: (1) the two types of player 1 face the same gross cost c for producing or using the more polite form $S^\star$; and (2) if the second player takes action a, the high type gets some extra payoff d that adds to the base payoff $(= 1)$ that both types get if player 2 takes a.

This game can be seen as a discrete variant of Milgrom and Roberts's (1986) model of advertising as a signal for product quality and Grafen's (1990) formalization of the handicap principle. A feature of this game needs explanation, namely that the high type has a higher base payoff (the extra d) only if the second player takes the "good" action a. In both Milgrom and Robert's model and Grafen's model, which are both stated for continuous strategy spaces, this assumption appears in the form of conditions on the derivatives of the payoff functions. Milgrom and Roberts base this assumption on the idea that a high quality product, if consumed once, will attract more consumption in the future, and therefore the firm providing it will profit more from a first sale than a firm with a lower quality product. In the context of the handicap principle, where payoffs stand for *fitness*, these assumptions also have to be understood as being grounded in an implicit dynamic argument: an individual with high background fitness profits more from an act of reproduction than an individual with lower background fitness—because his offspring too will have a higher background fitness and therefore a higher chance to reach reproduction age and therefore reproduce himself. A similar argument can be advanced in the context of politeness: the type who really cares about player 2 has more to win than the type who does not if player 2 takes the "good" action a, because then player 1 and 2 can develop their relationship, which gives player 1 a chance to prove himself to be of the high type and they can enjoy the fruits of their having aligned interests; while they cannot exploit the possibilities of their having aligned interests if the request of the first player is rebuffed in the first place. Combined with the assumption that player 2's taking $\bar{a}$ is somewhat neutral in that it gives neither a positive nor negative payoff to any of player 1's types, one arrives at a game like that in figure 2.

---

the "bad" action, they both have a base payoff of 0; exactly as in the game in figure 1. Van Rooy's game differs from that in figure 1 only with respect to the payoffs of the second player, but in an inessential way; structurally the two games are identical.

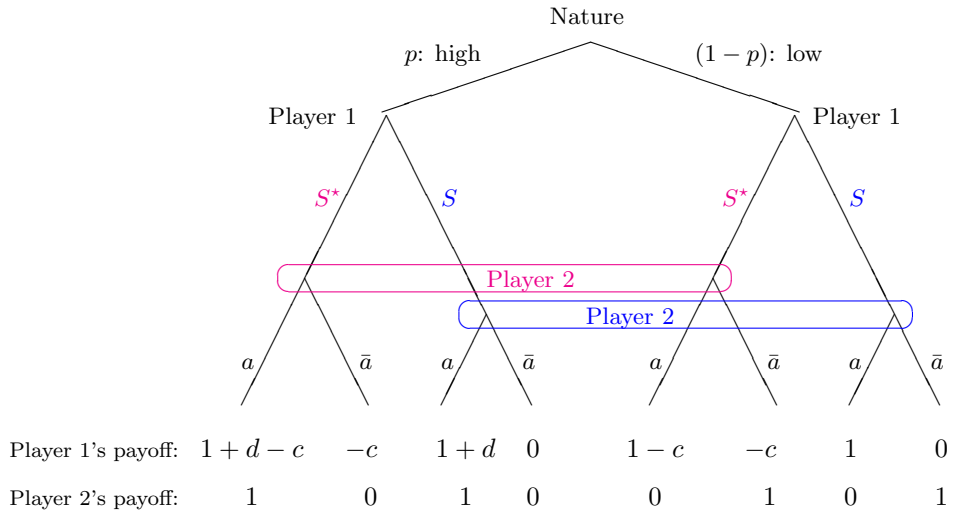


Figure 2a. Uniform costs $0 < c < 1$, and differential benefits, $0 < d \leq 1$.

	aa	$a\bar{a}$	$\bar{a}a$	$\bar{a}\bar{a}$
$S^\star S^\star$	$1-c+pd,\ p$	$1-c+pd,\ p$	$-c,\ 1-p$	$-c,\ 1-p$
$S^\star S$	$1+p(d-c),\ p$	$p(1+d-c),\ 1$	$-pc+(1-p),\ 0$	$-pc,\ 1-p$
$SS^\star$	$1+pd-(1-p)c,\ p$	$(1-p)(1-c),\ 0$	$p(1+d)-(1-p)c,\ 1$	$-(1-p)c,\ 1-p$
SS	$1+pd,\ p$	$0,\ 1-p$	$1+pd,\ p$	$0,\ 1-p$

Figure 2b. The normal form of the game in figure 2a.

The game in figure 2—and this is what makes its explanatory potential—is *structurally equivalent* to the game in figure 1. To make this more precise, one can define as the *net cost* of a signal for type t the payoff difference that results for type t from sending the signal and not sending it if the signal does not have the desired effect on the receiver (if the receiver takes $\bar{a}$ anyway) over the payoff difference that results for this type if he sends the signal and the receiver in response to the signal takes or does not take the desired action:

$$\text{net cost of } S^\star \text{ for type } t \;=\; \frac{\pi_t(S,\bar{a}) - \pi_t(S^\star,\bar{a})}{\pi_t(S^\star,a) - \pi_t(S^\star,\bar{a})}.$$

In the game in figure 1, the net cost of the marked signal $S^\star$ for the high type is c_1, and for the low type c_2; in the game in figure 2, the net cost of the marked signal $S^\star$ for the high type is $c/(1+d)$, and for the low type c. In both games then, the two types have *differential net costs of the marked form* $S^\star$. And, this is all that matters as far as the qualitative predictions of the model are concerned: the game in figure 2 has the *same equilibrium structure* as that in figure 1: the numerical values defining the equilibria and the payoffs in equilibrium can be obtained by those of the game in figure 1 by replacing c_1 by $c/(1+d)$ and c_2 by c.[16] In addition to that, these equilibria will have the *same refinement properties*: If $p < 1/2$, the equilibrium outcome

[16]Similarly as for the game in figure 1, perfectly separating equilibria do not exist if $c < 1$ but only if $c \geq 1$.

P1, in which both types use the unmarked signal and the receiver never takes a, can be discarded on the grounds of "divinity," forward induction, and Lewisian counterfactual proofsness, and as a consequence the partially separating equilibrium PS1 remains as the only prediction of the model. If $p > 1/2$, instead, none of the three equilibrium outcomes can be discarded, by none of these criteria. The arguments work out exactly as those for the game in figure 1a (table 2).

The structural equivalence of the two games is an important observation because it shows that differential costs in producing the signal and differential net costs of the signal emanating from differential benefits from using the signal if the signal has the desired effect on the receiver (the handicap principle) are not two rival mechanisms for explaining signaling phenomena but work in the same direction. In some applications both factors might come in. If this is the case, then one does not need to weigh off one factor against the other.

That this equivalence is not obvious, becomes more evident, when one compares the game in figure 2 to that in figure 3, in which the high type has an extra payoff of d irrespective of the reaction of player 2 (and not only if player 2 takes a). In this game, the net cost of the marked polite form $S^\star$ will be c for both types; the costly signal is *not* of differential net costs for the two types. Such a game will, on the surface, have the same equilibria as those of the game in figure 1 and respectively 2 (one gets them from those of the game in figure 1 by substituting c for both c_1 and c_2).[17] However, and here is where the structural difference is, these equilibria *will not have the same refinement properties*: for the game in figure 3, no matter what the prior p, none of the equilibria can be rejected on the basis of "divinity," forward induction, and Lewisian counterfactual proofness. This has consequences for the explanatory potential of the model. The game in figure 3 has less force than the games in figure 1 and respectively 2 to account for signaling phenomena when the supposed role of the signal is to push up the belief of player 2 that player 1 is of the high type. It is true that the game in figure 3, if $p < 1/2$, has an equilibrium in which the signal is used and at least partially informative, but this equilibrium is no more credible as a solution of the game than the outcome in which no signaling happens.

Another valuable comparison is between the game in figure 2 with $0 < c < 1$, as assumed, and $c = 0$, which I have excluded. If $c = 0$, the game in figure 2 is a so-called *cheap-talk game*. The only equilibria that exist in this case are equilibria in which both types of player 1 use $S^\star$ with some probability $0 \leq x \leq 1$ and S with the complementary probability $1 - x$, and player 2 acts on her prior belief. In these equilibria, because both types use $S^\star$, and respectively S, with the same probability, none of these signals will carry any information. Equilibria of this type are therefore often referred to as *babbling equilibria*. None of these equilibria can be excluded by any

[17] That the extra d is irrelevant for determining the equilibria of the game, is straightforward to see from the normal-form representation of the game (figure 3b), for the extra d multiplied by the probability of the high type p is added to player 1's payoff everywhere.

belief-based refinement.

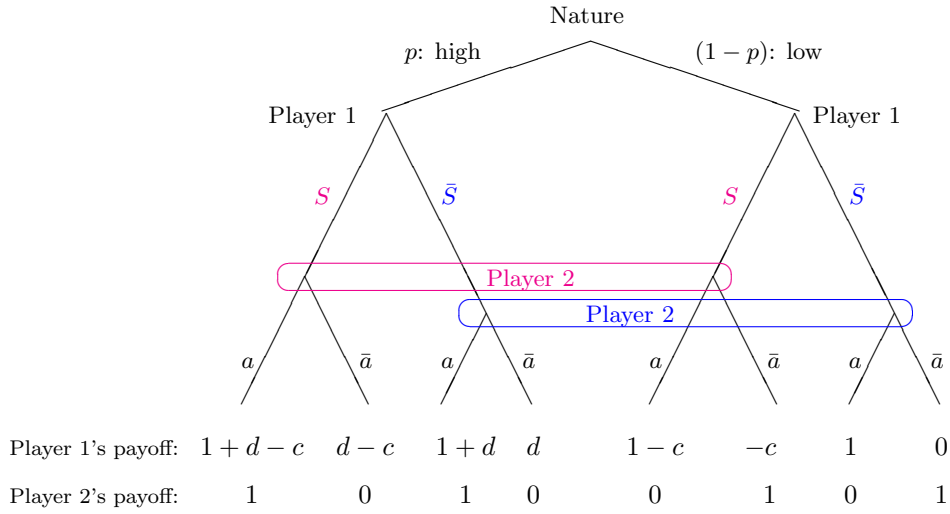


Figure 3a. $0 < c < 1$.

	aa	$a\bar{a}$	$\bar{a}a$	$\bar{a}\bar{a}$
SS	$1-c+pd,\ p$	$1-c+pd,\ p$	$pd-c,\ 1-p$	$pd-c,\ 1-p$
$S\bar{S}$	$1+p(d-c),\ p$	$p(1+d-c),\ 1$	$p(d-c)+(1-p),\ 0$	$p(d-c),\ 1-p$
$\bar{S}S$	$1+pd-(1-p)c,\ p$	$pd+(1-p)(1-c),\ 0$	$p(1+d)-(1-p)c,\ 1$	$pd-(1-p)c,\ 1-p$
$\bar{S}\bar{S}$	$1+pd,\ p$	$pd,\ 1-p$	$1+pd,\ p$	$pd,\ 1-p$

Figure 3b. The normal form of the game in figure 3a.

5.5 More specific conclusions with respect to politeness in language

Which conclusion can one draw from this investigation more specifically with respect to the explanatory potential of signaling games for implicatures driven by politeness?

The point of view that underlies the strategic stability program is perfectly suited to embed a model of politeness: a given equilibrium is taken as a convention that governs social interaction, which here governs how the marked form $S^\star$ and the unmarked form S are used. But not an arbitrary convention: it is a convention that is compatible with Bayesian rational behavior, and one that survives in a sort of meta thought experiment under which all possible conventions have to pass the test that things that never happen under this convention are compatible with a certain way of drawing inferences from the use and not use of $S^\star$.

How does players' counterfactual reasoning play out in the explanation of the politeness implicatures as modeled here? Take first the case that $p < 1/2$, which is, of course, just a numerical value resulting from the specific payoffs that I have assumed, which in general stands for a situation in which without any further information the exchange that it is all about—that player 2 takes a—will never happen because player 2's belief that the first player is of the "good type" is

too low. The partially separating equilibrium PS1, which is the one selected by "divinity," forward induction, and Lewisan counterfactual proofness, has the property that the unmarked form S perfectly reveals the "low" type. The marked, "polite" form $S^\star$, instead, does not fully reveal the "high type" but merely induces player 2 to push up her belief that player 1 is of the "high" type, namely to such an extent that she is indifferent between a and $\bar{a}$. This can be interpreted as mimicking a feature of politeness as it appears in language use: to hear the polite form invites the addressee of that speech act to do an inference, but not an extreme inference, rather an inference that keeps her neutral with respect to how to respond to the polite request, taking a or $\bar{a}$. *Not the hear the polite form*, instead, lets her know for sure that the speaker is not very well meaning, and she knows that $\bar{a}$ is the right response.

The alternative extensive form in figure 1c, in which player 1's strategy SS ("never use the marked form") plays the role of an outside option, and in which, under a low prior $p < 1/2$, only the partially separating equilibrium can be sustained as a sequential Bayesian Nash equilibrium, has a meaningful interpretation in the context of politeness, or pragmatics more generally. The inference that player 2 makes from observing $S^\star$ can be interpreted in the sense of a *communicative implicature* as proposed by Grice (1975): to never use the costly signal in principle (playing according to SS) is like talking according to the Maxims postulated by Grice. Making instead the choice to use the marked form $S^\star$ is to flaut the Maxims; it means to push player 2 "into the game" (into the subgame, to be more precise), into the thought experiment by which she will come to understand that $S^\star$ means a belief about player 1 that makes her is indifferent between a and $\bar{a}$—the implicature.

The case that $p > 1/2$ might look less interesting first because none of the three equilibrium outcomes that exist under this prior can be sorted out on the basis of the criterion in which coincide "divinity," forward induction, and Lewisian counterfactual proofness. One might wonder if this a negative result, a result that shows the limits of the theory? Such a view rests on the assumption that a theory of equilibrium refinement always has to single out a unique equilibrium. But the multiplicity of different solutions—all equally plausible, justifiable on rational grounds—might mimic reality. If a theory predicts under some conditions uniqueness of the solution, and under some other conditions multiplicity of the solution, this should not be held against the theory but rather be seen as part of its explanatory potential in that it can identify the conditions under which uniqueness or multiplicity of the solution prevails.

The case that $p > 1/2$ is often neglected, if not completely ignored, in biological studies of costly signaling. The reason for this might be that it is considered not important because one might think that in this case the signaling mechanism is not needed (because player 2 is ready to take a on his prior belief) and that one therefore can safely ignore signaling effects. The game-theoretic analysis shows that this is not the case: it shows that if some costly, marked form $S^\star$ is

potentially considered a signal, and this is common knowledge, the second player might condition her action on the observation of that signal, and if (for whatever reason) this is the case, then everybody will be obliged to use that signal even though it does not transport any additional information beyond what is already commonly known. The availability of a signaling mechanism can be the source of a social dilemma: Use routinely the polite form or not? The co-existence of the two equilibria P2 (everybody uses the polite form) and P3 (nobody uses the polite form) and their persistence under any refinement criterion might serve as an explanation why one can indeed observe both: societies which, when it is known that the prior is high, nevertheless go with the more polite form—go with "overstatement"; and societies that go with the unmarked, less polite form—go with "understatement."

The co-existence of the two equilibrium outcomes P2 and P3 in case of a high prior contributes also to the explanatory potential of the model it that it provides the basis for a phenomenon sometimes referred to as "countersignaling" (Feltovich et al. 2002), which is the idea that it is exactly the absence of the signal that comes to stand for a "very high type," a type who is sure to carry some other trait that makes him recognizable as a high type (or at least push the probability that he is of the high type sufficiently up) and therefore does not need to signal, as oppose to just a "high type," who is not sure to carry that other trait and therefore needs to make use of the costly signal to distinguish himself from the "low" type, which in the end produces the phenomenon that the "low type" and the "very high type" use the same signal. This too is a familiar phenomenon for researchers who study politeness: in a context where the conversational partners can be sufficiently sure that their interests coincide, a less polite form might be used, and the usage of that less polite form will exactly express that closeness—in fact, the *common knowing of that closeness*.

5.6 What is meaning making in a signaling game?

The discussion of "countersignaling" points to a more fundamental aspect of signaling games that is highly relevant in applications to language but that I have so far not explicitly pointed out: namely that because the equilibria depend on the prior, which by assumption is commonly known, the prior feeds into the "meaning" of a signal. *What is meaning making in a signaling game is not the signal alone but the combination of a signal with the prior.* In a signaling game, the meaning of signals has to be understood as a *system of meaning* that is structured by the prior. Clearly, the absence of the marked form in presence of a low prior means something else than the absence of the marked form in presence of a high prior. The prior, if one wishes, is part of the commonly known "context." This allows one to investigate aspects of meaning that go beyond information transmission concerning the state of nature (high or low type). What is really communicated, for instance, in speech acts that thrive on countersignaling, is not any information about the state of nature; rather the speech acts becomes a public display of a high prior on a specific state of nature

(the very high type).

That meaning and belief "play interlocking and complementary roles in the interpretation of speech" (Davidson 1974) has long been understood in the philosophy of language. Game theory might be a valuable tool for researchers who study language in that it allows to see this dependency arise in a simple mathematical model—not because it has been put into the model as an assumption, but because is arises as part of the solution on the game: equilibrium and its refinements.

References

[1] Banks, J. S. & Sobel, J. (1987). Equilibrium selection in signaling games. Econometrica, 55(3), 647–661.

[2] Brown, B. & Levinson, C. S. (1987). Politeness: Some Universals in Language Usage. Cambridge, New York: Cambridge University Press.

[3] Cho, I-K. & Kreps, D. M. (1987). Signaling games and stable equilibria. Quarterly Journal of Economics 102(2), 179–221.

[4] Dawkins, R. & Krebs, J. R. (1978). Animal signals: information and manipulation. In: Krebs J. R. & Davies N. B. (Eds.), Behavioral Ecology: An Evolutionary Approach (pp. 282–309). Oxford: Blackwell.

[5] Davidson, D. (1974). Belief and the basis of meaning. Synthese 27, 309–323.

[6] Elmes, S. & Reny, P. J. (1994). On the strategic equivalence of extensive form games. Journal of Economic Theory, 62, 1–23.

[7] Feltovich, N., Harbaugh, R. & To, T. (2002). Too cool for school: Signalling and countersignalling. *RAND Journal of Economics*, 33(4), 630–649.

[8] Govindan, S. & Wilson, R. (2009). On forward induction. Econometrica, 77(1), 1–28.

[9] Grafen, A. (1990). Biological signals as handicaps. Journal of Theoretical Biology, 144(4), 517–546.

[10] Grice, H. P. (1975). Logic and conversation. In Cole, P. & Morgan, J.L. (Eds.), Syntax and Semantics, Vol. 3, Speech Acts (pp. 41–58). New York: Academic Press.

[11] Harsanyi, J. C. (1967). Games with incomplete information played by 'Bayesian' players. Management Science, 14(3), 159–182.

[12] Hillas, J. & Kohlberg , E. (2002). Foundations of strategic equilibrium. In Aumann, R. J & Hart, S. (Eds.) Handbook of Game Theory Vol. 3 (pp. 1597–1663). Amsterdam, New York: Elsevier.

[13] Kohlberg, E. & Mertens J.-F. (1986). On the strategic stability of equilibria. Econometrica, 54(5), 1003–1037.

[14] Kreps, D. M. & Sobel, J. (1994). Signalling. In Aumann, R. J. & Hart, S. (Eds.) Handbook of Game Theory, Vol. 2 (pp. 849–867). Amsterdam, New York: Elsevier.

[15] Kreps, D. M. & Wilson, R. (1982). Sequential Equilibria. Econometrica, 50(4), 863–894.

[16] Kuhn, H. W. (1953). Extensive games and the problem of information. In Kuhn, H. W. & Tucker, A. W. (Eds.), Contributions to the Theory of Games, Vol. II (pp. 193–216). Princeton: Princeton University Press.

[17] Lewis, D. (1969). Convention: A Philosophical Study. Cambridge MA: Harvard University Press.

[18] Lewis, D. (1973). Counterfactuals. Cambridge MA: Harvard University Press.

[19] Lewis, D. (1978). Truth in fiction. American Philosophical Quarterly, 15(1), 37–46.

[20] McLennan, A. (1985). Justifiable beliefs in sequential equilibrium. Econometrica, 53(4), 889–904.

[21] Mailath, G., Okuno-Fujiwara, M. & Postlewaite, A. (1993). Belief-based refinements in signaling games. Journal of Economic Theory, 60, 241–276.

[22] Milgrom P. & Roberts, J. (1986). Price and advertising signals of product quality. Journal of Political Economy, 94(4), 796–821.

[23] Nash, J. (1950). Equilibrium points in n-person games. Proceedings of the National Academy of Sciences, 36, 48–49.

[24] Nash, J. (1951). Non-cooperative games. The Annals of Mathematics, 54(2), 286–295.

[25] Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. International Journal of Game Theory, 4(1), 25–55.

[26] Sobel, J. (2009). Signaling Games. In R. Meyers (Ed.) Encyclopedia of Complexity and System Science (pp. 8125–8139). New York: Springer.

[27] Spence, M. (1973). Job market signaling. Quarterly Journal of Economics, 87(3), 355–374.

[28] Számadó, S. (2011). The cost of honesty and the fallacy of the handicap principle. Animal Behavior, 81, 3–10.

[29] Van Rooy, R. (2003). Being polite is a handicap: towards a game theoretical analysis of polite linguistic behavior. Proceedings of TARK 9.

[30] Zahavi, A. (1975). Mate selection—a selection for a handicap. Journal of Theoretical Biology, 53(1), 205–214.

[31] Zollman, K. J. S., Bergstrom, C. T. & Huttegger, S. M. (2013). Between cheap and costly signals: the evolution of partially honest communication. Proceedings of the Royal Society London B, 280, 20121878.