The Evolutionary Dynamics of Costly Signaling

Josef Hofbauer[1] and Christina Pawlowitsch[2*]

[1]Faculty of Mathematics, University of Vienna, Oskar–Morgenstern–Platz 1, Vienna, 1090, Austria.
[2*]Laboratoire d'Économie Mathématique et de Microéconomie Appliquée, University Paris–Panthéon–Assas, 4 rue Blaise Desgoffe, Paris, 75006, France.

*Corresponding author(s). E-mail(s): christina.pawlowitsch@u-paris2.fr;
Contributing authors: josef.hofbauer@univie.ac.at;

Abstract

The theory of costly signaling (Spence 1973) is a well-established paradigm in economics and theoretical biology, where it is also known as the Handicap Principle (Zahavi 1975). Nevertheless, while costly-signaling games have been extensively studied in classical game theory (focused on Nash equilibrium and its refinements), evolutionary dynamics in costly-signaling games are relatively unexplored. This paper gives a comprehensive account of evolutionary dynamics in two canonical classes of games with two states of nature, two signals, and two possible reactions to signals: a model with differential signaling costs (similar to Spence's model) and a model with differential benefits from success (similar to Milgrom and Roberts's 1986, respectively Grafen's 1990, model). We first use index theory to give a necessary condition for the dynamic stability of the equilibria in these games. Then, we study the replicator dynamics and the best-response dynamics. Along the way, we relate our findings to classical equilibrium refinements that test for the plausibility of beliefs off the equilibrium path.

Keywords: Costly-signaling games, Handicap Principle, index theory, replicator dynamics, periodic orbits, best-response dynamics, equilibrium refinement, never-a-weak-best-response criterion, 'divinity,' intuitive criterion.

1 Introduction

"The term 'market signaling' is not exactly a part of the well-defined, technical vocabulary of the economist," Michael Spence writes in 1973 as the opening phrase of his now famous article. "As a part of the preamble, therefore," Spence adds, "I feel I owe the reader a word of explanation about the title." Today, fifty years later, caveats of this kind are obsolete: The theory of costly signaling is part of the well-defined, technical vocabulary of the economist—thanks to Spence's contribution.

Costly-signaling theory, or the *Handicap Principle*, as it is known in theoretical biology (Zahavi 1975), provides a rationale for the phenomenon that an observable variable of choice that comes at a cost (educational credentials, advertising) or an observable trait representing a 'handicap' (a prominent tail, elaborate plumage) indicates some unobservable characteristics, such as performance, quality, or reproductive fitness. The theory explains the informational content of such signals in terms of *differential costs* of the variable or trait that functions as a signal. Applications span over a wide range of phenomena studied in the social and natural sciences: education as a signal for productivity in the job market (Spence 1973), dividend payments as a signal for a firm's fundamentals (Miller and Rock 1985), advertising as a signal for product quality (Milgrom and Roberts 1986), 'handicaps' as signals for high-fitness types in mate selection (Zahavi 1975), predator-prey (Caro 1986a, 1986b, Bergstrom and Lachmann 2001), or parasite-host interaction (Archetti 2008), the begging of offspring as a signal for their need (Godfray 1991, Maynard Smith 1991), the practice of inefficient foraging strategies, embodied handicaps (Bliege Bird et al. 2001, Bliege Bird and Smith 2005), or politeness in language (Van Rooy 2003) as signals in social relationships.

Still and all, while costly-signaling games have been extensively studied in classical game theory (focused on Nash equilibrium and its refinements), the analysis of evolutionary dynamics in costly-signaling games is relatively unexplored—leaving not only some of the applications of costly-signaling theory without mathematical foundation (notably in the social sciences, where equilibria are often implicitly understood as patterns of behavior emerging in a population) but possibly also part of the explanatory potential of these models unexploited.

The idea to explain the emergence of equilibria in costly-signaling models by some dynamics in a population of agents is not new. Already Spence (1973) appeals to a dynamic argument as a justification for the signaling equilibria that he considers in the context of market interactions. In his thesis, later published as a book, Spence (1974) embeds a simplified version of his model in a dynamic, discrete-time process operating on a finite state space (analyzed in more detail by Nöldeke and Samuelson 1997). Spence's original modeling framework, however, is *not* game theory. It is in some sense not fully closed as a game-theoretic model and bears the traits of partial market equilibrium analysis, notably because the uninformed party, the employer, is by definition supposed to pay the expected marginal product (instead of endowing that player with a payoff function and considering their choice of action an endogenous strategic variable determined in equilibrium). Furthermore, the informed party, the job candidate, is by definition assumed to have a unique signal as the best response to the reaction of the uninformed party, and signaling equilibria are ex-ante assumed to fully

reveal the type of the informed party, which amounts to an exclusion of mixed-strategy equilibria that are partially revealing and partially pooling.

It was only later, in the context of the equilibrium-refinement literature of the 1980s, that costly-signaling games have been explicitly formulated in the language of game theory. In this line of research, the robustness of sequential Bayesian Nash equilibria is tested by restricting beliefs *off the equilibrium path*, that is, in the hypothetical case that a signal was observed that is actually never used in the equilibrium under study. Cho and Kreps (1987), for instance, show that in a game-theoretic reformulation of Spence's model, their *intuitive criterion* discards the no-signaling equilibrium outcome and selects the fully revealing equilibrium. The focus on fully revealing equilibria and the disregard of partially revealing equilibria persist in this literature (for a review, see Kreps and Sobel 1994).

The dynamic stability of equilibria in costly-signaling games, naturally, has elicited more attention in theoretical biology. In this literature, too, for a long time, researchers have focused on fully revealing equilibria—'honest' signaling equilibria as is also said—with parameters of the models chosen in such a way as to ensure their existence (see, notably, Grafen 1990 and Maynard Smith 1991). The criterion used to test the stability of equilibria in this line of research first has been that of an *evolutionarily stable strategy* (ESS), after Maynard Smith and Price (1973). The ESS criterion, which relyies on payoff comparisons between resident and mutant strategies, gives a first broad result: Fully revealing equilibria are strict Nash equilibria, and these trivially satisfy the ESS criterion. Later researchers have been attentive to the fact that the conditions guaranteeing the existence of fully revealing equilibria are for many applications overly restrictive. They have pointed out that under fairly plausible parameter constellations, partially revealing equilibria in which the high type expresses the costly signal for sure while the low type expresses it in some frequency—*hybrid equilibria* as is also said—might exist (see, for instance, Bergstrom and Lachmann 1997).

Building on these observations, researchers in theoretical biology have turned to the study of specific evolutionary dynamics in costly-signaling games in which hybrid equilibria appear. Pioneering work has been done by Huttegger and Zollman (2010), Wagner (2013), and Zollman, Bergstrom, and Huttegger (2013). In this line of research, authors have concentrated on showing that hybrid equilibria can have some form of local stability under standard evolutionary dynamics. Zollman, Bergstrom, and Huttegger (2013), for instance, show that in a discrete variant of Spence's, respectively Grafen's, model with two states of nature, two signals, and two reactions to signals, under the replicator dynamics, the hybrid equilibrium is surrounded by closed orbits in its supporting two-dimensional face, which, in turn, attracts an open set of nearby states. Their analysis, however, is restricted to certain parameter constellations (notably to the case that the frequency on the high type is *below* a certain value) and leaves critical aspects in the development of results unexplored.

The purpose of this article is to complement these results: first, by extending the range of cases covered, notably by covering all possible cases for the prior probability of types; second, by underpinning the dynamic analysis with a more detailed derivation of crucial steps involved and by studying global convergence; thrid, by investigating also the best-response dynamics. Furthermore, we relate the dynamic stability analysis

of equilibria to two other robustness concepts: index theory and 'classical' refinements of Bayesian Nash sequential equilibrium that rely on testing the plausibility of beliefs *off the equilibrium path*.

We conduct our study in the two classes of 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 to signals ('accept' and 'do not accept') studied also by Zollman, Bergstrom, and Huttegger (2013):

Class I: a game in which the production of the costly signal is of *different costs for the two types*—a discrete variant of Spence's (1973) model; and

Class I: a game in which the production of the costly signal is of the same cost for the two types, but the two types have *different benefits if the signal has the desired effect*, which can be considered a simplified version of Milgrom and Robert's (1986) model of advertising and, to some extent, Grafen's (1990) formalization of the Handicap Principle.

To be more precise: We study Class I first and then show that Class II can be derived from Class I under appropriate parameter substitutions.

Section 2 serves to introduce the model and analyze the equilibrium structure of the games: We give an account of the equilibria in Class I—both in terms of the Nash equilibria in the normal form as well as the sequential Bayesian Nash equilibria in the extensive form—by making a case distinction along two lines:

(1) whether the cost of the signal for the low type is (i) below, (ii) equal to, or (iii) strictly higher than the benefit from being accepted; and
(2) whether the probability of the high type p is below (the case considered by Zollman, Bergstrom, and Huttegger), above, or equal to the probability at which player 2 is indifferent between accepting or not.

All in all, this leads to nine subclasses. The distinction made under (1) amounts to splitting up Spence's model, which has a continuous signaling space, into three paradigmatic cases with different equilibrium patterns; the distinction made under (2) exhausts all possible equilibrium structures for any of the three cases defined under (1). Such a detailed case distinction allows us to expose (1) under which conditions regarding signaling costs fully revealing equilibria exist and (2) how the 'meaning' of a signal changes as a function of the prior probability distribution over types.

In Section 3, as a first step into the dynamic analysis, we make use of *index theory* (Shapley 1974, Hofbauer and Sigmund 1988, 1998, Ritzberger 1994, 2002, Demichelis and Ritzberger 2003) to get a necessary condition, namely having an index of $+1$, for the asymptotic stability of the respective equilibrium component under a wide range of evolutionary dynamics in these games.

Then, we study in detail the replicator dynamics and the best-response dynamics. Building on Gaunersdorfer, Hofbauer, and Sigmund (1991), respectively Cressmann (2003), we show that for the general class of signaling games with two types, two signals, and two actions, the replicator dynamics in the two-player (two-population) normal-form game, which gives rise to a six-dimensional system, is foliated into a two-parameter family of four-dimensional invariant manifolds and that on the central

invariant manifold—sometimes referred to as the *Wright manifold*—it coincides with the dynamics in the four-player (four population) game defined by the behavior strategies in the extensive form (Proposition 1). Then, for each of the nine subclasses, we determine the rest points of the replicator dynamics, study the qualitative behavior of the dynamics near them, and investigate convergence on the central invariant manifold (Propositions 2–12). The most general result emerging from this investigation is this: For each of the nine subclasses, all interior orbits converge to some Nash-equilibrium component or the union of the two-dimensional faces containing them. For the conspicuous case that the cost of the signal for the low type is below the benefit from being accepted and the prior probability of the high type is below the critical value at which player 2 is indifferent between accepting or not (the case also centrally studied by Zollman, Bergstrom, and Huttegger), we recover the periodic orbits around the partially revealing, hybrid equilibrium (which has index +1) with its supporting face attracting an open set of nearby states, making the equilibrium (locally) stable but not asymptotically stable.

For the best-response dynamics, we show a projection result (Proposition 13) dual to the invariant-foliation result concerning the replicator dynamics, which allows us to reduce the two-population best-response dynamics to a system in lower dimensions, for which we then study convergence and local stability (Propositions 14–16). Our results show in particular that for the conspicuous case that the cost of the signal for the low type is below the benefit from being accepted and the prior probability of the high type below the critical value, under the best-response dynamics, the partially revealing, hybrid equilibrium is not only stable but asymptotically stable. More generally, components with index +1 that are stable but not asymptotically stable under the replicator dynamics are asymptotically stable under the best-response dynamics. All in all, in relation to the index, our study of the two evolutionary processes shows the following: Equilibrium components with index +1, which notably include fully revealing, 'honest,' and partially revealing, hybrid equilibria (structurally the same equilibrium component under different parameter constellations), whenever they exist, are at least stable under the replicator dynamics and asymptotically stable under the best-response dynamics, while all other equilibrium components, including no-signaling–no-acceptance equilibrium outcomes, are unstable under both dynamics.

In Section 4, we show how our results for Class I (differential costs of producing the signal) translate to Class II (differential benefits from being accepted) by a simple parameter substitution. In Section 5, we relate our findings to classical equilibrium-refinement methods, focusing on three prominent criteria: the *never-a-weak-best-response criterion* (Kohlberg and Mertens 1986), *'divinity'* (Banks and Sobel 1987), and the *intuitive criterion* (Cho and Kreps 1987). In Section 6, we summarize and comment on our results.

2 The model

Costly-signaling theory starts from a problem of asymmetric information. 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 shares or a

product, the male), but only if the other player is by nature of a certain type, namely, of high productivity, high quality, high performance, high fitness, etc. The type of that other player—the state of nature—is not directly observable. Therefore, the player who has to make the choice of whether to hire, buy, mate, etc., cannot condition her choice on the other player's type. Of course, whether the player under consideration should accept depends on the gains that she has from accepting or not, as a function of the other player's type, and the probability that she attributes to the other player's types. So far, then, this is simply a problem of choice under uncertainty or 'game against nature.'

One immediately realizes what the social dilemma emanating from such a game against nature might be: The informed party might effectively be of the high type, but if the probability attributed to that type (the frequency of that type in the population) is too low, the right choice of the uninformed party might be not to accept. In other words, the social exchange in question might not happen due to an informational problem in society.

In such a situation, the player whose type is uncertain naturally has an interest in making the other player think that he is of the 'high' type and will try to communicate that. He will, in that aim, try to send a *signal* to the other player. However, if that signal is of no cost, the possibility of sending such a signal will *not* enable the involved parties to escape the unfortunate situation of no exchange. To see why, assume that indeed only the high type uses this signal and that the other player at observing the signal accepts and in the absence of the signal does not accept. If that were so, then the low type would also be better off using the signal, and therefore this way of interacting *cannot* constitute an equilibrium. The argument is intuitive: If talk is cheap, the player whose type is uncertain will always say "I am of the high type," "I am truly motivated," "I truly want this job," "This really is a high-quality product," etc. As Spence (1973, p. 356) remarks: "If the incentives for veracity in reporting anything by 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 *costly signals*.

2.1 Class I: Differential costs of producing the signal

This section presents a parametrized family of games, which can be seen as discretized versions of Spence's (1973) model.

The extensive form of the game is shown in the top panel of Figure 1: There are two players, 1 and 2, and two possible states of nature, 'high' and 'low,' referring to the types of player 1. Player 1 (the job candidate, the firm offering its shares or a product, the male) knows the state of nature, namely if he is of the high or low (productivity, quality, or fitness) type, but not the second player (the employer, the potential buyer, the female), who however has to take an action that affects the payoff of both players, namely whether to accept (a) or not to accept ($\bar{a}$) a certain productive exchange with player 1 (hire, buy, mate). Before player 2 takes her action, though, player 1, no matter what his type, has the possibility to send a *costly signal* s, that is, to express a certain variable of choice or trait that can be observed by the second player and that comes with a cost for the first player. In the game tree in Figure 1, the

uncertainty that player 2 faces about player 1's type is represented by a random move of nature at the root of the tree, which nature takes with probability p for the high type and $1-p$ for the low type. When player 2 comes to move, after having observed the costly signal s, or its absence $\bar{s}$, she still does not know the realization of nature's random move (indicated by putting the two respective nodes that player 2 cannot distinguish in the same *information set*, the ovals in Figure 1), but she can condition her choice on the observed signal. Most solution concepts in classical game theory build on the assumption that the probabilities of player 1's types are *common knowledge*. In an evolutionary interpretation, where each player's position is interpreted as a population of players, the two types of player 1 represent subpopulations of the player-1 population, with p and $1-p$ their frequencies. These two interpretations, of course, do not exclude each other but can be seen as complementary.

In the game in Figure 1, which we refer to as Class I, following Spence's original idea, it is assumed that it is the very production of the signal s that is of different costs for the two types of player 1. More specifically, in our game, the payoffs of player 1's types can be understood as the sum of two components:

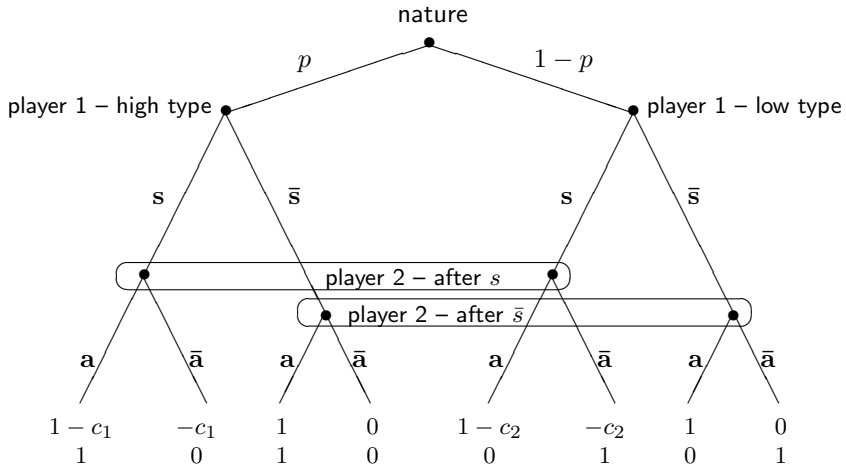
(1) a *background payoff*, which is identical for the two types, namely 1 if the second player takes action a, and 0 if the second player takes $\bar{a}$ (translating the assumption that the first player always wants to be accepted, no matter what his type), and,
(2) the *cost of the signal* s, which is deducted from the background payoff and which is a function of the type: c_1 for the high type, and c_2 for the low type, with $0 \leq c_1 < c_2$.

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

Pure strategies for player 1:	Pure strategies for player 2:
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, of course, 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, mixed strategies of the normal-form game can be interpreted as resulting from *behavior strategies*, that is, plans of action giving 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 player 1, for instance, would be: "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 is induced by a mixed strategy of $s\bar{s}$ and $\bar{s}\bar{s}$, with a probability of



	aa	**aā**	**āa**	**āā**
ss	$1-pc_1-(1-p)c_2,\ p$	$1-pc_1-(1-p)c_2,\ p$	$-pc_1-(1-p)c_2,\ 1-p$	$-pc_1-(1-p)c_2,\ 1-p$
ss̄	$1-pc_1,\ p$	$p(1-c_1),\ 1$	$-pc_1+(1-p),\ 0$	$-pc_1,\ 1-p$
s̄s	$1-(1-p)c_2,\ p$	$(1-p)(1-c_2),\ 0$	$p-(1-p)c_2,\ 1$	$-(1-p)c_2,\ 1-p$
s̄s̄	$1,\ p$	$0,\ 1-p$	$1,\ p$	$0,\ 1-p$

Fig. 1 Class I. At the top, the game in extensive form; at the bottom, the game in normal form resulting from that extensive-form game.

60% on the first and 40% on the second. A behavior strategy for player 2, for instance, would be: "If you observe the costly signal s, take a for sure; if you do not observe it, take a with a probability of 50% (and do not take it with the complementary 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 on behavior strategies—are, at least as to what concerns the existence of Nash equilibria, equivalent (Kuhn 1950, 1953). We denote behavior strategies as follows:

$$\begin{array}{ll} \underline{\text{Behavior strategies for player 1:}} & \underline{\text{Behavior strategies for player 2:}} \\ (x_h, x_\ell)\text{: } x_h = \text{prob}(s \mid \text{high}), & (y, y')\text{: } y = \text{prob}(a \mid s), \\ \qquad\quad x_\ell = \text{prob}(s \mid \text{low}) & \qquad\quad y' = \text{prob}(a \mid \bar{s}) \end{array}$$

A profile of behavior strategies, then, can be written in the form

$$(x_h, x_\ell, y, y').$$

This allows us to represent profiles of behavior strategies in the hypercube $[0,1]^4$, as we use it in Figures 2, 4, 6, and 8.

2.2 Nash equilibria in the normal-form game

Under the assumption that players evaluate payoffs as expected payoffs given the probabilities of the states of nature, 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 of the states of nature.

The Nash equilibria of this game, obviously, depend on the specific values of the cost parameters, c_1 and c_2, and the prior p. In the following, we first isolate three paradigmatic classes of differential signaling costs, namely, whether the cost of the signal for the *low* type c_2 is (i) smaller, (ii) equal to, or (iii) larger than 1. Then, within each of these classes, we make a case distinction according to three cases of the prior probability of the high type p, namely, $p < 1/2$, $p > 1/2$, and $p = 1/2$, which, for each of the classes of differential signaling costs, exhaust all possible equilibrium structures.

Class I.i: Costs of the signal for both types strictly below the benefit from being accepted: $\mathbf{0 \leq c_1 < c_2 < 1}$

- If $0 < p < \frac{1}{2}$, there is:
  - E1, an equilibrium in which player 1 uses a mixed strategy with a probability of $\frac{p}{1-p}$ on ss and the complementary probability on $s\bar{s}$, and player 2 uses a mixed strategy with a probability of c_2 on $a\bar{a}$ and the complementary probability on $\bar{a}\bar{a}$, as well as
  - P1, an equilibrium component in which player 1 takes $\bar{s}\bar{s}$, and player 2 uses a mixed strategy with some probability in $[0, c_1]$ on $a\bar{a}$ and the complementary probability on $\bar{a}\bar{a}$.
- If $\frac{1}{2} < p < 1$, there is:
  - E2, an equilibrium in which player 1 uses a mixed strategy with a probability of $1 - \frac{1-p}{p}$ on $s\bar{s}$ and the complementary probability on $\bar{s}\bar{s}$, and player 2 uses a mixed strategy with a probability of $1 - c_1$ on aa and the complementary probability on $a\bar{a}$,
  - P2, an equilibrium component in which player 1 takes ss and player 2 uses a mixed strategy with some probability in $[0, 1 - c_2]$ on aa and the complementary probability on $a\bar{a}$, and
  - P3, an equilibrium component in which player 1 takes $\bar{s}\bar{s}$ and player 2 any mixed strategy 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, a component spanned by the four vertices $ss \times \mathbf{y}$ with $\mathbf{y} = (0, 1, 0, 0), (1 - c_2, c_2, 0, 0), (0, c_2, 0, 1 - c_2), (0, 1 + c_2, 1 - c_2, 0)/2)$), and
  - 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, 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).

Class I.ii: Cost of the signal for the low type equal to the benefit from being accepted: $0 \leq c_1 < c_2 = 1$

Nash equilibria in the normal form are as in class I.i, only with the following substitutions:

- For $0 < p < \frac{1}{2}$: E1 is replaced by E*-E1, an equilibrium component in which player 1 mixes between ss and $s\bar{s}$ with some probability in $[0, p/(1-p)]$ on ss and player 2 takes $a\bar{a}$.
- For $\frac{1}{2} \leq p < 1$: P2 and E1'-P2 are replaced by E*-E1'-P2, an equilibrium component in which player 1 takes any mix between ss and $s\bar{s}$ and player 2 takes $a\bar{a}$.

Class I.iii: Cost of the signal for the low type higher than the benefit from being accepted: $0 \leq c_1 < 1 < c_2$

Nash equilibria in the normal form are as in class I.i, only that E1, P2, and E1'-P2 are replaced by E*, a *fully revealing equilibrium* in which player 1 takes $s\bar{s}$ and player 2 takes $a\bar{a}$.

These Nash equilibria can easily be verified by use of the game matrix.

2.3 Sequential Bayesian Nash equilibria in the extensive form

For many applications, reasoning about the game in extensive form, in terms of behavior strategies, is the more intuitive approach. This is the theoretical framework in which signaling games are usually discussed in classical game theory, and it also underlies our study of the dynamics on the central invariant manifold. We, therefore, include a discussion of the sequential Bayesian Nash equilibria in the extensive form of the game.

For a game in extensive form, a *sequential Bayesian Nash equilibrium* (Kreps and Wilson 1982) is a profile of behavior strategies together with a *vector of beliefs* (a probability distribution over the states of nature for each information set) such that:

(1) players' choices of actions at information sets where they potentially come to move are a best response to the other players' actions and their beliefs over the states of nature from that information set onward, and
(2) the beliefs assigned to information sets are:

(2.1) compatible with *Bayes' law along the path being played* (given the prior probability distribution p over the states of nature and players' strategies), and
(2.2) *'consistent' off the path being played*, in the sense that they can be deduced from Bayes' law after a small perturbation of the behavior strategies.

For signaling games as we consider them here, the condition that off the equilibrium path beliefs be consistent (2.2) is always fulfilled. This is easy to verify: Let $(p, 1-p)$

be the initial prior for (high, low). Suppose that in equilibrium a specific signal is never sent and let $(p^*, 1-p^*)$ be player 2's belief off the equilibrium path when she was to receive that signal. Suppose that player 1 perturbs his behavior strategies as follows: the high type sends the signal that in the original equilibrium 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 $(p^*, 1-p^*)$. The conditions of sequential Bayesian Nash equilibrium therefore reduce to (1) and (2.1) above.

For the game in Figure 1, each of the Nash equilibria in the normal-form game has indeed a translation into behavior strategies that constitutes a sequential Bayesian Nash equilibrium. This is shown in the following for each of the nine cases.

Class I.i: $0 \leq c_1 < c_2 < 1$

- For the case $0 < p < \frac{1}{2}$:
  - E1 translates to $(1, \frac{p}{1-p}, c_2, 0)$: the high type uses s for sure, while the low type uses it with probability $x_\ell = \frac{p}{1-p}$; player 2, in case that s is observed, takes a with probability $y = c_2$, and in case that it is not observed, does not take a. It is straightforward to verify that this profile of behavior strategies constitutes a sequential Bayesian Nash equilibrium: Given that the high type uses s, the probability with which the low type uses s is precisely such that at the observation of s, player 2's Bayesian updated belief is $1/2$:

$$p(h \mid s) = \frac{p}{p + (1-p) \cdot x_\ell} = \frac{1}{2} \quad \Leftrightarrow \quad x_\ell = \frac{p}{1-p}.$$

    At this belief, player 2 is indifferent between a and $\bar{a}$ and therefore ready to mix between the two. If s is not observed, player 2's updated probability of the high type will be 0, and to this belief, there is a unique best response: not to accept. These choices of player 2 are precisely such as to make player 1's *low* type indifferent between s and $\bar{s}$, which is needed to make him willing to use a mix between these two strategies, and the *high* type strictly better off using s. In E1, the absence of the costly signal fully reveals the low type, while the presence of the costly signal s pushes player 2's belief that player 1 is of high type *up to* $1/2$. We, therefore, characterize E1 as *partially revealing with partial pooling in* s. It is what in the literature in theoretical biology is often referred to as a *hybrid* equilibrium. Figure 2 shows E1 in the hypercube: it is an isolated equilibrium that sits in the 2-dimensional face given by $(1, *, *, 0)$.
  - P1 translates to $(0, 0, y, 0)$, $y \in [0, c_1]$: player 1 never uses s, no matter what his type; player 2, in the counterfactual event that s is observed, takes a with a probability not higher than c_1, and when s is not observed, does not take a. Every point in this equilibrium component maps to the same *equilibrium outcome*, that is, probability distribution over end nodes of the game tree. 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 it, the updated belief is equal to the prior, $p < 1/2$, and therefore player 2 has to choose $\bar{a}$. In the counterfactual event that player 2 observes s, a situation off the equilibrium path,

Bayes' law is not defined and hence imposes no restrictions. To make player 2's choice of taking a with a probability $y \in [0, c_1]$ compatible with sequential Bayesian Nash equilibrium, it therefore suffices to find *some* belief to which this is a best response. And there are many such beliefs: For any belief on the high type strictly smaller than $1/2$, player 2's best response will be to take a with 0 probability. If the belief is equal to $1/2$, then player 2 will be indifferent between a and $\bar{a}$, and hence taking a with some $y \in [0, c_1]$ is a best response. Equilibria of this form, in which all types use the same signal, are often referred to as *pooling equilibria*. Hence the symbol: P1. Figure 2 shows the position of P1 in the space of behavior strategies: it reaches from $(0, 0, 0, 0)$ to $(0, 0, c_1, 0)$, marked by -P1 in the figure.

- For the case $\frac{1}{2} < p < 1$:
  - E2 translates to $(1 - \frac{1-p}{p}, 0, 1, 1 - c_1)$, an equilibrium that is *partially revealing with partial pooling in* $\bar{s}$: the high type uses s with probability $x_h = 1 - \frac{1-p}{p}$, while the low type never uses it, which is such that player 2 in the absence of s will have an updated belief that will make her indifferent between a and $\bar{a}$:

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

    Player 2, if she observes s, will choose a for sure (which will be the best response to her updated belief, which is equal to 1), and if she does not observe 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 s, while ensuring that not using s is a best response for player 1's *low type*. Here the costly signal s fully reveals the *high type*, while the absence of the costly signal ($\bar{s}$) brings player 2's belief *down to* $1/2$.
  - P2 translates to $(1, 1, 1, y')$, $y' \in [0, 1 - c_2]$: both types of player 1 use s ('pooling' in s); player 2, when s is observed, will have the same belief as the prior, $p > 1/2$, and will therefore take a, and in the absence of s, 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* s, no matter what his type ('pooling' in $\bar{s}$); player 2, in the absence of s, will have the same belief as the prior, $p > 1/2$, and hence will choose a, and in the counterfactual event that s is observed can have any belief and best respond to it.
    Figure 4 shows E2, P2, and P3 in the hypercube.
- For the knife-edge case $p = \frac{1}{2}$:
  - E1'-P2 translates to $(1, 1, y, y')$, $y \in [c_2, 1]$, $y' \in [0, y - c_2]$: 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' $= (1, 1, c_2, 0)$. That is, a continuum of equilibrium outcomes, in which player 1 always uses s, no matter what his type ('pooling' in s), and player 2, when she observes s, will have the same belief as the prior $1/2$, at which she is 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 to deviate from s. In particular, when $y = c_2$, then $y' = 0$ (similarly as in E1); and when $y = 1$, then $y' \in [0, 1 - c_2]$ (as in P2).

– P1-E2'-P3 translates to $(0, 0, y, y')$, $y \in [0, y' + c_1]$, if $y' + c_1 \leq 1$, $y \in [0, 1]$ if $y' + c_1 > 1$; $y' \in [0, 1]$: 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 $(0, 0, 0, 1)$. That is, a continuum of equilibrium outcomes, in which player 1 never uses s, no matter what his type ('pooling' in $\bar{s}$), and player 2, in the absence of s, will have the same belief as the prior $1/2$ 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 $y' + c_1 > 1$. In particular, $y' = 0$ 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).
Figure 6 shows E1'-P2 and P1-E2'-P3 in the hypercube.

Class I.ii: $0 \leq c_1 < c_2 = 1$

- For $0 < p < \frac{1}{2}$, E*-E1 translates to $(1, x_\ell, 1, 0)$, $x_\ell \in [0, \frac{p}{1-p}]$, a continuum of equilibrium outcomes reaching from a fully revealing equilibrium E*, in which the high type always and the low type never uses s ($x_\ell = 0$), to an equilibrium that is partially revealing/partially pooling in s like E1 ($x_\ell = p/(1-p)$)). In any equilibrium belonging to this continuum, player 1's high type uses s and player 1's low type uses it with some probability x_ℓ, sufficiently low (possibly 0), such that if player 2 observes s, her updated belief will guarantee that choosing a is a best response, which will be the case if:

$$p(h \mid s) = \frac{p}{p + (1 - p) \cdot x_\ell} \geq \frac{1}{2} \quad \Leftrightarrow \quad 0 \leq x_\ell \leq \frac{p}{1 - p}.$$

  The absence of the costly signal ($\bar{s}$) fully reveals 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 s and $\bar{s}$.
- For $\frac{1}{2} \leq p < 1$, E*-E1'-P2 translates to $(1, x_\ell, 1, 0)$, $x_\ell \in [0, 1]$, a continuum of equilibrium outcomes reaching from the *fully revealing equilibrium* E* ($x_\ell = 0$), over partially revealing/partially pooling equilibria similar to E1, to an equilibrium in the style of P2, in which both types use s ($x_\ell = 1$). In any equilibrium belonging to this continuum, after s, player 2's 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

Table 1 Equilibrium structure for class I.i: $0 \leq c_1 < c_2 < 1$

Prior p	**Equilibrium component**	**Index**	**Rep. dyn.**	**BR dyn.**	**NWBR, 'divinity'**	**Intuitive**	**Payoffs** h ℓ 2
$< \frac{1}{2}$	E1: *part. revealing/ part. pooling in s:* $(1, \frac{p}{1-p}, c_2, 0)$	$+1$	stable	asympt. stable	yes	yes	$c_2 - c_1$ 0 $1-p$
	P1: *pooling in* $\bar{s}$*:* $(0,0,y,0)$, $y \in [0, c_1]$	0	unstable	unstable	no	yes	0 0 $1-p$
$> \frac{1}{2}$	E2: *part. revealing/ part. pooling in* $\bar{s}$*:* $(1 - \frac{1-p}{p}, 0, 1, 1-c_1)$	-1	unstable	unstable	yes	yes	$1-c_1$ $1-c_1$ p
	P2: *pooling in s:* $(1,1,1,y')$, $y' \in [0, 1-c_2]$	$+1$	stable	asympt. stable	yes	yes	$1-c_1$ $1-c_2$ p
	P3: *pooling in* $\bar{s}$*:* $(0,0,y,1)$, $y \in [0,1]$	$+1$	asympt. stable	asympt. stable	yes	yes	1 1 p
$= \frac{1}{2}$	E1'-P2: *pooling in s:* $(1,1,y,y')$, $y \in [c_2, 1]$, $y' \in [0, y - c_2]$)	$+1$	stable	asympt. stable	yes	yes	$[c_2-c_1, 1-c_1]$ $[0, 1-c_2]$ $\frac{1}{2}$
	P1-E2'-P3: *pooling in* $\bar{s}$*:* $(0,0,y,y')$, $(y,y') \in [0,1]^2$ $y \leq y' + c_1$	0	unstable	unstable	only when $y' \in [1-c_1, 1]$	only when $y' \in [0, 1-c_2] \wedge$ $y' \in [1-c_1, 1]$	$[0,1]$ $[0,1]$ $\frac{1}{2}$

be the unique best response. Any belief that puts a probability of at least 1/2 on player 1's low type after $\bar{s}$ supports this equilibrium.

Class I.iii: $0 \leq c_1 < 1 < c_2$

- E*, which exists under any prior, translates to $(1,0,1,0)$, a *fully revealing* or, as is also said, *honest signaling 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 equal to 1; the absence of s sets the belief to 0.

Tables 1, 2, and 3 give an overview of the equilibrium structure for each subclass, for each of the three cases concerning the prior p. Figure 8 shows the equilibrium components in the hypercube for each of the nine cases.

2.4 Excursion: Qualitative properties of the equilibrium structure and 'meaning'

The distinction of the three cases concerning the cost parameters (subclasses i–iii) and the three cases concerning the prior p ($<$, $>$, $= 1/2$) allows us to see how each parameter acts on qualitative properties of the equilibrium structure of the game.

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

Prior p	**Equilibrium component**	**Index**	**Rep. dyn.**	**BR dyn.**	**NWBR, 'divinity'**	**Intuitive**	**Payoffs** h ℓ 2
$< \frac{1}{2}$	E*-E1: *fully to part. revealing/part. pool. in s:* $(1, x_\ell, 1, 0)$, $x_\ell \in [0, \frac{p}{1-p}]$	$+1$	stable	asympt. stable	yes	yes	$1-c_1$ 0 $[1-p, 1]$
	P1: *pooling in $\bar{s}$:* $(0, 0, y, 0)$, $y \in [0, c_1]$	0	unstable	unstable	no	yes	0 0 $1-p$
$> \frac{1}{2}$	E2: *part. revealing/ part. pooling in $\bar{s}$:* $(1 - \frac{1-p}{p}, 0, 1, 1-c_1)$	-1	unstable	unstable	yes	yes	$1-c_1$ $1-c_1$ p
	E*-E1-P2: *fully revealing to pooling in s:* $(1, x_\ell, 1, 0)$, $x_\ell \in [0, 1]$	$+1$	stable	asympt. stable	yes	yes	$1-c_1$ 0 $[p, 1]$
	P3: *pooling in $\bar{s}$:* $(0, 0, y, 1)$, $y \in [0, 1]$	$+1$	asympt. stable	asympt. stable	yes	yes	1 1 p
$= \frac{1}{2}$	E*-E1-P2: *fully revealing to pooling in s:* $(1, x_\ell, 1, 0)$, $x_\ell \in [0, 1]$	$+1$	stable	asympt. stable	yes	yes	$1-c_1$ 0 $[\frac{1}{2}, 1]$
	P1-E2'-P3: *pooling in $\bar{s}$:* $(0, 0, y, y')$, $(y, y') \in [0, 1]^2$ $y \leq y' + c_1$	0	unstable	unstable	only when $y' \in [1-c_1, 1]$	only when $y' \in [0, 1-c_2] \wedge$ $y' \in [1-c_1, 1]$	$[0, 1]$ $[0, 1]$ $\frac{1}{2}$

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

Prior p	**Equilibrium component**	**Index**	**Rep. dynam.**	**BR dynam.**	**NWBR, 'divinity'**	**Intuitive**	**Payoffs** h ℓ 2
$< \frac{1}{2}$	E*: *fully revealing:* $(1, 0, 1, 0)$	$+1$	asympt. stable	asympt. stable	yes	yes	$1-c_1$ 0 1
	P1: *pooling in $\bar{s}$:* $(0, 0, y, 0)$, $y \in [0, c_1]$	0	unstable	unstable	no	no	0 0 $1-p$
$> \frac{1}{2}$	E2: *part. revealing/ part. pooling in $\bar{s}$:* $(1 - \frac{1-p}{p}, 0, 1, 1-c_1)$	-1	unstable	unstable	yes	yes	$1-c_1$ $1-c_1$ p
	E*: *fully revealing:* $(1, 0, 1, 0)$	$+1$	asympt. stable	asympt. stable	yes	yes	$1-c_1$ 0 1
	P3: *pooling in $\bar{s}$:* $(0, 0, y, 1)$, $y \in [0, 1]$	$+1$	asympt. stable	asympt. stable	yes	yes	1 1 $2:p$
$= \frac{1}{2}$	E*: *fully revealing:* $(1, 0, 1, 0)$	$+1$	asympt. stable	asympt. stable	yes	yes	$1-c_1$ 0 1
	P1-E2'-P3: *pooling in $\bar{s}$:* $(0, 0, y, y')$, $(y, y') \in [0, 1]^2$ $y \leq y' + c_1$	0	unstable	unstable	only when $y' \in [1-c_1, 1]$	only when $y' \in [0, 1-c_2] \wedge$ $y' \in [1-c_1, 1]$	$[0, 1]$ $[0, 1]$ $\frac{1}{2}$

First, the existence of the fully revealing—'honest'—equilibrium E* depends on *the cost of the signal for the low type* but not on the prior:

- E* *exists only if the cost of the signal for the low type is at least as high as the benefit that he gets from being accepted*, that is, $c_2 \geq 1$ (subclasses ii, Table 2; and iii, Table 3). 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 theoretical biology, this observation is sometimes expressed by saying that it is the 'cost of cheating' that sustains honest communication (see, for instance, Számadó 2011). When E* exists, it exists for any prior p (whereas the existence of other equilibria depends on the prior).
- E* and the partially revealing equilibrium E1 represent the same equilibrium component in different games belonging to the same family of games generated by varying the cost of the signal for the low type, c_2. Whenever E* and E1 co-exist in the same game (subclass ii, $c_2 = 1$, Table 2), they belong to the same equilibrium component.
- E* *is never the unique equilibrium.* Whenever it exists, there are also other equilibria, notably *no-signaling* equilibrium components, in which none of player 1's types expresses the costly signal and the second player acts on her prior, which, depending on the prior, can make her not accept (such as in P1) or accept (such as in P2).

The prior p, on the other hand, acts on what could be referred to as the *meaning* of a signal (in equilibria other than E*), if by the meaning of a signal one understands what it *does*: what is its effect on the updated belief of player 2 over player 1's types and what it therefore makes player 2 do.[1] Focusing on the two generic cases for p:

- When the prior on the high type is *below* the critical value $p < 1/2$, the costly signal s, if used in equilibrium, has the function of 'pushing up' the belief of player 2. How far, depends on c_2: when $c_2 < 1$, up to the critical value $p = 1/2$, which makes player 2 indifferent between accepting and not accepting (E1, subclasses i, Table 1); (ii) when $c_2 = 1$, up to the critical value $p = 1/2$ or above, possibly up to 1 (equilibria in E1-E*, subclasses ii, Table 2). And, when $c_2 > 1$, up to 1, making player 2 accept (E*, subclasses iii, Table 3). The absence of the signal, $\bar{s}$, instead, always makes that player 2 will not accept—no matter which of the two co-existing equilibrium components, E1 (respectively E1-E* or E*) or the no-signaling–no-acceptance P1 prevails.
- When the prior is *above* the critical value $p > 1/2$, the costly signal, if used in equilibrium, has the function of keeping the prior above the critical value (P2) or pushing it even higher, possibly up to 1 (equilibria in E*-E1-P2 other than P2, E* and E2) and therefore make player 2 accept. The meaning of the absence of the costly signal, instead, critically depends on which of the co-existing equilibrium outcomes prevails: in E2, it lowers the belief of player 2 to the critical value $p = 1/2$ and hence makes player 2 indifferent between accepting or not; in E* respectively equilibria in E*-E1-P2 other than P2, it makes the belief drop to 0 and hence makes player 2 not accept; in P3 it leaves the prior belief, which is already above 1/2, unchanged and hence makes player 2 accept.

[1] "The meaning of a word is its use in the language," Wittgenstein famously writes in §43 of *Philosophical Investigations*.

In terms of the welfare properties of equilibria, again focusing on the two generic cases of the prior p, we have the following:

- When the prior on the high type is *below* the critical value, $p < 1/2$, no matter whether the cost of the signal for the low type c_2 is below, equal to, or larger than 1 (Tables 1, 2, 3), the equilibrium component in which the costly signal is at least partially informative, namely, E1 respectively E*-E1 or E*, is better, in the sense of Pareto, than the co-existing no-signaling–no-acceptance equilibrium component P1: In E1, 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, different from E1, and in E*, player 2 is also made better off. Hence, the possibility to use a costly signal has the potential to increase social well-being over a situation in which such a costly signal is not available.
- When the prior on the high type is *above* the critical value, $p > 1/2$, then payoff comparisons between equilibria depend on the cost of the signal for the low type: When $c_2 < 1$ (Table 1), the equilibrium component P3, in which none of player 1's types uses the costly signal and player 2 accepts, Pareto *dominates* the two other equilibrium components E2 and P2 (both types of player 1 strictly prefer P3 over P2 and E2, while player 2 is indifferent between all three equilibrium outcomes). Hence, the possibility of using a costly signal can result in a social tragedy, namely when players 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 would not accept—which in the end has the effect that the costly signal does not carry any information. When c_2 is equal to or larger than 1 (Tables 2 and 3), equilibria can no longer be completely ranked according to the Pareto criterion: Player 2 prefers outcomes in the component E*-E1'-P2, different from P2, and E*, over E2 and P3; while both types of player 1 strictly prefer P3 over E2 and E*-E1'-P2 respectively E*. In other words, there is a potential conflict of interest between player 1 and player 2 over the co-existing equilibrium outcomes.

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 instance, male and female or predator and prey. If a player can be of two different *types*, these 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.

For an equilibrium to be a good prediction of the model from an evolutionary point of view, the corresponding state of the system has to be resistant to evolutionary shocks, that is, random drift among strategies already present in the population and newly appearing variation in the form of mutant strategies.

Theorists have approached the question of evolutionary stability on three different levels:

(1) 'static' criteria, such as, most prominently, Maynard Smith and Price's (1973) notion of *evolutionarily stable strategy* (ESS), which relies on payoff comparisons between mutant and resident strategies;
(2) the study of specific evolutionary dynamics defined on the respective game—a research program that has aimed at establishing relations between static ESS criteria and stability properties of the associated fixed point under specific dynamics (Taylor and Jonker 1978, Hofbauer et al. 1979, Hofbauer and Sigmund 1988, 1998); and
(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*.

We first turn to this last approach as it has the additional advantage of providing a complete system of classification. Then, we study in more detail the replicator dynamics and the best-response dynamics for our classes of games.

3.1 The index of equilibria: a necessary condition for evolutionary stability

Already Shapley (1974), in his description of the Lemke-Howson algorithm, associated an index, $+1$ or -1, to each *regular* equilibrium (in a 2-person game, an equilibrium is regular if and only if it is isolated and quasistrict, that is, unused strategies do strictly worse) 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.

Von Stengel (2021) gives a modern exposition of this approach.

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

Ritzberger (1994, 2002) extends 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 with 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 all equilibria in the closure of U are already in C. 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). The set 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 of determining 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:
(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 and asymptotically stable under some evolutionary dynamics, then its index is +1.

In Class I, subclass i ($0 \leq c_1 < c_2 < 1$), we get the following characterization of equilibria in terms of the index:

- 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, again 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 that P1-E2'-P3 has index 0 also by robustness of the index: 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 isolated and quasistrict—hence regular—and both players mix between two strategies. If we discard the unused strategies, the 2×2 restricted game is a coordination game, with two strict equilibria, and E2. Since strict equilibria have index +1, E2 has index −1. As a consequence, 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. As required by the robustness property

of the index, the index of the component P1-E2'-P3 (0) is, on the one hand, the same as that of P1 and, on the other hand, the same as the sum of the indices of E2 and P3. Table 1 summarizes these results. It follows from these results, by Demichelis and Ritzberger's theorem, that P1, E2, and P1-E2'-P3 *cannot* be asymptotically stable for any reasonable dynamics, while E1, P2, and P3, and E1'-P2 are candidates for asymptotic stability, at least for *some* evolutionary dynamics.

Due to the robustness property of the index, with the appropriate substitutions, these results extend to the two other subclasses capturing variations of the c_2 parameter: For Class I, subclass ii ($c_2 = 1$), E1 'turns into' E*-E1, and P2 and E1'-P2 into E*-E1'-P2; for Class I, subclass iii ($c_2 > 1$), E1, P2, and E1'-P2 turn into E* (see Section 2.2). Certainly this has to be so, because varying the cost parameters means to perturb the payoffs—to look at games 'nearby'—and what defines the index is precisely that it is robust under such perturbations. Tables 2 and 3 cover these cases.

Looking at the index of equilibrium components across all three subclasses, i–iii (Tables 1, 2, and 3), note in particular that the fully revealing equilibrium E*, whenever it exists, always sits in a component with index +1; and, curiously, equilibrium components with an index $\neq +1$ are those that *do not* change across the three subclasses.

3.2 Replicator dynamics for the game in normal form

In general, the replicator dynamics for an $n^1 \times n^2$ two-population game is given by the following system of differential equations:

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

where u_i^k is the payoff of player k playing strategy i, and $\bar{u}^k$ the average payoff of player k, and as usually, the dot-notation $\dot{x}_i$ and $\dot{y}_j$ refers to the derivatives with respect to time.

For our game, with the notation $\mathbf{y} = (y(aa), y(a\bar{a}), y(\bar{a}a), y(\bar{a}\bar{a}))$,

$$\begin{aligned} y &= y(aa) + y(a\bar{a}), \\ y' &= y(aa) + y(\bar{a}a), \end{aligned} \tag{2}$$

we can write the payoffs for player 1 against a mixed strategy $\mathbf{y}$ of player 2 as follows:

$$\begin{aligned} u_1^1 &= u^1(ss, \mathbf{y}) = y - pc_1 - (1-p)c_2, \\ u_2^1 &= u^1(s\bar{s}, \mathbf{y}) = p(y - c_1) + (1-p)y', \\ u_3^1 &= u^1(\bar{s}s, \mathbf{y}) = (1-p)(y - c_2) + py', \\ u_4^1 &= u^1(\bar{s}\bar{s}, \mathbf{y}) = y'. \end{aligned} \tag{3}$$

Note that:

$$u^1(ss) + u^1(\bar{s}\bar{s}) = u^1(s\bar{s}) + u^1(\bar{s}s). \tag{4}$$

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:

$$\begin{aligned} u_1^2 &= u^2(aa, \mathbf{x}) = p, \\ u_2^2 &= u^2(a\bar{a}, \mathbf{x}) = px_h + (1-p)(1-x_\ell), \\ u_3^2 &= u^2(\bar{a}a, \mathbf{x}) = p(1-x_h) + (1-p)x_\ell, \\ u_4^2 &= u^2(\bar{a}\bar{a}, \mathbf{x}) = 1-p, \end{aligned} \tag{5}$$

and

$$u^2(aa) + u^2(\bar{a}\bar{a}) = u^2(a\bar{a}) + u^2(\bar{a}a). \tag{6}$$

We point out that (4) and (6) hold for any normal-form game derived from a game tree as given in Figure 1 (for any specification of payoffs at the end nodes of the tree). These special features allow us to reduce the replicator dynamics to a smaller dimension.

Lemma 1. *Let*

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

be the replicator equations for a 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) (similarly as in Gaunersdorfer, Hofbauer, and Sigmund 1991, and Cressman 2003). □

Proposition 1 (Foliation of the replicator dynamics)**.** *For the normal-form game in Figure 1: $\frac{x(ss)x(\bar{s}\bar{s})}{x(s\bar{s})x(\bar{s}s)}$ 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 (1), which is to say that 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*

$$x(ss)x(\bar{s}\bar{s}) = x(s\bar{s})x(\bar{s}s), \quad y(aa)y(\bar{a}\bar{a}) = y(a\bar{a})y(\bar{a}a), \tag{10}$$

which is sometimes called the Wright manifold *(see, for example, Cressman 2003), the replicator dynamics simplifies to:*

$$\begin{aligned} \dot{x}_h &= x_h(1-x_h)(y-c_1-y')p, \\ \dot{x}_\ell &= x_\ell(1-x_\ell)[y-c_2-y'](1-p), \\ \dot{y} &= y(1-y)[px_h-(1-p)x_\ell], \\ \dot{y}' &= y'(1-y')[p(1-x_h)-(1-p)(1-x_\ell)], \end{aligned} \tag{11}$$

with the state space of this system the hypercube $[0,1]^4$.

Proof. By Lemma 1 and equations (4) and (6). Equation (11) results from inserting (3) and (5) into (8). □

3.3 Replicator dynamics for behavior strategies

Instead of looking at the two-population dynamics in the normal-form game with four pure strategies for each population (payoff matrix at the bottom of Figure 1), one can also look at the *replicator dynamics for behavior strategies*: the four-population dynamics as given by the agents in the extensive-form game with two pure strategies for each agent. The second coincides with the first on the central invariant manifold.

Proposition 2 (Replicator dynamics on Wright manifold—in behavior strategies)**.** *The system of differential equations (11) on the hypercube* $[0,1]^4$ *can be derived directly from the extensive form, as the* replicator dynamics for behavior strategies.

Proof. For this purpose, we interpret $x_h = \text{prob}(s|\text{high})$, $x_\ell = \text{prob}(s|\text{low})$, $y = \text{prob}(a|s)$, 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 $\dot{x} = x(1-x)[u(A)-u(B)]$. When we apply this to the 4-player game defined by the extensive-form game in Figure 1, this leads to (11). The factors p and $1-p$ in the first two equations come from the probabilities of nature's draw. □

In other words: the system of differential equations (11) is the replicator equation for a binary four-person game with linear incentives, with the hypercube $[0,1]^4$ as state space. In the following, we analyze this dynamics for each of the three subclasses concerning the cost parameters and, within each of these, the three relevant cases regarding p.

Class I.i: $0 \le c_1 < c_2 < 1$

$\mathbf{0 < p < \frac{1}{2}}$: All 2^4 corners of the hypercube as well as the Nash equilibrium E1 $= (1, \frac{p}{1-p}, c_2, 0)$, and the edges $(0,0,*,0)$, which includes the Nash-equilibrium component P1, $(0,0,*,1)$, $(1,1,0,*)$, and $(1,1,1,*)$ are rest points of (11).

Proposition 3 (Replicator dynamics near E1)**.** *For the replicator dynamics (11) on the hypercube* $[0,1]^4$*, E1 is Lyapunov stable. More precisely, it is surrounded by closed orbits in its supporting boundary face* $(1,*,*,0)$ *(lower front square in Figure*

2; lower right square in Figure 3), and there is an open neighborhood of E1 in $[0,1]^4$ *from where the orbits converge to the boundary face* $(1,*,*,0)$*, approaching one of the periodic solutions near E1. Every periodic solution in the face* $(1,*,*,0)$ *attracts a three-dimensional manifold of nearby orbits. However, the closed face* $(1,*,*,0)$ *is not stable, due to the orbit on the edge from* $(1,0,0,0)$ *to* $(0,0,0,0)$*, see Figure 2 or Figure 3.*

Proof. In the supporting boundary face of E1, $(1,*,*,0)$, which in Figure 2 corresponds to the lower front square, we have:

$$\begin{aligned} \dot{x}_\ell &= x_\ell(1-x_\ell)[y-c_2](1-p), \\ \dot{y} &= y(1-y)[p-(1-p)x_\ell], \end{aligned} \tag{12}$$

which is the replicator dynamics for a cyclic 2×2 game, with closed orbits around the equilibrium E1 (see Figure 3, lower right square). Linearization at E1 in direction of the missing strategies gives

$$\dot{x}_h = (1-x_h)(c_2-c_1)p, \quad \dot{y}' = y'(2p-1). \tag{13}$$

Hence E1 is a quasistrict Nash equilibrium, and it has a two-dimensional stable manifold, which intersects the open hypercube $(0,1)^4$ in a quarter 'plane' consisting of all interior orbits converging to E1. Moreover, there is an open neighborhood of E1 in $[0,1]^4$ from where the orbits converge to the boundary face $(1,*,*,0)$, approaching one of the periodic solutions near E1. This follows from center manifold theory and the reduction principle (see, e.g., Kuznetsov 2004, chapter 5.1): The two-dimensional boundary face $(1,*,*,0)$ is the center manifold at the equilibrium E1. Hence, by Kuznetsov (2004, Theorem 5.2), the flow near E1 is locally topologically equivalent to the partially linearized flow of (12) together with (13). In particular, E1 is Lyapunov stable. Actually, every periodic solution in the face $(1,*,*,0)$ attracts a three-dimensional manifold of nearby orbits. Indeed, the two external eigenvalues (Floquet or Lyapunov exponents) along such a periodic solution (with period T) are given by

$$\frac{1}{T}\int_0^T (c_1-y(t))p\,dt = (c_1-c_2)p < 0, \quad \text{and} \quad \frac{1}{T}\int_0^T -(1-p)(1-x_\ell(t))\,dt = -(1-2p) < 0$$

(by the averaging property of the replicator dynamics, see Hofbauer and Sigmund 1998, Exercise 10.4.1, or more generally, Theorem 7.6.4), and are equal to the two external eigenvalues at the equilibrium E1 appearing in (13). Hence, by applying the reduction principle (Kuznetsov 2004, Theorem 5.3) to the Poincaré return map at any of the periodic orbits in this two-dimensional face, one obtains an open set in $[0,1]^4$ containing the (relatively open) face $(1,*,*,0)$, where orbits are attracted by one of the periodic orbits in $(1,*,*,0)$. □

Proposition 4 (Replicator dynamics near P1)**.** *For the replicator dynamics* (11) *on the hypercube* $[0,1]^4$*, the equilibrium component P1 is unstable. Nevertheless, it has a basin of attraction with nonempty interior.*

Proof. Near the rest points $(0,0,y,0)$ we have the linearized dynamics:

$$\begin{aligned} \dot{x}_h &= (y-c_1)p\ x_h \\ \dot{x}_\ell &= (y-c_2)(1-p)\ x_\ell \\ \dot{y} &= 0 \\ \dot{y}' &= (2p-1)\ y' < 0. \end{aligned} \tag{14}$$

The rest points $(0,0,y,0)$ with $0 \le y \le c_1 < c_2$ are therefore Nash equilibria. For $0 \le y < c_1$, all three external eigenvalues are negative, hence the corresponding point is a quasistrict Nash equilibrium and attracts a 3-dimensional stable manifold (as a consequence of the stable manifold theorem). The basin of attraction of the whole component P1 has nonempty interior. This is again a consequence of the reduction principle, see, e.g., Kuznetsov (2004, Theorem 5.2), as the line of rest points $(0,0,*,0)$ forms the center manifold. But the study of the behavior of the dynamics near the end point of P1, which we denote by -P1 $= (0,0,c_1,0)$, shows that the component is unstable: -P1 has a 2-dimensional stable manifold and a 2-dimensional center manifold, the latter contained in the 2-dimensional face $(*,0,*,0)$ with dynamics

$$\begin{aligned} \dot{x}_h &= x_h(1-x_h)(y-c_1)p, \\ \dot{y} &= y(1-y)px_h. \end{aligned} \tag{15}$$

This is the replicator dynamics of a nongeneric 2×2 game shown in the top right square of Figure 3. There is one orbit converging to the endpoint -P1$= (0,0,c_1,0)$, and one orbit with -P1 as α–limit which converges to $(1,0,1,0)$, a corner of the face of E1. This shows that the endpoint -P1 is unstable (unlike all other Nash equilibria in the component P1) and hence the component P1 itself is unstable. □

Proposition 5 (Convergence, Class I.i, $0 < p < 1/2$)**.** *All orbits in the interior of the hypercube converge to the set* $\{(x_h, x_\ell, y, y') : (x_h = 1 \text{ or } x_\ell = 0) \text{ and } y' = 0\}$*, which is to say that in the long run, either the high type sends the costly signal s or the low type does not send it, and in the absence of the costly signal s, player 2 never accepts.*

Proof. To prove this, we use two Lyapunov functions. From the first two equations of (11) we see that

$$\frac{\dot{x}_h}{px_h(1-x_h)} - \frac{\dot{x}_\ell}{(1-p)x_\ell(1-x_\ell)} = c_2 - c_1 > 0. \tag{16}$$

Therefore:

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

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 that

$$(1-x_h)x_\ell\to 0, \tag{17}$$

which implies that all interior orbits converge to the union of the two facets $x_h=1$ (in Figure 2, the bottom cube) and $x_\ell=0$ (the inner cube). Similarly, since $p<\frac{1}{2}$, we obtain from the last two equations of (11)

$$[\log y-\log(1-y)+\log y'-\log(1-y')]^{\cdot}=\frac{\dot y}{y(1-y)}+\frac{\dot y'}{y'(1-y')}=2p-1<0, \tag{18}$$

and hence

$$yy'\to 0,$$

which implies that all interior orbits converge to the union of the two facets $y=0$ and $y'=0$. The ω–limit sets must then be contained in the union of the following four two-dimensional faces, shown in Figure 3:

- $(1,*,0,*)$, the lower left square in Figures 3 and 2, on which all interior orbits converge to $(1,0,0,0)$;
- $(1,*,*,0)$, the lower right square in Figure 3 and the lower front square in Figure 2, that is, the face containing E1 and the periodic solutions;
- $(*,0,0,*)$, the top left square in Figure 3, and the inner left square in Figure 2, on which all interior orbits converge to $(0,0,0,0)$; and
- $(*,0,*,0)$, the top right square in Figure 3, and the inner front square in Figure 2, which contains the equilibrium component P1 in an edge.

Actually, we can show that $y'\to 0$. This can be done in two ways. One is by eliminating dominated strategies in the original normal form game with payoff matrix in Figure 1. The other way is to directly show that an interior orbit cannot have ω–limit points with $y=0$ and $y'>0$. The ω–limit set of any orbit is a closed, connected, invariant, and *internally chain transitive* (ICT) subset (i.e., any two points in it can be connected by pseudo-orbits), see Benaïm (1999, Corollary 5.6). Note that the union of all four squares in Figure 3 is an ICT set. And many subsets are ICT. Even the whole state space $[0,1]^4$ is an ICT set. So this concept by itself is not enough to prove the result. Suppose the ω–limit set Ω of some interior orbit contains a point with $y'>0$ (i.e., in the left half of Figure 3). Then it must contain a rest point in the edge $(1,1,0,*)$, and even a continuum $\{(1,1,0,y'):y'\in[0,\bar y]\}$ of such rest points. Linearization at these rest points shows that there are two positive eigenvalues (in direction x_h and x_ℓ) and one negative eigenvalue (in direction y). Therefore the center manifold is only one-dimensional and coincides with the edge of rest points. By the reduction principle, there is an invariant foliation with 3-dimensional leaves transverse to this edge. However, since Ω contains a continuum of such rest points, the orbit must move slowly and close along this continuum, a contradiction. So ω–limits are contained in the union of the two squares on the right of Figure 3. □

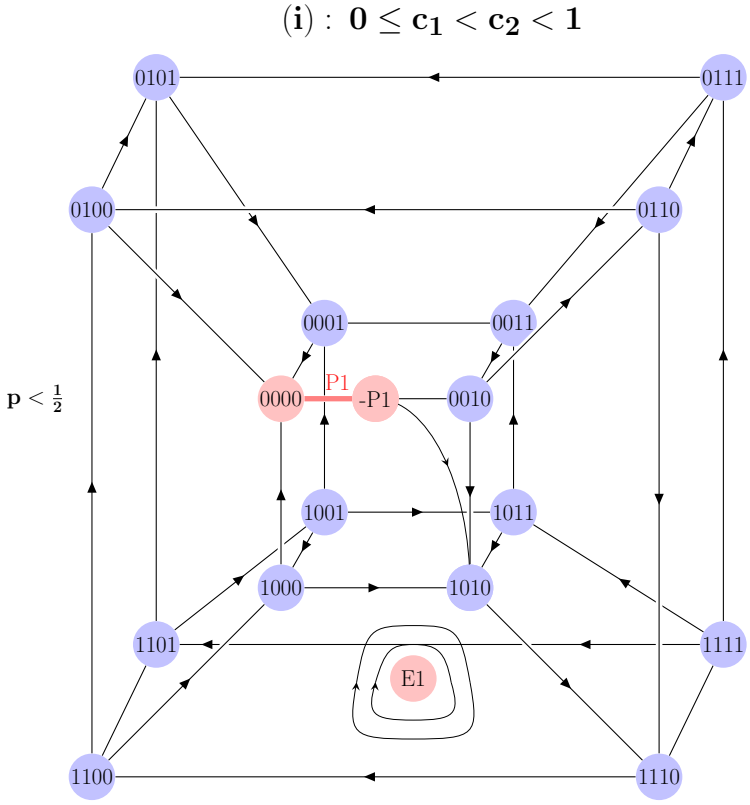


Fig. 2 Nash equilibria and replicator dynamics for Class I, i ($0 \le c_1 < c_2 < 1$), $0 < p < 1/2$: the partially revealing equilibrium E1, which sits in the face $(1, *, *, 0)$ (for a close-up of this face see Figure 3, bottom right square), and the continuum of equilibria P1, which stretches from the vertex $(0, 0, 0, 0)$ to the point $(0, 0, c_1, 0)$, marked as -P1 in Figure 3 (top right square). Arrows on the edges show the direction of the flow of the replicator dynamics (11). Edges without arrows consist of rest points. Also shown: periodic orbits around E1 and the connecting orbit from -P1 to $(1, 0, 1, 0)$.

Remark 1. *We conjecture that (almost) all orbits converge either to the supporting face of E1, in fact to one of the periodic orbits in the supporting face of E1, or to the component P1. However, this does not follow from the above arguments. There are many ICT sets in the union of the two squares (right half of Figure 3), e.g., the heteroclinic cycle* $1100 \to 1000 \to -P1 \to 1010 \to 1110 \to 1100$. *It is not obvious whether such heteroclinic cycles can attract orbits from the interior of the hypercube or not.*

$\mathbf{\frac{1}{2} < p < 1}$: In this case, (11) has the following rest points: all 2^4 corners of the hypercube, the edges $(1, 1, 0, *)$ and $(1, 1, 1, *)$, the latter containing the Nash-equilibrium

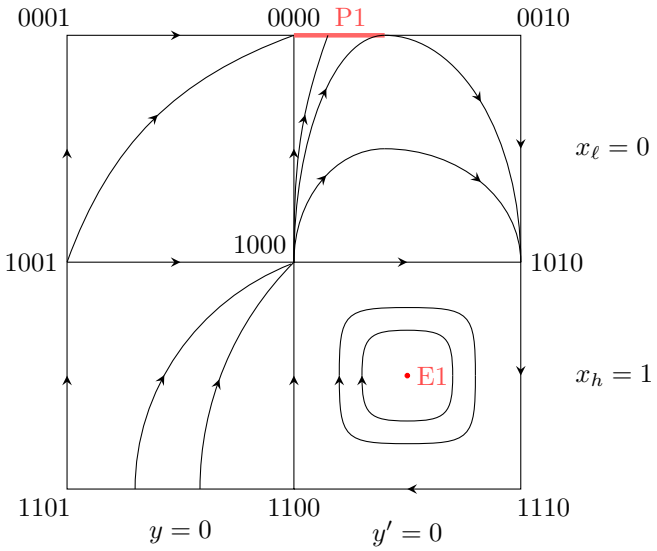


Fig. 3 The four two-dimensional faces that attract all interior orbits of the replicator dynamics (11) for $0 < p < 1/2$. Actually, the two on the right attract all of them.

component P2, as well as the edges $(0, 0, *, 0)$ and $(0, 0, *, 1)$, the latter coinciding with the Nash-equilibrium component P3, and the Nash equilibrium E2 $= (1 - \frac{1-p}{p}, 0, 1, 1 - c_1)$; see Figure 4.

Proposition 6 (Replicator dynamics near E2)**.** *For the replicator dynamics* (11) *on the hypercube* $[0, 1]^4$*, E2 is a saddle point within the face* $(*, 0, 1, *)$ *(lower left square of Figure 5), and hence unstable. It has a 3-dimensional stable and 1-dimensional unstable manifold in* $[0, 1]^4$*.*

Proof. E2 is a quasistrict Nash equilibrium, since there, $\frac{\dot{x}_\ell}{x_\ell} = (c_1 - c_2)p < 0$ and $\frac{(1-y)^{\cdot}}{1-y} = 1 - 2p < 0$. We know already from the analysis of the index that E2 is a saddle point within the face $(*, 0, 1, *)$ (lower left square of Figure 5). It therefore has a 3-dimensional stable and 1-dimensional unstable manifold in $[0, 1]^4$. □

Proposition 7 (Replicator dynamics near P2)**.** *For the replicator dynamics* (11) *on the hypercube* $[0, 1]^4$*, the equilibrium component P2 is stable but not asymptotically stable. Its basin of attraction has nonempty interior.*

Proof. Near the rest points $(1, 1, 1, y')$ we have the linearized dynamics:

$$\begin{aligned} \dot{x}_h &= (1 - x_h)(1 - y' - c_1)p \\ \dot{x}_\ell &= (1 - x_\ell)(1 - y' - c_2)(1 - p) \\ \dot{y} &= (1 - y)(2p - 1) \\ \dot{y}' &= 0. \end{aligned} \tag{19}$$

The rest points $(1,1,1,y')$ with $0 \leq y' \leq 1-c_2$, given that $1-c_2 < 1-c_1$, are therefore Nash equilibria. For $0 \leq y' < 1-c_2$, all three external eigenvalues are negative, hence the corresponding point is a quasistrict Nash equilibrium and, as a consequence of the stable manifold theorem, attracts a 3-dimensional stable manifold. The basin of attraction of the whole component P2 contains an open set from the hypercube, which is again a consequence of the reduction principle, see, e.g., Kuznetsov (2004, Theorem 5.2), as the line of rest points $(1,1,1,*)$ forms the center manifold. Let us now study the behavior near the end point of P2, which we denote by -P2 $= (1,1,1,1-c_2)$. This point has a 2-dimensional stable manifold and a 2-dimensional center manifold, the latter contained in the 2-dimensional face $(1,*,1,*)$ with dynamics

$$\begin{aligned} \dot{x}_\ell &= x_\ell(1-x_\ell)(1-c_2-y')(1-p), \\ \dot{y}' &= y'(1-y')(1-p)(x_\ell - 1). \end{aligned} \tag{20}$$

This is the replicator dynamics of a nongeneric 2×2 game shown in the bottom right square of Figure 5. Hence P2 is stable (in the 2-dimensional face $(1,*,1,*)$ as well as in the hypercube, by the reduction principle), and all interior orbits starting close to P2 converge to one of the Nash equilibria in P2, again by the reduction principle. However, P2 is not asymptotically stable for the replicator dynamics, since the whole edge $(1,1,1,*)$ consists of rest points (see the lower right square in Figure 5). □

Proposition 8 (Replicator dynamics near P3)**.** *For the replicator dynamics* (11) *on the hypercube* $[0,1]^4$*, the equilibrium component P3 is asymptotically stable. Its basin of attraction is an open neighborhood of P3.*

Proof. Analogously to (19), one can show that each equilibrium in P3 is quasistrict. So the edge P3 is a curb set ('closed under rational behavior,' see Ritzberger, 2002, section 5.2), and closed under better replies (Weibull, 1995, section 5.7), and hence a strict NE set. By the reduction principle (or a simple Lyapunov function argument), P3 is asymptotically stable. □

Proposition 9 (Convergence, Class I.i, $1/2 < p < 1$)**.** *Every orbit in the interior of the hypercube converges to a Nash equilibrium. (On the boundary, orbits may also converge to one of the rest points.) The 3-dimensional stable manifold of E2 separates the basins of attraction of the two equilibrium components P2 and P3.*

Proof. We use the same two Lyapunov functions as in the proof of Propositon 5. However, the expression in (18) 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 in Figure 4) and $y' = 1$ (the cube in the back). Recall (17), 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). Therefore, the ω–limit set of any interior orbit is contained in the union of the following four two-dimensional faces, shown in Figure 5:

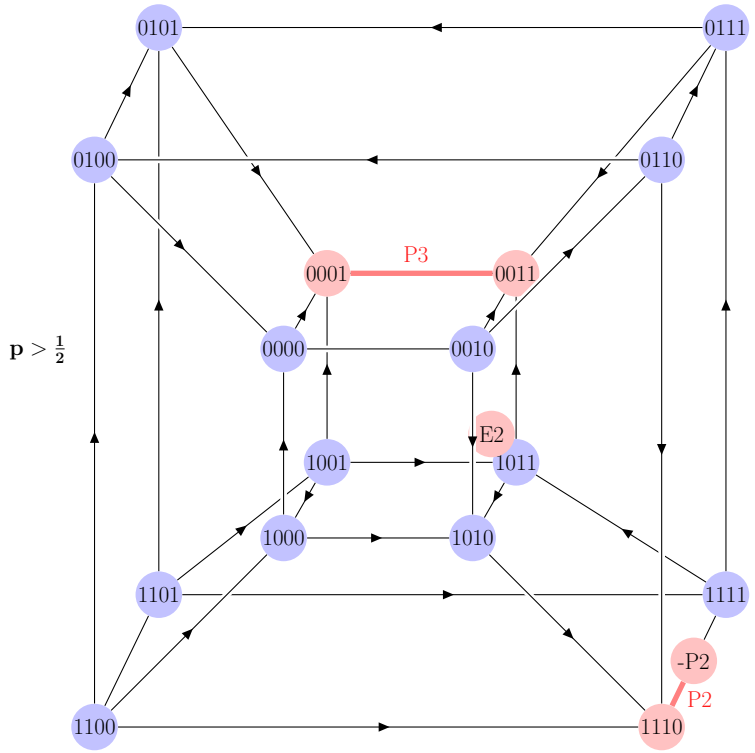


Fig. 4 Nash equilibria and replicator dynamics for Class I. i ($0 \leq c_1 < c_2 < 1$), $1/2 < p < 1$: the partially revealing equilibrium E2, which sits in the face $(*, 0, 1, *)$ (Figure 5, bottom left square); the continuum of equilibria P2, which stretches from the vertex $(1, 1, 1, 0)$ to the point $(1, 1, 1, 1 - c_2)$ marked as -P2 (Figure 5, bottom right square); and the continuum of equilibria P3, covering the entire edge from $(0, 0, 0, 1)$ to $(0, 0, 1, 1)$.

- $(1, *, 1, *)$, the lower right square in Figures 5 and 4, which contains the edge of rest points $(1, 1, 1, *)$, and on which interior orbits converge to one of the Nash equilibria $(1, 1, 1, y')$, $0 < y' < 1 - c_2$, in the equilibrium component P2;
- $(1, *, *, 1)$, the upper right square in Figure 5, and the lower back square of the hypercube Figure 4, on which interior orbits converge to the corner (1,0,1,1);
- $(*, 0, 1, *)$, the lower left square in Figure 5, and the inner right square of Figure 4, which contains the saddle point E2, and on which almost all orbits converge to $(0, 0, 1, 1) \in$ P3 or to $(1, 0, 1, 0)$, with E2 on the separatrix, i.e., the manifold separating the two basins of attraction. Note that $(1, 0, 1, 0)$ is unstable along the edge $(1, *, 1, 0)$, along which there is a connection to $(1, 1, 1, 0) \in$ P2.
- $(*, 0, *, 1)$, the upper left square in Figure 5, and the inner back square of Figure 4, which contains the edge of rest points $(0, 0, *, 1) =$ P3, and on which interior orbits converge to one of the Nash equilibria in P3.

The ω–limit set of an orbit is a closed, connected, invariant, and internally chain transitive (ICT) subset (i.e., any two points in it can be connected by pseudo-orbits), see Benaïm (1999), Corollary 5.6. The maximal ICT sets in the four squares in Figure 5 are

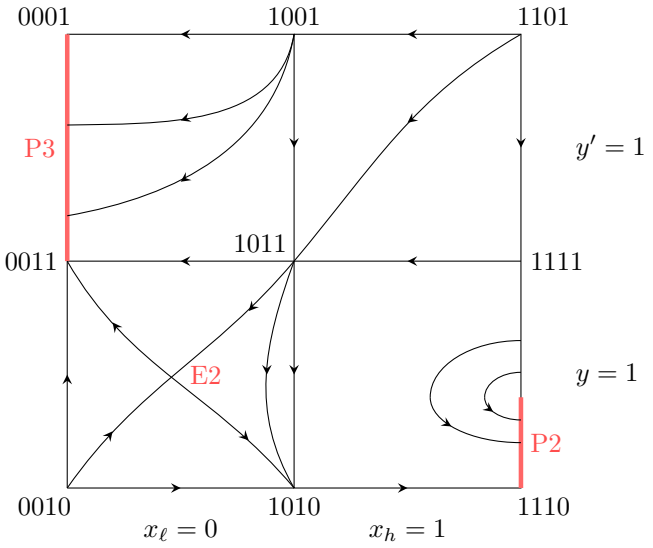


Fig. 5 The four two-dimensional faces that attract all interior orbits of the replicator dynamics (11) for $1/2 < p < 1$. Actually, as shown in Proposition 3, interior orbits converge to either P3, or E2, or P2.

the three corners 0010, 1001, 1101, the edge $(0, 0, *, 1)$ = P3, and the 'house–shaped' pentagon (let's call it H) spanned by the square $(1, *, 1, *)$, E2, and the orbits connecting 1011 with E2 and 1010. The corners $(0, 0, 1, 0)$, $(1, 0, 0, 1)$, $(1, 0, 1, 0)$, $(1, 0, 1, 1)$, $(1, 1, 0, 1)$ cannot attract orbits from the interior of the hypercube, since they are not Nash equilibria. If an ω–limit set Ω of an interior orbit satisfies $\Omega \subseteq$ P3, then the reduction principle implies that $\omega(x)$ is one of the NE in P3. If $\Omega \subseteq H$ (and $\Omega \neq \{\text{E2}\}$), then it must contain an equilibrium from P2, and then by the reduction principle, it is one of the NE $(1, 1, 1, y')$ with $0 < y' < 1 - c_2$. □

$\mathbf{p = \frac{1}{2}}$: In this case, (11) has the following rest points: all 2^4 corners of the hypercube (Figure 6), the square $(0, 0, *, *)$ containing the Nash-equilibrium component P1-E2'-P3, and the square $(1, 1, *, *)$ containing the Nash-equilibrium component E1'-P2. In this case, we first show that every orbit converges to a Nash equilibrium, and then we discuss the behavior near the two equilibrium components.

Proposition 10 (Convergence, Class I.i, $p = 1/2$)**.** *Each orbit in the interior of the hypercube converges to a Nash equilibrium, either in P1-E2'-P3 or in E1'-P2.*

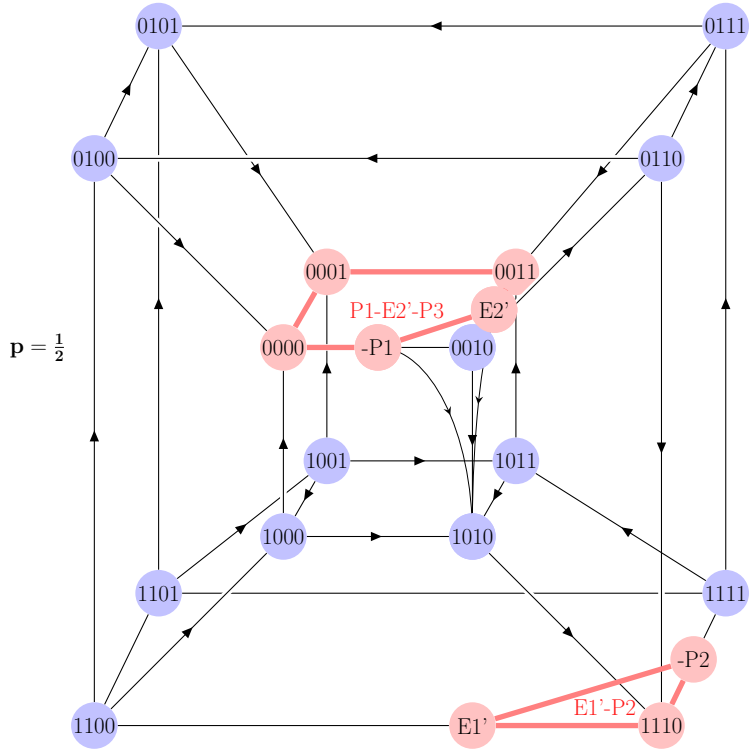


Fig. 6 Nash equilibria and replicator dynamics for Class I.i ($0 \leq c_1 < c_2 < 1$), $p = 1/2$: the equilibrium component P1-E2'-P3, which corresponds to the pentagon in the face $(0, 0, *, *)$ given by the convex hull of $(0, 0, 0, 0)$, -P1 $= (0, 0, c_1, 0)$, E2' $= (0, 0, 1, 1 - c_1)$, $(0, 0, 1, 1)$, and $(0, 0, 0, 1)$; and the equilibrium component E1'-P2, which corresponds to the triangle in the face $(1, 1, *, *)$ given by the convex hull of E1' $= (1, 1, c_2, 0)$, $(1, 1, 1, 0)$, and -P2 $= (1, 1, 1, 1 - c_2)$. Also shown: two orbits leading from the component P1-E2'-P3 to $(1, 0, 1, 0)$.

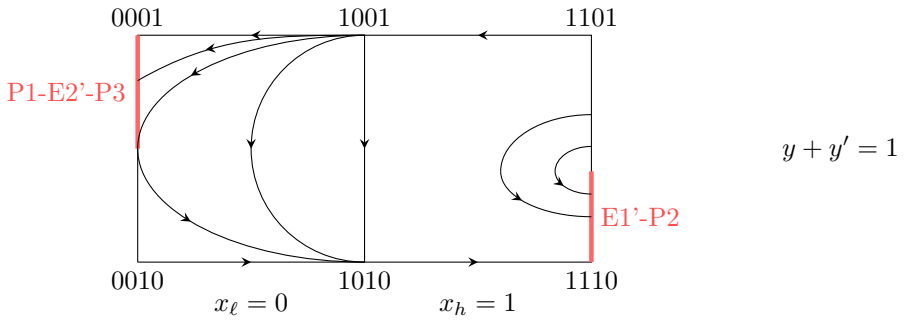


Fig. 7 The dynamics for $p = 1/2$ on the intersection of the two cubes $x_h = 1$ and $x_\ell = 0$ with the invariant diagonal $y + y' = 1$.

Proof. For this case, after omitting the common factor $\frac{1}{2}$, the replicator dynamics for behavior strategies (11) is given by:

$$\begin{aligned} \dot{x}_h &= x_h(1-x_h)(y-y'-c_1), \\ \dot{x}_\ell &= x_\ell(1-x_\ell)(y-y'-c_2), \\ \dot{y} &= y(1-y)[x_h - x_\ell], \\ \dot{y}' &= y'(1-y')[-x_h + x_\ell]. \end{aligned} \tag{21}$$

From (16) we get with $\phi(x) = \log \frac{x}{1-x}$ (where $\phi : (0,1) \to \mathbb{R}$ is strictly increasing and bijective)

$$(\phi(x_h) - \phi(x_\ell))^{\cdot} = c_2 - c_1 > 0$$

and hence, for some constant C_0,

$$\phi(x_h(t)) - \phi(x_\ell(t)) = (c_2 - c_1)t + C_0$$

Therefore, along each interior solution, there is a time t_0 s.t. $\phi(x_h(t_0)) = \phi(x_\ell(t_0))$, and hence $x_h(t_0) = x_\ell(t_0)$. Then for $t > t_0$: $\phi(x_h(t)) > \phi(x_\ell(t))$, hence $x_h(t) > x_\ell(t)$, and from (21), $\dot{y}(t) > 0$ and $\dot{y}'(t) < 0$. Thus $y(t)$ and $y'(t)$ are ultimately monotone, hence they converge. Therefore, again from (21), $x_h(t)$ and $x_\ell(t)$ are ultimately monotone and hence they converge. So the limit of each interior solution exists, and by the folk theorem of evolutionary game theory (Hofbauer and Sigmund 1998, Theorem 7.2.1 (b) or Cressman 2003, Theorem 2.5.3 ii), it must be a Nash equilibrium. □

Remark 2. *From the last two equations of (21), we get a constant of motion:*

$$[\log y - \log(1-y) + \log y' - \log(1-y')]^{\cdot} = \frac{\dot{y}}{y(1-y)} + \frac{\dot{y}'}{y'(1-y')} = 0 \tag{22}$$

and hence, with $C > 0$ constant,

$$yy' = C(1-y)(1-y'). \tag{23}$$

This provides a foliation of the hypercube into 3d invariant manifolds.

Note also that (21) is symmetric w.r.t. $(y, y') \mapsto (1-y', 1-y)$. In particular, the 3d set $y + y' = 1$ (a 'diagonal' of the hypercube) is invariant under the dynamics. This corresponds to the choice $C = 1$ in (23). In Figure 7, we visualize the dynamics on the intersection of this invariant diagonal with the cubes $x_h = 1$ and $x_\ell = 0$ (which together attract all interior orbits, as a consequence of (17) again, which holds for all $p \in (0,1)$).

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

$$\text{E1'-P2} : x_h = x_\ell = 1, \quad y' \leq y - c_2,$$

$$\text{P1-E2'-P3}: \; x_h = x_\ell = 0, \quad y' \geq y - c_1.$$

Proposition 11 (Replicator dynamics near E1'-P2)**.** *For the replicator dynamics (11) on the hypercube* $[0,1]^4$*, E1'-P2 is stable, but not asymptotically stable. Its basin of attraction has nonempty interior.*

Proof. 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 (see Figure 6). 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 $[0,1]^4$. In order to prove that the component is stable we proceed as in the proof of Proposition 7. At a boundary equilibrium $x_h = x_\ell = 1, y' = y - c_2$ the stable manifold has dimension 1. Inspection of the dynamics on the 3-dimensional centermanifold (which is contained in the facet $x_h = 1$) shows that no orbit moves away from the component E1'-P2 and hence it is stable, compare Figure 7. Since the whole face $x_h = x_\ell = 1$ consists of rest points, the component E1'-P2 cannot be asymptotically stable. □

Proposition 12 (Replicator dynamics near P1-E2'-P3)**.** *For the replicator dynamics (11) on the hypercube* $[0,1]^4$*, P1-E2'-P3 is unstable. Nevertheless, it has a basin of attraction with nonempty interior.*

Proof. The component P1-E2'-P3 is the convex hull of P1, E2' $= (0, 0, 1, 1 - c_1)$ and P3, see Figure 6. It is a pentagon with three right angles and a line of symmetry. All equilibria in this component 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, the component P1-E2'-P3 is unstable. Indeed, the vertex E2' is unstable: On $(*, 0, 1, *)$ (the inner right square), there is an orbit from E2' down to $(1, 0, 1, 0)$ (see Figure 6), 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 $(1, 0, 1, 0)$. It looks like a waterfall converging to one point. □

To summarize: For Class I.i ($c_2 < 1$): 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.ii: $0 < c_1 < c_2 = 1$

From (11) 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).

$\mathbf{0 < p < \frac{1}{2}}$: Here the equilibrium E1 (from Class I.i) moves from a 2-dimensional face onto the edge $(1, *, 1, 0)$ (the right lower front edge connecting the outside to the inner cube): E1 $= (1, \frac{p}{1-p}, 1, 0)$. The whole edge $(1, *, 1, 0)$ consists of rest points of the replicator dynamics, and these are Nash equilibria if and only if $x_\ell \leq \frac{p}{1-p}$. In other words, E1 is now the endpoint of a one-dimensional component of Nash equilibria, bounded by E1 and E* $= (1, 0, 1, 0)$, the fully revealing equilibrium. The equilibrium component E*-E1, and every single equilibrium in it, is stable under the replicator dynamics, but not asymptotically stable, as it is part of an edge of rest points. Its basin of attraction has nonempty interior. 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*.

$\mathbf{\frac{1}{2} \leq p < 1}$: The components P2 respectively E1'-P2, which exist in Class I.i, 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, denoted by E*-E1-P2, is stable under the replicator dynamics, but not asymptotically stable, as the entire edge $(1, 1, 1, *)$ (as in Class I.i) consists of rest points. The other components behave as in Class I.i.

Class I. iii: $\mathbf{0 < c_1 < 1 < c_2}$

From (11) we get $\dot{x}_\ell / x_\ell < 0$ and hence $\dot{x}_\ell \downarrow 0$ whenever $x_\ell < 1$. Now the fully revealing equilibrium E* = (1,0,1,0) is a strict Nash equilibrium, and therefore asymptotically stable under the replicator 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 Class I.i ($c_2 < 1$).

Figure 8 provides an overview of the Nash equilibria and the replicator dynamics in the hypercube for all three subclasses of Class I. It allows us to follow the change in the equilibrium structure and the flow of the replicator dynamics on the hypercube along changes in both parameters: c_2 (horizontal) and p (vertical dimension of the 'group picture'). From the first row ($p < 1/2$), we see that the no-signaling–no-acceptance equilibrium component P1 (with index 0) and its surrounding dynamics are unchanged through the three values of c_2. What changes is the 'signaling' component (with index +1): from E1 over E*-E1 to E*. Similarly, for the second row ($p = 1/2$), the component P1-E2'-P3 (with index 0) is unchanged through the three values of c_2, and what changes is the 'signaling' component (with index +1): from E1'-P2, over E*-E1-P2, to E*. For the third row ($p > 1/2$), the no-signaling–accept component P3 (with index 1) and the partially revealing E2 (with index −1, a saddle in the dynamics) are unchanged through the three values of c_2, while the 'signaling' component (with index +1) changes from P2, over E*-E1-P2, to E*. In terms of global convergence, the most general result emerging from our investigation is this: For each of the nine subclasses, all interior orbits of the replicator dynamics converge to some Nash-equilibrium component or the union of the two-dimensional faces containing them.

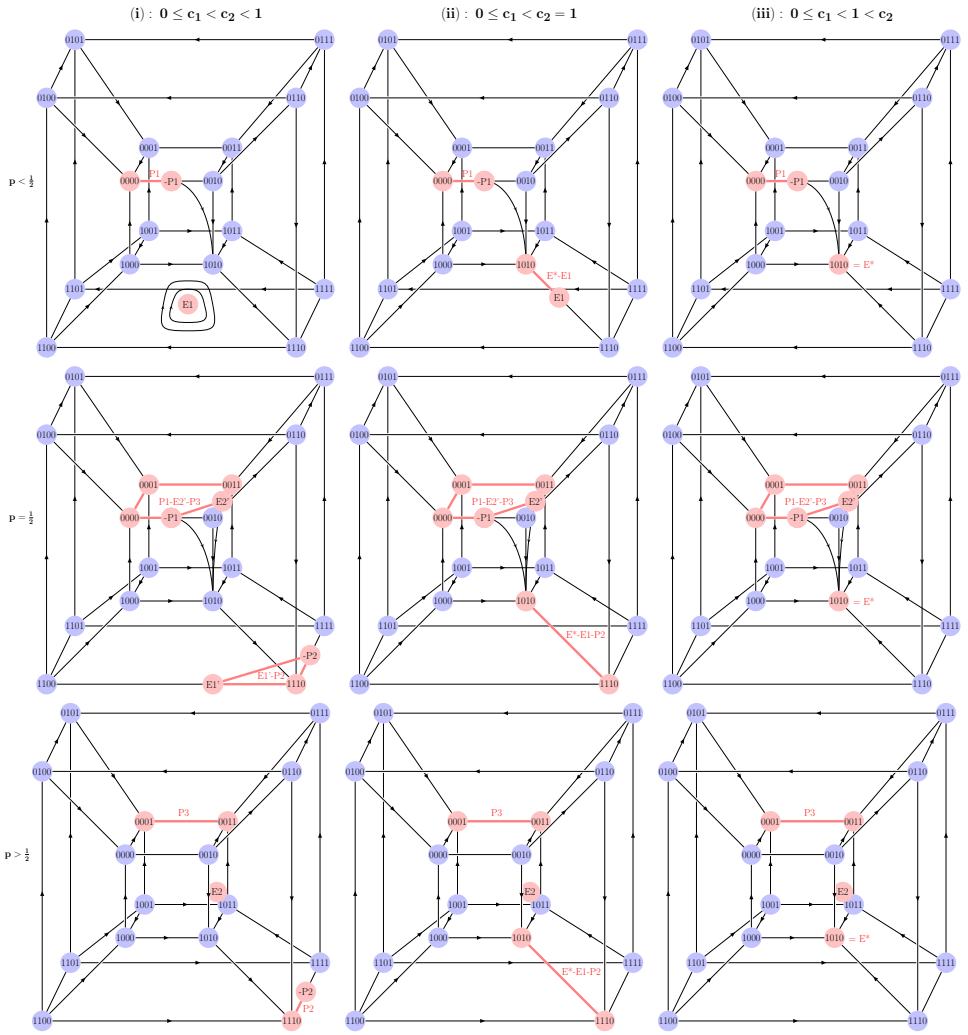


Fig. 8 Replicator dynamics in the hypercube for all three subclasses of Class I: The first column shows the class already seen: subclass i, $0 \leq c_1 < c_2 < 1$, for all three cases of p (Figures 2, 4, 6); the second column shows subclass ii, $0 \leq c_1 < c_2 = 1$; and third subclass iii, $0 \leq c_1 < 1 < c_2$, each for all three cases of p.

3.4 Best-response dynamics

The best-response dynamics for a two-population game is given by the system of differential inclusions (see Cressman 2003):

$$\begin{aligned} \dot{\mathbf{x}} &\in \mathrm{BR}^1(\mathbf{y}) - \mathbf{x}, \\ \dot{\mathbf{y}} &\in \mathrm{BR}^2(\mathbf{x}) - \mathbf{y}. \end{aligned} \tag{24}$$

In analogy to Lemma 1 on the replicator dynamics (foliation into invariant manifolds), we have the following projection result for the best-response dynamics.

Lemma 2. *Let $u : \Delta_4 \to \mathbb{R}^4$ be a payoff function that satisfies $u_1+u_4 = u_2+u_3$. Then $\mathbf{x} \in \Delta_4$ is a best reply, i.e., $\mathbf{x} \in \operatorname{Argmax}_{\mathbf{z}\in\Delta_4} \sum_{i=1}^4 z_i u_i$, if and only if $(x, x') \in [0,1]^2$ is a best response, i.e., it maximizes $(x, x') \mapsto x(u_1 - u_3) + x'(u_1 - u_2) = x(u_2 - u_4) - x'(u_2 - u_1) = \dots$, where $x = x_1 + x_2, x' = x_1 + x_3$.*

Proof. $(1,0,0,0) \in \Delta_4$ is best response iff $u_1 \geq u_2$ and $u_1 \geq u_3$ iff $x = x' = 1$ is best response in $[0,1]^2$. And, $(0,1,0,0) \in \Delta_4$ is best response iff $u_2 \geq u_1$ and $u_2 \geq u_4$ iff $x = 1, x' = 0$ is best response in $[0,1]^2$, and similarly for the two other cases. □

Lemma 2 is related to the behavior of the best-response dynamics in role games (compare Berger 2001 and Cressman 2003). Together with equations (4) and (6), Lemma 2 reduces (24) to a best-reponse dynamics on the hypercube $[0,1]^4$.

Let

$$H(u) = \begin{cases} 1 & \text{if } u > 0 \\ 0 & \text{if } u < 0 \\ [0,1] & \text{if } u = 0 \end{cases}$$

denote the set-valued version of the Heaviside function.

Proposition 13 (Projection of the BR dynamics)**.** *For the normal-form game in Figure 1, the best-response dynamics (24) simplifies to:*

$$\begin{aligned} \dot{x}_h &\in H(y - c_1 - y') - x_h, \\ \dot{x}_\ell &\in H(y - c_2 - y') - x_\ell, \\ \dot{y} &\in H(px_h - (1-p)x_\ell) - y, \\ \dot{y}' &\in H(p(1-x_h) - (1-p)(1-x_\ell)) - y', \end{aligned} \tag{25}$$

with the state space of this differential inclusion being the hypercube $[0,1]^4$.

Proof. This follows from Lemma 2 together with equations (4) and (6). □

Class I.i: $0 \leq c_1 < c_2 < 1$

Proposition 14 (BR dynamics, $0 < p < \frac{1}{2}$)**.** *All orbits of (24) converge to one of the two Nash-equilibrium components, either to E1 or P1. E1 is asymptotically stable. P1 is unstable, but its basin of attraction has nonempty interior.*

Proof. The square $\{(x_h, x_\ell) \in [0,1]^2\}$ is divided into three regions: $A = \{(x_h, x_\ell) \in [0,1]^2 : px_h - (1-p)x_\ell > 0\}$, $B = \{(x_h, x_\ell) \in [0,1]^2 : 2p - 1 < px_h - (1-p)x_\ell < 0\}$, and $C = \{(x_h, x_\ell) \in [0,1]^2 : p(1-x_h) - (1-p)(1-x_\ell) > 0\}$. For $x \in A$, y increases and y' decreases, more precisely, (y, y') moves straight towards $(1,0)$. In B, (y, y') moves towards $(0,0)$, and in C towards $(0,1)$. Similarly, the square $\{(y, y') \in [0,1]^2\}$ splits

into three regions: $D = \{(y, y') : y' < y - c_2\}$ where (x_h, x_ℓ) moves straight towards $(1, 1)$; $E = \{(y, y') : y - c_1 > y' > y - c_2\}$ where (x_h, x_ℓ) moves straight towards $(1, 0)$; and $F = \{(y, y') : y' > y - c_1\}$ where (x_h, x_ℓ) moves straight towards $(0, 0)$. Now, the region BF is positively invariant, i.e., an orbit that starts there, will stay there in positive time, and converge straight towards $(0, 0, 0, 0) \in$ P1.

But also every other equilibrium $(0, 0, y, 0) \in$ P1 (with $y \leq c_1$) attracts solutions: Start at any point in $(\overline{A} \cap \overline{B}) \times F$, and move within the set $px_h = (1 - p)x_\ell$ straight towards this equilibrium. Note that the component P1 is unstable, since there is a solution starting at -P1 $= (0, 0, c_1, 0)$ and heading straight towards $(1, 0, 1, 0)$, until it reaches $(\frac{c_2 - c_1}{1 - c_1}, 0, c_2, 0)$ (from where it will finally converge to E1).

It is easy to check that all orbits not converging to the component P1 must ultimately cycle between the four regions BE, AE, AD and BE, thereby moving towards the 2-dimensional face $x_h = 1, y' = 0$. This face corresponds to the cyclic 2×2 game that supports the quasistrict equilibrium E1. It follows (as in Berger, 2001, Lemma 3) that all these orbits converge to E1, and E1 is asymptotically stable. Note that orbits starting from the 2-dimensional set given by $px_h = (1 - p)x_\ell$ and $y' = y - c_2$ head straight towards E1.

□

Proposition 15 (BR dynamics, $1/2 < p < 1$)**.** *All orbits of (24) converge to one of the three Nash-equilibrium components, either E2, P2, or P3. E2 is unstable; P2 and P3 are asymptotically stable.*

Proof. 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 positively invariant (that is, invariant in the positive time direction) under the best-response dynamics, and orbits move straight towards the Nash equilibrium $(1, 1, 1, 0)$ in P2. In the positively 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 positively invariant region $\{(x_h, x_\ell, y, y') \in [0, 1]^4 : px_h - (1-p)x_\ell < 0, y-c_1-y' < 0\}$ 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 orbit converges to the set of Nash equilibria, and every Nash equilibrium is the limit of some orbit from the interior. Starting in the three-dimensional set $\{(x_h, x_\ell, y, y') \in [0, 1]^4 : 0 < px_h - (1-p)x_\ell \leq 2p-1, y' = y - c_1\}$ allows for the orbit to go straight towards E2 (but there are other orbits with the same initial condition heading towards P2 or P3). □

Proposition 16 (BR dynamics, $p = \frac{1}{2}$)**.** *All orbits of (24) converge to one of the two Nash-equilibrium components, either E1'-P2 or P1-E2'-P3. E1'-P2 is asymptotically stable. P1-E2'-P3 is unstable, but its basin of attraction has nonempty interior.*

Proof. The region $\{(x_h, x_\ell, y, y') \in [0, 1]^4 : x_h < x_\ell, y - c_1 - y' < 0\}$ is positively invariant under the best-response dynamics, and orbits move straight towards the Nash equilibrium $(0, 0, 0, 1)$ in P1-E2'-P3. The region $\{(x_h, x_\ell, y, y') \in [0, 1]^4 : x_h > x_\ell\}$ is also positively invariant. Orbits starting there enter the subset where $y - y' > c_2$ where they move straight towards the Nash equilibrium $(1, 1, 1, 0)$ in E1'-P2. It is easy

to check then that every best-response orbit converges to the set of Nash equilibria, and for every Nash equilibrium E one can find an orbit starting on the set $x_h = x_\ell$ converging straight to E. Furthermore: In a neighborhood of E1'-P2, $\dot{x}_h > 0$. Therefore, every orbit starting close enough to this component will reach the region $\{(x_h, x_\ell, y, y') \in [0,1]^4 : x_h > x_\ell\}$ while $y - y'$ is still larger than c_1, and then, as stated above, will converge back to $(1,1,1,0)$, showing that E1'-P2 is asymptotically stable. But there are orbits connecting the component P1-E2'-P3 to the component E1'-P2: Start at a NE on the line segment -P1-E2', i.e., $(0,0,y,y')$ with $y - y' = c_1$. Then the high type is indifferent, and there is an orbit heading straight for $(1,0,1,0)$. This orbit enters the region $\{(x_h, x_\ell, y, y') \in [0,1]^4 : x_h > x_\ell, c_1 < y - y' < c_2\}$, and then enters $\{(x_h, x_\ell, y, y') \in [0,1]^4 : x_h > x_\ell, y - y' \geq c_2\}$, where it converges straight towards $(1,1,1,0)$, connecting the component P1-E2'-P3 to the component E1'-P2. This implies that P1-E2'-P3 is unstable. □

These results carry over to the two other cases regarding c_2, namely $c_2 = 1$ (ii), and $c_2 > 1$ (iii), when we identify, for each value of p, the respective +1 equilibrium components: for $0 < p < 1/2$, E1 (i) 'turns' into E*-E1 (ii) respectively E* (iii); for $1/2 < p < 1$, P2 'turns' into E*-E1-P2 (ii) respectively E* (iii); for $p = 1/2$, E1'-P2 'turns' into E*-E1-P2 (ii) respectively E* (iii).

3.5 Dynamics in the normal form vs. extensive form

The classical interpretation of the replicator dynamics is that the game is played repeatedly by players drawn at random from a large population, or two large populations in the case of asymmetric games, with the average payoff of a strategy representing the fitness over the lifetime of an individual who carries it and the carriers of these strategies reproducing proportionally to their fitness based on the biological transmission of strategies. Alternatively, the replicator dynamics emerges as the limit of various processes based on imitation of strategies that perform well (see, for instance, Weibull 1995) or a simple model of reinforcement learning (Börgers and Sarin 1997).

For our games, the difference between the replicator dynamics in terms of strategies in the normal form and behavior strategies in the extensive form of the game is related to the question of what are the populations within which strategy replication takes place, and as a consequence what is the form of the behavioral program that is transmitted. For the dynamics in the normal-form game, the interpretation is straightforward: There are two populations, the player-1 and the player-2 population, within each of which strategy transmission takes place. What is transmitted are the programs how to behave in each of the possible roles of the respective player, the two different types (high and low) for player 1, and the two different information sets (after the expression or the absence of the costly signal) for player 2. This mechanism fits well to a scenario of biological transmission of strategies: For individuals of the player-1 population, after strategies have been transmitted, some random mechanism decides which type they are going to be (high or low), and then, over their lifetime, they apply the action (s or $\bar{s}$) that their inherited strategy prescribes for the respective type. Similarly for individuals of the player-2 population, only that it is not a random mechanism but the distribution of strategies in the player-1 population that

decides with which probability they find themselves in which role, that is, in front of a player 1 who does respectively does not express the costly signal, and conditional on that situation, they execute the inherited behavioral program.

For the replicator dynamics in terms of behavior strategies, strategies are replicated within the four subpopulations defined by the agents in the extensive form (the two types for player 1, and the two agents at different information sets for player 2). For the player-1 population that is to say that the type is decided first and that replication takes place within the two subpopulations. Similarly, for the player-2 population, replication takes place within the two roles, after observation of s respectively $\bar{s}$. This 4-population dynamics is easier to interpret in terms of a story of imitation.

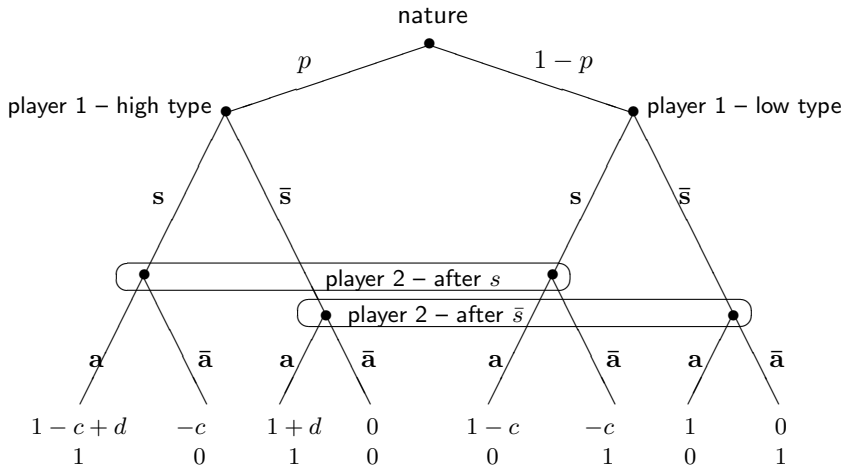
4 Variants of the model

The typical application of Class I, characterized by differential costs of producing the signal, in the spirit of Spence (1973), is educational credentials as a signal for performance or productivity—the underlying hypothesis being that obtaining a certain degree is less costly in terms of effort and time for the more productive type.

However, the assumption that the two types face differential costs in *producing* the signal is hard to justify in some applications of costly-signaling theory that have been advanced. This difficulty comes out most clearly when the cost of the signal is some fixed monetary value. For example: Placing an ad 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 that buys the ad. And indeed, models of advertising as a costly signal (see, for instance, Milgrom and Roberts 1986) usually do not turn on the assumption of differential costs in producing the signal but are grounded in the idea that different types have *differential benefits* when the signal induces the desired action. Class II captures this mechanism in the context of our simple modeling framework with a single costly signal.

4.1 Class II: Differential benefits in case of success

In Class II, the production of the signal is of the same cost $c > 0$ for the two types of player 1, 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 and simplified version of Milgrom and Roberts's (1986) model of advertising as a signal for product quality, and to some extent, Grafen's (1990) formalization of the *Handicap Principle*. In Milgrom and Roberts's model, 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. In Grafen's model, the argument of differential payoffs for different types of player 1, when player 2 accepts, appears implicitly in the form of assumptions on the derivatives of the payoff functions.



	aa	**aā**	**āa**	**āā**
ss	$1-c+pd,\ p$	$1-c+pd,\ p$	$-c,\ 1-p$	$-c,\ 1-p$
ss̄	$1-pc+pd,\ p$	$p(1-c)+pd,\ 1$	$-pc+(1-p),\ 0$	$-pc,\ 1-p$
s̄s	$1-(1-p)c+pd,\ p$	$(1-p)(1-c),\ 0$	$p-(1-p)c+pd,\ 1$	$-(1-p)c,\ 1-p$
s̄s̄	$1+pd,\ p$	$0,\ 1-p$	$1+pd,\ p$	$0,\ 1-p$

Fig. 9 Class II. At the top, the game in extensive form; at the bottom, the game in normal form resulting from that extensive-form game.

4.2 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, games in Class II have the *same equilibrium structure as those in Class I*:

(i) If $0 < c < 1$, the equilibrium structure of Class II is as that of Class I when $c_1 < c_2 < 1$;
(ii) if $c = 1$, as that of Class I when $c_1 < c_2 = 1$; and
(iii) if $1 < c \leq 1 + d$, as that of Class I when $c_1 \leq 1 < c_2$.

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: Let us call the *relative net cost of s* for type t (relative to not using s) the payoff of type t when he does not use s and player 2 in response does not accept minus his payoff when he uses s and player 2 in response, nevertheless, does not accept, $\pi_t(\bar{s},\bar{a}) - \pi_t(s,\bar{a})$, over the payoff difference when he uses s and player 2 does respectively does not accept, $\pi_t(s,a) - \pi_t(s,\bar{a})$:

$$\textit{relative net cost of } s \textit{ for type } t \;=\; \frac{\pi_t(\bar{s},\bar{a}) - \pi_t(s,\bar{a})}{\pi_t(s,a) - \pi_t(s,\bar{a})}. \tag{26}$$

Then, in Class I, the relative net cost of s for the high type is c_1, and for the low type c_2; and, in Class II, the relative net cost of s for the high type is $c/(1+d)$, and for the low type c. Both Class I and Class II then can be said to be characterized by *differential relative net costs of the signal s*.

The structural equivalence of the two classes can be explained in the following way:[2] Once the parameter changes indicated above are made, the game given a low type is the same in Class I and Class II, and the game given a high type in Class II is a rescaling of the game given a high type in Class I, namely if in the payoff function for the high type c_1 is replaced by $c/(1+d)$ and the entire function is premultiplied by $(1+d)$. This explains why, while games of Class II, as a whole, are strictly speaking not rescaled versions of games of Class I, they nonetheless have the same equilibrium structure.

4.3 The replicator dynamics for Class II

For Class II, the payoffs for player 1 against mixed strategies of player 2 are given by

$$\begin{aligned} u^1(ss, \mathbf{y}) &= (1+pd)y - c, \\ u^1(s\bar{s}, \mathbf{y}) &= -pc + p(1+d)y + (1-p)y', \\ u^1(\bar{s}s, \mathbf{y}) &= (1-p)(y-c) + p(1+d)y', \\ u^1(\bar{s}\bar{s}, \mathbf{y}) &= (1+pd)y'. \end{aligned} \tag{27}$$

Again (4) holds. For player 2 the payoffs are the same as in (5). Thus, the analog of (11), i.e., the replicator dynamics for behavior strategies, is now given by

$$\begin{aligned} \dot{x}_h &= x_h(1-x_h)[(1+d)(y-y') - c]p, \\ \dot{x}_\ell &= x_\ell(1-x_\ell)[y-y'-c](1-p), \\ \dot{y} &= y(1-y)[px_h - (1-p)x_\ell], \\ \dot{y}' &= y'(1-y')[p(1-x_h) - (1-p)(1-x_\ell)]. \end{aligned} \tag{28}$$

This is essentially the same as the replicator dynamics for Class I with $c_1 = \frac{c}{1+d}$ and $c_2 = c$, except that the speed of x_h is multiplied by $(1+d)$.

4.4 Class I or II, or a combination?

In some phenomena, both the conditions of Class I and of Class II 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 wealth or a signal for future performance and commitment. It is as if the prospective 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 is dress as a costly signal: Having a good suit or dress or 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).

[2] We would like to thank a reviewer of this journal for having shared this observation with us.

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, c_1 and c_2, and an extra payoff d for the high type if player 2 takes the desired action. In such a combined model, the relative net cost of s for the high type will be $c_1/(1+d)$, and for the low type c_2. The equilibrium structure will then be as in Class I only with c_1 replaced by $c_1/(1+d)$.

The advantage of considering Class I and II as two separate models first is both of practical as well as analytical nature: It is certainly more convenient to work with Class I first, instead of starting out with the combined model and conducting the analysis within that framework; not least because it is much easier to carry c_1 through all numerical expressions, tables, and graphs, than $c_1/(1+d)$. But, besides that, we also gain deeper insight into the mechanism of costly signaling by considering Class I and Class II in isolation, as it shows that *each of the two classes represents minimal conditions under which costly-signaling phenomena can be accounted for*. If the assumptions of both classes apply, the resulting equilibrium structure will have the same qualitative properties as those of Class I and Class II in isolation, and hence the same explanatory potential. But we do not need both of these assumptions to get these qualitative properties of the model: One of them, different costs in production or differential benefits from acceptance for the high type, is sufficient.

5 Classical belief-based equilibrium refinements in signaling games

The multiplicity of equilibria in signaling games and the question of which equilibrium outcomes should be considered plausible predictions of the model have been extensively discussed in classical game theory. In view of a unification of the field, we find it useful to compare our results to this approach of equilibrium refinement.

Sequential Bayesian Nash equilibrium (Kreps and Wilson 1982) requires that players update their beliefs over the possible states of nature according to Bayes' law *along the equilibrium path*. 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, an information set that could in principle be reached but that is not reached in the equilibrium under study—a counterfactual situation. In signaling games, an information set *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 consideration. In the games studied here, this concerns equilibrium outcomes in which both types of player 1 use the same signal, such as P1 (both types using $\bar{s}$), P2 (s), and P3 ($\bar{s}$).

Classical refinements of sequential Bayesian Nash equilibrium operate on the principle of imposing restrictions on players' beliefs off the equilibrium path. Such restrictions, so to say, come to complement Bayes' law 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, there is an entire family of such refinement concepts. Prominent in the literature are: the *never-a-weak-best-response* criterion (Kohlberg and Mertens 1986), *'divinity'* (Banks and Sobel 1987), and the *intuitive criterion* (Cho and Kreps 1987).

When using this kind of refinement to select equilibria, one point should be noted from the beginning: Fully revealing equilibria such as E* and partially revealing equilibria such as E1 or E2 trivially survive any refinement based on restrictions on beliefs off the equilibrium path—simply, because there is no signal off the equilibrium path.

5.1 The never-a-weak-best-response criterion and 'divinity'

For the class of games studied here, the *never-a-weak-best-response criterion* (Kohlberg and Mertens 1986) requires that after a signal off the equilibrium path, the support of the belief of the player acting at this information set 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. It is straightforward to check that for games with two states, two signals, and two possible reactions to signals, 'divinity' as defined by Banks and Sobel (1987) coincides with this criterion.

By this requirement, the equilibrium outcome P1, in which both types take $\bar{s}$, as well as outcomes in P1-E2'-P3 for which $y' \in [0, 1 - c_1)$ (the part of the component that stretches from outcomes 'similar' to P1 to outcomes 'similar' to E2', excluding those 'similar' to E2') are discarded. The argument for P1 is this: Within P1 there is one equilibrium, namely the one where player 2 in response to s takes a with a probability of exactly c_1 (the endpoint of that component, marked -P1 in Figure 2), 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 a probability of 1 has to be attributed to the *high* type. But then, 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. Hence: P1 is *not robust under the never-a-weak-best-response criterion.* A similar argument holds for outcomes in P1-E2'-P3 for which $y' \in [0, 1 - c_1)$.

All other equilibrium outcomes satisfy the never-a-weak-best-response criterion:

- The fully revealing equilibrium E* as well as partially revealing equilibria of the form of E1 or E2 survive any refinement based on restrictions on beliefs off the equilibrium path—trivially, because there is no signal off the equilibrium path.
- For P2 the argument, briefly sketched, is this: After $\bar{s}$, which here is off the equilibrium path, the never-a-weak-best-response criterion discards the high type and hence imposes a belief of 1 on the low type. But this is perfectly in line with the behavior strategies of player 2 that support the equilibrium outcome under study, which all require that in response to $\bar{s}$ player 2 takes $\bar{a}$ with a probability of c_2 at least!
- The equilibrium outcome P3 is stable under any refinement that restricts beliefs off the equilibrium path, for the simple reason that any reaction of player 2 to the off-the-equilibrium path signal s supports the equilibrium outcome. And, a similar argument applies to equilibrium outcomes in the component P1-E2'-P3 for which $y' \in [1 - c_1, 1]$ (the part of the component that stretches from outcomes 'similar' to E2' to outcomes 'similar' to P3).

5.2 The intuitive criterion

The *intuitive criterion* (Cho and Kreps 1987), probably the most prominent refinement of sequential Bayesian Nash equilibrium for signaling games, 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. This is generally less restrictive than the never-a-weak-best-response criterion, respectively 'divinity,' and, for the games studied here, has indeed less selective force. It is straightforward to check that when $c_2 < 1$ (Class I.i) or $c_2 = 1$ (Class I.ii), P1 and equilibrium outcomes in P1-E2'-P3 for which $y' \in [0, 1 - c_2]$ *survive* under the intuitive criterion (because the low type could profit from a deviation if player 2 were to accept in case she observes the signal). Only when $c_2 > 1$ (Class I.iii), that is, when the cost of the signal for the low type is *strictly higher* than the benefit from being accepted, the intuitive criterion will discard the no-signaling–no acceptance equilibrium outcome P1, respectively the outcome in P1-E2'-P3 for which $y' = 0$ (because then the low type cannot possibly get a higher payoff from deviating from $\bar{s}$ to s). This last case reflects the selection result that Cho and Kreps achieve with the intuitive criterion in their variant of Spence's model.

Tables 1, 2, and 3 summarize these results, which all carry over to Class II.

5.3 Comparison: Index vs. classical refinements

Comparing equilibrium-selection results based on the necessary condition for evolutionary stability of an index equal to $+1$ with belief-based refinements (see Tables 1, 2, and 3), one gets the following implication: *Whenever an equilibrium outcome is discarded by the never-a-weak-best-response criterion, respectively 'divinity,' then the equilibrium component in which it sits has an index of* 0*, hence different from* $+1$*, and therefore cannot be asymptotically stable under any evolutionary dynamics.* This concerns two cases: First, P1, which exists in any of the subclasses i–iii (Tables 1–3) when the prior is below the critical value $p < 1/2$. Second, *some* of the equilibrium outcomes in the component P1-E2'-P3, which exist in any of the subclasses i–iii in the knife-edge case $p = 1/2$, namely those for which $y' \in [0, 1 - c_1)$, casually speaking, outcomes reaching from those similar to P1 to those similar to E2', excluding those similar to E2'.

But there are equilibrium outcomes sitting in an equilibrium component with an index $\neq +1$, hence a component that *cannot* be asymptotically stable under any evolutionary dynamics, that do satisfy the never-a-weak-best-response criterion, respectively 'divinity.' This concerns two cases: First, the partially revealing equilibrium E2, with index -1, which exists in any of the three subclasses i–iii when the prior is above the critical value $p > 1/2$. Second, in the knife-edge case $p = 1/2$, the rest of the outcomes in P1-E2'-P3, namely those for which $y' \in [1 - c_1, 1]$, casually speaking, outcomes reaching from those similar to E2' to those similar to P3.

All in all then, the necessary condition for asymptotic stability of an index equal to $+1$ has a bit more selection force than the belief-based refinements considered here: It allows us to discard all equilibrium outcomes that are also discarded by the never-a-weak-best-response criterion respectively 'divinity,' namely, P1 and the outcomes

in the component P1-E2'-P3 for which $y' \in [0, 1 - c_1)$, and, in addition to that, E2 and the rest of the outcomes in P1-E2'-P3. To put this into perspective, one should keep in mind that an equilibrium like E2 cannot possibly be discarded by any of the belief-based refinements considered here—simply because there is no signal off the equilibrium path, and that the case in which P1-E2'-P3 exists, namely p exactly equal $1/2$, is in some sense not generic—at least in the sense that it produces equilibrium *components* that harbor different equilibrium *outcomes*. More specifically, P1-E2'-P3 shows the peculiar case that an equilibrium component, which has one index only (here 0), harbors different equilibrium outcomes that *do not agree* on the belief-based refinements that they fulfill.

In terms of the qualitative properties of the equilibria selected, these results imply the following: For the case that the prior probability of the high type is *below* the critical value, $p < 1/2$, both the necessary condition for asymptotic stability of having an index of $+1$ and the strongest belief-based refinements considered here (the never-a-weak-best-response criterion and 'divinity') coincide and select equilibrium outcomes in which the costly signal is partially revealing or fully revealing (depending on the case regarding c_2) over the co-existing, Pareto-inferior, no-signaling–no-acceptance equilibrium outcome P1 (see section 2.4 for a discussion of the welfare properties of equilibria).

For the case that the prior probability of the high type is *above* the critical value, $p > 1/2$, neither the index, nor the belief-based criteria considered here discriminate between the all-signaling–accept equilibrium outcome P2 (respectively outcomes in E*-E1-P2 or E*, depending on the case regarding c_2) and the co-existing, Pareto-dominant, no-signaling–accept equilibrium outcome P3: both have index $+1$ and survive under all belief-based refinements considered here.

5.4 Comparison: Evolutionary stability vs. index

The family of costly-signaling games studied here illustrates that having an index of $+1$ is indeed only a necessary but not sufficient condition for the respective equilibrium component to be asymptotically stable under a specific dynamics: E1 and E*-E1 (structurally the same component in subclasses i and ii, for the case that $p < 1/2$; see Tables 1 and 2), P2 and E*-E1-P2 (structurally the same component in subclasses i and ii, for the cases that $p > 1/2$) and E'-P2, and E*-E1-P2 (structurally the same component in subclasses i and ii, for the cases that $p = 1/2$) are only Lyapunov stable under the replicator dynamics but asymptotically stable under the best-response dynamics. In contrast to that, the fully revealing equilibrium E*, which exists only in subclass iii, but then for all values of p (Table 3), and P3 (which exists in all subclasses i–iii when $p > 1/2$) are asymptotically stable under both the replicator and the best-response dynamics.

On the other hand, equilibrium components with index 0, namely P1, which exists for all subclasses i–iii, when $p < 1/2$, and P1-E2'-P3, which exists for all subclasses i–iii, when $p = 1/2$, or with index -1, namely E2, which exists for all subclasses i–iii, when $p > 1/2$, are unstable—under both the replicator and the best-response dynamics.

A qualitative property of the replicator as well as the best-response dynamics that deserves special attention (because the wording suggests otherwise) is that not only the locally stable components P2 and E*-E1-P2 but also the unstable components P1 and P1-E2'-P3 have basins of attraction with nonempty interior, which is to say that there is a non-negligible set of initial conditions for which the dynamics converges to these components. What distinguishes unstable components like P1 and P1-E2'-P3 from stable but not asymptotically stable components like P2 and E*-E1-P2 is that there is at least one state from where the dynamics leads away from the component: for P1, this is the endpoint of the component marked as -P1 (Figure 2); for P1-E2'-P3, it is the boundary segment connecting the point marked as -P1 to the point marked as E2' (Figure 6). The usual interpretation of this form of instability is that the respective component is not stable under random drift among strategies already present in the population, that is, shifts inside the component. But this is to say that equilibrium components showing this form of instability cannot be completely excluded from an evolutionary point of view—at least not in the short or medium run when forces of random drift are weak relative to the pressures of selection.

The only equilibrium that can effectively be ruled out from an evolutionary point of view is the partially revealing equilibrium E2 with partial pooling in $\bar{s}$, which is a saddle. Remarkably, this equilibrium cannot possibly be excluded by any refinement that relies on restrictions on beliefs off the equilibrium path—because all signals are 'on the path.'

6 Summary and conclusions

This paper analyzes evolutionary dynamics in a family of costly-signaling games with two types, two signals, and two actions in response to signals, extending on results by Zollman, Bergstrom, and Huttegger (2013). Tables 1, 2, 3 and Figure 8 provide an overview and quick access to our main findings.

Our study is an extension of previous studies in several respects: First, as far as the cases covered go: Within the two basic classes considered also by Zollman, Bergstrom, and Huttegger, differential costs in producing the signal (Class I; Section 2) and differential benefits when the signal has the desired effect (Class II; Section 4), we distinguish three paradigmatic cases of the cost of the signal for the low type (subclasses i–iii); and, within each of these subclasses, three exhaustive cases regarding the prior probability distribution over types: p smaller than, equal to, and larger than $1/2$, leading all in all to nine cases. Second, in terms of the detail of the analysis: For both the replicator and the best-response dynamics, we explicitly relate the two-population dynamics in the normal-form game to the four-population dynamics based on behavior strategies in the extensive form, showing that the second corresponds to the first on a specific invariant manifold, which we study in detail for each of the nine cases, discussing local stability around the fixed points as well as global convergence. Finally, we connect the analysis of the dynamics to other methods of equilibrium refinement: index theory (Section 3.1) and the important research program on belief-based refinements in signaling games (Section 5).

The aim of such an integrative approach is theoretical clarification but also applicability. The detailed case distinction allows researchers exploring applications to quickly identify the class (or classes) relevant for the problem at hand and to make use of the results.

The workhorse of our investigation is subclass i (Table 1), the case where the cost of the signal for the low type is strictly between 0 and 1 (what he gains when accepted), where, as a consequence, fully revealing equilibria do not exist: a case often neglected in the literature (particularly under the assumption of a high prior, $p > 1/2$). As a way of summarizing, we give an outlook on the explanatory potential of our results for that case.

6.1 Beyond fully revealing equilibria: periodic orbits around partially revealing equilibria

Partially revealing equilibria in the style of E1, which exist in the case that the cost of the signal for the low type is strictly between 0 and 1 (subclass i, Table 1) and the prior on the high type is below the critical value 1/2, show an interesting pattern when it comes to explaining features of social meaning systems: The costly signal does not perfectly reveal the sender's type but still pushes the belief that it is the high type up to a certain level, in precisely such a way as to leave the receiver *indifferent* between accepting and not accepting. The costly signal, so to say, functions as a means to 'tune' the belief of the other.

This kind of equilibrium could be used, for instance, as a model of *indirect—'off-record'—speech*, that is, speech acts the intended meaning of which diverges from its literal meaning and that rely on the hearer's interpretation to get the speaker's purpose conveyed (Brown and Levinson, 1987). Pinker and coauthors (2007, 2008) suggest that the function of indirect speech is to avoid common knowledge of the type of the speaker while transferring the responsibility to accept or not the desired relationship change to the receiver (in our model, to get accepted, hired, etc.), which gives the speaker the chance to achieve the desired relationship change at least sometimes. Equilibria of the form E1 mimic this feature quite accurately: Using the costly signal avoids giving player 2 sure knowledge about player 1's type, and hence prevents player 1's type from being commonly known, however not because it would leave player 2 in ambiguity about player's type, but because it sets player 2's belief about player 1's type equal to a certain value, here 1/2, such that player 2 is indifferent between her possible actions. In such a situation, player 1 effectively puts it into the hands of player 2 how to react: to take the responsibility to accept or to decline. Player 2, then—because we are in equilibrium—at hearing the costly signal accepts with a certain probability, namely such as to make the low type of player 1 indifferent between using the costly signal and not using it. The *absence* of the costly signal, instead, perfectly reveals the low 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.

How well can such a partially revealing equilibrium like E1, in which the probabilistic strategies (frequencies of the strategies in the population), defining it have to

hold exactly, mimic reality? One might question if players ever 'hit' the right proportions. What might be more realistic is that players get these proportions approximately right—that the frequencies in the population are approximately right—and that they cycle around them. Interestingly, this is the pattern that the replicator dynamics shows close to that equilibrium: E1 is surrounded by periodic orbits (see Figure 2). Once the replicator dynamics has come close to it, it will land on a closed orbit—a 'cycle'—around it in its supporting face. The best-response dynamics mimics a process of tâtonnements around that equilibrium, ultimately converging towards it.

6.2 Co-existence of no-signaling and all-signaling equilibria

The case that in subclass i (Table 1) the prior on the high type is *above* the critical value at which player 2 accepts 1/2 is hardly ever considered in studies of costly signaling. Unjustly, though, because it shows an interesting equilibrium pattern too: the coexistence of the all-signaling–accept P2 and the no-signaling–accept P3, both stable under belief-based refinements as well as under the replicator and the best-response dynamics. Such a multiplicity of solutions is not necessarily a shortcoming of the model, or the refinement methods employed, but might mimic reality. The study of social meaning systems provides numerous illustrations for the phenomenon that different conventions co-exist in different societies. Politeness is a case in point. The equilibrium outcome P2 could model a society where conventionally everybody uses the polite form to make some social exchange happen: overstatement; while P3 could stand for a society where conventionally nobody uses the polite form to make that same exchange happen: understatement.

Besides that, the co-existence of the two equilibrium outcomes P2 and P3 could also serve as explanation of some form of indirect discrimination based on costly signaling, namely when these two signaling conventions are in place for two different subgroups within the same community that are defined by some observable trait that is *not* a matter of choice (such as assigned sex or skin tone) and that does *not* affect the prior probability of the unobservable trait in question, which, however, makes it possible that the action of player 2 is conditioned on it.

In a biological context, the co-existence of P2 and P3 could explain why certain handicaps that transmit no information at all (because the entire population expresses them) might persist in one population (P2) but not in another (P3).

All in all, subclass i, the case where signaling costs for both types are strictly below the gains of acceptance, is a useful tool for investigating expressions of costly-signaling mechanisms—and their dynamics—that manifest themselves through signaling patterns other than fully revealing equilibria.

References

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

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

[3] Benaïm, M. 1999. *Dynamics of stochastic approximations.* Séminaire de Probabilités. Lectures Notes in Mathematics, Vol 1709, pp. 1–68.

[4] Berger, U. 2001. Best response dynamics for role games. *International Journal of Game Theory* 30: 527–538.

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

[10] Börgers, T., Sarin, R. 1997. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory* 77, 1–14.

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

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

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

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

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

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

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

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

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

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

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

[22] 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.

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

[23] 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.

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

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

[26] 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.

[27] Kuznetsov, Y. A. 2004. *Elements of Applied Bifurcation Theory*, 3rd edition. New York: Springer.

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

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

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

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

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

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

[34] Nöldeke, G., Samuelson, L. 1997. A dynamic model of equilibrium selection in signaling games. *Journal of Economic Theory* 73 (1): 118–156.

[34] Pinker, S. 2007. *The Stuff of Thought. Language as a Window into Human Nature.* New York: Viking.

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

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

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

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

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

[38] Spence, M. 1974. *Market signaling: Informational Transfer in Hiring and Related Screening Processes.* Cambridge, MA: Harvard University Press.

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

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

[41] Von Stengel, B. 2021. Finding Nash equilibria of two-player games. arXiv: 2102.04580.

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

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

[44] Weibull, J. 1995. *Evolutionary Game Theory.* Cambridge, MA: MIT Press.

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

[46] 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.