Meaning in Costly-Signaling Games

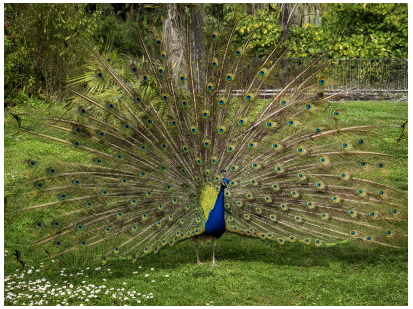
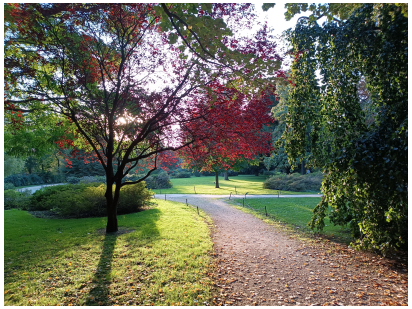
Christina Pawlowitsch

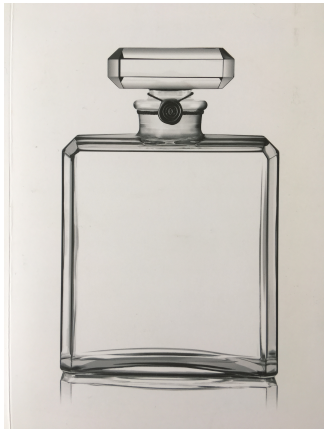
Université Paris–Panthéon–Assas

LINGUAE seminar

Institut Jean Nicod, June 6, 2024

Costly-signaling theory: wide range of applications




Spence (1973): educational credentials as a costly signal
Miller and Rock (1985): dividend payments as a costly signal
Milgrom and Roberts (1986): advertising as a costly signal
Zahavi (1975): "The Handicap Principle" $\rightarrow$ Grafen (1990): formal model
Dawkins and Krebs (1978): "Animal Signals: Information and Manipulation"
Caro (1986): costly signals in predator–prey; Archetti (2008): parasite-host interaction
Bliege Bird and Smith: inefficient foraging strategies, communal sharing
Van Rooy (2003): "Politeness is a Handicap"

Precursor: Veblen (1899): *Theory of the Leisure Class.*

"Evolutionary Dynamics in Costy Signaling Games"

(Hofbauer and Pawlowitsch, working paper)

$\longrightarrow$ Minimalist model: 2 types ("high" and "low"), 2 signals (s and $\bar{s}$), 2 actions in response to signals ("accept" and "do not accept")

$\longrightarrow$ Class I: differential costs of the signal as function of type

Class II: uniform costs of the signal, differential gains of types when accepted

- vary cost parameters (3 cases: cost for low type $<$, $=$, $>$ gain)
- vary prior beliefs (3 exhaustive cases) $\rightarrow$ 9 subclasses

$\longrightarrow$ For each subclass:

- analyze entire equilibrium structure: in the game matrix (Nash equilibrium), and in the game tree (Bayesian Nash equilibrium)
  - classical refinements of Bayesian-Nash equilibrium, based on restrictions on beliefs "off the equilibrium path" ("counterfactual situation")
  - index theory and evolutionary dynamics: replicator and best-response dynamics

This talk/current project:

What is "meaning" in these games?

These games allow us to say something about "meaning," the emergence of meaning, attached to a signal (or its absence) as a function of:

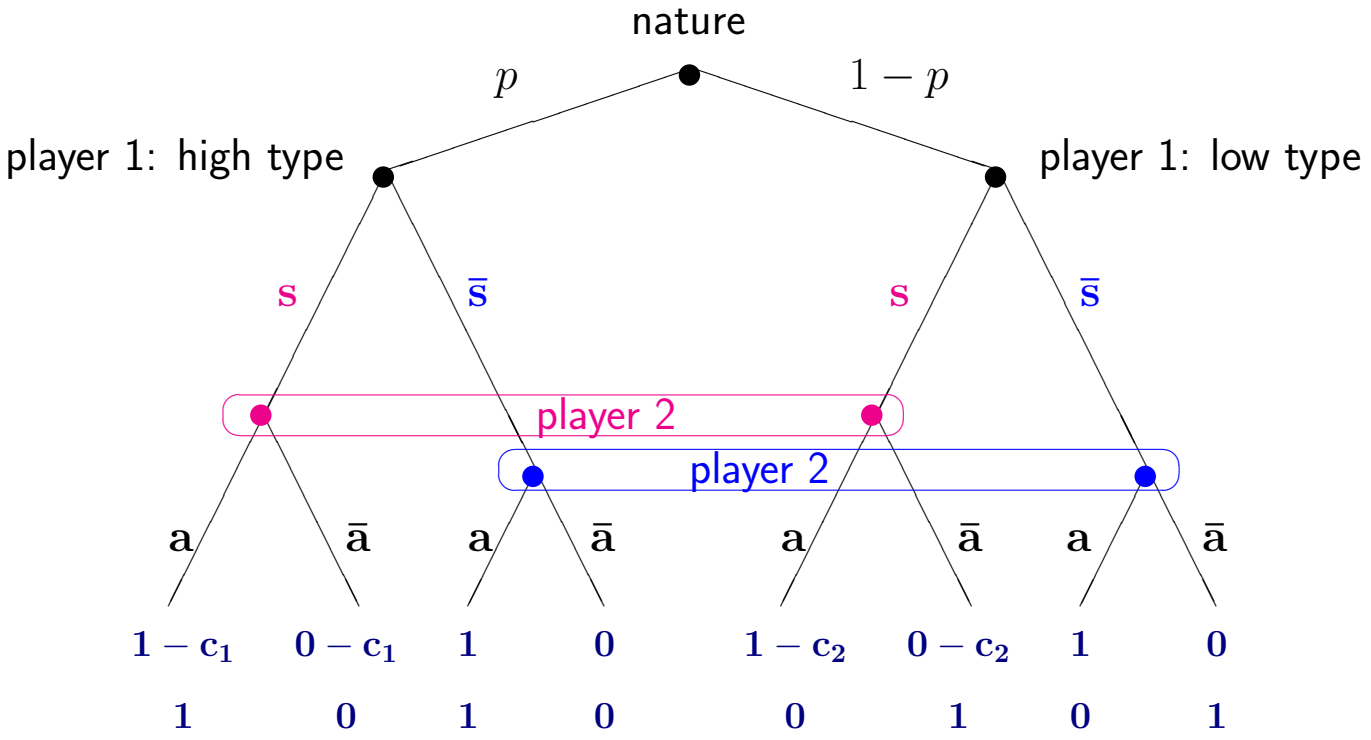
- the costs of the signal carried by various types
- the prior probability distribution over types

Explore potential for applications in the study of language

$\longrightarrow$ First: overview of results.

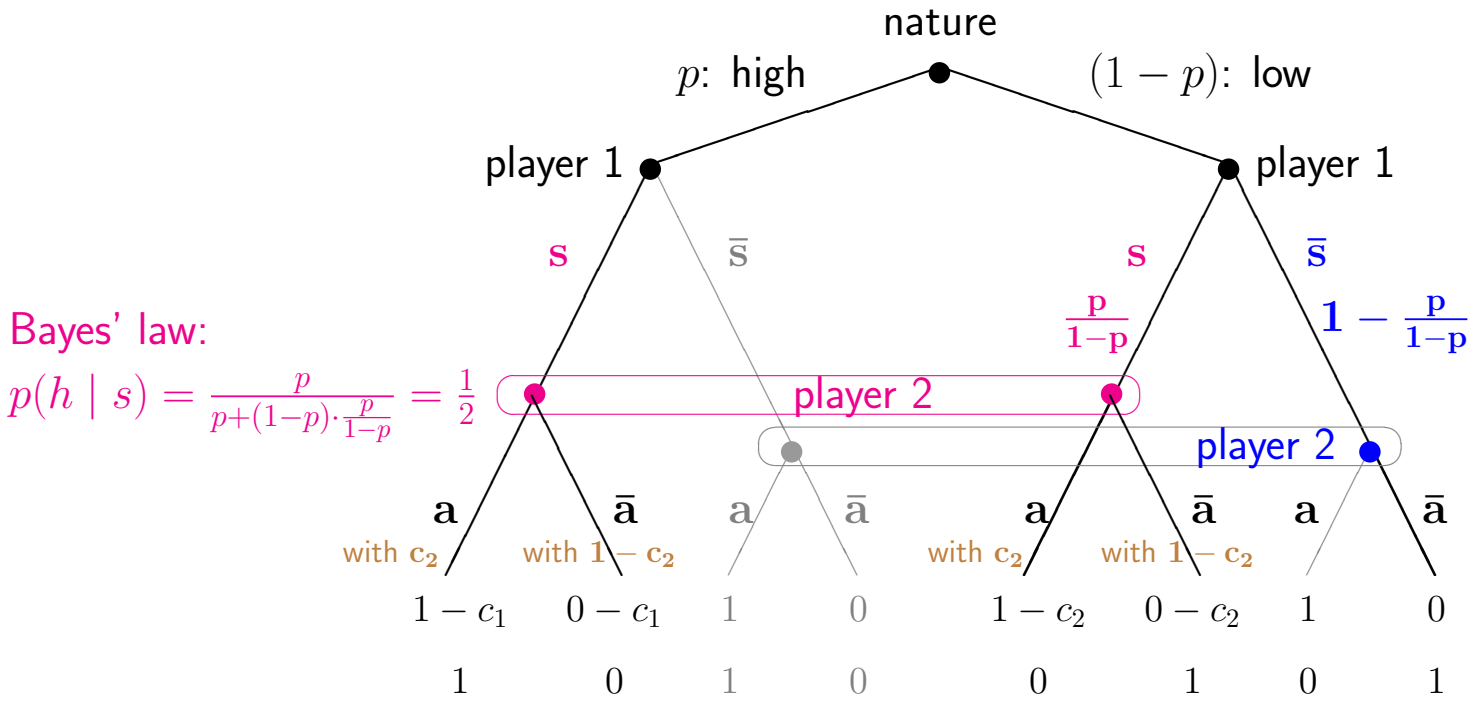
PART I
The model

Costly-signaling game, Class I (discrete version of Spence 1973)



	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	$1,\ p$	$0,\ 1-p$	$1,\ p$	$0,\ 1-p$

Case $0 \leq c_1 < c_2 < 1$, $p < 1/2$: E1 partially revealing equilibrium



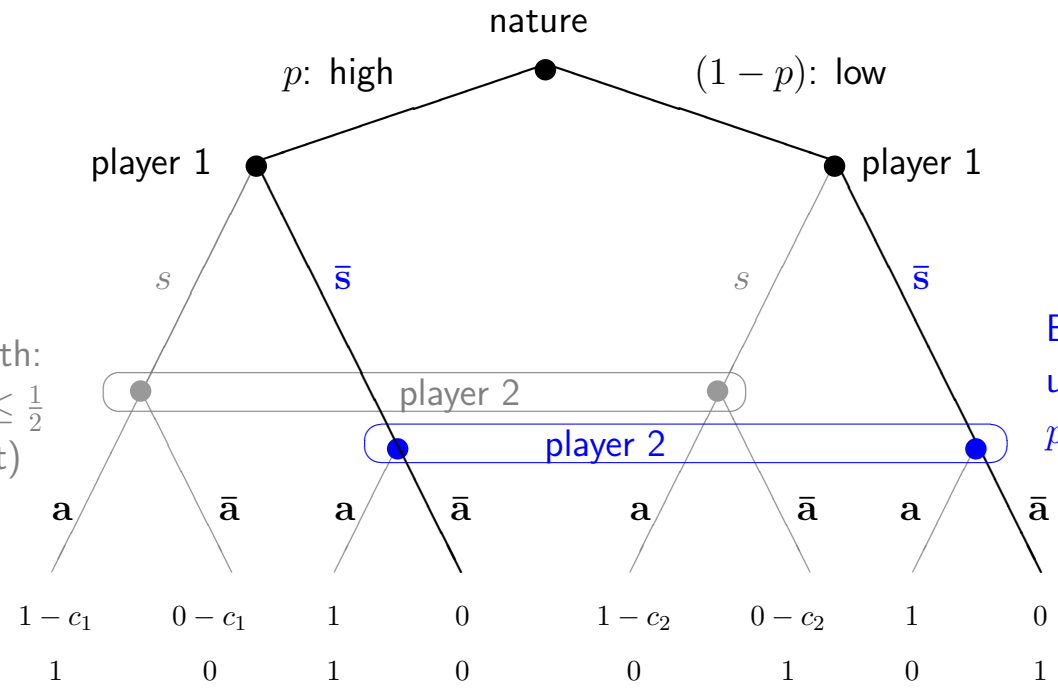
Bayes' law:

$p(h \mid s) = \frac{p}{p+(1-p)\cdot\frac{p}{1-p}} = \frac{1}{2}$

		aa	aā (c_2 ↓ aā)	āa	āā ($1-c_2$ ↓ āā)
$\frac{p}{1-p} \to$	ss	$1-pc_1-(1-p)c_2,\ p$	$\mathbf{1-pc_1-(1-p)c_2},\ p$	$-pc_1-(1-p)c_2,\ \mathbf{1-p}$	$-pc_1-(1-p)c_2,\ \mathbf{1-p}$
$1-\frac{p}{1-p} \to$	sā	$1-pc_1,\ p$	$p(1-c_1),\ \mathbf{1}$	$-pc_1+(1-p),\ 0$	$-pc_1,\ 1-p$
	ās	$1-(1-p)c_2,\ p$	$(1-p)(1-c_2),\ 0$	$p-(1-p)c_2,\ \mathbf{1}$	$-(1-p)c_2,\ 1-p$
	āā	$\mathbf{1},\ p$	$0,\ \mathbf{1-p}$	$\mathbf{1},\ p$	$\mathbf{0},\ \mathbf{1-p}$

- E1: 1 mixes between ss and $s\bar{s}$ with $\frac{p}{1-p}$ on first; 2 between $a\bar{a}$ and $\bar{a}\bar{a}$, with c_2 on first.

Case $0 \leq c_1 < c_2 < 1$, $p < 1/2$: P1 "no-signaling" equilibrium outcome



Off equilibrium path: must have belief $\leq \frac{1}{2}$ (a with c_1 at most)
"Plausible"?

By Bayes' law, updated belief = prior: $p(h \mid \bar{s}) = p < \frac{1}{2}$

	aa	with $y \in [0, c_1] \to$ **aā**	**āa**	with $1-y \to$ **āā**
ss	$1-pc_1-(1-p)c_2,\ p$	$\mathbf{1-pc_1-(1-p)c_2},\ p$	$-pc_1-(1-p)c_2,\ \mathbf{1-p}$	$-pc_1-(1-p)c_2,\ \mathbf{1-p}$
sš	$1-pc_1,\ p$	$p(1-c_1),\ \mathbf{1}$	$-pc_1+(1-p),\ 0$	$-pc_1,\ 1-p$
šs	$1-(1-p)c_2,\ p$	$(1-p)(1-c_2),\ 0$	$p-(1-p)c_2,\ \mathbf{1}$	$-(1-p)c_2,\ 1-p$
šš	$\mathbf{1},\ p$	$0,\ \mathbf{1-p}$	$\mathbf{1},\ p$	$\mathbf{0},\ \mathbf{1-p}$

- P1: No-signaling: 1 takes $\bar{s}\bar{s}$; 2 mix between $a\bar{a}$ and $\bar{a}\bar{a}$ with $y \in [0, c_1]$ on first.

Table 1. Equilibrium structure of the game in Figure 1: $0 \leq c_1 < c_2 < 1$

Prior	Equilibrium component	Index	Replicator dynamics	Best-response dynamics	Classical refinements	Invariance criterion	Payoffs:
$p < \frac{1}{2}$	(E1): *partially revealing/ partially pooling in* s*:* $(1, \frac{p}{1-p}, c_2, 0)$	$+1$	stable	as. stable	yes	—	$h : c_2 - c_1$ $\ell : 0$ $2 : 1 - p$
	(P1): *pooling in* $\bar{s}$*:* $(0, 0, y, 0)$, $y \in [0, c_1]$	0	unstable	unstable	no	not invariant	h: 0 ℓ: 0 $2 : 1 - p$
$p > \frac{1}{2}$	(E2): *partially revealing/ partially pooling in* $\bar{s}$*:* $(1 - \frac{1-p}{p}, 0, 1, 1 - c_1)$	-1	unstable	unstable	yes	—	h: $1 - c_1$ ℓ: $1 - c_1$ $2 : p$
	(P2): *pooling in* s*:* $(1, 1, 1, y')$, $y' \in [0, 1 - c_2]$	$+1$	stable	as. stable	yes	—	h: $1 - c_1$ ℓ: $1 - c_2$ $2 : p$
	(P3): *pooling in* $\bar{s}$*:* $(0, 0, y, 1)$, $y \in [0, 1]$	$+1$	as. stable	as. stable	yes	—	h: 1 ℓ: 1 $2 : p$
$p = \frac{1}{2}$	(E1'-P2): *pooling in* s*:* $(1, 1, y, y')$, $y \in [c_2, 1]$, $y' \in [0, y - c_2]$)	$+1$	stable	as. stable	yes	—	h: $[c_2 - c_1, 1 - c_1]$ ℓ: $[0, 1 - c_2]$ $2 : \frac{1}{2}$
	(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 $y' \in [1 - c_1, 1]$	not all	h: $[0, 1]$ ℓ: $[0, 1]$ $2 : \frac{1}{2}$

Classical refinements of Bayesian Nash equilibrium: restricting beliefs "off the equilibrium path"

In signaling games: "off the equilibrium path" = point after an unused signal (counterfactual situation)

"Strategic robustness test": based on the idea that there is a ruling equilibrium (convention), which is tested against thought experiments: "What would player 2 think (about types of player 1), if she sees a signal that so far is unused?"

- Cho and Kreps (1987): never-a-weak-best-response criterion
- Banks and Sobel (1987): "divinity"
- Govindan and Wilson (2009): "forward induction"

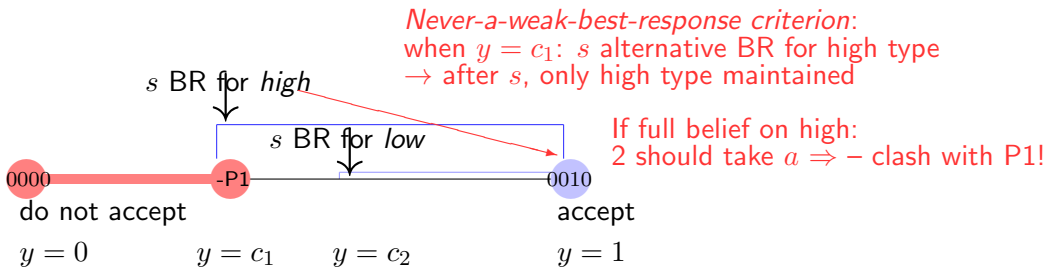
In our game: all 3 coincide. Different ways of formalizing the idea that it is more plausible that the high type deviates to using the costly signal when in the ruling equilibrium (convention) the costly signal is not produced
$\rightarrow$ discard the no-signaling equilibrium outcome P1.

The formal argument

$p < 1/2$: P1 $(\bar{s}\bar{s} \to \bar{a})$

Graph:

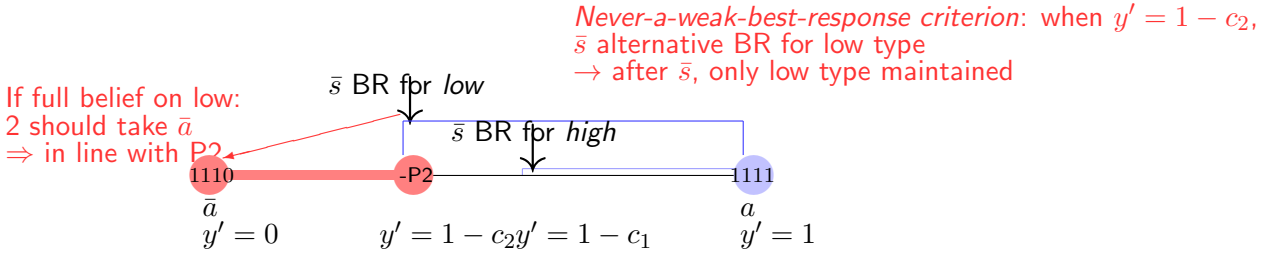
responses of player 2 to the off-the-equilibrium-path signal $\mathbf{s}$:



$p > 1/2$: P2 $(ss \to a)$

Graph:

responses of player 2 to the off-the-equilibrium-path signal $\bar{\mathbf{s}}$:



Invariance

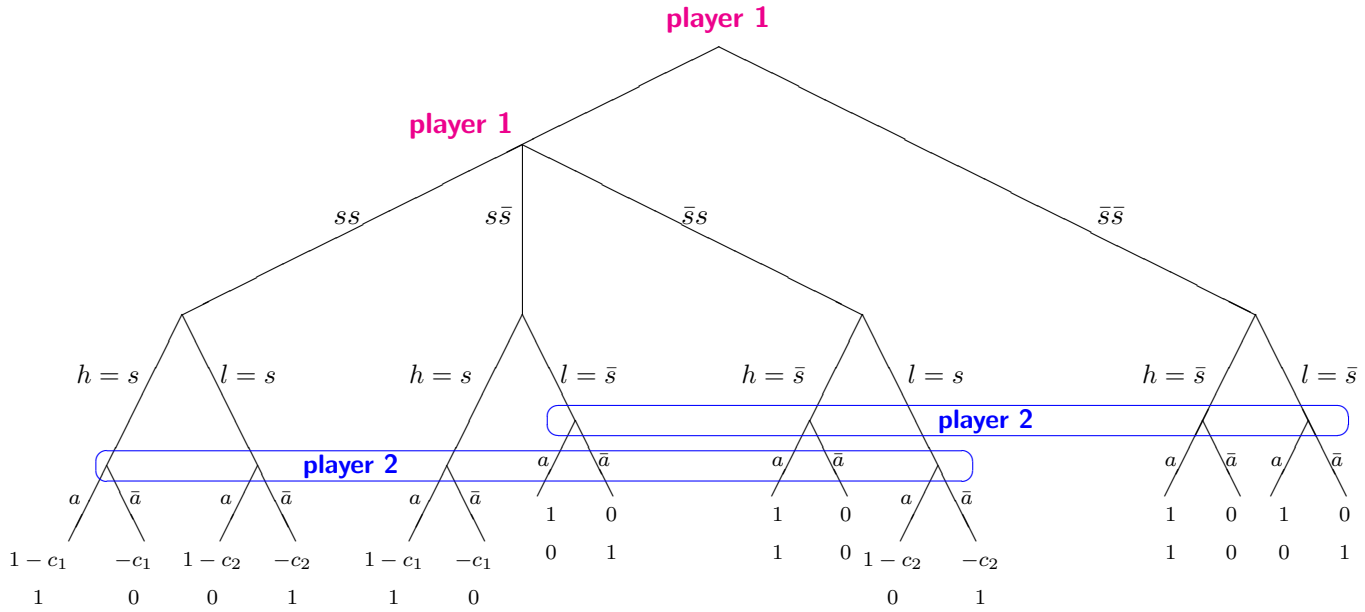
Kohlberg and Mertens (1986): a Nash equilibrium should be selected only if it corresponds to a sequential Bayesian Nash equilibrium in every extensive-form game that maps to the same (reduced) normal form.

Govindan and Wilson (2009):

$$\text{invariance} \Rightarrow \text{forward induction}$$

$$\text{not invariance} \Leftarrow \text{not forward induction}$$

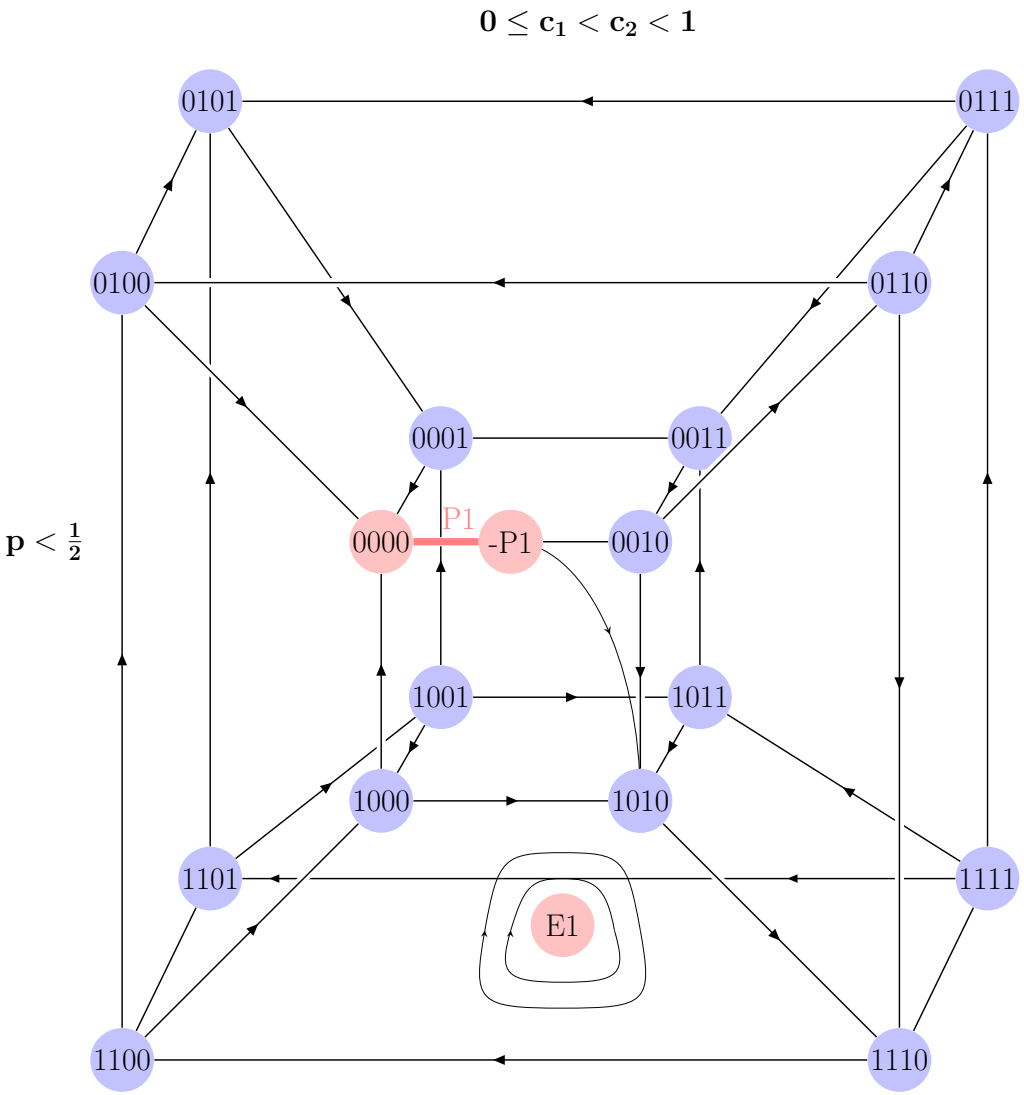
$\mathbf{p < 1/2}$: P1 (index 0) not forward induction $\Rightarrow$ not invariant.
Alternative extensive form – tree – in which P1 fails to be a sequential Bayesian Nash equilibrium:



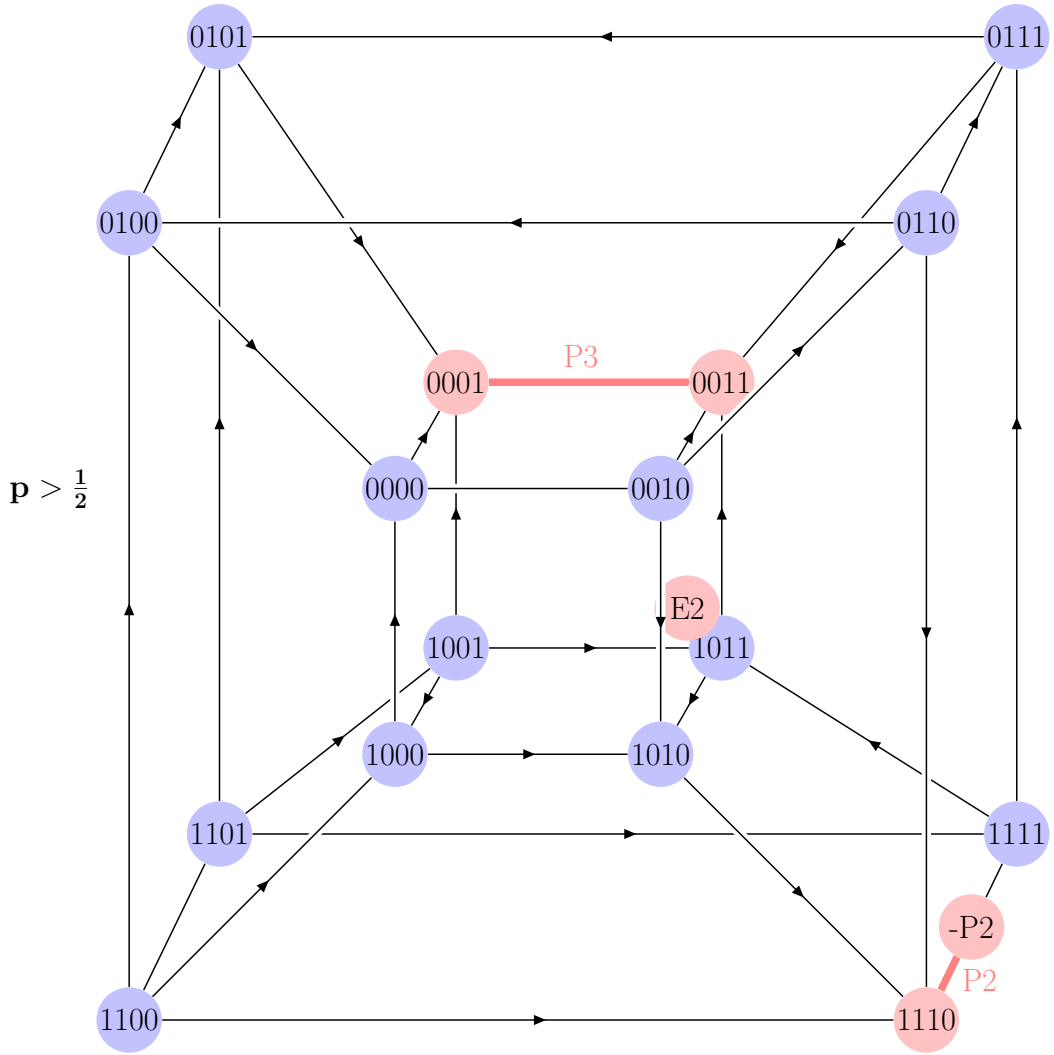
	aa	**aā**	**āa**	**āā**
ss	$1 - pc_1 - (1-p)c_2,\ p$	$1 - pc_1 - (1-p)c_2,\ p$	$-pc_1 - (1-p)c_2,\ 1-p$	$-pc_1 - (1-p)c_2,\ 1-p$
ss̄	$1 - pc_1,\ p$	$p(1-c_1),\ 1$	$-pc_1 + (1-p),\ 0$	$-pc_1,\ 1-p$
s̄s	$1 - (1-p)c_2,\ p$	$(1-p)(1-c_2),\ 0$	$p - (1-p)c_2,\ 1$	$-(1-p)c_2,\ 1-p$

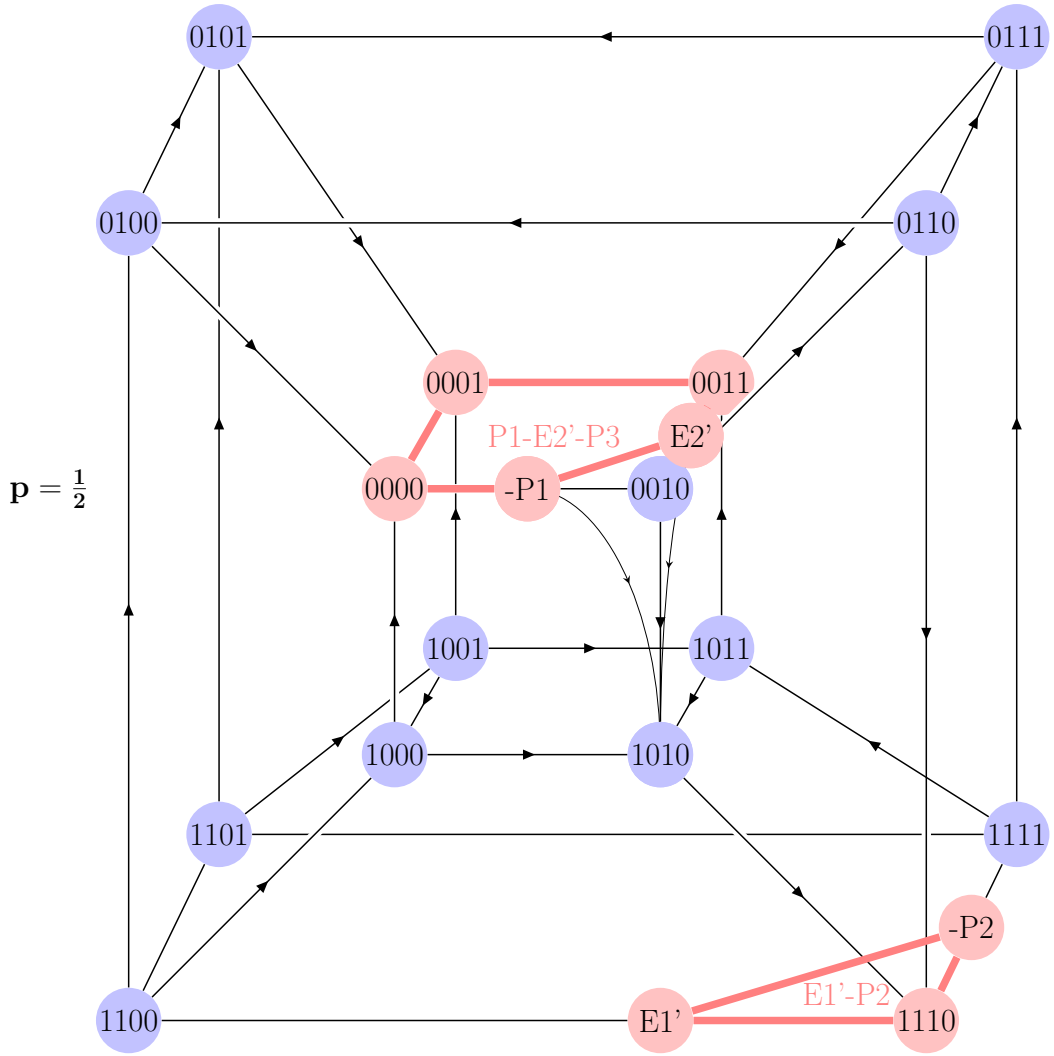
	aa	**aā**	**āa**	**āā**
s̄s̄	$1,\ p$	$0,\ 1-p$	$1,\ p$	$0,\ 1-p$

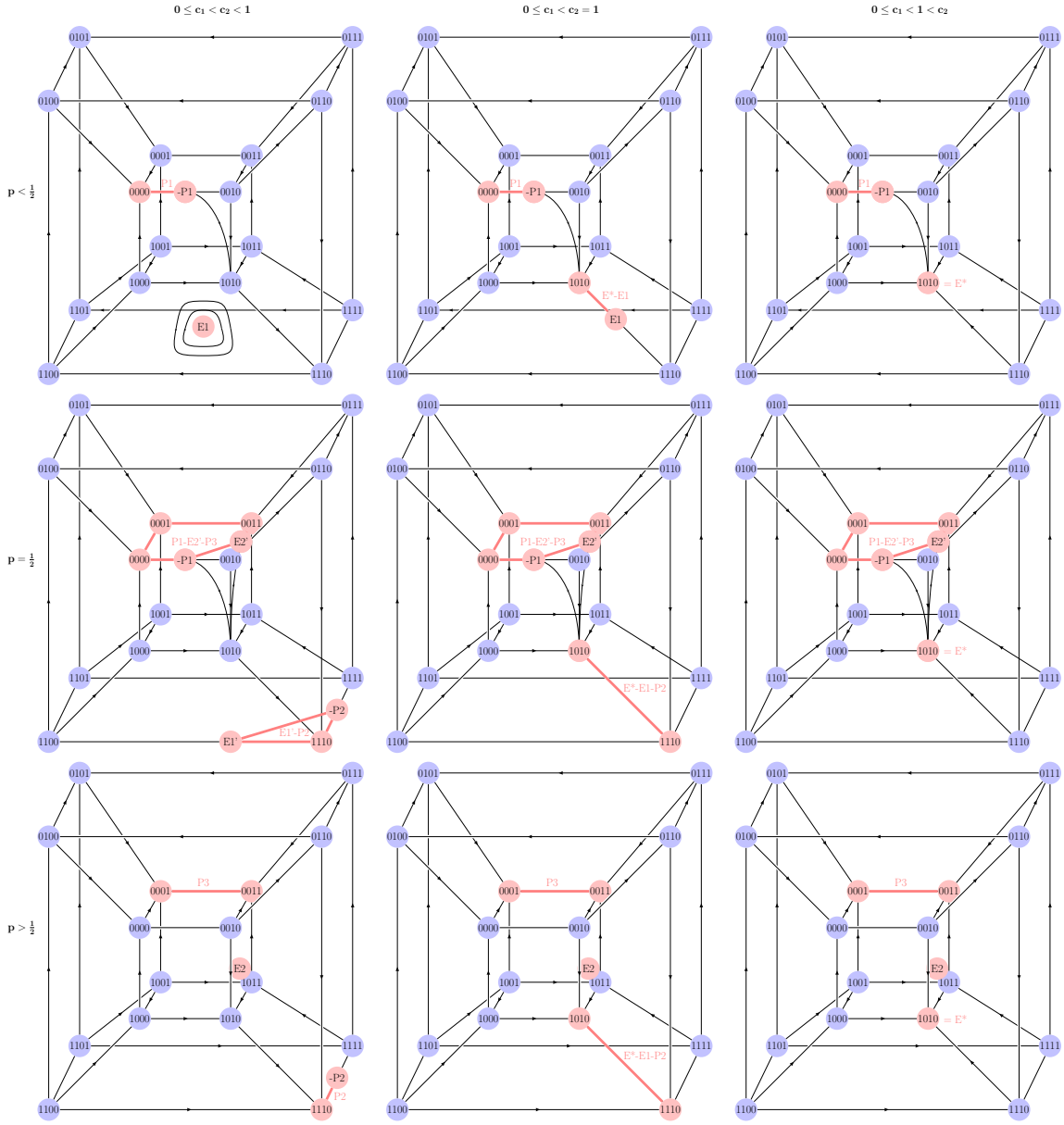
Replicator dynamics: state space: (high,low,after s, after $\bar{s}$)



E1 (partially revealing), surrounded by periodic orbits, sits on the stairs.
P1 ("no signaling") stretches along the inner front edge.


0101
0111
0100
0110
P3
0001
0011
$\mathbf{p} > \frac{1}{2}$
0000
0010
E2
1001
1011
1000
1010
1101
1111
-P2
P2
1100
1110


0101
0111
0100
0110
0001
0011
P1-E2'-P3
E2'
$\mathbf{p} = \frac{1}{2}$
0000
-P1
0010
1001
1011
1000
1010
1101
1111
-P2
E1'-P2
1100
E1'
1110


$0 \le c_1 < c_2 < 1$
$0 \le c_1 < c_2 = 1$
$0 \le c_1 < 1 < c_2$
$p < \frac{1}{2}$
$p = \frac{1}{2}$
$p > \frac{1}{2}$
0101
0111
0100
0110
0001
0011
0000
0010
1001
1011
1000
1010
1101
1111
1100
1110
P1
-P1
E1
E*-E1
= E*
P1-E2'-P3
E2'
-P2
E1'-P2
E*-E1-P2
P3
E2
P2

Moving up the cost of the signal

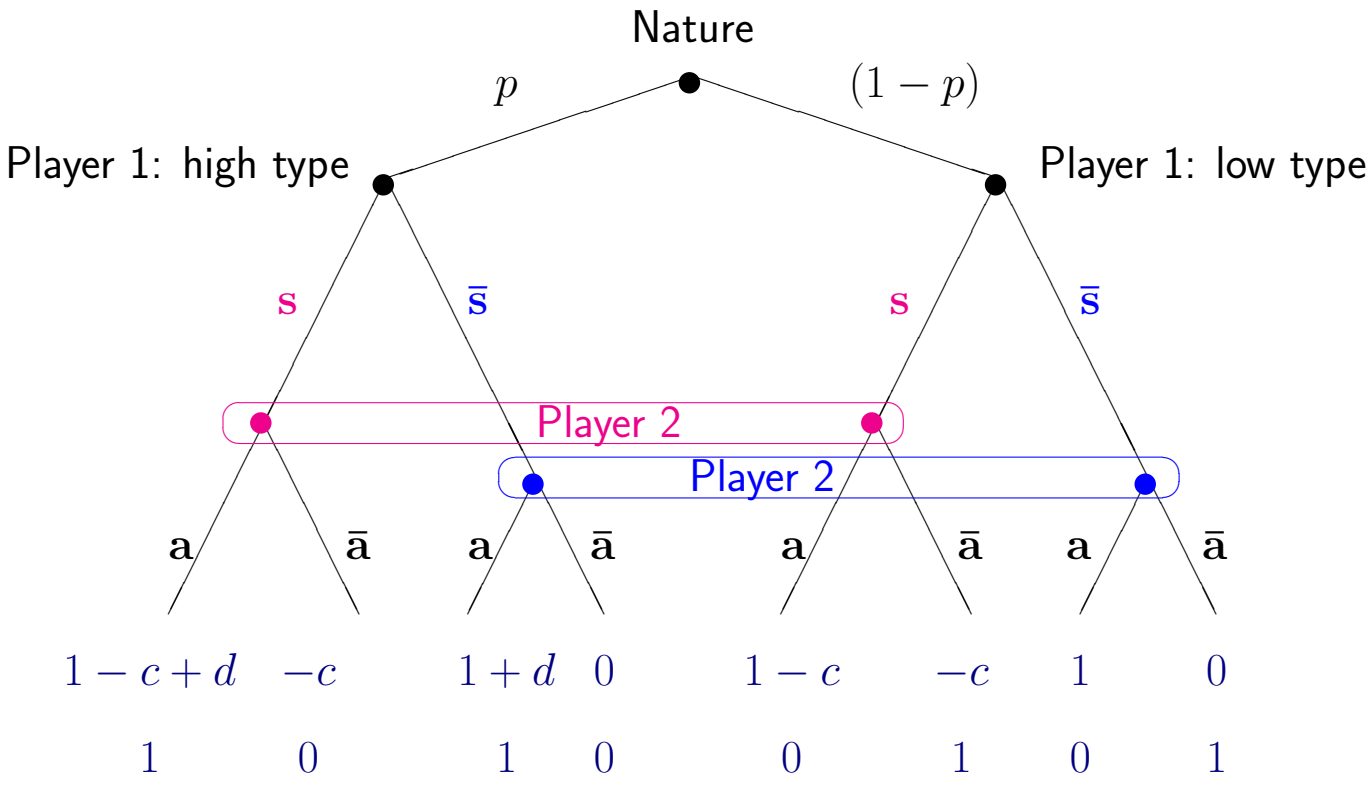
Class I: differential costs of high and low type for producing the signal:

- A fully revealing equilibrium will exist only if the cost of the signal for the low type is at least as high as the benefit from being accepted: $c_2 \geq 1$. Very restrictive!
- Then, this fully revealing equilibrium will exist for any value of the prior p.
- But in addition to it, there will always also exist no-signaling equilibria.

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

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

Class II: uniform costs, differential gains



	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	$1+pd,\ p$	$0,\ 1-p$	$1+pd,\ p$	$0,\ 1-p$

Class II

Same equilibrium structure as class I: for numerical determination of equilibria only need to

$$\text{replace } c_1 \text{ by } \frac{c}{1+d}$$

Combination of class I and II:

$$\text{replace } c_1 \text{ by } \frac{c_1}{1+d}$$

Structural equivalance of class I and II

Important for applications:
both Class I and Class II represent sufficient, minimal, conditions to account for costly-signaling phenomena. We need one or the other. If both apply – "the better"; in a more "stable" way the phenomenon can b accounted for.

Hypothesis:
In human interaction – in an evolutionary cultural dimension – particularly stable as costly signals are variables of choice or traits

- that combine these two "mechanisms": physical cost and cost in terms of some social equivalent ("money"), and
- that are permanently put on display.

Examples:

Education: costly in the sense of Class I, effort to get the degree, but often also costly in terms of a fixed monetary cost (independent of type).

Dress: difficult to wear, but also costly in terms of money.

Educational credentials, in many ways, transported by
$\rightarrow$ language (accents, ability to speak in different languages, switch between different languages)

PART II

Applications in the study of language
Inherent linguistic questions

What is meaning in a costly signaling game?

No ex-ante, conventional meaning

"Meaning" of the costly signal (or its absence) arises in equilibrium as a function of:

- the costs of the signal carried by various types
- the prior probability distribution over types

Can there by "lying"?

Question comes up in the debate of the Handicap Principle.

Fully revealing equilibria are reffered to as "honest" equilibria. The validity of the theory is sometimes quated with the questions whether such "honest" equilibria exist.

Position cannot be found in Zahavi's (1975) paper.

Dawkins and Krebs (1978): "Animal Signals: Information and Manipulation"

Applications in the study of language:

Language: ideal carrier of costly signals, as constantly put on display.

- Politeness:

Costly signal:	Absence of costly signal:
$\hat{S}$	S
"marked – polite – form"	"unmarked form"

Strong component of Class II:
Polite form is objectively more costly to produce – as a function of the form itself (longer, more complex). Only secondarily correlated with the type (maybe in terms of psychological effort to be polite).

- Accents, dialects:

> *THOSE OF US who move from the provinces pay a toll at the city's gate, a toll that is doubled in the years that follow as we try to find a balance between what was so briskly discarded and what was so carefully, hesitantly, slyly put in its place. [...] Did I know, they asked, that my accent and tone, indeed my entire body language, had changed when I met their maid? I was almost a different person. Was I aware that I had, in turn, changed back to the person they had met in Egypt once I was alone with them again? I asked them, did they not speak in different ways to different people? No, they insisted, they did not. Never! They looked at me as if I was the soul of inauthenticity. And then I realized that those of us who move from the periphery to the center turn our dial to different wavelengths depending on where we are and who else is in the room.*
>
> (Colm Tóibín, New York Review of Books, July 13, 2017)

Costly signal:	Absence of costly signal:
$\hat{S}$	S
"marked form"	"unmarked form"
"standard"	"regional dialect/vernacular"

.. it's a cost imposed – not on those coming from the center (as "standard" as it is their "natural" language), but on those coming from the periphery.

Strong component of Class I: it is a real, physical effort to learn the standard, and how well one manages that might well correlate with other unobserved social abilities that are thereby being "signaled" (like attentiveness, willingness and ambition to integrate).

Ilustration: Costly signals in language as a narrative device

SS Colonel Hans Landa is the central character of Quentin Tarantino's *Inglorious Basterds*. Landa is smart. (The tension of the movie comes to a large part from that.) How do we know?

From the first scene—when Landa makes his appearance in front of Mister LaPadite's property and in an effortlessly fluent and elegantly cut-out French invites himself into LaPatite's house, where, after some polite exchange, he switches into an equally refined English. The performance is impressive, not just within the fiction. Tarantino is reported to have been short of calling the movie off because he could not find the right cast for Landa:

> Landa is a linguistic genius, and the actor who played him needed the same facility with language or he would never be what he was on the page. I told my producers, I might have written a part that was unplayable. I said, I don't want to make this movie if I can't find the perfect Landa, I'd rather just publish the script than make a movie where this character would be less than he was on the page. When Christoph came in and read the next day, he gave me my movie back."[1]

[1] Fleming, Michael (May 17, 2009), "Tarantino Reflects On 'Basterds,'" *Variety*.

Structural patterns explained:

- Shaping beliefs. "Indirect speech." When prior is low, $p < 1/2$, partially revealing equilibrium E1 (high always costly signal; low with some probability): Costly signal becomes a means to shape, to the belief of the other, to "push it up," such as to make the other (player 2) indifferent between accepting and not.

Examples:

Police officer: "Isn't it a wonderful day today!"
(= Put your belief that I can be bribed up.)

The car driver, then, has to take the risk of offering a bribe or not.

A: "Would you like to come up for a coffee?"
(= Put your belief that I want to move to a more intimate kind of relationship up.)

B., then, has to take the risk of making a physical move or not.

Remarks:

- E1 welfare-improving over "no-signaling" equilibrium outcome P1. Costly signal has a "positive" social function here.
- Classical refinements and evolutionary dynamics both offer good explanations for stability, or "emergence," of E1:
  - Classical refinements: P1 is not strategically stable under the thought experiment what player 2 should think if "out of the blue" she observes the costly signal, given that under "normal" conditions the costly signal is not used: E1 emerges then.
  - Replicator dynamics: it is a rather strong assumption that players have the probabilities they ought to use in the partially revealing equilibrium E1 exactly right. But it is quite plausible that they circle around the right probabilities (periodic orbits surrounding E1).

- Over- and understatement. When prior is high, $p > 1/2$, both routinely using the costly signal (P2) and routinely not using the costly signal (P3) are strategically and evolutionarily stable equilibrium outcomes.

  P2 represents a social tragedy: everybody needs to signal, but signal carries no information! No-signaling convention (P3) is socially more beneficial here.

  $\longrightarrow$ Possible explanation for why some communities go conventionally by overstatement (P2) and some by understatement (P3).

  Multiplicity of equilibria is not a deficiency of the model, but source of its explanatory potential!

- Indirect discrimination by costly signal. When coordination on P2 or P3 is linked to some other observable characteristic (gender, skin tone): possible source of discrimination.

- Counter-signaling. P3 (no signal – accept), which exists when prior high, in contrast to E1 (no signal – no accept), which exist when prior low, can also be interpreted as "countersignaling."

A linguistic example of counter-signaling:

Chers tous,

J'espère que la reprise n'est pas trop rude !

Il n'y aura pas de conseil de département lundi prochain faute d'un ordre du jour suffisamment étoffé. La DRH nous demande toutefois de faire formellement approuver par le bureau du département le classement des candidats sur le poste LRU en mathématiques que nous avons publié en urgence au mois de juillet. Cette approbation permettra, après avis favorable des conseils centraux, à la personne recrutée de débuter son service au mois d'octobre.

Le classement a été réalisé au mois de juillet par une commission inter-centres présidée par N [...] H [...] Voici le classement : [...]

Amitiés,
Bertrand C [...]

Chers tous,
Je vous souhaite une bonne rentrée. Avis favorable évidemment. Je rajoute qu il y avait 5 candidats et que nous avons auditionné et classé les trois cités. Les deux premiers avaient un poste d'ATER (non renouvelable) chez nous.
Amitiés,
Naila

Chers tous,
avis favorable également!
Bonne journée,
Lucie

Chers tous,
Merci à la commission inter-centres pour ce travail!
Avis favorable,
Claudine

OK
AB

Bonjour à tous,
Je suis favorable également.
Amicalement,
Maria

Bonjour à tous,
avis favorable également. Bonne reprise à tous !
Amicalement,
Fabienne

```
Cher Bertrand (cc:  Chers tous),
Avis favorable également,
Amitiés,
Ali
```

```
Merci à tous pour votre réponse rapide !
Amitiés,
Bertrand
```

Among the seven respondants, one, besides being a member of the board of the department, is the vice-president of the university. Who is that person?

Meaninig in costly-signaling games arises as a function of the entire equilibrium structure—not just one isolated equilibrium.

Appendix A: The index of equilibria

Shapley (1974): Index, $+1$ or -1, to every regular equilibrium

- Strict equilibrium has index $+1$.
- Removing or adding unused strategies does not change the index.
- *Index Theorem*: the sum of the indices of all equilibria is $+1$.

Hofbauer and Sigmund (1988, 1998): index as the sign of the determinant of the negative Jacobian of the replicator dynamics

Ritzberger (1994, 2002): extends this to equilibrium components:

- Index as an integer, such that the sum over all components is again $+1$
- Index robust under payoff perturbations.

Demichelis and Ritzberger (2003):

- If an equilibrium component is asymptotically stable under some evolutionary dynamics, then its index equals its Euler characteristics.If it is convex or contractible, then its index is $+1$.

In our game (based on Hofbauer and Pawlowitsch 2023):

$p < 1/2$:

- E1: Isolated and quasistrict $\longrightarrow$ regular
  - removing unused strategies $\longrightarrow$ 2×2 cyclic game
  - in this game, E1 only equilibrium $\longrightarrow$ index $+1$

  $\Rightarrow$ candidate for asymptotically stable equilibrium
- P1: by Index Theorem $\longrightarrow$ index 0

  $\Rightarrow$ not asymptotically stable, under no evolutionary dynamics

$p > 1/2$:

- P2: by robustness $\longrightarrow$ index $+1$
- E2: Isolated and quasistrict $\longrightarrow$ regular
  - removing unused strategies $\longrightarrow$ 2×2 coordination game with 3 equilibria: E2 and two strict equilibria (index $+1$)
  - by Index Theorem $\longrightarrow$ index -1.
- P3: by Index Theorem $\longrightarrow$ index $+1$

Appendix B: Evolutionary dynamics in costly-signaling games

The Replicator Dynamics (Taylor and Jonker 1978; Hofbauer, Schuster, and Sigmund 1979)

Game played repeatedly in a large population. Growth rate of a strategy proportional to its fitness-difference relative to the average fitness in the population.

For a two-population game:

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

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

The Replicator Dynamics for our game in normal form

Payoffs

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

Where $\mathbf{y} = (y(aa), y(a\bar{a}), y(\bar{a}a), y(\bar{a}\bar{a}))$, a mixed strategy of player 2, and

$y = y(aa) + y(a\bar{a})$
$y' = y(aa) + y(\bar{a}a)$

We observe:

$$u^1(ss) + u^1(\bar{s}\bar{s}) = u^1(s\bar{s}) + u^1(\bar{s}s) \tag{2}$$

Similarly:

$$\begin{aligned} u^2(aa, \mathbf{x}) &= p \\ u^2(a\bar{a}, \mathbf{x}) &= px_h + (1-p)(1-x_\ell) \\ u^2(\bar{a}a, \mathbf{x}) &= p(1-x_h) + (1-p)x_\ell \\ u^2(\bar{a}\bar{a}, \mathbf{x}) &= 1-p \end{aligned} \tag{3}$$

$\mathbf{x} = (x(ss), x(s\bar{s}), x(\bar{s}s), x(\bar{s}\bar{s}))$,

$x_h = x(ss) + x(s\bar{s})$,
$x_\ell = x(ss) + x(\bar{s}s)$

And we observe also that:

$$u^2(aa) + u^2(\bar{a}\bar{a}) = 1 = u^2(a\bar{a}) + u^2(\bar{a}a) \tag{4}$$

Eqs. (2) and (4): for any game with the same extensive form.

Gaunersdorfer, Hofbauer, and Sigmund (1991):

If $u_1 + u_4 = u_2 + u_3$, then $\frac{x_1 x_4}{x_2 x_3}$ is a constant of motion for the replicator dynamics $\rightarrow$ foliation of state space $\Delta_4 \times \Delta_4$ into 4-dimensional invariant manifold.

The 'central' invariant manifold, given by $x_1 x_4 = x_2 x_3$, the *Wright manifold*, can be parameterized:

$$\begin{aligned} x_1 &= x x', \\ x_2 &= x(1 - x'), \\ x_3 &= (1 - x) x', \\ x_4 &= (1 - x)(1 - x'), \end{aligned}$$

with $(x, x') \in [0, 1]^2$: $x = x_1 + x_2, x' = x_1 + x_3$.

On this invariant manifold, the replicator dynamics can be written as:

$$\begin{aligned} \dot{x} &= x(1 - x)(u_1 - u_3) \\ \dot{x}' &= x'(1 - x')(u_1 - u_2) \end{aligned} \tag{5}$$

In our game:

On the 'central' invariant manifold:

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

with $x_h = x(ss) + x(s\bar{s})$, $x_\ell = x(ss) + x(\bar{s}s)$
and $y = y(aa) + y(a\bar{a}), y' = y(aa) + y(\bar{a}a)$:

$$\begin{aligned} \dot{x}_h &= x_h(1 - x_h)(y - c_1 - y')p \\ \dot{x}_\ell &= x_\ell(1 - x_\ell)[y - c_2 - y'](1 - p) \\ \dot{y} &= y(1 - y)[px_h - (1 - p)x_\ell] \\ \dot{y}' &= y'(1 - y')[p(1 - x_h) - (1 - p)(1 - x_\ell)] \end{aligned} \tag{6}$$

This system of differential equations on the hypercube $[0, 1]^4$ can be derived directly from the extensive form, as the

$\longrightarrow$ *replicator dynamics for behavior strategies*.