The evolutionary dynamics of costly signaling

Josef Hofbauer

University of Vienna, Department of Mathematics

and

Christina Pawlowitsch

Université Panthéon-Assas, Paris II, LEMMA–Laboratoire d'Economie Mathématique et de Microéconomie Appliquée

July 23rd, 2019

Abstract

Costly-signaling games have a remarkably wide range of applications, from education as a costly signal in the job market over handicaps as a signal for fitness in mate selection to politeness in language. While the use of game-theoretic equilibrium analysis in such models is often justified by some intuitive dynamic argument, the formal analysis of evolutionary dynamics in costly-signaling games has only recently gained more attention. In this paper, we study evolutionary dynamics in two basic classes of games with two states of nature, two signals, and two possible reactions in response to signals: a discrete version of Spence's (1973) model and a discrete version of Grafen's (1990) formalization of the handicap principle. We first use *index theory* to give a rough account of the dynamic stability properties of the equilibria in these games. Then, we study in more detail the replicator dynamics and to some extent the best-response dynamics. We relate our findings to equilibrium analysis based on classical, rationality-oriented methods of equilibrium refinement in signaling games.

Keywords: Replicator dynamics, best-response dynamics, index, sequential equilibrium, equilibrium refinement

Acknowledgements: Financial support from the ANR SIGNAL (ANR-19-CE26-0009) is gratefully acknowledged.

1 Introduction

Can an observable variable of choice like a degree from a good school or an advertisement in a fancy magazine or—when we look into the natural world—an observable trait like a prominent tail or elaborate plumage certify some unobservable characteristic like performance, product quality or reproductive fitness? Theories of *costly signaling* explain such phenomena in terms of costs associated to the variable or trait that functions as a signal (Spence 1973, Zahavi 1975). The problem of costly signaling, in the large sense, is a problem of *cooperation*: It is the question whether signaling costs are an effective mechanism to facilitate information transfer—and hence cooperation—in situations where one player faces uncertainty about the type of the other player, namely whether the other is "cooperative" (of high quality, high performance, high fitness) or not.

In economics, classical illustrations of costly signaling are education as a costly signal in the job market (Spence 1973), dividend payments as a signal for a firm's fundamentals (Miller and Rock 1985), and advertising as a costly signal for product quality (Milgrom and Roberts 1986). In biology, costly-signaling games have been used to formalize Zahavi's (1975) *handicap principle* (Grafen 1990), the hypothesis that traits that represent a handicap in relation to the ecological problems of the species come to function as a signal for high fitness types in mate selection. Other natural phenomena for which costly-signaling explanations have been proposed are signals in predator-prey or respectively parasite-host interaction (Caro 1986, Bergstrom and Lachmann 2001, Archetti 2008), and the begging behavior of offspring as a signal for their need directed to parents (Godfray 1991, Maynard Smith 1991). In anthropology and sociology, costly-signaling arguments have been suggested as alternative or complementary explanations to reciprocal altruism in accounting for certain forms of communal sharing and gift-giving (Bliege Bird et al. 2001). Other puzzling phenomena of social life for which explanations in terms of costly signaling have been advanced are the practice of inefficient foraging strategies, rituals, and embodied handicaps (Bliege Bird and Smith 2005). In the study of language, costly-signaling arguments have been evoked as a frame for politeness in language (van Rooy 2003). Veblen, whose *Theory of the Leisure Class* (1899) can be seen as a forerunner of the game-theoretic treatment of costly signaling, points out a number of social phenomena that can be interpreted as a manifestation of *conspicuous consumption* or *conspicuous leisure*, as he refers to the wasteful expenditure of goods and services: for example, charity, fashion (the wearing of high heels, the cylinder hat, the corset), courteous manners, a taste for art and culture, or a preference for antiquated forms in language.

The use of game-theoretic equilibrium analysis in such models is often justified by intuitive dynamic arguments. The formal analysis of evolutionary dynamics in costly signaling games, however, is relatively unexplored.

In economics, researchers have from the beginning pointed out that models of costly signaling

typically have multiple equilibria (Spence 1973, Banks and Sobel 1987, Cho and Kreps 1987, Kreps and Sobel 1994). This has inspired a rich literature on how to assess the plausibility of equilibria in costly-signaling games—how the Nash equilibrium notion should be *refined*—in order to select those equilibria that should be retained as the solution of the game. In this literature, the plausibility of equilibria has been tested mostly by asking whether players' beliefs (probability distributions) over the states of nature (for instance, the high and low productivity type) that support their equilibrium behavior are reasonable—that is, consistent with Bayes' law along the path being played and "plausible" off that path. There is a vivid debate on this in classical game theory (see, for example, Kohlberg and Mertens 1986, Banks and Sobel 1987, Cho and Kreps 1987, Govindan and Wilson 2009). We will, in chapter 4, give some insight into such methods and indicate their results for the games studied here.

In biology, the analysis of game dynamics in costly-signaling games has been delayed by a certain conceptual orientation: the *handicap principle* has often been interpreted in the strict sense, namely that the handicapping trait or behavior (the costly signal) shall perfectly reveal the "good" type (the high fitness, high energy, truly needy type), and its absence, the "bad" type. In formal accounts of the theory, researchers have therefore often focused on identifying conditions (parameter specifications of the model) under which perfectly revealing equilibria in which the "good" type expresses the costly signal and the "bad" type does not and the second player accepts when the signal has been expressed and does not accept when the signal has not been expressed—*honest signaling equilibria* as has been said—exist (Grafen 1990, Maynard Smith 1991). When such an "honest" signaling equilibrium exists, it will be a strict Nash equilibrium. Strict Nash equilibria are evolutionarily stable strategies. They are asymptotically stable under any kind of evolutionary dynamics that have been conceived. The question of the study of explicit evolutionary dynamics therefore seemed answered.

It was only in the second wave of game-theoretic studies of costly signaling in theoretical biology that researchers argued that the conditions under which "honest" signaling equilibria exist are for many applications overly restrictive: in discrete models, they require that for the low type the cost of the signal is at least as high as the benefit that he gets if the second player takes the desired action—the discrete version of a condition which in economics, for games with continuous cost and benefit functions, is known as the *single-crossing property* (see for example, Kreps and Sobel 1994). It has been pointed out that in standard models, under wide ranges of fairly plausible parameter specifications (namely that for the bad type the cost of the signal stays below the benefit that he gets if the second player takes the desired action), there are equilibria in which the costly signal does not perfectly reveal the good type, because the bad type uses it also sometimes, which nevertheless induces the second player to take the desired action (accept, mate) with a least some probability—*hybrid equilibria*, as they have been called (Bergstrom and

Lachmann 1997, Lachmann and Bergstrom 1998). Results of this kind have been interpreted as a critique of the handicap principle, which, on this take, has been identified with the position that a handicap always has to be perfectly revealing. In that line of study, researchers have turned to the formal analysis of evolutionary dynamics in costly-signaling games (Huttegger and Zollman 2010, Wagner 2013, Zollman et al. 2013, Huttegger and Zollman 2016). This literature has concentrated on bringing the proof that hybrid equilibria, when they exist, do have some form of local dynamic stability under standard evolutionary dynamic processes. Zollman et al. (2013), for instance, show that in a discrete version of Spence's respectively Grafen's model, the hybrid equilibrium, when it exists, is surrounded by closed orbits in its supporting 2-dimensional face, which in turn attracts an open set of nearby states in the replicator dynamics. By aid of computer simulations, Zollmann and coauthors compare the basin of attraction of this face, under parameter specifications in which the hybrid equilibrium exists, to the basin of attraction of the "honest" signaling equilibrium, under parameter specifications in which this last one exists, and find that the two are similar in size. Such a comparison has to be interpreted with care, because it is to compare equilibria that exist *in two different games* (under different parameter specifications in the same family of games).

A question that so far has received little attention in the study of evolutionary dynamics in costly-signaling games, is the *co-existence of equilibria in the same game* (under the same parameter specification) and hence the equilibrium refinement and selection problem stemming from this particularity. The co-existence of equilibria in standard costly signaling games, notably in discrete versions of these games, concerns not so much fully and partially revealing equilibria (for most parameter specifications either one or the other exists) but equilibria in which the signal transports some information (fully or respectively partially revealing equilibria) versus equilibria in which either nobody or everybody uses the costly signal, so-called *pooling equilibria*, and therefore, the costly signal (or respectively its absence) transports no information at all.

A second point that has been neglected in the discussion of costly signaling in biology is the role of the prior probability on the states of nature (the frequencies of types, such as high and low fitness types) in shaping the equilibrium structure. Researchers have focused on the case that the prior probability on the good type (its frequency in the population) is low, so low indeed that the second player a priori would not accept. However, once one writes down a costly-signaling interaction as a game, a game will be defined for any prior probability distribution over the types, and that game will have equilibrium solutions—even when the prior on the good type is already high. In fact, as we will see for the class of games studies here, the game will in this case have multiple equilibria with quite diverse signaling patterns, reaching from equilibria in which nobody uses the costly signal, over equilibria in which the high type "imitates" the low type and sometimes *does not* use the costly signal, to equilibria in which everybody has to use the costly signal in

order to make the second player accept with the effect that such a costly signal will transport no information at all—a social or ecological trap. To our mind, in order to capture the explanatory potential of costly-signaling games from an evolutionary point of view, one needs to study the entire equilibrium structure, not only for different values of the cost and benefits parameters but also for different values of the prior probability of the types (the states of nature), and then, for each set of fixed parameters, evolutionary dynamics on the entire state space.

In this paper, we give a complete structural analysis of the equilibria and of evolutionary dynamics in two basic classes of costly signaling games with two states of nature (*high* and *low*), two signals (a costly signal and the absence of that costly signal), and two possible reactions in response to signals (*accept* and *do not accept*):

(I) a game in which the production of the costly signal is of *different costs for different types*—a discrete version of Spence's (1973) model,

(II) a game in which the production of the costly signal is of the *same cost for different types*, but types have *different benefits if the signal has the desired effect*—a discrete version of Milgrom and Robert's (1986) model of advertising and Grafen's (1990) formalization of the handicap principle.

For each of these classes, we study different specification of the cost and benefit parameters (namely those that give rise to different equilibrium structures), and within each of the so-defined games, we distinguish three relevant cases concerning the prior probability. We first make use of the theory of the *index* of equilibria (Shapley 1974, Hofbauer and Sigmund 1988, 1998, Ritzberger 1994, 2002, Demichelis and Ritzberger 2003) to give a rough account of evolutionary dynamics in these games. Then, for each of these classes, we study in more detail the replicator dynamics and to some extent the best-response dynamics.

Why focus on such basic discrete games? First, because they are empirically relevant. In life, many signaling problems boil down to such a simple binary structure. Discrete signals are easier to discern, and often reactions are conditioned just on whether such a discrete costly signal has been expressed or not: Does the candidate have a degree from a certain school or not? Is a certain handicapping trait or behavior displayed or not? Is a certain polite form used or not? Similarly, the actions that the signal aims to precipitate are often binary in nature too: hire or not, buy or not, mate or not. In many applications, anyway, even if the game is defined with a continuous signaling space and continuous reaction functions to signals, in equilibrium, it eventually breaks down to a binary structure, where all that matters is if the level of signaling is above or below a certain threshold. Second, such simple binary games are an important theoretical reference point. Games with two signals, two states of nature, and two possible reactions to signals provide a basic grammar of signaling games—a simple model in which one can see in pure, minimal form

phenomena that arise in more complex models, in which it might be harder to discern which property of the solution is driven by which assumption. In applications involving richer signaling structures, in order to make the model tractable, researchers often focus on a specific class of equilibria—most importantly perfectly revealing equilibria—and leave unexplored the existence of other kinds of equilibria. The simple, binary games that we consider, instead, allow us to analyze the entire equilibrium structure—making it possible to see in a clear-cut way under which conditions which kind of equilibria can exist, and how the structure of equilibria changes as a function of changes in the basic parameters of the model.

2 The model

Costly-signaling theory begins with a problem of cooperation—a problem of cooperation *under uncertainty*. A player (the hiring firm, the potential buyer, the female) in principle wants to conclude an exchange with some other player (the job candidate, the firm offering its stock or a product, the male), but only if that other player is by nature of a certain *type*, namely, of high productivity, high quality, high performance, high fitness, ... Unhappily, the type of that other player (the state of nature) is not directly observable. Therefore, the player under consideration cannot condition her choice whether to accept the proposed exchange or not (hire, buy, mate) on the other player's type. Surely, given the gains from accepting or not as a function of other player's types, whether the player under consideration should accept or not depends on the probability that she attaches to the other player's type. The probability of types is assumed to be given by the environment. It might, of course, be determined by some other ecological game, but it is treated here as a parameter of the model. In our model, for simplicity, we assume that when the other player is of the "good" type (high productivity, high quality, high fitness), then the player under consideration, if she accepts, will have a payoff of 1, and if she does not accept, a payoff of 0; and the other way round if the other player is of the "bad" type: 0 if she accepts and 1 if she does not accept. Figure 1 represents that situation. For given payoffs, there will be a critical probability p of the "good" type below which the player under consideration should not accept and above which she should accept. For the payoffs that we assume, this critical probability will be $1/2$. So far then, this is a simple problem of *choice under uncertainty*, or *a game against nature*.

One realizes what the tragedy emanating from such a game against nature might be: the other player (the job candidate, the firm offering its stock or a product, the male) might effectively be of the good type (and *know* this), but if the frequency of that type in the population is too low, the right choice of the player under consideration (the hiring firm, the potential buyer, the female) might be not to accept. Cooperation might not take place due to an informational problem in the society. In the situation depicted in Figure 1, this happens when $p < 1/2$.

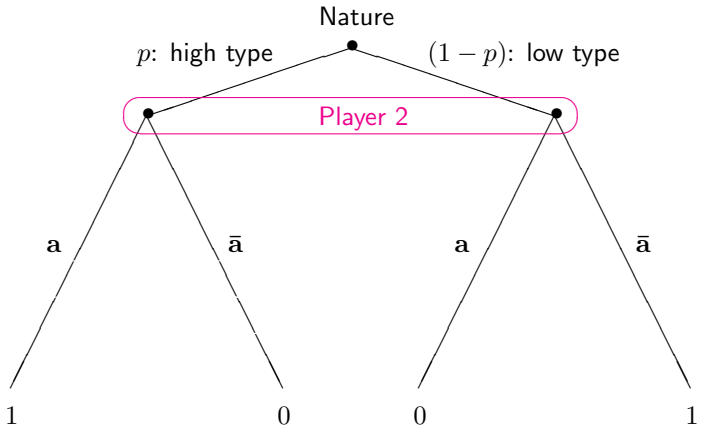


Figure 1: A social identification problem: a game against nature

In economic life the phenomenon can frequently be observed: in some regions (for instance, regions shaken by political or social crisis) investments might not take place because the expectations that the returns are good are, on average, too low. Akerlof (1970) has famously referred to this as the problem of "the market of lemons" ("lemon" for bad used car): Suppose a price corresponding to the average quality is payed. Owners of high quality cars know that their car is worth more than the price they could get by selling it in the market. They might therefore refrain from selling it, with the effect that only low quality cars will be in the market, possibly with a quality so low, that nobody wants to buy, and hence the market stays closed.

What could be a mechanism that lifts players out of such a situation of no exchange, no trade, no cooperation arising from an informational problem?

–*Signals*? Imagine that the player whose type is unknown can emit a signal, that is, take some action or express some trait that can be observed by the player who decides whether to accept or not. This transforms the model into a situation of strategic interaction, a *game* in the nontrivial sense, because the player who has to decide whether to accept or not can now condition this decision on observation of that signal. Figure 2 represents the sequential structure of such a game by a tree.

In the game tree in Figure 2, and in what follows, the player about whose type there is uncertainty and who can send a signal—because he is the one who moves first—is referred to as *player 1*; and the player who observes the signal and then has to decide whether to accept or not as *player 2*. The uncertainty about the type of player 1 (the state of nature) is represented by a move of Nature at the root of the tree. Nature makes her move with probability p for the high type and $1-p$ for the low type. Each of player 1's types (*high* and *low*) can emit the signal s or not (indicated by $\bar{s}$). Player 2, before she makes her move, observes whether player 1 has

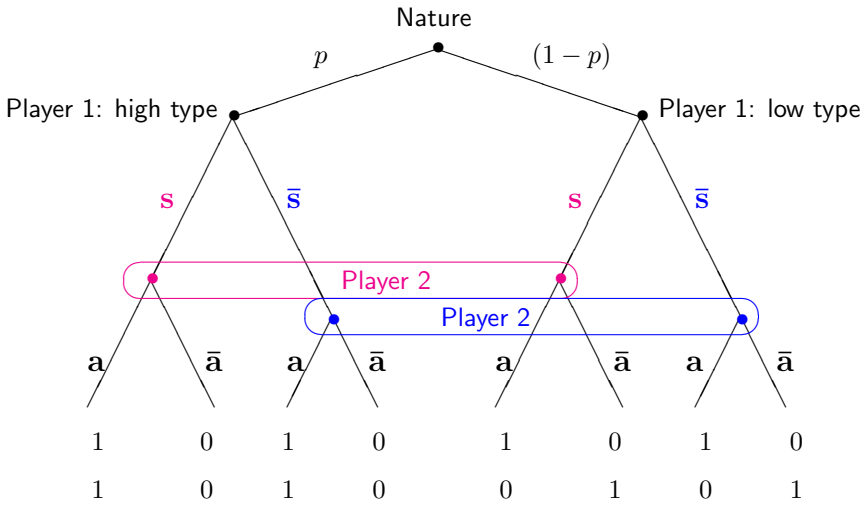
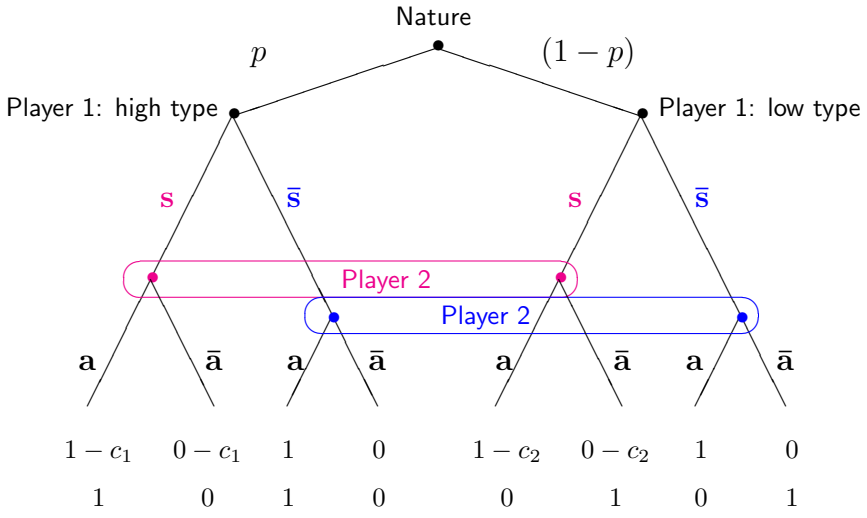


Figure 2: Giving player 1 the possibility to express a signal transforms the situation into a game in the nontrivial sense. Here is is assumed that expressing the signal is of no cost.

emitted the signal s or not, but a priori she still does not know of which type player 1 is. This is indicated in the game tree by gathering the node after *high type* and s and the node after *low type* and s in an oval, which represents an *information set* of player 2—nodes of the tree that player 2 cannot distinguish the moment she is called to make a move. And similarly for the nodes after the absence of the signal. After the signal or respectively its absence has been observed, player 2 makes her move, that is, either accepts (a) or not ($\bar{a}$).

A game tree alone does not define a game. To define a game, at each end node of the tree, payoffs for the players have to be specified. Once an end node is reached, payoffs materialize. Following standard convention, at each end node of the tree, payoffs are indicated with the first number giving the payoff of player 1, and the second that of player 2.

In Figure 2, the payoffs of players have been specified under the preparatory assumption that expressing the signal is of *no cost* for player 1, no matter what his type. Considering this situation first is instructive to understand the role of signaling costs. One quickly verifies that if signals are of no cost, with the payoffs from the identification problem in Figure 1 in the backdrop, the possibility to send a signal *does not* enable the players in the model to escape the unfortunate situation of no exchange if the prior probability on the good type is low. To see why, assume that in fact only the good type uses the signal s and that player 2 at observing the signal accepts, and in the absence of the signal does not accept. If this were so, then the bad type of player 1 would also be better off using the signal, and hence this assignment of behavioral strategies cannot be an equilibrium in the population. The argument is intuitive: if talk is cheap, player 1 will always say "I am of the good type": "Yes, I am highly motivated." "Sure, if you make me an offer, I will take the job." As Spence (1973) remarked: "If the incentives for veracity in reporting anything by



	aa	**a$\bar{a}$**	**$\bar{a}$a**	**$\bar{a}\bar{a}$**
ss	$1-pc_1-(1-p)c_2,\ p$	$1-pc_1-(1-p)c_2,\ p$	$-pc_1-(1-p)c_2,\ 1-p$	$-pc_1-(1-p)c_2,\ 1-p$
s$\bar{s}$	$1-pc_1,\ p$	$p(1-c_1),\ 1$	$-pc_1+(1-p),\ 0$	$-pc_1,\ 1-p$
$\bar{s}$s	$1-(1-p)c_2,\ p$	$(1-p)(1-c_2),\ 0$	$p-(1-p)c_2,\ 1$	$-(1-p)c_2,\ 1-p$
$\bar{s}\bar{s}$	$1,\ p$	$0,\ 1-p$	$1,\ p$	$0,\ 1-p$

Figure 3: Class I: at the top, the game given by a tree—the game in *extensive form*; at the bottom: the matrix game resulting from that extensive-form game.

means of a conventional signaling code are weak, then one must look for other means by which information transfers take place." Spence's fertile idea was to look at the effect of signaling costs.

2.1 Class I: different costs of producing the signal

In class I, a discrete version of Spence's (1973) model, it is the very *production*, or *expression*, of the signal s that is of different costs for the two types of player 1.[1] Following Spence's original idea, it is assumed that for the high type it is less costly to produce the costly signal than for the low type. We assume that $0 \leq c_1 < 1$ and $c_1 < c_2$, where c_1 is the cost of the signal for the high type and c_2 that of the low type, and that these costs are deducted from the background payoffs for the two types that are depicted in Figure 2. This gives the game in Figure 3.

For the game in Figure 3, a strategy for player 1 is a plan of action whether to emit the costly signal or not, that is, take s or $\bar{s}$, as a function of his type; and a strategy for player 2 is a plan of action, a or $\bar{a}$, conditional on which signal she has observed. Each player then has four possible *pure strategies*:

[1] In Spence's model, the space of signals (the level of education) and the space of possible reactions to signals (the offered wage) are continuous. In equilibrium, Spence's model breaks down to a discrete signaling structure where a certain level of education stands for the high type.

Pure strategies for player 1:	Pure strategies for player 2:
ss: If high, then s; if low, then s	aa: If s, then a; if $\bar{s}$, then a
$s\bar{s}$: If high, then s; if low, then $\bar{s}$	$a\bar{a}$: If s, then a; if $\bar{s}$, then $\bar{a}$
$\bar{s}s$: If high, then $\bar{s}$; if low, then s	$\bar{a}a$: If s, then $\bar{a}$; if $\bar{s}$, then a
$\bar{s}\bar{s}$: If high, then $\bar{s}$; if low, then $\bar{s}$	$\bar{a}\bar{a}$: If s, then $\bar{a}$; if $\bar{s}$, then $\bar{a}$

Players' strategies can also be *mixed*, that is, in terms of a probability distribution over their respective set of pure strategies. We write $x(ss)$, $x(s\bar{s})$, etc., for the probability attributed by a mixed strategy $\mathbf{x}$ to the pure strategies ss, $s\bar{s}$, etc. And similarly for player 2, using $\mathbf{y}$.

Given the sequential structure of the game (as represented by the game tree in Figure 3), there is however a different way of interpreting mixed strategies: Every mixed strategy induces at least one so-called *behavior strategy*, that is, a plan of action that gives for every node or information set of the respective player a probability distribution over the actions that he or she has available there. A behavior strategy for the first player is, for instance: "If you happen to be of the high type, send the costly signal s with a probability of 60% (and do not send it with the complementary probability of 40%); if you happen to be of the low type, do not send the costly signal." This particular behavior strategy, obviously, is induced by a mixed strategy of $s\bar{s}$ and $\bar{s}\bar{s}$, with a probability of 60% on the first and 40% on the second. A behavior strategy for the second player is, for instance: "If you see the costly signal s, take a for sure; if you do not see it, take a with a probability of 50%," which is induced by a mixed strategy of aa and $a\bar{a}$ with a probability of 50% on each of them. The two games—the one based on mixed strategies defined on complete contingent pure strategies and the other based or behavior strategies—are, at least as what concerns the existence of Nash equilibria, equivalent (Kuhn 1950, 1953). We denote behavior strategies as follows:

Behavior strategies for player 1:	Behavior strategies for player 2:
(x_h, x_ℓ): x_h prob. with which high type uses s, x_ℓ prob. with which low type uses s	(y, y'): y prob. with which 2 takes a after s, y' prob. with which 2 takes a after $\bar{s}$

A profile of behavior strategies then can be denoted in the form (x_h, x_ℓ, y, y'). This allows us to represent strategy profiles in the $[0,1]^4$ cube—the hypercube (see Figures 7, 8, 9).

Nash equilibria in the matrix game

The standard approach to solve games with uncertainty is by *Bayesian Nash equilibrium* (Harsanyi 1967), an extension of Nash's (1950, 1951) equilibrium concept to games under imperfect information, which operates on the assumption that players evaluate payoffs as expected payoffs given the probabilities of the states of nature. Under this assumption, the *normal form* of the game—

the game matrix—can be derived by considering all 4×4 combinations of pure strategies and evaluating the payoffs of players at the end nodes of the paths induced by the respective strategy combination, weighted by the probabilities with which these end nodes will be reached, given the prior probability on the states of nature. The Bayesian Nash equilibria of the game then are the Nash equilibria of the game given by that payoff matrix and can be determined in the usual way by searching for strategy combinations such that the strategy of one player is a best response to the strategy of the other player (the usual Nash-equilibrium condition).

Sequential Bayesian Nash equilibria in the extensive form—the tree representation—of the game

In the Nash equilibria of the matrix game, the sequential structure of the game—and how players might reason about it—has been lost, or is only implicitly present. When one translates the equilibrium strategies into behavior strategies, one can however check if players' choices of actions respect some sequential logic.

For a game given by a tree, a *sequential Bayesian Nash equilibrium* (Kreps and Wilson 1982) is a profile of behavior strategies, one for each player, together with a vector of beliefs (a probability distribution) over the states of nature, namely one for every node or information set of the game tree, such that:

(1) players' choices of actions at the nodes or information sets where they potentially come to move are a best response to (i) the beliefs over the states of the nature assigned to that node or information set and (ii) the other players' actions as given by their behavior strategies from that node onward, and

(2) the beliefs assigned to nodes or information sets are compatible with *Bayes' law along the path being played*, given the prior probability distribution p over the states of nature and players' equilibrium strategies.[2]

For the game tree in Figure 3, the two players will be in a sequential Bayesian Nash equilibrium if one player's plan of action (expressed as a behavior strategy) is a best response to the other player's

[2]Kreps and Wilson (1982), in their definition of sequential Bayesian Nash equilibrium, require also that beliefs *off the equilibrium path* (a node or information set never reached in the equilibrium under study) be *consistent* in the sense that they can be deduced from Bayes' law after a small perturbation of the behavior strategies. It is easy to see that for signaling games as we consider them here, the condition is always fulfilled: 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 she receives that signal. Suppose that player 1 perturbs his behavior strategies as follows: the high type sends the signal which in the original equilibrium outcome is never used 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' law, the updated belief is $(p^*, 1-p^*)$.

plan of action, and the second player's plan of action is *consistent with Bayes' law along the path being played*, that is, if the second player's decision to take a or $\bar{a}$ when she comes to move in any of her information sets is a best response to the belief about the first player's type attributed to that information set and if, in case that the information set is effectively reached along the path defined by the equilibrium strategies, the belief about the first player's types results from a Bayesian update of the prior probability p (given the probabilities with which the first player uses s or respectively $\bar{s}$ as a function of his type as prescribed by his equilibrium strategy). One recovers the typical "static" nature of Nash equilibrium here: the second player has to behave *as if she updated her prior belief as if she knew the first player's strategy* (that is, the probabilities with which the first player takes action a depending on his type), and the first player has to make his choice based on his expectations how the second player reacts to which signal. But how the players come to know these probabilities is not part of this equilibrium concept.

For games without moves of nature, sequentiality means *subgame perfectness* (Selten 1965, 1975), the requirement that a Nash equilibrium for the entire game has to induce a Nash equilibrium in every subgame (branch of the tree)—a notion which is usually considered as capturing the idea of *backward induction* for games of imperfect information.

2.2 Parameter specifications—a family of games

Which equilibria a game under imperfect information has, no matter if one looks at it in the matrix or the extensive form, depends on the value of the prior probability on the states of nature—here given by p, the prior probability on the high type. For the game in Figure 3, no matter what the values of c_1 and c_2, three cases are relevant: $p < 1/2$, $p > 1/2$, and the knife-edge case $p = 1/2$. *Why is $p = 1/2$ critical?* This comes from the decision problem of player 2: $1/2$ is the probability that player 2 has to attach to the high type, at the moment of making her choice, in order to be indifferent between a and $\bar{a}$. Since use of the signal induces no costs for player 2, the probability $1/2$ will be critical no matter which signal player 2 has observed. For the parametrized version of the game that we consider, the equilibria depend of course also on the specific values of c_1 and c_2. In the following, we distinguish first three paradigmatic cases concerning the cost parameters, and then, within each of these, we distinguish the three relevant cases concerning p.

Class I, case 1: $0 \leq c_1 < c_2 < 1$: Nash equilibria in the matrix game

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

  E1: an equilibrium in which player 1 mixes between ss and $s\bar{s}$ with a probability of $\frac{p}{1-p}$ on the first (and $1 - \frac{p}{1-p}$ on the second) and player 2 mixes between $a\bar{a}$ and $\bar{a}\bar{a}$ with a probability of c_2 on the first (and $1 - c_2$ on the second), and

P1: an equilibrium component in which player 1 takes $\bar{s}\bar{s}$, that is, both types of player 1 take $\bar{s}$, and player 2 mixes between $a\bar{a}$ and $\bar{a}\bar{a}$ with some probability in $[0, c_1]$ on the first (and the complementary probability on the second). Equilibria of this form, in which all types use the same signal, are sometimes referred to as *pooling equilibria* (hence the name P1).

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

E2: an equilibrium in which player 1 mixes between $s\bar{s}$ and $\bar{s}\bar{s}$ with a probability of $1 - \frac{1-p}{p}$ 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 ss and player 2 mixes between aa and $a\bar{a}$ with some probability in $[0, 1 - c_2]$ on the first, and

P3: an equilibrium component in which player 1 takes $\bar{s}\bar{s}$ and player 2 any mix between aa and $\bar{a}a$.

- In the knife-edge case $p = \frac{1}{2}$, there is:

E1'-P2: An equilibrium component in which player 1 takes ss and player 2 a mixed strategy in the 3-dimensional polyhedron determined by $y(a\bar{a}) \geq y(\bar{a}a) + c_2$. In other words, E1'-P2 is spanned by the four vertices $ss \times \mathbf{y}$ with $\mathbf{y} = (0, 1, 0, 0), (1-c_2, c_2, 0, 0), (0, c_2, 0, 1-c_2)$, and $(0, 1 + c_2, 1 - c_2, 0)/2$.

P1-E2'-P3: An equilibrium component in which player 1 takes $\bar{s}\bar{s}$ and player 2 a mixed strategy in the triangular frustum, determined by $y(a\bar{a}) \leq y(\bar{a}a) + c_1$. In other words, P1-E2'-P3 is the convex hull of the six vertices $\bar{s}\bar{s} \times \mathbf{y}$ with $\mathbf{y} = (1, 0, 0, 0), (0, 0, 1, 0), (0, 0, 0, 1)$ at the base and $\mathbf{y} = (1 - c_1, c_1, 0, 0), (0, c_1, 0, 1 - c_1), (0, 1 + c_1, 1 - c_1, 0)/2$ at the top.

Each of these Nash equilibria in the game matrix has a translation into behavior strategies (relative to the game tree in Figure 3) in which it appears as a sequential Bayesian Nash equilibrium—which we show below.

Class I, case 1: $0 \leq c_1 < c_2 < 1$: Sequential Bayesian Nash equilibria

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

E1: In terms of behavior strategies, E1 translates to $(1, \frac{p}{1-p}, c_2, 0)$: the high type uses the costly signal s for sure, the low type with probability $x_\ell = \frac{p}{1-p}$, and player 2, in case that she has observed s, takes a with probability $y = c_2$, and in case that she has not observed it, does not take a (takes $\bar{a}$ for sure). It is straightforward to check that this profile of behavior strategies is compatible with the requirement that players update

their beliefs along the path being played by Bayes' law: Given that the high type always uses s, the probability $x_\ell = \frac{p}{1-p}$ with which the low type uses s is precisely such that at observation of s, player 2's Bayesian updated belief p_s^* will be 1/2:

$$p_s^* = p(h \mid s) = \frac{p}{p + (1-p) \cdot x_\ell} = \frac{1}{2} \qquad \Longleftrightarrow \qquad x_\ell = \frac{p}{1-p}.$$

At this belief, player 2 is indifferent between taking a and $\bar{a}$ and hence ready to mix between the two, which is what she effectively does at observation of s in the equilibrium under study. At observation of $\bar{s}$, player 2's updated belief of the high type, if she updates her prior belief according to Bayes' law given player 1's behavior strategy, will be 0, and to this belief, there is a unique best response, namely not to accept (take $\bar{a}$), which is what she does in the equilibrium under study. In turn, given that in the absence of the costly signal, player 2 takes $\bar{a}$ for sure, player 2's choice after s to take a with probability $y = c_2$ is precisely such as to make the *low* type of player 1 indifferent between s and $\bar{s}$, which is needed to make this type be willing to use a mix between the two, which is what he does in the equilibrium under study; while the high type is strictly better off using s, which is what this type does in the equilibrium under study. In this equilibrium, the absence of the costly signal ($\bar{s}$) fully reveals the low type, whereas the presence of the costly signal s does not fully reveal the high type but still pushes player 2's belief that player 1 is of high type (which a priori is below 1/2) *up to* precisely 1/2. We refer to this kind of equilibrium, in which one signal fully reveals one type, as a *partially revealing equilibrium*. Figure 4 illustrates how this equilibrium and the beliefs that support it unfold in the game tree. In Figure 7, one can see its position in the space of behavior strategies represented by the hypercube: it is an isolated equilibrium point that sits in the 2-dimensional face given by $(1, *, *, 0)$.

P1: In terms of behavior strategies, the component P1 translates to $(0, 0, y, 0)$, $y \in [0, c_1]$: player 1 *never uses the costly signal*, no matter what his type; player 2, in case that she observes the costly signal s, takes a with a probability not higher than c_1, and when she does not observe it (observes $\bar{s}$), will not accept (take $\bar{a}$). In the game tree, any point in this component maps to the same *outcome*, that is, probability distribution over end nodes: the node after *high type* – $\bar{s}$ – $\bar{a}$ will be reached with probability p, and the node after *low type* – $\bar{s}$ – $\bar{a}$ will be reached with probability $1 - p$. It is straightforward to check that any point in P1 can be sustained as a sequential Bayesian Nash equilibrium: After $\bar{s}$, given that both of player 1's types use $\bar{s}$, the updated belief is the same as the prior belief, $p_{\bar{s}}^* = p < 1/2$, and therefore player 2, if she responds optimally to her beliefs, has to choose $\bar{a}$. In the event that player 2 observes s, which actually never happens in the equilibrium outcome under study—game theorists refer

to this as a situation *off the equilibrium path*—Bayes' law is not defined, and so imposes no restrictions. To make player 2's choice of taking a with a probability $y \in [0, c_1]$ in the counterfactual event that s is observed compatible with sequential Bayesian Nash equilibrium, it suffices therefore to find *some* belief (probability distribution over the types) for which taking a with a probability $y \in [0, c_1]$ is a best response. There are actually many such beliefs: for any belief on the high type strictly smaller than $1/2$, the best response will be to take a with 0 probability; if the belief after s is equal to $1/2$, then player 2 will be indifferent between a and $\bar{a}$, and so taking a with some $y \in [0, c_1]$ will be good. What is not compatible with this equilibrium outcome are beliefs after s that attribute to the high type a probability strictly higher than $1/2$, because then player 2 would have to take a for sure, and that would upset the equilibrium (because then both types would want to use s instead of $\bar{s}$). Figure 5 illustrates this equilibrium outcome and the beliefs that support it in the game tree. In Figure 7, one can see its position in the space of behavior strategies: it is an equilibrium component that reaches from $(0, 0, 0, 0)$ to $(0, 0, c_1, 0)$, the point marked -P1 in the figure. For the case that $c_1 = 0$, the component is reduced to the point $(0, 0, 0, 0)$.

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

  E2: translates to $(1 - \frac{1-p}{p}, 0, 1, 1 - c_1)$, a *partially revealing equilibrium* that is the mirror image of E1: the high type uses the costly signal s with probability $x_h = 1 - \frac{1-p}{p}$ and *does not* use it with probability $\frac{1-p}{p}$, while the low type never uses the costly signal (takes $\bar{s}$ for sure), which is such that player 2 in the absence of the costly signal will have an updated belief $p^*_{\bar{s}}$ that will make her indifferent between a and $\bar{a}$:

$$p^*_{\bar{s}} = \frac{p \cdot (1 - x_h)}{p \cdot (1 - x_h) + (1 - p)} = \frac{1}{2} \quad \Leftrightarrow \quad 1 - x_h = \frac{1 - p}{p}.$$

  Player 2, if she observes the costly signal s, will choose a for sure (which will be a best response to her updated belief $p^*_s = 1$), and if she does not see it, will choose a with probability $y' = 1 - c_1$, which is the probability that will make player 1's *high type* indifferent between using and not using the costly signal, while ensuring that not using the costly signal is a best response for player 1's *low type*. In this equilibrium, the costly signal s fully reveals the *high type*. The absence of the costly signal ($\bar{s}$) instead does not fully reveal the low type, but—and again this is the mirror effect of what s does in E1—will bring player 2's belief *down to* $p^*_{\bar{s}} = 1/2$.

  P2: translates to $(1, 1, 1, y')$, $y' \in [0, 1 - c_2]$: both types of player 1 use the costly signal s; player 2, when she observes s, will have the same belief as her prior belief, $p^*_s = p > 1/2$, and will therefore accept, and in the absence of the signal, which will be "off the

equilibrium path," either believes that player 1 is of the high type with a probability of less than 1/2, in which case she will choose $\bar{a}$, or believes that 1 is of the high type with a probability of 1/2 and will choose 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 s.

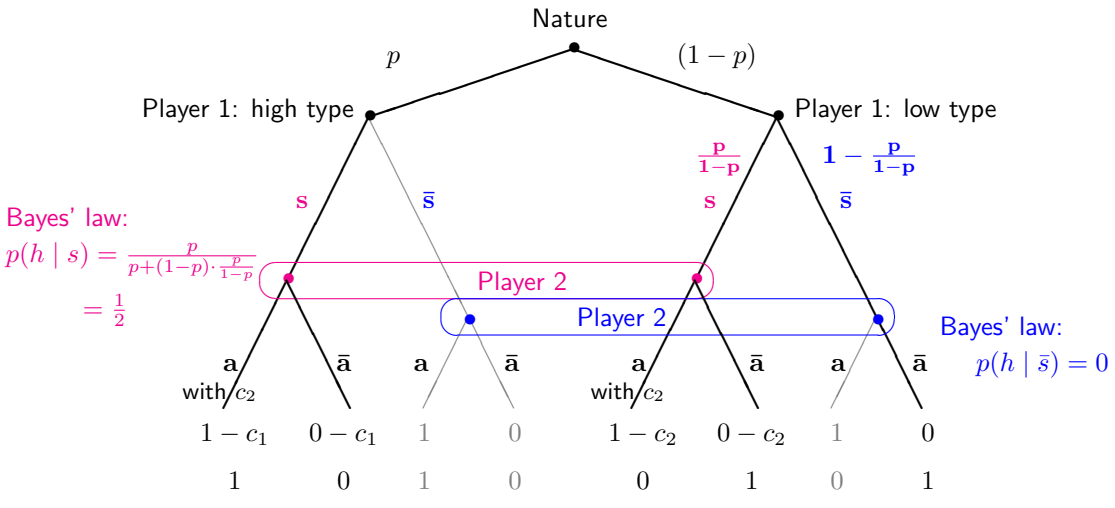
P3: translates to $(0, 0, y, 1)$, $y \in [0, 1]$: player 1 *never uses the costly signal*, no matter what his type, and player 2, in the absence of the costly signal, will have the same belief as her prior belief, $p^*_{\bar{s}} = p > 1/2$, and hence will choose a, and in case that the costly signal has been sent, which will be "off the equilibrium path," can have any belief and best respond to it.

Figure 8 shows E2, P2, and P3 in the hypercube.

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

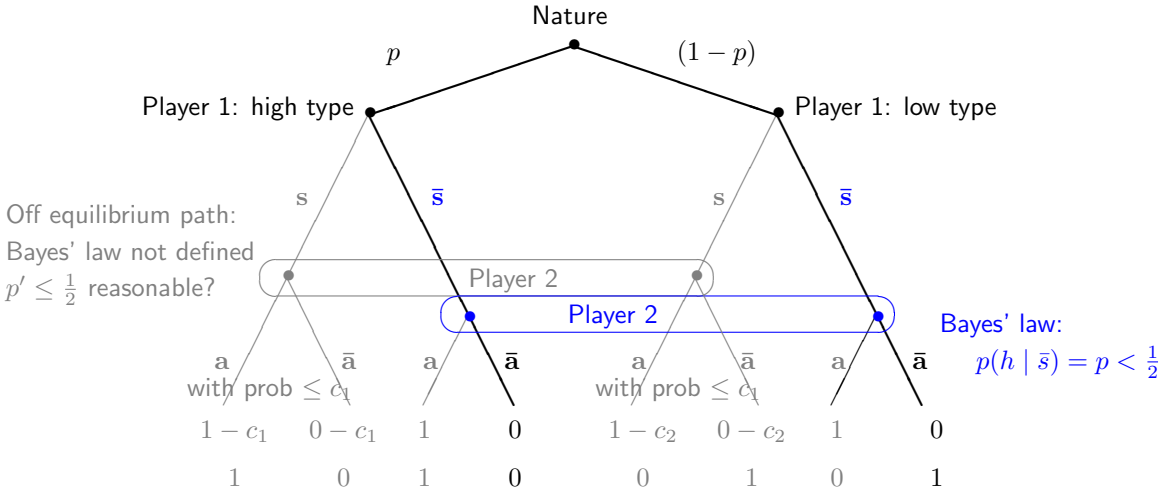
E1'-P2: translates to a 2-dimensional set of behavior strategies, an isosceles right triangle, spanned by $(1, 1, 1, 0)$, -P2= $(1, 1, 1, 1-c_2)$, and E1' = $(1, 1, c_2, 0)$ (see Figure 9). That is, player 1 always uses s no matter what his type, and player 2, when she observes s, will have the same belief as the prior 1/2 (will hence be indifferent between a and $\bar{a}$) and will take a with some probability $y \in [c_2, 1]$, and in response to the off-the-equilibrium-path signal $\bar{s}$ will take a with some probability $y' \in [0, y-c_2]$, which guarantees that both types of player 1 have no incentive not to use s. For example, when $y = c_2$, then $y' = 0$ (similarly as in E1); when $y = 1$, then $y' \in [0, 1-c_2]$ (as in P2). In the game tree this gives a continuum of outcomes in which both types use s with one outcome differing from another only in the reaction of player 2 to s.

P1-E2'-P3: translates to a 2-dimensional set of behavior strategies spanned by $(1, 0, 0, 0)$, -P1= $(0, 0, c_1, 0)$, and E2' = $(0, 0, 1, 1-c_1)$, $(0, 0, 1, 1)$, and $(0, 0, 0, 1)$ (see Figure 9). That is, player 1 always uses $\bar{s}$ (never uses the costly signal), no matter what his type, and player 2, when she does not observe the costly signal, will have the same belief as her prior belief 1/2 and will take a with some probability $y' \in [0, 1]$, and in response to the off-the-equilibrium-path signal s will take a with some probability $y \in [0, y' + c_1]$ if $y' + c_1 \leq 1$ and with some probability $y \in [0, 1]$ if $y' + c_1 > 1$. For example, $y' = 0$ is supported by any $y \in [0, c_1]$ (as in P1); $y' = 1 - c_1$ by any $y \in [0, 1]$ (similarly as in E2); and $y' = 1$ by any $y \in [0, 1]$ (as in P3). In the game tree this gives a continuum of outcomes in which both types use $\bar{s}$ with one outcome differing from another only in the reaction of player 2 to $\bar{s}$.



	aa	$\mathbf{c_2 \to a\bar{a}}$	$\bar{a}a$	$\mathbf{1-c_2 \to \bar{a}\bar{a}}$
$\frac{p}{1-p} \to$ **ss**	$1-pc_1-(1-p)c_2,\ p$	$\mathbf{1-pc_1-(1-p)c_2},\ p$	$-pc_1-(1-p)c_2,\ \mathbf{1-p}$	$-pc_1-(1-p)c_2,\ \mathbf{1-p}$
$1-\frac{p}{1-p} \to$ **s$\bar{s}$**	$1-pc_1,\ p$	$p(1-c_1),\ \mathbf{1}$	$-pc_1+(1-p),\ 0$	$-pc_1,\ 1-p$
$\bar{s}s$	$1-(1-p)c_2,\ p$	$(1-p)(1-c_2),\ 0$	$p-(1-p)c_2,\ \mathbf{1}$	$-(1-p)c_2,\ 1-p$
$\bar{s}\bar{s}$	$\mathbf{1},\ p$	$0,\ \mathbf{1-p}$	$\mathbf{1},\ p$	$\mathbf{0},\ \mathbf{1-p}$

Figure 4: Class I, $0 \le c_1 < c_2 < 1$, $p < \frac{1}{2}$: *partially revealing equilibrium* E1: player 1 mixes between ss and $s\bar{s}$ with $\frac{p}{1-p}$ on the first, player 2 between $a\bar{a}$ and $\bar{a}\bar{a}$ with c_2 on the first.



	aa	with $y \in [0,c_1] \to \mathbf{a\bar{a}}$	$\bar{a}a$	with $1-y \to \mathbf{\bar{a}\bar{a}}$
ss	$1-pc_1-(1-p)c_2,\ p$	$\mathbf{1-pc_1-(1-p)c_2},\ p$	$-pc_1-(1-p)c_2,\ \mathbf{1-p}$	$-pc_1-(1-p)c_2,\ \mathbf{1-p}$
$s\bar{s}$	$1-pc_1,\ p$	$p(1-c_1),\ \mathbf{1}$	$-pc_1+(1-p),\ 0$	$-pc_1,\ 1-p$
$\bar{s}s$	$1-(1-p)c_2,\ p$	$(1-p)(1-c_2),\ 0$	$p-(1-p)c_2,\ \mathbf{1}$	$-(1-p)c_2,\ 1-p$
$1 \to \bar{s}\bar{s}$	$\mathbf{1},\ p$	$0,\ \mathbf{1-p}$	$\mathbf{1},\ p$	$\mathbf{0},\ \mathbf{1-p}$

Figure 5: Class I, $0 \le c_1 < c_2 < 1$, $p < \frac{1}{2}$: *no-signaling* equilibrium component P1: player 1 takes $\bar{s}\bar{s}$, and 2 mixes between $a\bar{a}$ and $\bar{a}\bar{a}$ with $y \in [0, c_1]$ on first.

Class I, case 2: $0 \leq c_1 < c_2 = 1$: Nash equilibria in the matrix game

As case 1 only with the following substitutions:

- $p < 1/2$: E1 is replaced by

  E*-E1: an equilibrium component in which 1 mixes between ss and $s\bar{s}$ with some probability in $[0, p/(1-p)]$ on the first and 2 takes $a\bar{a}$.

- $p \geq 1/2$: P2 respectively E1'-P2 is replaced by

  E*-E1'-P2: an equilibrium component in which 1 takes any mix between ss and $s\bar{s}$ and 2 takes $a\bar{a}$.

Class I, case 2: $0 \leq c_1 < c_2 = 1$: Sequential Bayesian Nash equilibria

- $p < \frac{1}{2}$: E*-E1 translates to $(1, x_\ell, 1, 0)$, $x_\ell \in [0, \frac{p}{1-p}]$, an equilibrium component reaching from a fully revealing equilibrium E* ($x_\ell = 0$) to a partially revealing equilibrium like E1. In any point belonging to this component, player 1's high type uses the costly signal and player 1's low type uses it with some probability x_ℓ, sufficiently low (possibly 0), such that if player 2 observes the costly signal, her updated belief p_s^* will guarantee that choosing a is a best response, which will be the case if:

  $$p_s^* = \frac{p}{p + (1-p) \cdot x_\ell} \geq \frac{1}{2} \qquad \Longleftrightarrow \qquad 0 \leq x_\ell \leq \frac{p}{1-p}.$$

  In any point belonging to this component, the absence of the costly signal ($\bar{s}$) will fully reveal the low type, and hence player 2's best response is unique: $\bar{a}$. Given player 2's behavior strategy, player 1's high type is strictly better off using s and the low type is indifferent between a and $\bar{a}$. In the game tree, this component gives a continuum of equilibrium outcomes.

- $p > \frac{1}{2}$: E*-E1'-P2 translates to $(1, x_\ell, 1, 0)$, $x_\ell \in [0, 1]$, a component reaching from a *fully revealing equilibrium* E* ($x_\ell = 0$), over partially revealing equilibria similar to E1, to an equilibrium in the style of P2 in which both types use s ($x_\ell = 1$). In any point belonging to this component, after s, the updated belief is strictly above 1/2: taking a therefore is the best response. For any $x_\ell < 1$, $\bar{s}$ fully reveals the low type, and therefore $\bar{a}$ is the unique best response to $\bar{s}$. When $x_\ell = 1$ (P2), the updated belief after s will be the same as the prior, and because this is above 1/2, taking a will be the unique best response. This point is supported by beliefs that put a probability of at least 1/2 on player 1's low type after $\bar{s}$.

Class I, case 3: $0 \leq c_1 < 1 < c_2$: Nash equilibria in the matrix game

As case 1 only that E1, P2, and E1'-P2 are replaced by

E*: a *perfectly revealing equilibrium* in which player 1 takes $s\bar{s}$, and player 2 $a\bar{a}$.

Class I, case 3: $0 \leq c_1 < 1 < c_2$: Sequential Bayesian Nash equilibria

E*: translates to $(1, 0, 1, 0)$, a *perfectly revealing equilibrium* in which the high type uses s and the low type $\bar{s}$, and player 2 in reaction to s takes a and in reaction to $\bar{s}$ takes $\bar{a}$. The Bayesian update is trivial here: observation of s sets the belief to 1, $\bar{s}$ to 0. Each signal fully reveals one of the two types.

2.3 Properties of the equilibrium structure

A couple of observations are in place:

- The "honest" *perfectly revealing* or, as economists say, *perfectly separating* equilibrium E*, in which the high type uses the costly signal and the low type does not, and player 2 accepts if the costly signal has been expressed and does not accept if the costly signal has not been expressed, does not always exist. Whether it exists or not depends on the cost of the signal for the *low type*: for E* to exist, the cost of the signal for the *low type* has to be at least as high as the benefit that he gets if player 2 accepts, that is, $c_2 \geq 1$. This reflects a condition for continuous games, which in the economics literature is known as the *single-crossing property* (see, for example, Kreps and Sobel 1994). In the literature in theoretical biology, this observation has sometimes been expressed by saying that it is the "cost of cheating" that sustains honest communication (see, for instance, Számadó 2011). Note in particular that even if the signal is of no cost at all for the high type, $c_1 = 0$, but of some cost $c_2 < 1$ for the low type, a perfectly revealing equilibrium will still not exist. In the extreme case that $c_1 = 0$ and $c_2 \geq 1$, and E* obtains, then—though it is the cost of the signal s for the low type which makes that s, respectively its absence, perfectly reveals the type—actually nobody pays anything for expressing the signal: the high type expresses s, but it does not cost anything to him, and the low type does not express it.

- When c_2 is such that E* exists ($c_2 \geq 1$), then it exists for any prior p (whereas the existence of other equilibria, E1, P1, E2, P2, and P3, depends on the prior).

- Even if E* exists ($c_2 \geq 1$), for any prior, there are also other equilibria, notably equilibria in which nobody expresses the costly signal and the second player acts on her prior belief—*no-signaling equilibria* as one might say.

- Except for the case $c_2 = 1$, the perfectly revealing equilibrium E* and the partially revealing equilibrium E1 do not co-exist in the same game (that is, for a fixed set of parameters of the model: c_1 and c_2, and p) but in different games. Only when $c_2 = 1$, will E* and equilibria of the form E1 co-exist in the same game. But then they will belong to the same equilibrium component. Structurally, E* and E1 represent the same equilibrium component

in different games belonging to the same family of perturbed games (see the analysis of the *index* of equilibria in section 3). Their properties—notably their stability under evolutionary dynamics—have to be evaluated in comparison to other equilibria that exist in the same game.

On the basis of Bayesian Nash equilibrium (no matter if one looks at it as the Nash equilibria in the game matrix or as the sequential Bayesian Nash equilibria in the game tree in Figure 3), all equilibria are equally good predictions of the model. This raises the question whether one shouldn't demand more from a good prediction of the model than being a Nash equilibrium? In the following section, we approach this question by evolutionary dynamics.

3 Evolutionary dynamics

In an evolutionary context, Nash equilibria are interpreted as equilibria in a population of players. Games with two players are understood as models of interaction between two different populations; for example, male and female or predator and prey. Each player position then represents a population of individuals. If a player can be of two different *types*, these types represent subpopulations of the respective population with the frequencies given by the prior probability distribution of types. A state of the two-population system corresponds to a distribution of strategies for each of the player positions representing a population.

What matters for an equilibrium to be a good prediction of the model from an evolutionary point of view is whether the corresponding state of the system is resistant against evolutionary shocks, that is, drift when variation on the level of strategies is already present, and newly appearing variation in the form of mutant strategies.

Theorists have approached the question of evolutionary stability on three different levels : (1) comparative static criteria, as Maynard Smith and Price's (1973) notion of *evolutionarily stable strategy* (ESS), which rely on payoff comparisons between mutant and resident strategies; (2) the study of specific evolutionary dynamic processes defined on the respective game—a research program that has aimed to establish relations between static ESS criteria and stability properties of the associated fixed point under specific dynamic process (Taylor and Jonker 1978, Hofbauer et al. 1979, Hofbauer and Sigmund 1988, 1998), and finally (3) qualitative dynamic stability properties of equilibria under a wider range of dynamic processes based on topological properties of the respective equilibrium component, an approach related to *index theory*.

3.1 The index of equilibria: a rough guide to evolutionary stability

Already Shapley (1974), in his description of the Lemke-Howson algorithm, associated an index ($+1$ or -1) to each *regular* equilibrium[3] with the following properties:

(1) Every strict equilibrium has index $+1$.

(2) Removing or adding unused strategies does not change the index of a regular equilibrium.

(3) The sum of the indices of all equilibria, if they are all regular, is 1. This is often referred to as the *index theorem*, which implies the *odd number theorem*: In generic games, the number of equilibria is odd.

In Hofbauer and Sigmund (1988, 1998) an alternative approach to the index is given, based on the replicator dynamics and Brouwer's degree theory. Here the index of a regular equilibrium is the sign of the determinant of the negative Jacobian matrix.

Ritzberger (1994, 2002) has extended this approach and defines the index of components of Nash equilibria. Recall that in a finite game (finitely many players, each mixing among finitely many pure strategies), the set of Nash equilibria is semialgebraic, and hence consists of finitely many connected components. An index (which can now be an arbitrary integer) can be associated to any of these components, such that the sum over all components is again $+1$. This index is robust against payoff perturbations, in the following sense: Let C be a component and U an open neighborhood of C, such that there is no equilibrium on the boundary of U. A perturbation of the payoffs will in general change C. Now, let C^ε be the set of all equilibria of the perturbed game that lie in U (we assume that the perturbation is small enough, so that again no perturbed equilibrium lies on the boundary of U). This C^ε need not be connected, but it is the finite union of connected components $C_1^\varepsilon, \ldots, C_k^\varepsilon$. Brouwer's degree theory then implies that the sum of the indices of $C_1^\varepsilon, \ldots, C_k^\varepsilon$ equals the index of C. It might happen that C^ε is empty—but only if C has index 0. Using these simple properties, one can easily compute the index of any Nash equilibrium component.

For practical matters, there are three efficient ways to determine the index of an equilibrium component or a degenerate, that is, nonregular, equilibrium:

(a) Perturb the game so that all perturbed equilibria are regular: the index of an equilibrium component of the original game is then the sum of the indices of the corresponding nearby equilibria in the perturbed game—the *robustness property of the index*.

(b) Use the index theorem (if the indices of all other components are known). And finally:

[3] In a 2-person game, an equilibrium is regular if and only if it is isolated and quasistrict (unused strategies do strictly worse).

(c) If an equilibrium component C is asymptotically stable for some evolutionary dynamics, then its index equals its Euler characteristic.

The last property, which is of particular interest here because it establishes the connection to evolutionary dynamics, is given by a beautiful theorem by Demichelis and Ritzberger (2003). An important special case is: If an equilibrium component is convex or contractable (for example, a singleton) and asymptotically stable under some reasonable dynamics, then its index is +1.

In class I, case 1, $0 < c_1 < c_2 < 1$:

- When $0 < p < \frac{1}{2}$, the partially revealing equilibrium E1 is an isolated and quasistrict—hence regular—equilibrium in which both players mix between two strategies. Omitting the strategies that are unused at this equilibrium leads to a cyclic 2×2 game, similar to a matching pennies game. E1 is the only equilibrium in this restricted game. By the index theorem, then, its index is +1. Therefore, in the full (4×4) game, in turn by the index theorem, the only other component P1 must have index 0.

- At $p = \frac{1}{2}$, there are still two components, E1'-P2 and P1-E2'-P3. By the robustness property of the index, E1'-P2 has index +1: in any perturbed game that comes to lie in the case $p < 1/2$, E1 corresponds to the component E1'-P2, which implies that the two have the same index. By the index theorem, the component P1-E2'-P3 then has index 0. (Note also that by robustness, P1-E2'-P3 has index 0: in any perturbed game that comes to lie in the case $p < 1/2$, P1 corresponds to the component P1-E2'-P3 and they therefore have to have the same index.)

- When $\frac{1}{2} < p < 1$, there are three components: By robustness, P2 (which corresponds to the component E1'-P2) has index +1. The partially revealing equilibrium E2 is again isolated and quasistrict, hence regular, and both players mix between two strategies. However, if we discard the unused strategies, the 2×2 restricted game is now a coordination game, with two strict equilibria, and E2. Since strict equilibria have index +1, E2 has index −1. Hence in the full game, by the index theorem, the third component, P3, has index +1.

In other words, as p increases through the critical value $\frac{1}{2}$, the equilibrium component P1 splits into the two components E2 and P3. (Note that, as required by the robustness property of the index, the index of the component P1-E2'-P3 (0) is the same as that of $P1$, on the one hand, and as the sum of the indices of E2 and P3, on the other hand.) Note also that at $p = \frac{1}{2}$, the component P1-E2'-P3 is substantially bigger than the limit of the three, but still disjoint from the E1'-P2 component. This should just remind us of the fact that the set of Nash equilibria is upper semicontinuous against payoff perturbations, but in general not continuous. These results are indicated also in Table 1.

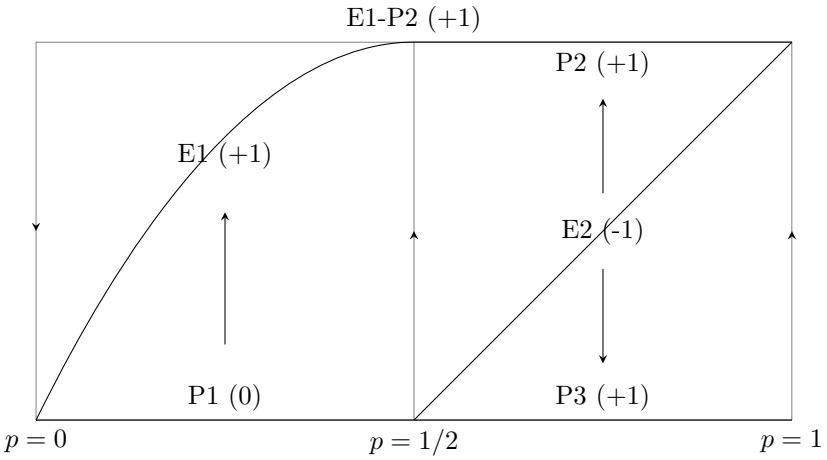


Figure 6: The index of equilibria: class I, case 1: $0 \leq c_1 < c_2 < 1$ for $0 \leq p \leq 1$. For $p = 0$ there is a unique component E1-P1. For $p = 1$ there are two components: P3 (with index +1 and E2-P2 (with index 0).

Table 1. Equilibrium structure, class I, case 1: $0 \leq c_1 < c_2 < 1$

Prior	Equilibrium component	Strategies		Payoffs	
$p < \frac{1}{2}$	(E1) : *partially revealing*	$h \longrightarrow s$	$s \longrightarrow a$ with c_2	h: $c_2 - c_1$	2: $1 - p$
	Index: +1, forward induction	$\ell \longrightarrow s$ with $\frac{p}{1-p}$	$\bar{s} \longrightarrow \bar{a}$	ℓ: 0	
	(P1): *both use* $\bar{s}$	$h \longrightarrow \bar{s}$	$s \longrightarrow a$ with $y \leq c_1$	h: 0	2: $1 - p$
	Index: 0, not fwd. induction	$\ell \longrightarrow \bar{s}$	$\bar{s} \longrightarrow \bar{a}$	ℓ: 0	
$p > \frac{1}{2}$	(E2) : *partially revealing*	$h \longrightarrow \bar{s}$ with $\frac{1-p}{p}$	$s \longrightarrow a$	h: $1 - c_1$	2: p
	Index: -1, forward induction	$\ell \longrightarrow \bar{s}$	$\bar{s} \longrightarrow a$ with $1 - c_1$	ℓ: $1 - c_1$	
	(P2): *both use* s	$h \longrightarrow s$	$s \longrightarrow a$	h: $1 - c_1$	2: p
	Index: +1, forward induction	$\ell \longrightarrow s$	$\bar{s} \longrightarrow a$ with $y' \leq 1 - c_2$	ℓ: $1 - c_2$	
	(P3): *both use* $\bar{s}$	$h \longrightarrow \bar{s}$	$s \longrightarrow a$ with any y	h: 1	2: p
	Index: +1, forward induction	$\ell \longrightarrow \bar{s}$	$\bar{s} \longrightarrow a$	ℓ: 1	
$p = \frac{1}{2}$	(E1'-P2): *both use* s	$h \longrightarrow s$	$s \longrightarrow a$ with $y \in [c_2, 1]$	h: $[c_2 - c_1, 1 - c_1]$	2: $\frac{1}{2}$
	Index: +1, all forward induction	$\ell \longrightarrow s$	$\bar{s} \longrightarrow a$ with $y' \in [0, y - c_2]$	ℓ: $[0, 1 - c_2]$	
	(P1-E2'-P3): *both use* $\bar{s}$	$h \longrightarrow \bar{s}$	$s \longrightarrow a$ with $y \in [0, \min\{y' + c_1, 1\}]$	h: $[0, 1]$	2: $\frac{1}{2}$
	Index: 0, not all fwd. induction	$l \longrightarrow \bar{s}$	$\bar{\mathbf{s}} \longrightarrow a$ with $y' \in [0, 1]$	ℓ: $[0, 1]$	

This implies: P1 and E2 *cannot* be asymptotically stable for any reasonable dynamics, while E1, P2, and P3 are candidates for asymptotic stability, at least for *some* reasonable dynamics (but certainly not for all dynamics). For example, for the replicator dynamics, as we will see in the following section, E1 is not asymptotically stable, only stable (since in the supporting 2-dimensional face it is surrounded by periodic solutions). However, for the best response dynamics, E1 is indeed asymptotically stable.

These results extend to the two other cases ($c_2 = 1$ and $c_2 > 1$), due to the robustness property of the index. Table 2 and Table 3 indicate these results. Note in particular that the fully revealing equilibrium E* always sits in a component with index +1.

Table 2. Equilibrium structure, class I, case 2: $0 \leq c_1 < c_2 = 1$

Prior	Equilibrium component	Strategies		Payoffs	
$p < \frac{1}{2}$	(E*-E1):	$h \longrightarrow s$	$s \longrightarrow a$	h: $1 - c_1$	2: [1-p,1]
	Index: +1, forward induction	$\ell \longrightarrow s$ with $\leq \frac{p}{1-p}$	$\bar{s} \longrightarrow \bar{a}$	ℓ: 0	
	(P1):	$h \longrightarrow \bar{s}$	$s \longrightarrow a$ with $y \leq c_1$	h: 0	2: $1 - p$
	Index: 0, not fwd. induction	$\ell \longrightarrow \bar{s}$	$\bar{s} \longrightarrow \bar{a}$	ℓ: 0	
$p > \frac{1}{2}$	(E2):	$h \longrightarrow \bar{s}$ with $\frac{1-p}{p}$	$s \longrightarrow a$	h: $1 - c_1$	2: p
	Index: -1, forward induction	$\ell \longrightarrow \bar{s}$	$\bar{\mathbf{s}} \longrightarrow \mathbf{a}$ with $1 - c_1$	ℓ: $1 - c_1$	
	(E*-E1'-P2):	$h \longrightarrow s$	$s \longrightarrow a$	h: $1 - c_1$	2: $[p, 1]$
	Index: +1, forward induction	$\ell \longrightarrow s$ with any x_ℓ	$\bar{s} \longrightarrow \bar{a}$	ℓ: 0	
	(P3): *both use* $\bar{s}$	$h \longrightarrow \bar{s}$	$s \longrightarrow a$ with any prob	h: 1	2: p
	Index: +1, forward induction	$\ell \longrightarrow \bar{s}$	$\bar{s} \longrightarrow a$	ℓ: 1	
$p = \frac{1}{2}$	(E*-E1'-P2):	$h \longrightarrow s$	$s \longrightarrow a$	h: $1 - c_1$	2: $[\frac{1}{2}, 1]$
	Index: +1, all forward induction	$\ell \longrightarrow s$ with any x_ℓ	$\bar{s} \longrightarrow \bar{a}$	ℓ: 0	
	(P1-E2'-P3): *both use* $\bar{s}$	$h \longrightarrow \bar{s}$	$s \longrightarrow a$ with $y \in [0, \min\{y' + c_1, 1\}]$	h: $[0, 1]$	2: $\frac{1}{2}$
	Index: 0, not all fwd. induction	$\ell \longrightarrow \bar{s}$	$\bar{s} \longrightarrow a$ with $y' \in [0, 1]$	ℓ: $[0, 1]$	

Table 3. Equilibrium structure, class I, case 3: $0 \leq c_1 \leq 1 < c_2$

Prior	Equilibrium component	Strategies		Payoffs	
$p < \frac{1}{2}$	(E*): *perfectly revealing*	$h \longrightarrow s$	$s \longrightarrow a$	h: $1 - c_1$	2: 1
	Index: +1, forward induction	$\ell \longrightarrow \bar{s}$	$\bar{s} \longrightarrow \bar{a}$	ℓ: 0	
	(P1) : *both use* $\bar{s}$	$h \longrightarrow \bar{s}$	$s \longrightarrow a$ with $y \leq c_1$	h: 0	2: $1 - p$
	Index: 0, not fwd. induction	$\ell \longrightarrow \bar{s}$	$\bar{s} \longrightarrow \bar{a}$	ℓ: 0	
$p > \frac{1}{2}$	(E2): *partially revealing*	$h \longrightarrow \bar{s}$ with $\frac{1-p}{p}$	$s \longrightarrow a$	h: $1 - c_1$	2: p
	Index: -1, forward induction	$\ell \longrightarrow \bar{s}$	$\bar{s} \longrightarrow a$ with $1 - c_1$	ℓ: $1 - c_1$	
	(E*) : *perfectly revealing*	$h \longrightarrow s$	$s \longrightarrow a$	h: $1 - c_1$	2: 1
	Index: +1, forward induction	$\ell \longrightarrow \bar{s}$	$\bar{s} \longrightarrow \bar{a}$	ℓ: 0	
	(P3) : *both use* $\bar{s}$	$h \longrightarrow \bar{s}$	$s \longrightarrow a$ with any y	h: 1	2: p
	Index: +1, forward induction	$\ell \longrightarrow \bar{s}$	$\bar{s} \longrightarrow a$	ℓ: 1	
$p = \frac{1}{2}$	(E*) : *perfectly revealing*	$h \longrightarrow s$	$s \longrightarrow a$	h: $1 - c_1$	2: 1
	Index: +1, all forward induction	$\ell \longrightarrow \bar{s}$	$\bar{s} \longrightarrow \bar{a}$	ℓ: 0	
	(P1-E2'-P3): *both use* $\bar{s}$	$h \longrightarrow \bar{s}$	$s \longrightarrow a$ with $y \in [0, \min\{y' + c_1, 1\}]$	h: $[0, 1]$	2: $\frac{1}{2}$
	Index: 0, not all fwd. induction	$\ell \longrightarrow \bar{s}$	$\bar{s} \longrightarrow a$ with $y' \in [0, 1]$	ℓ: $[0, 1]$	

3.2 Replicator dynamics and best-response dynamics

Replicator dynamics for the normal form

The replicator dynamics for a two-population game is given by:

$$\begin{aligned} \dot{x}_i &= x_i(u_i^1 - \bar{u}^1), \quad i = 1, \dots n^1 \\ \dot{y}_j &= y_j(u_j^2 - \bar{u}^2), \quad j = 1, \dots n^2, \end{aligned} \tag{1}$$

where u_i^k is the payoff of player k playing strategy i, and $\bar{u}^k$ is the average payoff of player k.

For our game, with the notation $\mathbf{y} = (y(aa), y(a\bar{a}), y(\bar{a}a), y(\bar{a}\bar{a}))$, $y = y(aa) + y(a\bar{a})$, $y' = y(aa) + y(\bar{a}a)$, we can simplify the payoffs for player 1 against a mixed strategy $\mathbf{y}$ of player 2:

$$\begin{aligned} u_1^1 &= u^1(ss, \mathbf{y}) = y - pc_1 - (1-p)c_2 \\ u_2^1 &= u^1(s\bar{s}, \mathbf{y}) = p(y - c_1) + (1-p)y' \\ u_3^1 &= u^1(\bar{s}s, \mathbf{y}) = (1-p)(y - c_2) + py' \\ u_4^1 &= u^1(\bar{s}\bar{s}, \mathbf{y}) = y' \end{aligned} \tag{2}$$

Note that

$$u^1(ss) + u^1(\bar{s}\bar{s}) = u^1(s\bar{s}) + u^1(\bar{s}s) \tag{3}$$

Similarly, with $\mathbf{x} = (x(ss), x(s\bar{s}), x(\bar{s}s), x(\bar{s}\bar{s}))$, $x_h = x(ss) + x(s\bar{s})$ and $x_\ell = x(ss) + x(\bar{s}s)$, we can express the payoffs for player 2 against a mixed strategy $\mathbf{x}$ of player 1 as

$$\begin{aligned} u_1^2 &= u^2(aa, \mathbf{x}) = p \\ u_2^2 &= u^2(a\bar{a}, \mathbf{x}) = px_h + (1-p)(1 - x_\ell) \\ u_3^2 &= u^2(\bar{a}a, \mathbf{x}) = p(1 - x_h) + (1-p)x_\ell \\ u_4^2 &= u^2(\bar{a}\bar{a}, \mathbf{x}) = 1 - p \end{aligned} \tag{4}$$

Note again that

$$u^2(aa) + u^2(\bar{a}\bar{a}) = u^2(a\bar{a}) + u^2(\bar{a}a) \tag{5}$$

We point out that (3) and (5) hold for any normal-form game derived from a game tree as given in Figure 3 (for any specification of payoffs at the end nodes of the tree). These special features allow us to reduce the replicator dynamics to smaller dimension: As shown in the Proposition, equations (3) and (5) imply that $\frac{x(ss)x(\bar{s}\bar{s})}{x(s\bar{s})x(\bar{s}s)}$ and $\frac{y(aa)y(\bar{a}\bar{a})}{y(a\bar{a})y(\bar{a}a)}$ are constants of motion for the replicator dynamics of the normal form games, see Gaunersdorfer, Hofbauer, and Sigmund (1991), Cressman (2003). So the 6-dimensional state space $\Delta_4 \times \Delta_4$ is foliated into a two parameter family of 4-dimensional invariant manifolds. On the "central" invariant manifold given by

$$x(ss)x(\bar{s}\bar{s}) = x(s\bar{s})x(\bar{s}s), \quad y(aa)y(\bar{a}\bar{a}) = y(a\bar{a})y(\bar{a}a) \tag{6}$$

which is sometimes called the *Wright manifold* (see, for example, Cressman, 2003), the replicator dynamics simplifies to (10) below, as we will show now.

Proposition. Let

$$\dot{x}_i = x_i(u_i - \bar{u}), \quad i = 1, \dots 4 \tag{7}$$

be the replicator equations for one population whose payoff function $u : \Delta_4 \to \mathbb{R}^4$ satisfies $u_1 + u_4 = u_2 + u_3$. Then $\frac{x_1 x_4}{x_2 x_3}$ is a constant of motion for (7). The invariant manifold $x_1 x_4 = x_2 x_3$ can be parameterized by $x_1 = xx', x_2 = x(1-x'), x_3 = (1-x)x', x_4 = (1-x)(1-x')$ with $(x, x') \in [0,1]^2$ where conversely, $x = x_1 + x_2, x' = x_1 + x_3$. On this invariant manifold, (7) can be written as

$$\begin{aligned} \dot{x} &= x(1-x)(u_1 - u_3) \\ \dot{x}' &= x'(1-x')(u_1 - u_2) \end{aligned} \tag{8}$$

Proof. Applying the quotient rule to (7) yields:

$$\left(\frac{x_1 x_4}{x_2 x_3}\right)^{\cdot} = \left(\frac{x_1 x_4}{x_2 x_3}\right)(u_1 + u_4 - u_2 - u_3) = 0. \tag{9}$$

By $x_1 x_4 = x_2 x_3$ one obtains (8). □

Applying (8) to (2) and (4) yields the replicator dynamics on the Wright manifold:

$$\begin{aligned} \dot{x}_h &= x_h(1-x_h)(y - c_1 - y')p \\ \dot{x}_\ell &= x_\ell(1-x_\ell)[y - c_2 - y'](1-p) \\ \dot{y} &= y(1-y)[px_h - (1-p)x_\ell] \\ \dot{y}' &= y'(1-y')[p(1-x_h) - (1-p)(1-x_\ell)] \end{aligned} \tag{10}$$

Replicator dynamics for behavior strategies

The above system of differential equations on the hypercube $[0,1]^4$ can be derived directly from the extensive form, as the *replicator dynamics for behavior strategies*. For this purpose we interpret $x_h = \text{Prob}(s|\text{high})$, $x_l = \text{Prob}(s|\text{low})$, $y = \text{Prob}(a|s)$, and $y' = \text{Prob}(a|\bar{s})$. Recall that in a binary choice game, with alternatives A and B, and frequencies x and $1-x$, the replicator dynamics reads $\dot{x} = x(1-x)[u(A) - u(B)]$.

For the costly-signaling game in Figure 3 (Class I) this leads exactly to (10). The factors p and $1-p$ in the first two equations come from the probabilities of Nature's draw. Equation (10) is like the replicator equation for a binary 4-person game with linear incentives, with the hypercube $[0,1]^4$ as state space.

We now analyze the dynamics (10) for class I, for each of the three relevant cases concerning the cost parameters and, within each of these, the three relevant cases regarding p.

Class I, case 1, $0 < \mathbf{c_1} < \mathbf{c_2} < 1$:

$\mathbf{p < \frac{1}{2}}$: All 2^4 corners of the hypercube are rest points of (10), and additionally also the Nash equilibrium E1 $= (1, \frac{p}{1-p}, c_2, 0)$, and the edges $(0,0,*,0), (0,0,*,1), (1,1,0,*), (1,1,1,*)$.

Dynamics near E1:
E1 is a quasistrict Nash equilibrium, since the external eigenvalues are $\frac{(1-x_h)^{\cdot}}{1-x_h} = (c_2 - c_1)p < 0$ and $\frac{\dot{y'}}{y'} = 2p - 1 < 0$. In the supporting boundary face $(1,*,*,0)$, which in Figure 7 corresponds to the lower front square, we have

$$\begin{aligned} \dot{x}_\ell &= x_\ell(1-x_\ell)[y - c_2](1-p) \\ \dot{y} &= y(1-y)[p - (1-p)x_\ell] \end{aligned} \tag{11}$$

which is the replicator dynamics for a cyclic 2×2 game, with closed orbits around the equilibrium E1. Since for each of these periodic solutions, the two external eigenvalues (Floquet exponents) equal the above two external eigenvalues at the equilibrium E1 (by the averaging property of replicator dynamics), these attract each a three-dimensional manifold of solutions. Altogether the boundary face $(1,*,*,0)$ attracts an open set of initial conditions from $[0,1]^4$.

Dynamics near the edge containing P1, $(0,0,*,0)$:
Near the rest points $(0,0,y,0)$, for the transversal directions, we have the linearized dynamics

$$\begin{aligned} \dot{x}_h/x_h &= (y - c_1)p \\ \dot{x}_\ell/x_\ell &= (y - c_2)(1-p) \\ \dot{y'}/y' &= p - (1-p) < 0 \end{aligned} \tag{12}$$

so these are Nash equilibria for $0 \le y \le c_1 < c_2$. For $0 \le y < c_1$, all three external eigenvalues are negative, hence this is a quasistrict Nash equilibrium and attracts a 3-dimensional stable manifold. The basin of attraction of the whole component P1 contains an open set from the hypercube. Now we study the behavior near the end point of P1, -P1$= (0,0,c_1,0)$. This point has a 2-dimensional stable manifold and a 2-dimensional center manifold, the latter contained in the 2-dimensional face $(*,0*,0)$ with dynamics

$$\begin{aligned} \dot{x}_h &= x_h(1-x_h)[y - c_1]p \\ \dot{y} &= y(1-y)px_h \end{aligned} \tag{13}$$

This is the replicator dynamics of a degenerate/nongeneric 2×2 game shown in the left panel of Figure 10. There is one orbit converging to the endpoint -P1, and one orbit with -P1 as α–limit which converges to the corner $(1,0,1,0)$ (this corner is unstable in the x_ℓ direction and hence is not a Nash equilibrium). This shows that the endpoint -P1 is unstable (in contrast to all other

Nash equilibria in the component P1) and hence the component P1 itself is unstable: There is an orbit connecting to the corner $(1,0,1,0)$, sitting on the face of E1.

Convergence

We show that all orbits in the interior of the hypercube converge to either the supporting face of E1 or to the component P1. On the boundary, orbits may also converge to one of the corners:

From the first two equations of (10) we see that

$$\frac{\dot{x}_h}{px_h(1-x_h)} - \frac{\dot{x}_\ell}{(1-p)x_\ell(1-x_\ell)} = c_2 - c_1 > 0 \tag{14}$$

and hence

$$\frac{1}{p}[\log x_h - \log(1-x_h)]^{\cdot} - \frac{1}{1-p}[\log x_\ell - \log(1-x_\ell)]^{\cdot} = c_2 - c_1 > 0$$

and

$$\left[\frac{x_h}{1-x_h}\right]^{1-p}\left[\frac{1-x_\ell}{x_\ell}\right]^{p} \uparrow \infty$$

Since the numerators are bounded, we infer

$$(1-x_h)x_\ell \to 0 \tag{15}$$

so that all interior orbits converge to the union of the two facets $x_h = 1$ (in Figure 7, the bottom cube) and $x_\ell = 0$ (the inner cube).

Similarly, we obtain from the last two equations of (10)

$$[\log y - \log(1-y) + \log y' - \log(1-y')]^{\cdot} = \frac{\dot{y}}{y(1-y)} + \frac{\dot{y}'}{y'(1-y')} = 2p - 1 < 0 \tag{16}$$

and, since $p < \frac{1}{2}$,

$$yy' \to 0$$

so that all interior orbits converge to the union of the two facets $y = 0$ and $y' = 0$. All in all, the ω–limit sets must be contained in the union of four 2-dimensional faces:

$(1,*,0,*)$ — there, all orbits converge to $(1,0,0,0)$,

$(1,*,*,0)$ — this is the face containing E1 and the periodic solutions (Figure 10, top left panel),

$(*,0,0,*)$ — there, all orbits converge to $(0,0,0,0)$, and

$(*,0,*,0)$ — the dynamics on this face, the inner front square in Figure 7, which contains the equilibrium component P1 in an edge, was described above (Figure 10, top right panel).

Best-reponse dynamics

The best-reponse dynamics for a two-population game is given by:

$$\begin{aligned} \dot{\mathbf{x}} &= \mathrm{BR}^1(\mathbf{y}) - \mathbf{x} \\ \dot{\mathbf{y}} &= \mathrm{BR}^2(\mathbf{x}) - \mathbf{y} \end{aligned} \tag{17}$$

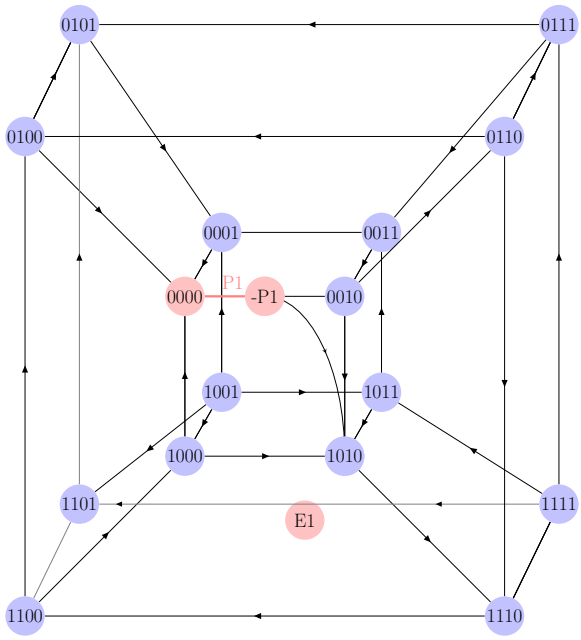


Figure 7: The hypercube: Nash equilibria for class I, case 1 ($0 < c_1 < c_2 < 1$), $p < 1/2$. Arrows on the edges show the direction of the flow of the replicator dynamics (10). Edges without arrows consist of rest points. Nash equilibria are coloured red. Also shown: the connecting orbit from -P1 to 1010.

All orbits converge to one of the Nash equilibria: either to E1, or to P1. This follows, for instance, from Berger (2005), since we can reduce the 4×4 game to a 3×2 game (for $p < \frac{1}{2}$) in the following way: For all $p \in (0, 1)$, the counter intuitive strategy of player 1 $\bar{s}s$ (don't use the costly signal if high, use it if low) is strictly dominated:

$$u^1(\bar{s}s) < (1-p)u^1(ss) + pu^1(\bar{s}\bar{s}).$$

If $p < \frac{1}{2}$, then for player 2, aa is strictly dominated by $\bar{a}\bar{a}$, and (after $\bar{s}s$ is eliminated) also $\bar{a}a$ is dominated by $\bar{a}\bar{a}$ (except at ss, that is, $x_h = x_\ell = 1$). Therefore, the game is reduced to the 3×2 game, or where $x_h \geq x_\ell$ and $y' = 0$. (This would also give an alternative proof for the replicator dynamics that $y' \to 0$.)

E1 is asymptotically stable, the component P1 is not. Still and all, both components attract big open sets. Most orbits converging to P1 converge to the corner $(0, 0, 0, 0)$.

$\mathbf{p > \frac{1}{2}}$: Here (10) has the following rest points: all 2^4 corners of the hypercube (Figure 8), the edges $(1, 1, 0, *)$ and $(1, 1, 1, *)$ where player 1 uses the costly signal in both of his types (the latter containing the Nash-equilibrium component P2), the edges $(0, 0, *, 0)$ and $(0, 0, *, 1)$ where player 1 never uses the costly signal, in none of his types (which is the Nash-equilibrium component P3),

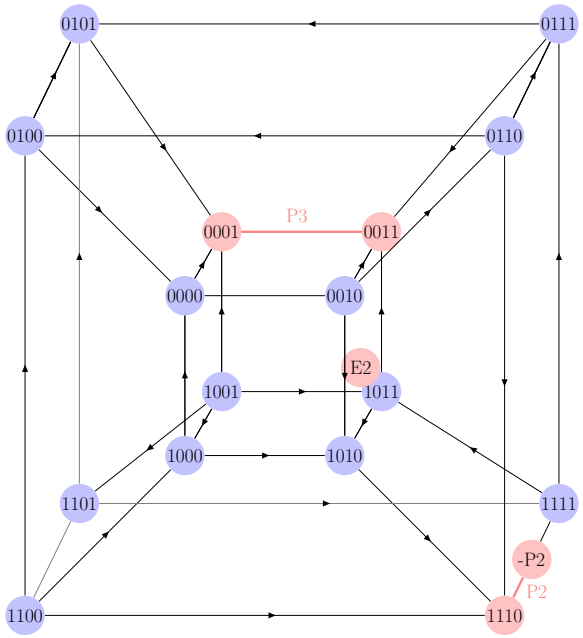


Figure 8: The hypercube: Nash equilibria for class I, case 1, $p > 1/2$.

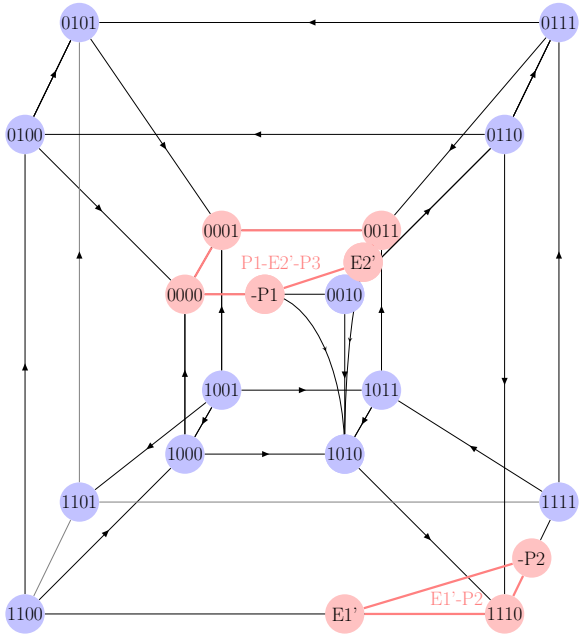


Figure 9: The hypercube: Nash equilibria for class I, case 1, $p = 1/2$. Also shown: two orbits leading from the component P1-E2'-P3 to 1010.

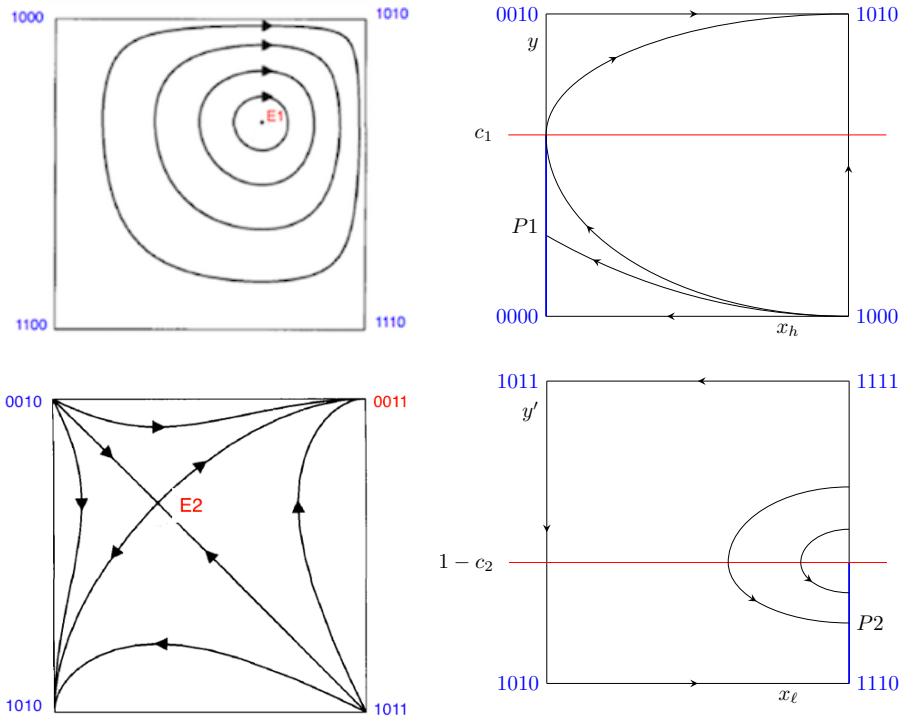


Figure 10: Phase portraits of the replicator dynamics. At the top for the case $p < 1/2$: left, on the face $(1, *, *, 0)$ containing E1; right, on the face $(*, 0, *, 0)$ containing P1. At the bottom for the case $p > 1/2$: left, on the face $(*, 0, 1, *)$ containing E2; right, on the face $(1, *, 1, *)$ containing P2.

and the Nash equilibrium at E2 $= (1 - \frac{1-p}{p}, 0, 1, 1 - c_1)$.

The expression in (16) is now positive, because $p > \frac{1}{2}$, and hence

$$(1 - y)(1 - y') \to 0.$$

This means that all orbits converge to the union of the two facets $y = 1$ (the cube at the right) and $y' = 1$ (the cube in the back). Together with (15), which holds for all $p \in (0, 1)$ and shows convergence to the union of $x_h = 1$ (the bottom cube) and $x_\ell = 0$ (the inner cube), the ω–limit sets must be contained in the union of the following four 2-dimensional faces:

$(1, *, 1, *)$ — this face (the lower right square) contains the edge of rest points $(1, 1, 1, *)$; interior orbits in this face converge to one of the Nash equilibria $(1, 1, 1, y')$ in P2 (with $0 < y' < 1 - c_2$); see Figure 8 and lower right panel in Figure 10.

$(1, *, *, 1)$ — interior orbits in this face (the lower back square) converge to the corner (1,0,1,1), see Figure 8.

$(*, 0, 1, *)$ — this face (the inner right square) contains the isolated equilibrium E2. Most orbits in this face converge to $(0, 0, 1, 1) \in$ P3 or to $(1, 0, 1, 0)$. The face itself is unstable along the edge $(1, *, 1, 0)$ along which there is a connection to $(1, 1, 1, 0) \in$ P2. The saddle point E2 lies on the separatrix, i.e., the manifold separating the two basins of attraction; see Figure 8 and lower left panel in Figure 10.

$(*, 0, *, 1)$ — this face (the inner back square) contains the edge of rest points $(0, 0, *, 1)$ which is exactly the equilibrium component P3. Interior orbits in this face converge to one of the Nash equilibria in P3.

Behavior near P3. In an analogous way to (12), one can show that each equilibrium in P3 is quasistrict. Therefore, P3 is asymptotically stable.

Behavior near P2. P2 is stable and interior attracting (Cressman 2003), but not asymptotically stable, since the whole edge spanned by P2 consists of rest points.

Best-response dynamics

The region $\{(x_h, x_\ell, y, y') \in [0, 1]^4 : p(1 - x_h) - (1 - p)(1 - x_\ell) < 0, y - c_2 - y' > 0\}$ is forward invariant under the best-response dynamics, and orbits move straight towards the Nash equilibrium $(1, 1, 1, 0)$ in P2. In the forward invariant region $\{(x_h, x_\ell, y, y') \in [0, 1]^4 : 0 < px_h - (1 - p)x_\ell < 2p - 1, y - c_1 - y' < 0\}$ orbits move straight towards the Nash equilibrium $(0, 0, 1, 1)$ in P3. And in the forward invariant region $\{(x_h, x_\ell, y, y') \in [0, 1]^4 : 0 > px_h - (1 - p)x_\ell, y - c_1 - y' < 0\}$ orbits move straight towards the Nash equilibrium $(0, 0, 0, 1)$ in P3. Furthermore, it is easy to check that both P2 and P3 are asymptotically stable, every best-response path converges to the set of Nash equilibria, and that every Nash equilibrium is the limit of some orbit from the interior.

$\mathbf{p = \frac{1}{2}}$: The replicator dynamics for behavior strategies (10) is now (after omitting the common factor $\frac{1}{2}$) given by

$$\begin{aligned} \dot{x}_h &= x_h(1-x_h)(y-y'-c_1) \\ \dot{x}_\ell &= x_\ell(1-x_\ell)(y-y'-c_2) \\ \dot{y} &= y(1-y)[x_h - x_\ell] \\ \dot{y}' &= y'(1-y')[-x_h + x_\ell] \end{aligned} \tag{18}$$

From the last two equations we get a constant of motion:

$$[\log y - \log(1-y) + \log y' - \log(1-y')]^{\cdot} = \frac{\dot{y}}{y(1-y)} + \frac{\dot{y}'}{y'(1-y')} = 0 \tag{19}$$

and hence, with $C > 0$ constant,

$$yy' = C(1-y)(1-y').$$

Recall that the argument leading from (14) to (15) is valid for all $p \in (0,1)$. Hence

$$(1-x_h)x_\ell \to 0 \tag{20}$$

so that all interior orbits converge to the union of the two facets $x_h = 1$ (the bottom cube) and $x_\ell = 0$ (the inner cube).

The set of Nash equilibria splits into two connected components, each of them 2-dimensional:

$$x_h = x_\ell = 0, \quad y' \geq y - c_1$$

$$x_h = x_\ell = 1, \quad y' \leq y - c_2$$

The first is the component P1-E2'-P3 which is exactly the convex hull of P1, E2' $= (0,0,1,1-c_1)$ and P3 (Figure 9). It is a pentagon with 3 right angles and a line of symmetry. All equilibria with $x_h = x_\ell = 0, y' > y - c_1$ are quasistrict and attract a 2-dimensional stable manifold, together an open set of orbits in $[0,1]^4$. However, this component P1-E2'-P3 is unstable, in agreement with its index being 0. Indeed the vertex E2' is unstable (as it is for $p > \frac{1}{2}$): On $(*,0,1,*)$ (the inner right square), there is an orbit from E2' down to $(1,0,1,0)$ (see Figure 9) and from there to $(1,1,1,0)$ in the component E1'-P2. Similarly, every point on the line segment $(0,0,y'+c_1,y') : 0 \leq y' \leq 1-c_1$ (the edge of the pentagon connecting E2' with the endpoint -P1 of the component P1) is unstable. From each of these points there is a connecting orbit to $(1,0,1,0)$.

The other component E1'-P2 is stable (but not asymptotically stable) under the replicator dynamics. Since E1' $= (1,1,c_2,0)$ and P2 is the line segment from $(1,1,1,1)$ to $(1,1,1,1-c_2)$, the component E1'-P2 is the convex hull of E1 and P2, a triangle. All equilibria with $x_h = x_\ell =$

$1, y' < y - c_2$ are quasistrict and attract a 2-dimensional stable manifold, together an open set of orbits in $[0,1]^4$.

Best-response dynamics

The region $\{(x_h, x_\ell, y, y') \in [0,1]^4 : x_h < x_\ell, y - c_1 - y' < 0\}$ is forward invariant under the best-response dynamics, and orbits move straight towards the Nash equilibrium $(0,0,0,1)$ in P3. In the forward invariant region $\{(x_h, x_\ell, y, y') \in [0,1]^4 : x_h > x_\ell\}$ orbits move towards the Nash equilibrium $(1,1,1,0)$ in E1'-P2. Furthermore, it is easy to check that every best-response path converges to the set of Nash equilibria. If we start on the set $x_h = x_\ell$ we can reach any Nash equilibrium.

We remark that for $p = \frac{1}{2}$, both dynamics on the hypercube are symmetric w.r.t. $(y, y') \mapsto (1 - y', 1 - y)$.

To summarize, how does the flow on the hypercube change, as p goes through $\frac{1}{2}$? The flow on $x_h = x_\ell = 0$ (the upper inner square) switches in the y' direction from $\downarrow$ to $\uparrow$, thus replacing the attractor P1 with the attractor P3. The flow on $x_h = x_\ell = 1$ (the bottom outer square) switches in the y direction from $\leftarrow$ to $\rightarrow$. All the other arrows on the one-dimensional skeleton of the hypercube stay the same!

Class I, case 2: $\mathbf{0 < c_1 < c_2 = 1}$

From (10) we get $\dot{x}_\ell < 0$ in $(0,1)^4$ and $\dot{x}_\ell = 0$ if $y = 1$ and $y' = 0$. Hence the ω–limit of every interior orbit is contained in the union of $(*, *, 1, 0)$ (the front right square) and $(*, 0, *, *)$ (the inner cube). This is an example of a weakly dominated strategy that is not eliminated under the replicator dynamics.

$\mathbf{p < \frac{1}{2}}$: Here the equilibrium E1 moves from a 2-dimensional face onto an edge (the right lower front edge connecting the outside to the inner cube): E1 $= (1, \frac{p}{1-p}, 1, 0)$. Therefore, this whole edge $(1, *, 1, 0)$ consists of rest points of the replicator dynamics, and these are Nash equilibria if and only if $x_\ell \leq \frac{p}{1-p}$. So E1 is now the end point of a one-dimensional component of Nash equilibria, bounded by E1 and E* $= (1, 0, 1, 0)$, the perfectly revealing equilibrium. This equilibrium component E*-E1 (and every single equilibrium in it) is stable under the replicator dynamics. The component is even asymptotically stable under the best-response dynamics. But E* is the only point in this component which is stable under the best-response dynamics.

The other component P1 is again unstable: there is an orbit in $(*, 0, *, 0)$ (the inner front square) connecting the endpoint of P1 to E*.

$\mathbf{p \geq \frac{1}{2}}$: The components P2 and E1'-P2 shrink to the singleton $(1,1,1,0)$ as $c_2 \uparrow 1$. But for $c_2 = 1$ the whole edge $(1, *, 1, 0)$ connecting E* $= (1, 0, 1, 0)$ with $(1, 1, 1, 0)$ consists of Nash equilibria. This component is again stable under the replicator dynamics and asymptotically stable under

the best-response dynamics. The other components behave as in the case $c_2 < 1$.

Class I, case 3, $0 < c_1 < 1 < c_2$

From (10) we get $\dot{x}_\ell / x_\ell < 0$ in $[0,1]^4$ and hence $\dot{x}_\ell \downarrow 0$ whenever $x_\ell < 1$. Now the perfectly revealing equilibrium E* = (1,0,1,0) is a strict Nash equilibrium, and therefore asymptotically stable for the replicator and the best-response dynamics. As c_2 increases from the value 1 to values larger than 1, the one-dimensional component on the edge from E* to $(1,1,1,0)$ shrinks suddenly to the strict equilibrium E*. The other components behave as in the case $c_2 < 1$.

Summary:

In Class I, the components with index +1, that is, E1, E*-E1, E*, P2, E*-E1'-P2, and P3, whenever they exist, are stable under the replicator dynamics and asymptotically stable under the best-response dynamics. All other components are unstable.

3.3 Applications

The typical application of class I are educational credentials as a signal for performance or productivity as Spence (1973) has suggested it—the underlying hypothesis being that obtaining a certain degree is less costly in terms of effort and time for the more productive type.

The *education-as-a-costly-signaling* hypothesis has a corollary for phenomena related to language, for language competences (in one's own or a foreign language) often seem to function as the carriers of such educational signals. To speak with a certain twist of tongue, to express oneself elaborately or in a certain tone is often taken as correlating with a certain level of education, up to standing for a certain school or type of school. Bourdieu (1982, 1991) prominently describes such phenomena. Similarly as what concerns foreign language competences: having more foreign languages on one's CV usually is considered to give one an edge in the job market. This hypothesis might provide insight into the economics of languages. It might explain, for example, why workers who have competences in foreign languages that are *not used* in a given work environment still have a higher wage (a phenomenon reported, for instance, by Ginsburgh and Pietro-Rodriguez 2011). More generally, the bare ability to speak and write grammatically correct might function as a signal of certain social abilities, such as the ability to abide to certain rules, to understand and adapt to different social environments, which are not only valuable qualities in the work place, but which more broadly testify of our being reliable and predictable members of society. Language competences are an ideal carrier of such qualities because they are permanently put on display. In that perspective, Class I might be a good model for phenomena studied in sociolinguistics, such as the social meaning of certain accents or dialects, but also the bare ability to switch between different such *styles* (see, for example, Eckert and Rickford 2001).

Costly-signaling arguments that anthropologists have advanced to explain certain seemingly wasteful foraging strategies also seem to fall into the pattern captured by class I—*differential costs in producing the signal*. Bliege Bird and co-authors (Bliege Bird et al. 2001, Bliege Bird and Smith 2005), for example, have found that the Meriam, a Melanesian people, engage in certain forms of hunting, namely, spearfishing and collaborative turtle hunting, that are inefficient in terms of calories and macronutrients with respect to other viable foraging strategies, most importantly the collecting of shellfish. Bliege Bird et al. notably argue that spearfishing, which is practiced mostly by young men, comes to function as a signal precisely because its rate of success is a function of the individual performing it, while the collecting of shellfish, in which everybody participates, has constant outcomes over individuals. What's signaled by such inefficient foraging strategies, so Bliege Bird at al., are unobservable physical qualities and cognitive skills, such as strength, agility, precision, and risk-taking, and, in the case of turtle hunting, also leadership skills and generosity (the hunt is organized in groups under a hunt leader and proceeds of the hunt are provided for public feasts), which increase social status and give advantages in mate choice.

On the other hand, the assumption of Class I that the two types face different costs in *producing* the signal might be hard to justify in some applications. This comes out most clearly when the cost of the signal is some fixed monetary value. For example: placing an add in a newspaper has a price, but that price usually is a fixed rate and not a function of the quality of the company or institution who buys the ad. And quite similarly for advertising in the animal world by the display of a *handicap*: while having a colorful coat plausibly can be considered a cost in terms of the chances of survival, because an individual who carries it will be spotted earlier or more likely by a predator, it is less clear that that cost should be different for different types. After all, the augmented probability to be seen is a function of the observable trait and not necessarily a function of the unobservable trait. And, indeed, formal models of advertising or respectively the *handicap principle* do not turn on the assumption of different costs in producing the signal, but are grounded in the idea that different types have a *different background payoff*, or *fitness*, from which the cost of the signal, possibly uniform across types, is deducted. Class II (section 5) captures this mechanism.

4 Belief-based refinements of sequential Bayesian Nash equilibrium

Sequential Bayesian Nash equilibrium (Kreps and Wilson 1982) requires that players update their beliefs over the possible states of nature (here player 1's types) according to Bayes' rule *along the equilibrium path*, that is, the path through the game actually taken in the equilibrium under

study as determined by the players' equilibrium strategies. However, it does not—at least not for the class of games to which belong signaling games—impose any restrictions on beliefs "off the equilibrium path," that is, a situation that could in principle happen, but that does not happen in the equilibrium under study—a counterfactual situation, one could say. This is important because what can be an equilibrium outcome depends on what players would do "off the equilibrium path."

In signaling games, a situation off the equilibrium path is one after a signal that is in principle part of the game but that is not used in the equilibrium under study. In the games studied here, this concerns equilibrium outcomes in which both types use the same signal, such as P1, P2, and P3. Take, for instance the equilibrium outcome P1 (which exists for $p < 1/2$), in which both types of player 1 use $\bar{s}$, and player 2 in response to $\bar{s}$ takes $\bar{a}$. Relative to this equilibrium outcome, the situation that the costly signal s is observed is "off the equilibrium path." Certainly, for any of player 1's types, whether s or $\bar{s}$ is a best response to player 2's strategy depends not only on what player 2 does in the absence of the signal (on the equilibrium path), but also on what player 2 were to do off the equilibrium path, in the event that the costly signal s were observed. In any equilibrium that belongs to the component P1, player 2 in response to s takes a with a probability in $[0, c_1]$, that is, in no case higher than c_1, which by assumption is strictly below 1. Imagine that contrary to that player 2 in response to s were to take a for sure, that is, with a probability of 1. Then, for player 1, no matter if he is of the high or low type, using $\bar{s}$ would no longer be a best response. He should use s instead. The equilibrium would break down.

In a rationality-oriented game-theoretic perspective, players' equilibrium strategies have to be supported by their beliefs. Let us look at P1 again: player 2's equilibrium strategy which in response to the off-the-equilibrium-path signal s has her take a with a probability of c_1 *at most* implies that after s player 2 attributes to the high type a probability of $1/2$ *at most* (for if she were to attribute to the high type a probability of more than $1/2$, she would have to take a for sure). One could wonder whether that is plausible, because s is less expensive for the high type. In the extreme case that $c_1 = 0$ this appears particularly implausible: the high type pays nothing for the signal, but when the signal is expressed, one should think that it came from the low type? Bayes' law, we should be reminded, does not help us here, because it is not defined (see Figure 5).

Classical refinements of sequential Bayesian Nash equilibrium take such considerations as a starting point: they operate on the principle of imposing restrictions on players' beliefs "off the equilibrium path." Such restrictions, so to say, come to complement Bayes' rule where it is not defined, and thereby *refine* the Bayesian Nash equilibrium notion. Depending on what is considered a plausible restriction on beliefs off the equilibrium path (how one thinks that Bayesian rational players should think when Bayes' rule does not apply), there is an entire family of such refinement concepts. Some of those concepts, for instance, the *never-a-weak-best-response* criterion (Kohlberg and Mertens 1986), a criterion called *divinity* (Banks and Sobel 1987), and *forward*

induction as defined by Govindan and Wilson (2009) indeed discard the no-signaling equilibrium outcome P1. We give below the argument for Govindan and Wilson's forward-induction criterion, which, to our mind, is the most fundamental of the three, because it is defined for any game in extensive form and has a foundation in certain decision-theoretic requirements, and for the class of games that we study, conveniently, coincides with the fairly simple to check never-a-weak-best-response criterion.

Forward induction after Govindan and Wilson (2009) (the *never-a-weak-best-response criterion*) requires that after a signal off the equilibrium path the support of the belief should not contain types for whom that off-the-equilibrium-path signal is *never* (that is, for no reaction of player 2 to the off-the-equilibrium-path signal that supports the equilibrium outcome under study) an alternative best response relative to the signal used in the equilibrium under consideration. By this rule, the equilibrium outcome P1 is indeed ruled out: Within P1 there is one equilibrium point, namely the one where player 2 in response to s were to take a with a probability of exactly c_1 (the endpoint of that component), in which for the *high* type taking s is indeed an alternative best response relative to taking $\bar{s}$. For the low type there is no such point. Hence, after s, the low type has to be discarded from the support of the belief, and therefore full belief (a probability of 1) has to be put on the high type. But then, as we saw above, after s, player 2 should take a for sure (and not with a probability of c_1 at most), and this will upset the equilibrium outcome under study: P1 is *not robust under forward induction (the never-a-weak-best-response criterion).*

For class I, when $c_2 < 1$ (case 1), for $p \neq 1/2$, the no-signaling equilibrium outcome P1 is in fact the only equilibrium outcome that can be discarded by Govindan and Wilson's notion of forward induction (the never-a-weak-best-response criterion). All other equilibrium outcomes satisfy it. Notice that equilibrium outcomes in which every signal is used with at least some probability by some type, such as E1, E2 and E*, are trivially robust under any belief-based refinement (because there is no signal off the equilibrium path). In the knife-edge case $p = 1/2$, in the component E1'-P2, all outcomes are stable under the never-a-weak-best-response criterion; in the component P1-E2'-P3, some outcomes (namely those that lie between P1 and E2', including P1) are discarded by the never-a-weak-best-response criterion.

Banks and Sobel's *divinity* criterion gives the same results. Another prominent refinement of sequential Bayesian-Nash equilibrium for signaling games is the *intuitive criterion* (Cho and Kreps 1987). The intuitive criterion is less restrictive than the never-a-weak-best-response criterion: it discards a type from the support of the belief after an off-the-equilibrium path signal only if for *every possible reaction* of player 2 to the off-the-equilibrium path signal that type is strictly worse off than in the equilibrium outcome under study. Under this criterion, in P1, none of the types is discarded after s, and hence P1 survives.

Comparing equilibrium-refinement results based on forward induction with those based on the

index, one gets a fairly close overlap. In the games studied, whenever an equilibrium outcome is discarded by forward induction in the sense of Govindan and Wilson (the never-a-weak-best-response criterion), then the equilibrium component in which it sits has an index of 0, and hence cannot be asymptotically stable under any standard evolutionary dynamics. Table 1 provides an overview of the equilibrium structure of class I for the case $c_2 < 1$, indicating for each equilibrium component its index as well as whether the outcomes that belong to it satisfy forward induction in the sense of Govindan and Wilson or not. The results extend to the two other cases regarding c_2, $c_2 = 1$ and $c_2 > 1$ (Tables 2 and 3).

Refinements of sequential Bayesian Nash equilibrium that rely on imposing restrictions on beliefs off the equilibrium path can be seen as a form of *strategic* stability or robustness test, because the equilibrium outcome under study is tested in light of what a rationally reasoning player ought to believe in case that they observe a deviation from the equilibrium outcome under study, and such a deviation can be seen as another player's deliberate deviation from the strategy that they are supposed to use in the equilibrium outcome under study. (Hence also the term *forward induction*: it is as if the deviating player were counting on another player who moves further down the tree to draw a certain inference from that deviation.) Important to note is that these criteria do not require that in order to destabilize the equilibrium outcome under study, the strategy profile resulting from these deviations would itself constitute an equilibrium outcome. They truly are robustness criteria only. The deviations involved should not be thought of as being acted out, rather they should be thought of as a thought experiment that takes place in the minds of the players in the game, and a strategically stable equilibrium outcome, so the underlying idea, should be robust under this kind of thought experiment. Certainly, such criteria are relevant in applications where it is about human interaction, where the players in the game are reason-inspired social individuals. It is good to know that an equilibrium outcome that can be discarded on such rational, plausibility-of-beliefs grounds, will also be one that can be discarded on evolutionary grounds.

5 Variants of the model

5.1 Class II: same costs, different benefits in case of success: the Handicap Principle

In class II, the production of the signal is of the same cost $c > 0$ for the two types, but the high type gets an extra payoff of $d > 0$ *if the second player takes action a*. The game is shown in Figure 9. This model 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 Zahavi's (1975) handicap

principle. In Milgrom and Roberts's model (1986), the idea is 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. The argument seems to us particularly well fitted for scenarios where the function of advertising is not so much in generating a decision to buy but a *decision to inquire*, in the process of which the firm or individual with the high quality product or desired trait can provide more verifiable information, which finally will bring about the decision to buy or accept (we think, for instance, of long-term consumer goods, luxury goods, art). In Grafen's model, the argument appears implicitly in the form of assumptions on the derivatives of the fitness function.

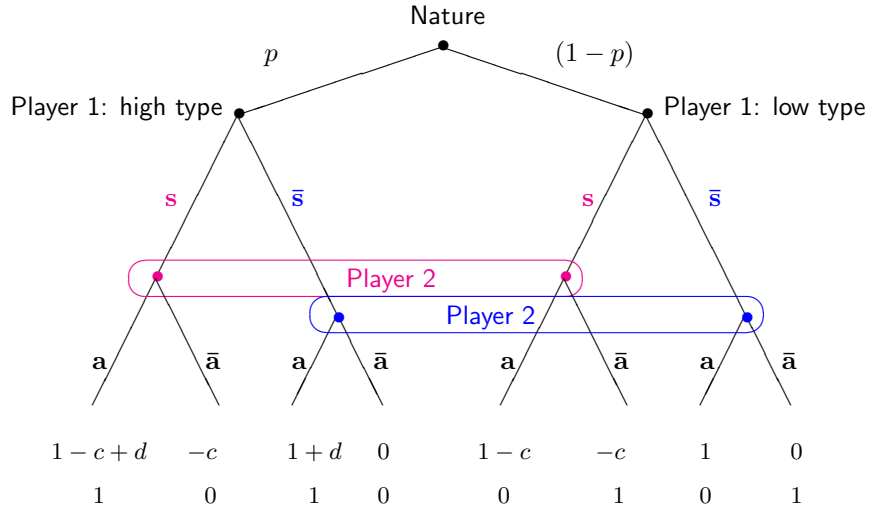
In Zahavi's (1975) original exposition of the handicap principle, which is purely verbal, it is not clear if the argument is to be understood in the sense of class I or class II. We would argue that it has to be understood in the sense of class II: payoffs are in terms of reproductive fitness, which over the lifetime of an individual has to be understood as composed of several variables, notably the success with which an individual gets mates and the chances of survival. A certain trait, like prominent plumage, that represents an effective deduction from the fitness of the individual who carries it (because it will be more visible, less fast, etc.) comes to function as a signal between potential mates if the background fitness from which the cost of this signal is deducted differs for different types. The particular payoff structure of class II (uniform costs of producing the handicap but differential payoffs if the female accepts) arises then from an implicitly dynamic argument (similar to Milgrom and Robert's repeat sales): because payoffs are in terms of fitness, an individual with higher 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 himself reproduction age.

5.2 Class I and II are structurally equivalent

A convenient circumstance links class II to class I: Provided that c and d are positive (which we assume), the games in class II have the *same equilibrium structure as those in class I*:

- If $0 < c < 1$, the equilibrium structure will be as that in class I when $c_2 < 1$;
- if $c = 1$, as that in class I when $c_2 = 1$; and
- if $1 < c \leq 1 + d$, as that in class I when $0 \leq c_1 \leq 1 < c_2$.

The numerical values defining the equilibria of class II can be obtained by those of class I by substituting c_1 by $c/(1+d)$ and c_2 by c. These values can be interpreted in a meaningful way: they represent the *net cost* of the signal—the cost benefit-ratio of using the signal—for the respective type (Table 4). Both class I and II are characterized by *differential net costs of the signal* s for



	aa	**a$\bar{a}$**	**$\bar{a}$a**	**$\bar{a}\bar{a}$**
ss	$1-c+pd,\ p$	$1-c+pd,\ p$	$-c,\ 1-p$	$-c,\ 1-p$
s$\bar{s}$	$1-pc+pd,\ p$	$p(1-c)+pd,\ 1$	$-pc+(1-p),\ 0$	$-pc,\ 1-p$
$\bar{s}$s	$1-(1-p)c+pd,\ p$	$(1-p)(1-c),\ 0$	$p-(1-p)c+pd,\ 1$	$-(1-p)c,\ 1-p$
$\bar{s}\bar{s}$	$1+pd,\ p$	$0,\ 1-p$	$1+pd,\ p$	$0,\ 1-p$

Figure 11: Class I: in the top panel, the game in extensive form—the game given by a game tree; in the bottom panel: the matrix game induced by that game in extensive form.

the two types. This property guarantees that the equilibria in these two classes will also have the same robustness properties: as far as the index and the belief-based refinements discussed in the previous section go, everything goes through exactly as for class I.

Table 4: The net cost of a signal

We call the *net cost* of a signal for type t the payoff of type t when he does not use the costly signal and player 2 in the absence of the costly signal does not take the desired action $(\pi_t(\bar{s},\bar{a}))$ minus his payoff when he does use the costly signal and player 2 at observing the costly signal does not take the desired action $(\pi_t(s,\bar{a}))$ over the payoff difference for this type when he uses the costly signal but player 2 does or does not take the desired action $(\pi_t(s,a)-\pi_t(s,\bar{a}))$:

$$\textit{net cost of } s \textit{ for type } t \;=\; \frac{\pi_t(\bar{s},\bar{a})-\pi_t(s,\bar{a})}{\pi_t(s,a)-\pi_t(s,\bar{a})}.$$

Class I: net cost of s for the high type: c_1; for the low type: c_2.

Class II, the net cost of s for the high type: $c/(1+d)$; for the low type: c.

5.3 The replicator dynamics

Here the payoffs for player 1 against mixed strategies of player 2 are given by

$$\begin{aligned} u^1(ss, \mathbf{y}) &= (1+pd)y - c \\ u^1(s\bar{s}, \mathbf{y}) &= -pc + p(1+d)y + (1-p)y' \\ u^1(\bar{s}s, \mathbf{y}) &= (1-p)(y-c) + p(1+d)y' \\ u^1(\bar{s}\bar{s}, \mathbf{y}) &= (1+pd)y' \end{aligned} \tag{21}$$

Again (3) holds. For player 2 the payoffs are the same as in (4). Thus the analog of (10), i.e., the replicator dynamics for behavior strategies, is now given by

$$\begin{aligned} \dot{x}_h &= x_h(1-x_h)[(1+d)(y-y') - c]p \\ \dot{x}_\ell &= x_\ell(1-x_\ell)[y-y'-c](1-p) \\ \dot{y} &= y(1-y)[px_h - (1-p)x_\ell] \\ \dot{y}' &= y'(1-y')[p(1-x_h) - (1-p)(1-x_\ell)] \end{aligned} \tag{22}$$

This is essentially the same as Class I with $c_1 = \frac{c}{1+d}$ and $c_2 = c$.

5.4 Some cost of the signal is needed

If in class II producing the signal were of no cost at all ($c = 0$), which we have excluded by assumption, then the only equilibria that would exist are such that none of the signals pushes player 2's belief over the critical value 1/2 and, as a consequence, player 2 acts on her prior belief, no matter which signal she has observed. That is to say: some positive cost of the signal is necessary for the signal to be at least partially revealing or "informative," and hence enable cooperation (the desired exchange: hire, buy, mate) at least sometimes. To our mind, this is the essence of the *handicap principle* (and not the claim that costly signaling is always perfectly revealing).

5.5 Applications

As a general model of advertising, class II is extremely versatile. Not only firms and animals advertise. Individuals participating in human society, as already Veblen (1899) has pointed it out, advertise for themselves too. For example, by the houses we live in, the cars we drive, and the dresses we wear, but also by the degrees we earn and the language we speak. Differential background payoffs of different types can be considered to stand not only for differential pecuniary rewards but also for differential levels of emotional involvement, attachment, desire, or esteem, which further expands the range of possible applications in the social sciences, psychology, and

linguistics. Class II seems to be the right model when investigating communal sharing, gift-giving, and charity as costly signals for status or wealth.

When it comes to linguistic applications, class II is, we would argue, a good model for problems in linguistic pragmatics where a certain speech act may well be of some cost, but where it might be hard to argue why the *production* of that speech act should be of different costs for different types. Politeness strategies (Brown and Levinson 1987) are a good example: using a more polite form (expressing oneself more elaborately, writing a longer rather than a shorter letter, attenuating a face threat by an indirect speech act, etc.) is costly, but it might be disputed that this cost should differ for different types. "Please, could you pass me the salt" is certainly longer and hence more costly than "Pass the salt!" But that is so no matter who pronounces the phrase, be it a speaker who really means well with the addressee (a cooperative type) or not. It can however reasonably be assumed that different speakers have different background payoffs *if the addressee takes the desired action*, which can be understood, for example, as expressing different degrees by which the speaker cares about the addressee.

5.6 Class I or II, or a combination?

In some phenomena, both the conditions of class I (differential cost in producing the signal) and of class II (differential background payoffs in case that the second player takes the desired action) might come in. Education is a case in point. If a certain educational credential is costly not only in terms of effort but also in terms of money, it can also come to function as a signal in the sense of class II. Having been to a certain school then becomes a signal of status or a signal for future performance and commitment. It is as if the prospect employee were saying: "It pays off for me to have invested into my degree, because once I get hired, I know that I will perform well and therefore not lose my job quickly, and so the initial investment in my degree pays off for me." Another example are signals of dress: having a good suit or dress and shoes is expensive (a signal in the sense of class II), but wearing them might, under certain circumstances, also be a physical effort that different individuals might master in different degrees (a signal in the sense of class I).

The structural equivalence of class I and II is a powerful property. It tells us that when choosing the model, we can focus on the mechanism that is the dominant one for the problem at hand; that we do not need to disentangle the two effects, because they work "in the same direction," because the results do not change qualitatively if the other aspect comes in at the margin.

If, for a certain application, both aspects are relevant, and one is interested in a finer-grained analysis, one can set up a combined model with differential costs of producing the costly signal s and an extra payoff d for the high type if player 2 takes the desired action. In such a combined model, the net cost of s for the high type will be $c_1/(1+d)$, and for the low type c_2, and the equilibrium structure will be as in class I with c_1 replaced by $c_1/(1+d)$.

6 Interpretation

6.1 Costly signaling is not necessarily a waste of social resources

A thought that runs through costly-signaling theory in economics is that signaling in markets can lead to situations in which players "overinvest" in the economic variable that functions as a signal, with the effect that the private returns to the economic variable that functions as a signal exceed that variable's marginal contribution to productivity (for reviews see, for example, Kreps and Sobel 1994, Hörner 2006, Riley 2001, Spence 2002). Taking marginal productivity as a reference point is to compare the equilibrium in the game (under asymmetric information) to an equilibrium under perfectly competitive markets (under perfect information). From a game-theoretic point of view, this is problematic because these are two different games, two different worlds. What individuals do in a situation in which information is not complete—whether what they do is efficient or not—has to be evaluated not with respect to what would be possible in another (ideal) world without informational asymmetries, but with respect to what is possible given these informational asymmetries. The welfare properties of an equilibrium outcome of that game then should be compared to other equilibrium outcomes of that game. In order to do so, one needs, of course, a fully-closed game-theoretic model.

For the games discussed here, it is possible to define in a meaningful way a "no-signaling outcome" *within the game*, namely as an outcome in which both types *do not* use the costly signal and player 2, in the absence of the costly signal, acts on her prior belief: when $p < 1/2$, she will not accept, and when $p > 1/2$, she will accept.[4] For any prior p, the thereby defined no-signaling outcome, P1 respectively P3, constitutes an *equilibrium outcome* of the game. It is therefore possible to compare the social welfare of equilibria in which the costly signal is used at least sometimes by some type (partially revealing equilibria as E1 and E2, the perfectly revealing equilibrium E*, or an equilibrium outcome in which everybody uses the costly signal) to the respective no-signaling equilibrium outcome. Such a comparison (see the last column in Tables 1 – 3, which indicates the payoffs of the two types of player 1 and of player 2 for the respective equilibrium component) shows that costly signaling, at least in the classes of games considered here, is not necessarily wasteful on a social level. Instead, whether it is or not depends on the prior probability of the types of player 1:

- When the prior probability on the good type is below the critical value, $p < 1/2$, no matter whether the cost of the signal for the low type c_2 (respectively c in class II) is below, equal or larger than 1, the equilibrium component in which the costly signal is at least partially informative (E1 respectively E*-E1 or E*) is better, in the sense of Pareto, than the co-

[4] There are games, for which this is not so obvious; for instance, the so-called "beer-quiche" game (Cho and Kreps 1987), in which the two types of player 1 get differential positive payoffs from using the two different signals.

existing equilibrium component P1, in which none of player 1's types uses the costly signal and player 2 does not take the desired action: in an equilibrium of the form E1 (respectively E*), relative to P1, nobody is made worse off and at least someone, namely, the high type of player 1, is made strictly better off (in an equilibrium in the component E*-E1 that is different from E1 respectively in E*, player 2 is also made better off relative to P1). That is: when $p < 1/2$, the use of a costly signal improves social well-being over a situation where that signal is not used, or not available. This result is readily accessible to intuition, notably in an economic context: when confidence in the quality, performance, or productivity is low, and therefore a priori nobody would buy or invest, the availability of a costly signal makes it possible to get out of such a situation in which due to informational asymmetries the market would otherwise stay closed, and this increases overall social well-being. Remarkably—and from a social point of view that can be considered a positive result—both evolutionary dynamics and classical belief-based refinements of Nash equilibrium favor the emergence of E1, respectively an equilibrium in the component E*-E1 or E*, over that of the no-signaling equilibrium outcome P1. To which extent depends on the specific evolutionary dynamics, respectively belief-based criterion that one considers (see Section 4 and 5).

- When the prior probability on the good type is above the critical value, $p > 1/2$, then payoff comparisons depend on the cost of the signal for the low type.

Case 1: When c_2, respectively c, is below 1 (Table 1), the equilibrium component P3, in which none of player 1's types uses the costly signal and player 2 in the absence of the costly signal *takes the desired action*, Pareto dominates the two other equilibrium components that exist in this case and in which the signal is used at least sometimes by some type, E2 and P2: both types of player 1 strictly prefer P3 over P2, and the low type of player 1 even strictly prefers E2 over P2, while player 2 is indifferent in all three equilibrium outcomes. The possibility to use a costly signal can be harmful here. It can result in a social tragedy, namely when players, due to self-confirming expectations, get caught in the suboptimal equilibrium outcome P2, in which everybody is forced to express the costly signal—because everybody thinks that otherwise player 2 were not to accept—which in the end has the effect that the supposed signal does not carry any information. If the players in this game were to collectively step out of such expectations, and players in the player 2 position did in fact accept when they did not observe the costly signal (based on the fact that the prior is already sufficiently high), nobody would need to signal: society as a whole would be better off. However, both P2 and P3 are stable, under both evolutionary dynamics and belief-based, strategic stability criteria. That is to say: once players have coordinated on the unhappy equilibrium outcome P2, neither

evolution nor individuals' decentralized strategically rational reasoning will take them away from there.

Evolutionary dynamics, other than belief-based refinements of Nash equilibrium, at least discard the partially revealing equilibrium E2, which has index -1 and hence cannot be stable under any standard evolutionary dynamics (it is a saddle under both the replicator and the best-response dynamics). Equilibria in the style of E2, where the absence of the costly signal brings down the prior belief to some critical value, have rarely been considered. This can be taken as evidence that such equilibria are very unintuitive to the human mind. The fact that these equilibria are unstable under evolutionary dynamics might serve as an explanation for the presence of such an intuition.

Case 2 and 3: When c_2, respectively c, is larger than or equal to 1 (Tables 2 and 3), while E2 can still be discarded on evolutionary grounds, the remaining equilibrium components can no longer be ranked according to the Pareto criterion. Player 2 now strictly prefers outcomes in the component E*-E1'-P2 that do not put full weight on P2, respectively E*, over P3. Certainly, player 2 rather gets some information about player 1, as opposed to accepting throughout, which is the best she can do if nobody uses the costly signal. The underlying potential conflict of interest between player 1 and 2 resurfaces as diverging preferences over the possible equilibrium outcomes in the game. Here again, both relevant components are stable under both evolutionary dynamics and belief-based refinements of Nash equilibrium.

6.2 In defense of the "handicap principle"

It is by now widely agreed upon that the handicap principle cannot be maintained or understood in the narrow sense that only perfectly revealing—"honest"—signaling equilibria can evolve due to the fact that signals have to be costly. It is well understood that partially revealing—"hybrid"—equilibria in the style of E1, in which the costly signal is used in equilibrium by different types with different probabilities, and hence transmits partial information, are evolutionarily relavant (Lachmann and Bergstrom 1998, Huttegger and Zollman 2010, Számadó 2011, Zollman et al. 2013).

Dawkins and Krebs (1978, Krebs and Dawkins 1985) have strenuously argued that signaling, even in the animal world, is an exercise in mind-reading and manipulation and that therefore any signaling mechanism, once in place, tends to be corrupted or invites to "cheating," which can lead to situations in which signals are only partially informative. Dawkins and Krebs's account of animal signals has sometimes been opposed to Zahavi's (1975) theory of the *handicap principle*,

which, on this view has been identified with the claim that signaling always has to be "honest" due to the handicap principle. Remarkably, many of the later game-theoretic findings, notably, the existence of partially revealing equilibria in the style of E1, mimic the phenomena of "cheating" described by Dawkins and Krebs. Though it should be emphasized that from a game-theoretic point of view there is nothing "cheating" or "dishonest" about partially revealing equilibria. These equilibria simply have the property that the costly signal does not fully reveal the high type but rather provides the receiver with an indication as to how to evaluate the probabilities of which type his opponent is going to be. In equilibrium, these evaluations correctly reflect the distribution of the behavioral program of using the signal or not using it in the two subpopulations corresponding to the two types of player 1, and females' responses to the character in question balance out the advantages and costs of carrying it: the equilibrium conditions rooted in nature do not lie.

The game-theoretic analysis shows: whether signaling in equilibrium is perfectly or only partially revealing is not a matter of principle but of degree: it depends on the specific cost parameters associated to the signal for different sender types. More specifically, what matters for the signal to be potentially a carrier of information is *not* the cost of the signal actually incurred by the high type in equilibrium, but the cost of the signal for the low type. In a perfectly revealing equilibrium, the cost of the signal for the low type has to be so high that it prevents him from using the signal at all; in a partially revealing equilibrium (in the style of E1) it prevents him from using the signal more often. Class I, in which signaling phenomena are sustained by differences in the costs directly involved in producing the signal, exposes this aspect in absolute terms: a perfectly revealing equilibrium exists only when the cost of producing the signal for the low type c_2 is at least as high as the benefit that he gets when player 2 accepts, which here is equal to 1. In the special case that $c_2 \geq 1$ and the cost of the signal for the high type c_1 is 0, there is a perfectly revealing equilibrium, in which nobody bears any direct cost for producing the signal. On the other hand, the signal being of no cost at all for the high type (and of some cost for the low type) is not sufficient to guarantee the existence of a fully separating equilibrium. If $c_1 = 0$, as long as $c_2 < 1$, only a partially revealing equilibrium will exist. In class II, if the signal is of no cost for the high type, it will also be of no cost for the low type, and then the only equilibria that exist are such that player 2 acts on her prior belief. In class II then—which represents the mechanism of the handicap principle in pure form, namely uniform costs of the signal against differential background fitness—some positive cost of the signal is necessary to guarantee that the signal transmits at least partial information.

The observation that in a number of species one sex (often the male) displays handicaps, characters such as antlers, ornaments, or brilliant coloration that seem to have no function or to be in outright opposition to the ecological problems of the species goes back to Darwin. Darwin explained such phenomena to be the result of *sexual selection*, the hypothesis that females prefer-

ably mate with individuals who excel in the display of the character in question. What is not sufficiently explained by Darwin's theory is why females would evolve such a preference. Zahavi's theory aims at tracing sexual selection back to natural selection. As Darwin already remarked, and Zahavi in that straightforwardly builds on him: the effects of sexual selection have to be compatible with the existence of the species. But therefore—and here is the twist that Zahavi brings in—only those individuals who are best adapted to the selective pressure of the species can afford to carry the risk that comes with the handicap:

> I suggest that sexual selection is effective because it improves the ability of the selecting sex to detect quality in the selected sex. [...] Before mate selection achieved its evolutionary effect the organism was in equilibrium with the pressures of natural selection. If the selective pressure of mate preference, which has no value to the survival of the individual, is added to the variety of selective pressures, the effect must be negative. *The larger the effect of the preference the more developed the character and the larger the handicap imposed. Hence a character affected by sexual selection should be correlated to the handicap it imposes on the individual.* (Zahavi 1975, p. 207, our emphasis)

That is the handicap principle (and not the claim that the handicap always fully reveals the type). The correlation between the female preference for the handicap and the cost of the handicap that Zahavi hypothesizes appears in the equilibrium conditions of the game-theoretic analysis, most clearly in the partially revealing equilibrium E1: the female's willingness to accept when she observes the handicap (the costly signal s) is given by the net cost of the handicap for the low fitness type (c_2 respectively c). The higher that cost, the more likely the female is to accept: her willingness to accept, that is, her preference for the handicap, is correlated to its cost (for the low fitness type), but that does not imply that that cost has to be so high that the low fitness type population can in no measure afford to express the handicap. The low fitness type can express it in a measure precisely such as that observation of the handicap gives the female just as much information about the male so that she is indifferent between accepting or not. Some of the females then will accept and some will not, in a proportion, which in turn is such that the population of the low fitness male is indifferent between expressing the handicap and not expressing it. One sees from this discussion that focusing on fully revealing equilibria eventually is to focus on monomorphic equilibria. Exploring the theory under parameter values for which polymorphic signaling equilibria such as E1 exist, to our mind, does not invalidate the handicap principle as originally formulated.

Zahavi's handicap principle and Dawkins and Krebs's theory of mind-reading and manipulation are, we would like to defend from a game-theoretic point of view, not to be understood as two

opposing paradigms but as two cases emerging from different parameter specifications that can be accommodated under one coherent theory.

6.3 Phenomena explained: new applications in the study of language and meaning systems

6.3.1 Indirect speech

Partially separating equilibria in the style of E1, in which the good type always uses the costly signal and the bad type uses it with a certain probability have a property that makes them a potentially very productive model when it comes to explaining features of human language, or more generally social meaning systems: The costly signal s does not perfectly reveal the speaker's type but still gives the listener an indication about the speaker's type (it will push the belief that it is the good type up to a certain level) in precisely such a way as to leave the listener *indifferent* between accepting and not accepting. In such a situation, the speaker, so to say, puts it into the hands of the listener how to react: to take the responsibility to accept or to decline. In equilibrium, of course, the listener takes this decision (deliberates between accepting and declining) with a certain regularity, namely such that the bad type is indifferent between using the costly signal s and not using it ($\bar{s}$). The costly signal s, in such an equilibrium, one can say, functions as a means to *shape the belief* of the other player in a particular way. It is as if the costly signal were to come with a tag that says: "When you receive me, understand that your belief about the good type should be 1/2—and that you are hence indifferent between accepting and not accepting." This can be interpreted as some kind of *indirect speech*.[5] In such an equilibrium, the absence of the costly signal ($\bar{s}$), on the other hand, perfectly reveals the bad type, and hence frees player 2 of the responsibility to take any strategic decision in a non-trivial sense, because when she sees that the costly signal has not been expressed, her best response is unique: not to accept. Such a situation seems quite accurate for a number of scenarios in which politeness in language acts as a costly signal in negotiating social relationships. For example, to hear the polite form, "Could you please pass me the salt," "Thank you so much for coming ...," "You have a new haircut. It looks nice." etc., often does not tell the receiver much, in the sense that it really leaves her indifferent as to whether a change in the current relationship type that links her to the speaker

[5] Steven Pinker and coauthors (2007, 2008) advance the hypothesis that the function of *indirect speech* is to *avoid common knowledge* of the type of the speaker while giving the speaker the chance to achieve the desired relationship change at least sometimes (here that would be to get accepted, hired, etc.). Equilibria of the form E1 mimic this feature, at least in a certain way: using s avoids to give player 2 sure knowledge of the sender type—however not because it leaves her in complete ambiguity about the sender's type, but because it sets her belief about the sender to a certain value somewhere strictly between 0 and 1 (here: 1/2) such that she will be indifferent between here possible actions.

of the message is warranted or not. Instead, *not the hear the polite form*, "The salt!" or simply silence is a clear negative sign, and the answer should be accordingly (for example, downgrade the current relationship type or not move to a higher relationship type).

6.3.2 Cycles around the partially separating equilibrium

How well can an equilibrium like E1, in which the probabilistic strategies that define it have to hold in a very precise way, be thought of as mimicking reality? This is where the evolutionary analysis might be particularly insightful. We have seen that under the replicator dynamics, the equilibrium E1 is locally stable but not asymptotically stable because in its supporting 2-dimensional face it is surrounded by periodic orbits. But we have also seen that this supporting face attracts an open sets of states from the interior of the state space, which is to say that close to E1, the replicator dynamics converges to a situation in which the players will cycle around something quite similar to E1: The good type will always use s and player 2, in the absence of s, will not accept (in these two positions, players behave exactly as in E1, which is precisely what defines the supporting face of E1; see Figure 7). The bad type, instead, will express the costly signal with some probability and player 2, when she observes the costly signal, will accept with some probability. These probabilistic choices will not be exactly such as to make that altogether players are in equilibrium. As a consequence, players who imitate behaviors that did well, will still have incentives to adjust their behavior. But this imitating and adjusting will make them cycle around the partially revealing equilibrium E1. And this might quite well mimic phenomena of real life. It would be a very strong assumption to require that at observing the costly signal (for example, the polite form), the listener makes a perfect Bayesian update and then takes the desired action with precisely the probability that renders the bad type indifferent between expressing the costly signal and not expressing it. But to assume that players have it approximately right and take their actions with a probability that makes them cycle around an equilibrium like E1 seems rather realistic. And similarly for applications where the interacting players are not reason-inspired humans but animals species or other organisms: with the game-theoretic, dynamic analysis, we see that hybrid signaling patterns in the style of E1 are not completely away from equilibrium, but close to it, cycling around it, and that this may well be the outcome of evolution.

6.3.3 Coexistence of different signaling conventions

Another focus of our study is the so-far neglected case that the prior probability of the high type is already above the critical value at which player 2 is indifferent between accepting and not accepting. Is the coexistence of the two equilibrium components that exist in this case and that are both stable under evolutionary dynamics, for instance P2 and P3 when $c_2 < 1$, a shortcoming of the model? Or the methods that we use? Such a view, to our mind, 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 and justifiable on evolutionary and 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 respectively multiplicity of the solution prevails.

The study of language, or meaning systems in general, provides numerous illustrations for the phenomenon of the co-existence of different equilibrium conventions. Interpret, for instance, the game of class II as a model of politeness (Brown and Levinson 1987), with the costly signal s standing for the more polite, marked form, and $\bar{s}$ for an unmarked form. Social scientists and linguists have pointed out that there are societies that routinely (and *routinely* can be understood in the sense of "when the prior on the good type is sufficiently high") use the polite form to make some exchange happen—overstatement, while others routinely use the unmarked form—understatement. Such conventions are reflected in the two stable equilibrium outcomes P2 and P3. A similar phenomenon can be observed for signals of dress: if it is commonly known that within a certain group the probability that someone is of a certain social standing or identification is sufficiently high ($p > 1/2$), then both dressing up (P2) and dressing down (P3) can be the ruling convention. Different codes of dress might be in place, for example, in different professions, different companies, or different campuses of quite similar social composition.

At the same time, the co-existence of the two equilibrium outcomes P2 and P3, both stable under evolutionary and rationality-based criteria, can become to basis of a form of discrimination, namely when these two signaling conventions are in place for two different subgroups defined by some observable trait *that is not a matter of choice* (for example, the skin tone or gender a person grows up with) and that *does not affect the prior probability of the unobservable trait in question* (productivity, for example), which, however—because it is observable—makes it possible to condition the action of player 2 on that observable trait. Suppose, for example, that the prior of the high productivity type in both the female and the male population is above $1/2$. Because these two populations can be distinguished based on the observable variable "gender," in one population the decision to hire might be bound to the expression of some costly signal (P2) while in the other that might not be the case (P3). Spence (1973, chapter 6), in fact, already points to this sort of phenomenon.

6.3.4 Social tragedies, evolutionary traps, and co-adaptation

The existence of the equilibrium outcome P2 certainly carries in it a social tragedy: in P2 everybody has to express the costly signal—because expectations are such that if the costly signal were not expressed, player 2 would not to accept—but that signal carries no information, for the

very reason that everybody expresses it. Such situations are not only relevant in a social context, they can also represent an ecological or evolutionary trap. This kind of equilibrium outcome can explain, for instance, why certain handicaps that transmit no information at all (because the entire population expresses them) might persist. And this, in turn, might have some explanatory potential in a longer-run evolutionary perspective for phenomena of co-adaptation, where a handicap that is or has become without function in information transmission still survives in the population and then is recruited for some other function in a different game later on. Human language, one can speculate, might have evolved in this way.

This touches on a crucial question when it comes to applications: to analyze some trait or behavior as a costly signal is not identical with the claim that this functionality as a costly signal is why that trait or behavior has originally evolved. Language, again, is a good example: Languages vary naturally. Different variants (languages, dialects or accents) evolve due to neutral drift, migration, language contact, etc. But once such variants do exist, they can become functional beyond the transmission of conventionally encoded meaning—"in some other social game," so to say. For instance, how quickly an individual learns a new variant or competently navigates between different such variants (the ability of code switching) might become a signal for some social quality such as one's social alertness, willingness to adapt, etc. Or take food sharing, gift-giving, or ritualized forms of hunting: Such practices might have started for a multitude of reasons, and they might serve a multitude of functions (reciprocal altruism, for instance). But once they are a social practice, they might also become functional in transmitting information about the abilities of the individuals involved.

References

[1] Akerlof, G. A. 1970. The market for "lemons": quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84 (3): 488–500.

[2] Archetti, M. 2000. The origin of autumn colours by coevolution. *Journal of Theoretical Biology* 205: 652–630.

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

[4] Berger, U. 2005. Fictitious play in $2 \times n$ games. *Journal of Economic Theory* 120 (2): 139–154.

[5] Bergstrom, C. T., Lachmann M. 1997. Signalling among relatives I. Is costly signalling *too* costly? *Philosophical Transactions of the Royal Society London B*, 352: 609–617.

[6] Bergstrom, C. T., Lachmann M. 2001. Alarm calls as costly signals of anti-predator vigilance: the watchful babbler game. *Animal Behavior* 61: 535–543.

[7] Bliege Bird, R., Smith E. A. 2005. Signaling theory, strategic interaction and symbolic capital. *Current Anthropology* 46 (2): 221–248.

[8] Bliege Bird, R., Smith E. A., Bird, D. W. 2001. The hunting handicap: costly signaling in human foraging strategies. *Behavioral Ecology and Sociobiology* 50: 9–19.

[9] Bourdieu, P. 1982. *Ce que parler veut dire: l'économie des échanges linguistiques*, Paris: Fayard.

[10] Bourdieu, P. 1991 *Language and Symbolic Power*, ed. by J. B. Thompson, transl. by G. Raymond and M. Adamson. Cambridge, MA: Harvard University Press.

[11] Brown, B., Levinson C.S. 1987. *Politeness: Some Universals in Language Usage*. Cambridge/New York: Cambridge University Press.

[12] Caro, T. M. 1986a. The functions of stotting in Thomson's gazelles: a review of the hypotheses. *Animal Behavior* 34: 649–662.

[13] Caro, T. M. 1986b. The functions of stotting in Thomson's gazelles: some tests of the predictions. *Animal Behavior* 34: 663–684.

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

[15] Cressman, R. 2003. *Evolutionary Dynamics and Extensive Form Games*. Cambridge MA: MIT Press.

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

[17] Demichelis, S., Ritzberger K. 2003. From evolutionary to strategic stability. *Journal of Economic Theory* 113 (1): 51–75.

[18] Gaunersdorfer, A. Hofbauer J., Sigmund K. 1991. On the dynamics of asymmetric games, *Theoretical Population Biology*, 39: 345–357.

[19] Eckert, P., Rickford, J. R. (Eds.) 2001. *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press.

[20] FitzGibbon, C. D., Fanshawe, J H. (1988) Stotting in Thomson's gazelles: an honest signal of condition. *Behavioral Ecology and Sociobiology* 23: 69–74.

[21] Ginsburgh, V. A., Pietro-Rodriguez, J. 2011. Returns to foreign languages of native workers in the European Union. *Industrial and Labor Relations Review* 64 (3): 599–618.

[22] Godfray, H. C. J. 1991 Signaling of need by offspring to their parents. *Nature* 352: 328–330.

[23] Govindan, S., Wilson, R., 2009. On forward induction. *Econometrica* 77 (1): 1–28.

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

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

[26] Hörner, J. 2006. Signalling and screening. In: Steven N. Durlauf, S. N, Blume L. E. (Eds.), *The New Palgrave Dictionary of Economics*, Second Edition.

[27] Hofbauer, J. Sigmund, K. 1988. *The Theory of Evolution and Dynamical Systems*, Cambridge UK: Cambridge University Press.

[28] Hofbauer, J., Sigmund, K. 1998. *Evolutionary Games and Population Dynamics*, Cambridge UK: Cambridge University Press.

[29] Hofbauer, J., P. Schuster, Sigmund, K., 1979. A note on evolutionarily stable strategies and game dynamics. *Journal of Theoretical Biology* 81: 609–612.

[30] Huttegger, S. M., Zollman, K. J. S. 2010. Dynamic stability and basins of attraction in the Sir Philip Sidney game. *Proceedings of the Royal Society London B*, 277: 1915–1922.

[31] Huttegger, S. M., Zollman, K. J. S. 2016. The robustness of hybrid equilibria in costly signaling games. *Dynamic Games and Applications*, 6: 347–358.

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

[33] Krebs, J. R., Dawkins, R. 1984. Animal signals: mind-reading and manipulation. In: Krebs J. R. and Davies N. B. (Eds.) *Behavioral Ecology: An Evolutionary Approach*, 2nd Edition. Oxford: Blackwell, pp. 380–402.

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

[35] Kreps, D. M., Wilson, R. 1982. Sequential equilibria. *Econometrica*, 50 (4): 863–894.

[36] Kuhn, H. W. 1950. Extensive games. *Proceedings of the National Academy of Sciences*, 36: 570–576.

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

[38] Lachmann, M., Bergstrom, C. T. 1998. Signalling among relatives II. Beyond the Tower of Babel. *Theoretical Population Biology*, 54: 146–160.

[39] Lachmann, M., Bergstrom, C. T., Számado, S. 2001. Cost and conflict in animal signals and human language. *Proceedings of the National Academy of Sciences* 98 (23): 13189–13194.

[40] Maynard Smith, J., 1982. *Evolution and the Theory of Games*, Cambridge, UK: Cambridge University Press.

[41] Maynard Smith, J. 1991. Honest signalling: The Philip Sidney game. *Animal Behavior*, 42: 1034–1035.

[42] Maynard Smith, J., Price, G. R. 1973. The logic of animal conflict. *Nature*, 246: 15–18.

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

[44] Miller, M. H., Rock, K. 1985. Dividend policy under asymmetric information. *The Journal of Finance*, XL (4), 1031–1051.

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

[46] Nash, J. 1951. Noncooperative games. *The Annals of Mathematics*, 54 (2): 286–295.

[47] Pinker, S., Nowak, M. A., Lee, J.J. 2008. The logic of indirect speech. *Proceedings of the National Academy of Sciences* 105, 833–838.

[48] Riley, J. G. 2001. Silver signals: Twenty-five years of screening and signaling. *Journal of Economic Literature*, 39(2): 432–478.

[49] Ritzberger, K. 1994. The theory of normal form games from the differentiable viewpoint. *International Journal of Game Theory* 23: 207–236.

[50] Ritzberger, K. 2002. *Foundations of Non-Cooperative Game Theory*, Oxford University Press.

[51] Selten, R. 1965. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. *Zeitschrift für die gesamte Staatswissenschaft*, 121.

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

[53] Shapley, L. S. 1974. A note on the Lemke-Howson algorithm. *Mathematical Programming Study* 1: 175–189.

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

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

[56] Spence, M. 2002. Signaling in Retrospect and the Informational Structure of Markets. *The American Economic Review*, 92(3): 434–459.

[57] Számadó, S. 2011. The cost of honesty and the fallacy of the handicap principle. *Animal Behavior*, 81: 3–10.

[58] Taylor, P., Jonker, L., 1978. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences* 40, 145–156.

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

[60] Veblen, T. 1899. *The Theory of the Leisure Class: An Economic Study of Institutions*. New York: The Macmillan Company.

[61] Wagner, E. O. 2013. The dynamics of costly signaling. *Games*, 4: 163–181.

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

[63] 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.