Towards a Theory of Deception *

David Ettinger†and Philippe Jehiel‡

7th June 2004

Abstract

Deceiving an opponent about one own's cognitive abilities cannot be captured by standard approaches in which players understand the strategy of their opponent perfectly well. We introduce a framework with boundedly rational players to explain deception. Following Jehiel (2003) we assume that players partition the decision nodes of their opponents into analogy classes, and form expectations only about the average reaction function of their opponent over the various nodes of analogy classes; we further differentiate cognitive types according to whether or not the player can distinguish between the types of the opponent. An equilibrium concept is proposed for such environments. Deception arises in our setup because the updating of beliefs is made as if each cognitive type behaved in the same way in the various nodes of an analogy class, thus allowing rational players to induce the wrong belief that they are less sophisticated than they really are (by adopting a behavior that is typical of the average behavior of a less sophisticated type). We illustrate the phenomenon through a variety of applications, first in a zero-sum game, then in a simple concession game. We also show the implication of the approach in a monitoring game and in a negotiation about the pay rise between an employer and employee

1 Introduction

Being smart is helpful for problem solving, but it often hurts being thought of as too smart (such a belief generally triggers undesirable reactions). This

*We would like to thank.... All errors are ours.

†THEMA, Université de Cergy-Pontoise, 33 boulevard du Port, 95011 Cergy-Pontoise cedex, France ; ettinger@eco.u-cergy.fr

‡CERAS-ENPC, 48 boulevard Jourdan, 75014 Paris, France and ELSE-University College London ; jehiel@enpc.fr

very fact in turn creates incentives to build a reputation for being naive even if you are smart. But, the question arises as to whether the person(s) one is interacting with may or may not be deceived by such an (strategic) attitude.

Modern game theory has made significant progress in modelling the idea of reputation (see for example ch. 9 of Fudenberg and Tirole 1991 for an account of major contributions). Yet, by assuming that all agents are perfectly rational, the by now standard approach does not allow the players to be deceived. That is, you may convince your opponent that you will behave in a certain way. But, this does not mean that your opponent is being deceived: in standard equilibrium approaches, you do behave in that way; so the opponent makes no mistake. Besides, the notion of type in the standard approach cannot be interpreted in a cognitive way (unlike the wording "crazy" type might suggest): All types are perfectly rational, and they all understand perfectly well the strategy of the various types of their opponent.[1] Thus, deceiving your environment (including other fellows with whom you may interact) about your cognitive ability cannot be captured by the standard approach.[2]

This paper proposes a formalization of the idea of deception in a framework with boundedly rational players[3]. Specifically, we consider two-players dynamic games with incomplete information in which the type of a player -to be interpreted as a cognitive type- is defined by his ability to represent (or learn) the strategy of his opponent. Following Jehiel (2003) we assume that players partition the decision nodes of their opponents into analogy classes, and form only expectations about the average reaction function of the opponent over their various analogy classes. We further differentiate types

[1]Crazy types in Kreps et al. (1982) and following literature should be interpreted as having payoff specifications other than the ones of the original game, thus making their "mechanical" behavior optimal.

[2]Some might argue that mixed strategy equilibria provide a model of deception with the standard approach: if type A plays action a more frequently than type B does, by playing a type B deceives his opponent about the identity of his type. Yet, type B should play a infrequently for the argument to hold. Thus, deception may occur, but only infrequently and somehow by accident rather than as a result of a deliberate strategy (to justify mixing, player B should be indifferent between playing a and some other action). Crawford (2003) provides a recent illustration of this view of deception.

[3]We define deception as a successful attempt to look of a different cognitive type than one really is. Thisis to be contrasted with Vrij (2001) quoted in Gneezy (2004) who defines deception as a successful or unsuccessful deliberate attempt , without forewarning, to create in another a belief that the communicator considers to be untrue in order to increase the communicator's payoff at the expense of the other side. Vrij's definition puts an emphasis on the lie aspect of the deception whereas we put emphasis on the cognitive process through which the communicator manages to convey a false belief to the other side.

according to whether or not the player can distinguish between the types of the opponent. Thus, cognitive types may vary in two dimensions: A player may be more or less fine on the partition of the decision nodes of his opponent (the analogy part), and a player may or may not distinguish the behaviors of the various types of his opponent (we refer to it as the sophistication part).

We propose an equilibrium concept called the Analogy-based Perfect Bayesian Equilibrium to describe the interaction of players with such limited cognitive abilities. In equilibrium the analogy-based expectations correctly represent the average behavior in the various analogy classes, and players play best-responses to their analogy-based expectations and to their belief about the type of their opponent. As the game proceeds, players update their belief about the type of their opponent according to Bayes' rule as derived from their analogy-based expectations.[4]

A simple interpretation of the solution concept is in terms of learning: In the learning phase, players have a limited access to the database that record all behaviors of all subjects. The cognitive type of a player as defined by his analogy partition and his sophistication regarding the identification of other types' behaviors summarizes his information treatment at the learning stage. For example, a player with a coarse analogy partition keeps track only of the average behavior within each of his classes. And a sophisticated player keeps track of these (average) behaviors type by type. The Analogy-based Perfect Bayesian Equilibrium concept assumes that what subjects try to learn has been learned properly, i.e. it assumes the underlying learning process with information treatment as just described has converged.

Deception is possible in such a setup whenever you may be of several possible cognitive types and your opponent is a *sophisticated coarse* type, by which we mean that your opponent differentiates between the behaviors of your various possible types (the sophisticated part), but she does not distinguish the behaviors at each possible node separately, i.e. she uses a coarse analogy partition and she knows only the average behaviors of your various types in each analogy class (the coarse part).

Deception is possible because your opponent updates her belief as if your various cognitive types behaved in the same way in all nodes of every analogy class. Thus, when you are rational, by playing the action that mimics best the average behavior of the coarse type (this need not even be the action chosen by the coarse type at that period, see the monitoring example developed

[4]The literature in psychology suggests a number of biases related to belief updating (for example the so called base rate fallacy or the conjunction fallacy, see Khaneman et al. (1982) or Thaler (1991)). However, for our theory of deception only the qualitative features of Bayesian updating matter (more weight should be assigned to a type who is perceived to play more frequently according to the observed data).

below) you may induce your opponent to believe that you are coarser than you really are, thereby deceiving your opponent in a way that may turn out to be beneficial in the rest of the game.

Related literature:

The paper can be viewed as proposing a bridge between the literature on psychology and the game theory literature, especially the one interested in modeling the phenomenon of reputation. We have already pointed out the connection/difference with the literature on reputation. Regarding psychology, we note that following the lead of Simon (1956) many researchers have emphasized the role of heuristics as opposed to rational analysis (see Giegerenzer et al. 1989). But, most heuristics discussed in this literature concern behavioral heuristics like "Tit for Tat" or "Take the best" or the recognition heuristic etc (see Gigerenzer and Selten for an introductory presentation of these heuristics). To some extent, the cognitive types in our approach can be viewed as standing for a heuristic used by the players to understand the reaction of their environment; but, note that our cognitive types are better viewed as defining a learning heuristic rather than a behavioral heuristic.[5]

There are many works in economics and psychology that suggest a number of biases relative to the laws of probabilities (for example, the base rate and conjunction fallacies, the law of small numbers, the gambler's fallacy, overconfidence...).[6] Most of these biases are better understood as arising in non-repeated interactions.[7] By contrast, our theory of deception assumes that the underlying interaction is repeated sufficiently many times so that players have learned what their cognitive types allow them to. Our work should thus be viewed as complementary to those works analyzing biases that arise in non-repeated interactions.

The rest of the paper is organized as follows. We illustrate the phenomenon of deception using a simple two-period zero-sum game example in the next Section. In Section 3 we provide a formal framework to describe the Analogy-based Perfect Bayesian Equilibrium in general extensive-form games. In the following three sections of the paper we suggest a number of stylized applications in which the phenomenon may be of relevance: These include a concession game with mediation, a monitoring game and a wage bar-

[5] Selten (ch2 same book) mentions the need to develop theories of expectation formation. Our setup can be viewed as providing one such theory.

[6] While Khaneman et al. (1982) identify a number of these biases, Thaler (1991) also shows their significance in experimental economics. A number of economists have also developed theories motivated by these biases (see, for example, Mullainathan (2002) or Rabin (2002)).

[7] Subjects seem better at understanding frequencies - which arise with repetition - than at manipulating the laws of probabilities.

gaining game between an employer and an employee trying to cheat/deceive the employer about his willingness to make an effort to get another job. All missing proofs can be found in Appendix.

2 Deception in a simple Zero-Sum Game

Before we elaborate on the concept of Analogy-based Perfect Bayesian Equilibrium, we illustrate the idea of deception through a simple two periods example.

Two players, a Row player and a Column player, play twice, in two consecutive periods, a zero-sum stage game, G. In game G the Row player chooses an action U or D, the Column player chooses an action L or R, and stage game payoffs are as represented in Table 1. Players do not discount payoffs between the two periods.

	L	R
U	5, -5	3, -3
D	0, 0	7, -7

Table 1. The stage game G

When players are rational, they play the unique Nash equilibrium of the stage game in every period. The Row player plays U with probability 7/9 and D with probability 2/9 and the Column player plays L with probability 4/9 and R with probability 5/9. The value of the game is 70/9 for the Row player and $-70/9$ for the Column player.

In equilibrium players play in mixed strategies in order to avoid being predictable. But, note that no player is ever deceived by his opponent: Whatever players do in the first period they are expected to play according to the same mixed strategy in the second period, and players do behave according to that expected mixed strategy in period two.

The main contribution of this paper will be to provide a setup in which players may deceive their opponent in equilibrium as a result of the limited cognitive ability of the opponent.

Specifically, there are two types of Row players which are equally likely: The Row player may either be *Rational* or *Coarse* each with probability $\frac{1}{2}$. When he is *Rational* the Row player has a perfect understanding of the strategy of the Column player, as in the standard case. When he is *Coarse* he knows (or learns) only the average behavioral strategy of the Column player over all the various circumstances in which the Column player plays the stage game G. That is, the Row player when coarse bundles the two times periods in which the stage game G is being played and he has only an expectation over the average behavior of the Column player over these two periods.

The Column player can only be of one type. He is assumed to be *Sophisticated* in the sense that he distinguishes between the behaviors of the *Rational* Row player and the *Coarse* Row player. But, he is assumed to be *Coarse* in the sense that for each type of the Row player he knows (or learns) only the average behavior of this type over all games G that this type must play. In short, we say that the Column player is a *Sophisticated Coarse* player.

In the next Section we define formally a solution concept that describes the equilibrium interaction in such a setup. It is viewed as the limiting outcome of a learning process in which the expectation characterizing each cognitive type would eventually be correct. In the present context, we will now check that the following strategy profile is an Analogy-based Perfect Bayesian Equilibrium:

Rational Row Player: Play U in period 1. Play D in period 2.

Coarse Row Player: Play U both in periods 1 and 2.

Column Player (Sophisticated Coarse): Play L in period 1. Play R in period 2 if the Row player played U in period 2. Play L in period 2 if the Row player played D in period 1.

According to this strategy profile, (U, L) is played in period 1, and (D, R) and (U, R) are each played with equal probability in period 2 (depending on whether the Row player is Rational or Coarse).

Thus, the Coarse Row Player should expect the Column Player to play L and R with equal probability on average across the various stage games G.[8] Given his expectation, the Coarse Row player finds it optimal to play U whenever he has to move.[9]

It is also readily verified that the Rational Row player plays a best-response to the Column player' strategy, as he gets an overall payoff of $5 + 7 = 12$.[10]

It remains to explain the behavior of the *Sophisticated Coarse* Column player. Based on the strategy of the Row player, the Column player's expectation about the average behavior of the *Coarse* Row player should be that he plays U, and her expectation about the average behavior of the *Rational* Row player should be that he plays U and D with an equal frequency. Given these expectations, the Column player can update her belief about the type

[8] This is because in period 1, the Column player plays L and in period 2 on the equilibrium path the Column player plays R; thus on average he plays L and R with an equal frequency.

[9] This is because $\frac{1}{2}(5+3) > \frac{1}{2}(0+7)$.

[10] He would only get an overall payoff of $0 + 5$ at best if he were to play D in period 1, and he would obviously get a lower payoff by playing U in period 2.

of the Row player at the end of period 1 (after observing the action played by the Row player in period 1) as follows: When action D is being played in period 1, the Column player should believe that she faces the Rational player with probability 1;[11] When action U is being played in period 1, the Column player should believe that she faces the *Coarse* Row player with probability $\frac{1/2}{1/2+1/2\times 1/2} = \frac{2}{3}$.[12]

Given the above expectation and belief systems of the Column player, we now check why her behavior is optimal using backward induction. In period 2 after action D has been played in period 1, the column player believes that she faces the Rational player with probability 1 and her expectation about this player's behavior is that he plays U and D with an equal probability. It is then optimal for the Column player to play L (as $\frac{1}{2}(-5+0) > \frac{1}{2}(-3-7)$). In period 2 after action U has been played in period 1, the Column player believes that she faces the Coarse player with probability $\frac{2}{3}$ (and this type is expected to play U always) or the Rational player with probability $\frac{1}{3}$ (and this type is expected to play U and D each with probability $\frac{1}{2}$). So overall in period 2 after U has been played in period 1, the Column player expects the Row player to play U with probability $2/3 \times 1 + 1/3 \times 1/2 = 5/6$ and D with probability $1/6$. The Column player chooses optimally to play R in this case (since $\frac{5}{6}(-3) + \frac{1}{6}(-7) > \frac{5}{6}(-5)$). In period 1, the Column player believes it is equally likely that the Row player is Coarse or Rational. Thus, overall in period 1 the Column player expects the Row player to play U with probability $1/2 \times 1 + 1/2 \times 1/2 = 3/4$ and D with probability $1/4$. The Column player optimally chooses to play L in period 1.[13]

To summarize, a Rational Row player gets an overall payoff of 12, a Coarse Row player gets an overall payoff of 8 and the expected payoff of the Sophisticated Coarse Column player is -10. Note that both the Rational Row player and the Coarse Row player get an expected payoff higher than the value of the game which we is $70/9 < 8$. Observe also that the Rational Row player achieves a payoff higher than that the Coarse Row player. This is, of course, no coincidence, since in equilibrium a Rational player can always mimic the strategy of a less sophisticated player, thereby getting at least her payoff (see Proposition 2 below).

More interestingly, the finding that the Row player gets more than his value in the above example is related to the phenomenon of deception that

[11] This is because the Coarse Row player is expected to play U always and the Rational player is expected to play U and D with an equal probability.

[12] This results from the application of Bayes' rule assuming that the Coarse player plays U with probability 1 and the Rational player plays U and D each with probability 1/2.

[13] This is because $\frac{3}{4}(-3) + \frac{1}{4}(-7) < \frac{3}{4}(-5)$ and the strategy of the Row player in period 2 is not perceived by the Column player to depend on the first period action profile.

is caused by the nature of the Column player's cognitive type. Specifically, after action U is being played in period 1, the Column player makes the wrong inference that it is more likely that the Row player is Coarse. This is erroneous, since the correct updating should be that it is equally likely that the Row player is Coarse or Rational (both the Row and the Column player always play U in period 1 and $50-50$ is the prior belief). Yet, it is the updating that (rationally) follows from the limited knowledge of the Column player about the strategy of the Row players. Note that for such erroneous updating to arise, one needs to have cognitive types that distinguish the behavior of the various types of their opponent and yet do not distinguish the behaviors of these types in all possible nodes (as otherwise we would have the standard correct updating).

We believe this captures an essential aspect of the phenomenon of deception: In equilibrium, the Column player is being deceived by the Rational Row player whom by behaving in some way (typically, in a way that is as close as possible to the typical average behavior of the Coarse Row player) makes her believe that the Row player is of a type that he really is not.

Comment 1: Observe that if there were only one type for each player characterized by his analogy partition as in Jehiel (2003), then it would be impossible to reproduce the behavioral strategies as described above: For the Column player to play a different action in periods 1 and 2 she should treat separately the behavior of the Row in the two time periods, but then in period 1 she could not find it optimal to play L given that the Row player always plays U.

Comment 2: It would also be impossible to reproduce the behavioral strategies as described above assuming there is a small fraction of behavioral types behaving in a mechanical way whereas the rest of the players would behave rationally (the so called crazy type approach, see Kreps et al. (1982)). Obviously, in such a setup the Column player when rational could not choose to play L in the first period expecting the Row player to choose U irrespective of his type in this period.

3 A general framework

3.1 The class of games

We consider multi-stage games with complete information. We assume that actions are observable, and we restrict attention to games with two players $i = 1, 2$ (excluding Nature) and a finite number of stages such that, at every stage and for every player (including Nature), the set of pure actions is finite. This class of finite multi-stage games is referred to as Γ.

The standard representation of an extensive form game in class Γ includes the game tree Υ, and the VNM preferences u_i of every player i defined on lotteries over outcomes in the game.

A node in the game tree Υ is denoted by n, and the set of nodes is denoted by N. The set of nodes at which player i must move is denoted by N_i. For every such node $n \in N_i$, we let $A_i(n)$ denote player i's action space at node n. A node n will also be identified with the history h of play till node n. The set of players who must move after history h is denoted $I(h)$. A history following h is referred to as ha where $a \in \times_{i \in I(h)} A_i(h)$ is the action profile played by the players who must move at node h. The set of histories is denoted by H.

Cognitive types and analogy grouping:

We now introduce the idea of cognitive type parameterized by how fine the type knows (or learns) the strategy of his opponent. Specifically, each player i forms an expectation about the behavior of the other player by pooling together several contingencies in which the other player must move. Each such *pool* of contingencies is referred to as a *class of analogy*. Players are also differentiated according to whether or not they distinguish between the behaviors of the various types of their opponent.

Formally, a cognitive type θ_i of player i is characterized by (An_i, δ_i) where An_i stands for player i's analogy partition and δ_i is a dummy variable that stands for whether or not type θ_i distinguishes between the behaviors of the various types θ_j of player j. We let $\delta_i = 1$ when type θ_i distinguishes between types θ_j's behaviors and $\delta_i = 0$ otherwise.

Following Jehiel (2003), type θ_i 's analogy grouping An_i is defined as a partition of the set N_j of nodes where player j must move into subsets α_i referred to as analogy classes.[14] When n and n' are in the same analogy class α_i, it is required that $A_j(n) = A_j(n')$. That is, in two nodes n and n' that player i treats by analogy, the action space of player j should be the same. The common action space in the analogy class α_i will be denoted by $A(\alpha_i)$.

The set of types θ_i is denoted by Θ_i and the profile of type space is denoted by $\Theta = \Theta_1 \times \Theta_2$.[15]

Strategic environment:

[14] A partition of a set X is a collection of subsets $x_k \subseteq X$ such that $\bigcup_k x_k = X$ and $x_k \cap x_{k'} = \emptyset$ for $k \neq k'$.

[15] Observe that Θ is finite in our setup with finite node and action spaces.

A strategic environment in our setup is described by (Υ, u_i, p) where p denotes the prior joint distribution on the type space $\Theta = \Theta_1 \times \Theta_2$. To simplify notation we will assume that the types of the two players are independently distributed from each other, and we will refer to $p_i(\theta_i)$ as the prior probability that player i is of type θ_i.

3.2 Concepts

Analogy-based expectations:

An analogy-based expectation for player i of type θ_i is denoted by β_{θ_i}. It specifies for every analogy class α_i of type θ_i of player i a probability measure over the action space $A(\alpha_i)$ of player j. Types θ_j of player j are distinguished in this expectation according to whether $\delta_i = 1$ or 0. If $\delta_i = 1$, types θ_j are distinguished, and β_{θ_i} can be viewed as a function of θ_j and α_i where $\beta_{\theta_i}(\theta_j, \alpha_i)$ should be interpreted as type θ_i-player i's expectation about the average behavior in class α_i of player j when his type is θ_j. If $\delta_i = 0$, player i merges the behaviors of all types θ_j of player j, and β_{θ_i}can be viewed as a sole function of α_i where $\beta_{\theta_i}(\alpha_i)$ should be interpreted as player i's expectation about the average behavior in class α_i of player j (where the average is taken over all possible types). We let $\beta_i = (\beta_{\theta_i})_{\theta_i}$ denote the analogy-based expectation of player i for the various possible types $\theta_i \in \Theta_i$.

Strategy:

A behavioral strategy for any player assigned to the role of player i is denoted by s_i. It is a mapping that assigns to each node $n \in N_i$ at which player i must move a distribution over player i's action space at that node.[16] We let σ_{θ_i} denote the behavioral strategy of type θ_i, and for every $n \in N_i$ we let $\sigma_{\theta_i}(n) \in \Delta A_i(n)$ denote the distribution over $A_i(n)$ according to which player i of type θ_i selects actions in $A_i(n)$ when at node n. We let $\sigma_{\theta_i}(n)[a_i]$ be the corresponding probability that type θ_i plays $a_i \in A_i(n)$, and we let $\sigma_i = (\sigma_{\theta_i})_{\theta_i}$ denote the strategy of player i for the various possible types θ_i. We let σ denote the strategy profile of all players.

Belief system:

When player i distinguishes the types of player j, i.e. $\delta_i = 1$, he needs to hold a belief about the type of his opponent and this belief may vary from one node to another. Formally, we let μ_{θ_i} denote the belief system of player

[16]Mixed strategies and behavioral strategies are equivalent since we consider games of perfect recall.

i of type $\theta_i = (An_i, \delta_i)$. We let $\mu_{\theta_i}(\theta_j)(h)$ be the probability that player i of type θ_i assigns to the event "player j is of type θ_j" conditional on the history h being realized.

The above notion of belief system is indispensable when $\delta_i = 1$. For the case $\delta_i = 0$, it is not indispensable, but to save on notation when defining the notion of best-response (see below) we assume that in this case player i 's belief coincides with the prior p_j throughout the game. We call μ_i the belief system of player i for the various possible types θ_i, and we let μ be the profile of belief systems for all players.

Sequential rationality:

Based on his analogy-based expectation, player i of type θ_i constructs a strategy of player j. That strategy requires that in all nodes n of the analogy class α_i player j behaves according to the expectation hold by player i about the average behavior in class α_i. The constructed strategy may thus depend on the type θ_j of player j according to whether $\delta_i = 1$ or 0. At every node where he must play, player i is assumed to play a best-response to this constructed strategy of player j (where the best-response is defined relative to the belief system when $\delta_i = 1$).

Formally, we define the β_{θ_i}-perceived strategy of player j, $\sigma_j^{\beta_{\theta_i}}$, as

$$\begin{aligned} \text{If } \delta_i &= 1 \quad \sigma_{\theta_j}^{\beta_{\theta_i}}(n) = \beta_{\theta_i}(\theta_j, \alpha_i) \quad \text{for every } n \in \alpha_i \text{ and } \theta_j \in \Theta_j \\ \text{If } \delta_i &= 0 \quad \sigma_{\theta_j}^{\beta_{\theta_i}}(n) = \beta_{\theta_i}(\alpha_i) \quad \text{for every } n \in \alpha_i \text{ and } \theta_j \in \Theta_j \end{aligned}$$

Given the strategy s_i player i and given history h, we let $s_i\mid_h$ denote the continuation strategy of player i induced by s_i from history h onwards. We also let $u_i^h(s_i\mid_h, s_j\mid_h)$ denote the expected payoff obtained by player i when history h has been realized, and players i and j behave according to s_i and s_j, respectively.

Definition 1 *(Criterion) Player i's strategy σ_i is a sequential best-response to (β_i, μ_i) if and only if for all θ_i such that $p_i(\theta_i) > 0$, for all strategies s_i and all nodes $n \in N_i$,*[17]

$$\sum_{\theta_j \in \Theta_j} \mu_{\theta_i}(\theta_j)(n) u_i^h(\sigma_{\theta_i}\mid_n, \sigma_{\theta_j}^{\beta_{\theta_i}}\mid_n) \geq \sum_{\theta_j \in \Theta_j} \mu_{\theta_i}(\theta_j)(h) u_i^h(s_i\mid_n, \sigma_{\theta_j}^{\beta_{\theta_i}}\mid_n)$$

Consistency:

[17] Remember that the node n is identified with the history h that leads to node n.

In equilibrium, we require two notions of consistency: one that relates the analogy-based expectations to the strategy profile, and one that relates the belief systems to the analogy-based expectations.

We start with the consistency of the analogy-based expectations. We require the analogy-based expectations to correspond to the real average behaviors in every considered class and for every possible type (if types are differentiated) where the weight given to the various elements of an analogy class must itself be consistent with the real probabilities of visits of these various elements.

The consistency requirement should be thought of as a result of a learning process in which players would eventually manage to have correct analogy-based expectations (see below for a more detailed interpretation). To present formally the consistency idea, let $P^\sigma(\theta_i, \theta_j, n)$ denote the probability that node n is reached when players i and j are of types θ_i and θ_j respectively, and players play according to σ.

Definition 2 *Player i's analogy based expectation β_i is consistent with the strategy profile σ if and only if:*

- *For any $(\theta_i, \theta_j) \in \Theta$ such that $\delta_i = 1$: for all $\alpha_i \in An_i$,*

$$\beta_{\theta_i}(\theta_j, \alpha_i) = \frac{\sum_{(\theta_i',n) \in \Theta_i \times \alpha_i} p_{\theta_i'} P^\sigma(\theta_i', \theta_j, n) \cdot \sigma_{\theta_j}(n)}{\sum_{(\theta_i',n) \in \Theta_i \times \alpha_i} p_{\theta_i'} P^\sigma(\theta_i', \theta_j, n)}$$

  whenever there exist θ_i' and $n \in \alpha_i$ such that $P^\sigma(\theta_i', \theta_j, n) > 0$.

- *For any $\theta_i \in \Theta$ such that $\delta_i = 0$: for all $\alpha_i \in An_i$,*

$$\beta_{\theta_i}(\alpha_i) = \frac{\sum_{(\theta_i',\theta_j',n) \in \Theta \times \alpha_i} p_{\theta_i'} p_{\theta_j'} P^\sigma(\theta_i', \theta_j', n) \cdot \sigma_{\theta_j'}(n)}{\sum_{(\theta_i',\theta_j',n) \in \Theta \times \alpha_i} p_{\theta_i'} p_{\theta_j'} P^\sigma(\theta_i', \theta_j', n)}$$

  whenever there exist θ_i', θ_j' and $n \in \alpha_i$ such that $P^\sigma(\theta_i', \theta_j', n) > 0$.

The consistency of the analogy-based expectations should be thought of as a result of a learning process. Specifically, assume that there are populations of players i and j who are repeatedly and randomly matched to play the game. In the population of players i, there is a fraction $p_i(\theta_i)$ of players of type θ_i. After the end of a session, the behaviors of all the players and their types are revealed. All the information is gathered in a general data set, but players have different access to this data set depending on their types. A player i with type $\theta_i = (An_i, \delta_i)$ such that $\delta_i = 0$ has access to the average empirical distribution of behavior in every analogy class $\alpha_i \in An_i$ where the average is taken over all nodes $n \in \alpha_i$ and over the entire population of players j. A player with type $\theta_i = (An_i, \delta_i)$ such that $\delta_i = 1$ has access to the average empirical distribution of behavior in every $\alpha_i \in An_i$ for each subpopulation of types θ_j of players j.

Now suppose that the true pattern of behavior adopted by the players is that described by the strategy profile σ. A player i with type $\theta_i = (An_i, \delta_i)$ such that $\delta_i = 1$ will collect data about the average behavior of types θ_j in every class $\alpha_i \in An_i$ as soon as sometimes a player j with type θ_j reaches some node $n \in \alpha_i$ (according to σ). In the long run, every such statistics should converge (in the Cesaro' sense) and the limit point should be an average of what player j with type θ_j actually does in each of the nodes n where $n \in \alpha_i$, that is, $\sigma_{\theta_j}(n)$. The weighting of $\sigma_{\theta_j}(n)$ should also coincide with the frequency with which n is visited (according to σ) relative to other elements in α_i, hence the above expression for $\beta_{\theta_i}(\theta_j, \alpha_i)$. A similar argument applies when $\delta_i = 0$ for the expression of $\beta_{\theta_i}(\alpha_i)$.

Remark 1 *Definition 2 places no restrictions on player i's expectations about those analogy classes that are not reached according to σ. A stronger notion of consistency would require that the expectations in this case correspond to limits of expectations that would be consistent with small perturbations of σ. (Such a notion is in the spirit of sequential equilibria - see Kreps and Wilson 1982 - and is discussed in Jehiel (2003) in a simpler context. We have chosen to present the weaker notion of consistency for expositional purposes).*

Remark 2 *In the above learning story we have assumed that there was a common pool of data. An alternative specification would be that types θ_i of players i have access only to those plays where player i was of type θ_i. This would lead to alternative notions of consistency,[18] but the spirit of the examples discussed in the paper would continue to hold under these alternative specifications.*

[18] Specifically, if $\delta_i = 1$:

$$\beta_{\theta_i}(\theta_j, \alpha_i) = \frac{\sum_{n \in \alpha_i} P^\sigma(\theta_i, \theta_j, n) \cdot \sigma_{\theta_j}(n)}{\sum_{n \in \alpha_i} P^\sigma(\theta_i, \theta_j, n)}$$

whenever there exists $n \in \alpha_i$ such that $P^\sigma(\theta_i, \theta_j, n) > 0$.

If $\delta_i = 0$

$$\beta_{\theta_i}(\alpha_i) = \frac{\sum_{(\theta'_j, n) \in \Theta_j \times \alpha_i} p_{\theta'_j} P^\sigma(\theta_i, \theta'_j, n) \cdot \sigma_{\theta'_j}(n)}{\sum_{(\theta'_j, n) \in \Theta_j \times \alpha_i} p_{\theta'_j} P^\sigma(\theta_i, \theta'_j, n)}$$

whenever there exist θ'_j and $n \in \alpha_i$ such that $P^\sigma(\theta_i, \theta'_j, n) > 0$.

We now move on to the second consistency requirement that relates the belief systems of players to their analogy-based expectations. The analogy-based expectation β_{θ_i} of player i with type $\theta_i = (An_i, \delta_i)$, $\delta_i = 1$ allows him to distinguish between the behaviors of players j with different types. As the game proceeds, player i updates his belief about the type of player j using Bayes' rule (whenever it is applicable) and assuming that player j behaves according to $\sigma_{\theta_j}^{\beta_{\theta_i}}$ as defined above. Formally,

Definition 3 *Player i's belief system μ_i is consistent with the analogy based expectation β_i if and only if : for any $(\theta_i, \theta_j) \in \Theta$ such that $\delta_i = 1$*

$$\mu_{\theta_i}(\theta_j)(\emptyset) = p_j(\theta_j).$$

And for all histories h, ha

$$\mu_{\theta_i}(\theta_j)(ha) = \mu_{\theta_i}(\theta_j)(h) \text{ whenever } h \notin N_j$$

$$\mu_{\theta_i}(\theta_j)(ha) = \frac{\mu_{\theta_i}(\theta_j)(h)\sigma_{\theta_j}^{\beta_{\theta_i}}(h)[a_j]}{\sum_{\theta'_j \in \Theta_j} \mu_{\theta_i}(\theta'_j)(h)\sigma_{\theta'_j}^{\beta_{\theta_i}}(h)[a_j]}$$

whenever $h \in N_j$, there exists θ'_j s.t. $\sigma_{\theta'_j}^{\beta_{\theta_i}}(h)[a_j] > 0$ and player j plays a_j at h.

Comment: The consistency of the belief system μ_i with the analogy-based expectation β_i should be thought as a result of an introspective calculus of player i. Based on his representation of the strategy of his opponent for the various possible types he makes inference (using Bayes' law) as to the likelihood of the various possible types he is facing. This should be contrasted with our learning interpretation of the consistency requirement for the analogy-based expectations (see above Definition 2).

Solution concept:

In equilibrium, we require that, at every node, players play best-responses to their analogy-based expectations (sequential rationality) and that expectations and beliefs are consistent.

Definition 4 *A strategy profile σ is an Analogy-based Perfect Bayesian Equilibrium if and only if there exist analogy-based expectations β_i and belief systems μ_i such that for every player i:*

1. σ_i is a sequential best-response *to (β_i, μ_i),*
2. β_i is consistent *with σ and*
3. μ_i is consistent *with β_i.*

An Analogy-based Perfect Bayesian Equilibrium conceptually differs from a Perfect Bayesian Equilibrium with incomplete information in several important respects. In particular, the types in our setup are not characterized by their preferences and their information partitions, but by their cognitive abilities to understand (or learn) the strategy of their opponents. This is a totally new notion of types that cannot be interpreted with the standard approach.[19] In particular, note that the limited capability of players to learn the correct strategy of the various types of their opponent in turn leads them to make erroneous inferences about the identity of the types they are facing.[20] The exploitation of these erroneous inferences give rise to the possibility of deception (see Section 2 for an example of this). In the next Sections, we discuss a number of settings in which the phenomenon of deception arises.

3.2.1 Preliminary results

We first note that an equilibrium always exists:

Proposition 1 *In finite environments, there always exists at least one Analogy-based Perfect Bayesian Equilibrium.*

Proof 1 *The proof follows standard methods, first noting the existence of equilibria in which each player i is constrained to play any action $a_i \in A_i(n)$ at any node $n \in N_i$ with a probability no less than ε, and then by showing that the limit as ε tends to 0 of such strategy profiles is an Analogy-based Perfect Bayesian Equilibrium.* Q. E. D.

We next observe that in an environment in which some types are rational, these perform better than other cognitive types in equilibrium:

Proposition 2 *Consider an Analogy-based Perfect Bayesian Equilibrium of an environment in which one of the types of player i is rational*[21]*. Then this*

[19] Even when each player i can be of only one cognitive type characterized by his analogy partition, Jehiel (2003) notes that an Analogy-based Expectation Equilibrium cannot be viewed as a standard Bayes-Nash equilibrium of another game with modified information structure. A fortiori when there are several possible cognitive types, an Analogy-based Perfect Bayesian Equilibrium cannot be interpreted as a standard equilibrium of another game with a modified information structure.

[20] In this respect, observe that our notion of consistent belief system relies on Bayes' rule. The possibility of erroneous inferences would a fortiori hold if players did not use Bayes' rule.

[21] A rational type is characterized by an analogy partition that is finest. Whether he can or cannot differentiate the various types of his opponent ($\delta_i = 1$ or 0) is irrelevant when he has the finest analogy partition.

type of player i gets the highest equilibrium expected payoff among all types of player i.

Proof 2 *The rational type of player i may always mimic the behavior of any other type θ_i of player i, thereby ensuring that he can get at least the expected payoff obtained by any other type.* Q. E. D.

Comment: This should be contrasted with results suggesting that irrational types may perform better in equilibrium. Here it is a comparison of the equilibrium payoffs obtained by different types within the same equilibrium. It is not a comparison of equilibrium payoffs of the rational types vs the irrational ones when one switches from an environment with only rational types to an environment with only irrational types.

4 A Concession Game

4.1 Description of the game

Two risk-neutral agents negotiate over the division of a pie of size 400. The negotiation is represented by a series of reciprocal concessions with the possible intervention of an external mediator. A player, when it is his turn to make a decision, must choose between two options. He can either concede one fourth of the pie to the other player or ask for the intervention of an external mediator. If he chooses to concede and some share of the pie has not been conceded yet, in the next period, it is the other player's turn to face the same choice.

If a player asks for the intervention of the mediator, this mediator splits what has not been conceded yet into two equal shares given to the two players. Besides, the player who asked for his intervention must pay a commission fee, c, equal to 7% of the share of the pie that the mediator has to split.

The game ends when the pie is completely distributed among the two players either through reciprocal concessions, in the fourth period, or after the intervention of the mediator. When a player asks for the intervention of the mediator, we say that he *opts out.* Player 1 has to make decisions in periods 1 and 3, and player 2 in periods 2 and 4. If the game ends with player 1 opting out in period k, payoffs are equal to $(200 - 7(5 - k), 200)$. If it ends with player 2 opting out in period k, payoffs are equal to $(150, 250 - 7(5 - k))$. If players make reciprocal concessions until the complete allocation of the pie without the intervention of the mediator, payoffs are equal to $(200, 200)$. For simplicity, we normalize payoffs by subtracting 150 from player 1 's payoffs and 200 from player 2's payoffs (the end node corresponding payoffs are thus

$(50-7(5-k),0)$, $(0,50-7(5-k))$ and $(50,0)$). The game is represented in figure 1.

The concession game described in figure 1 is a variant of the centipede game (Rosenthal 1982). With rational players, the game has a unique subgame perfect Nash equilibrium. Both players opt out whenever they have the opportunity to. This is the worst possibility for players 1 and 2 in terms of the share of the pie given to the mediator. Player 1 pays the mediator 28, the highest possible commission fee. Player 1 does not concede in the first period since he anticipates that, if he does, player 2 will call the mediator in the next period.

4.2 The game played by boundedly rational agents

We consider boundedly rational agents who bundle nodes into analogy classes as follows. With probability $\frac{1}{3}$, player 1 is *Coarse* i.e. $An_1 = \{\{n_2, n_4\}\}$, he puts n_2 and n_4 in the same analogy class. Whether $\delta_1 = 0$ or $\delta_1 = 1$ does not matter since player 2 is assumed to be of only one type. With probability $\frac{2}{3}$, player 1 is *Rational,* i.e. $An_1 = \{\{n_2\}, \{n_4\}\}$.[22] With probability 1, player 2 is *Coarse,* i.e. $An_2 = \{\{n_1, n_3\}\}$, he puts n_1 and n_3 in the same analogy class, and *Sophisticated*, i.e. $\delta_2 = 1$, player 2 distinguishes the behaviors of the two possible types of player 1.

Proposition 3 *The following strategy profile is an Analogy-based Perfect Bayesian Equilibrium:*

Player 2 concedes in n_2 and opts out in n_4. Player 1 when rational concedes in n_1 and opts out in n_3. Player 1 when coarse concedes in both n_1 and n_3.

Remark. The strategy profile in which players opt out whatever their types whenever they have to move is also an equilibrium. Together with the equilibrium shown in Proposition 3 these are the only equilibria in pure strategies (see Appendix).

The equilibrium outcome can be summarized as follows: with probability $\frac{2}{3}$ the mediator is called at n_3; with probability $\frac{1}{3}$ he is called at n_4. The expected fee paid to the mediator is equal to $\frac{28}{3}$.

It is important to realize that for concessions to take place in the game of figure 1 it is vital that player 1 can be of several types and that player 1 distinguishes between the behaviors of the two types of player 1. If players

[22]It is irrelevant whether $\delta_1 = 1$ or 0 when the analogy partition is the finest.

1 and 2 were each of one type, it would be impossible to observe concessions in equilibrium however these types are specified.[23]

The crucial part of the equilibrium is about understanding the reasoning of the coarse player 1 at n_3. We will also elaborate on the reasoning of player 2 at node n_2.

The coarse player 1 at node n_3 believes that player 2 will concede with probability 3/4 at the next node and hence he concedes at n_3.[24]

Player 2 concedes at node n_2 based on the following reasoning. Player 1 when rational is perceived to concede with probability 1/2 on average.[25] Player 1 when coarse is perceived to concede always. Thus, when player 2 sees player 1 conceding at his first decision node he updates his belief about the type of player 1 is facing using Bayes' rule and assuming that the behaviors of the various types of player 1 follow the perceptions just described: his updated belief is that player 1 is coarse with probability $\frac{(1/3)1}{(1/3)1+(2/3)(1/2)} = \frac{1}{2}$. Combining with the perceptions of the various types' behaviors, player 2 believes at node n_2 that player 1 will concede with probability $(\frac{1}{2})(\frac{1}{2})+(\frac{1}{2})(1) = \frac{3}{4}$ in n_3. His perceived expected payoff if he concedes in n_2 is $(\frac{3}{4})43 = \frac{129}{4}$, which is strictly higher than 29, his payoff if he opts out in n_2. Conceding in n_2 is a best response to his belief.

Several observations are in order. First, note that the updated belief of player 2 about player 1's type at node n_2 is different from the correct Bayesian updated belief according to which at n_2 the probability that player 1 is coarse should coincide with the prior 1/3 (all types concede at n_1). The rational player 1 deceives player 2 by conceding at node n_1. He makes player 2 wrongly believe that he is more likely to be the coarse type than he really is. This is turn induces player 2 to concede at node n_2 because he is sufficiently confident that player 1 is coarse, thereby always conceding.

[23]It is easy to see that no equilibrium in pure strategy with concessions can be sustained whatever the analogy partitions. The types most favorable to concessions are the coarsest types. But, whatever his belief player 2 opts out at n_4. Thus, in a pure strategy equilibrium, the most optimistic belief that player 1 may have is that player 2 concedes with probability 1/2 on average. With such a belief, player 1 chooses to opt out at n_3. But, then player 2's belief about 1 is that he concedes at best with probability 1/2. This in turn induces player 2 to opt out at n_2 (because $29 > 43/2$). Hence, player 1 's belief is that player 2 opts out always, and he must opt out at n_1. It is left to the reader to check that there is no equilibrium in mixed strategy either.

[24]At the equilibrium, player 2 always concedes in n_2 reached with probability 1 and opts out in n_4 reached with probability $\frac{1}{3}$ (because rational player 1 opts out at n_3 and the probability that player 1 is rational is 2/3).

[25]At node n_1 he concedes, at node n_3 he opts out and these two nodes are met with the same frequency.

Second, observe that if player 2 did not distinguish between the types of player 1 (i.e. $\delta_2 = 0$) he could not be deceived about player 1's type, and also at node n_2 he would opt out rather than conceding.[26]

Taking advantage of this latter observation, consider now the effect of adding a small proportion of Coarse non Sophisticated players 2. It is readily verified that the following strategy profile would be an equilibrium. For all players other than the coarse non-sophisticated player 2, they would behave as described in Proposition 3. The coarse non-sophisticated player 2 would opt out at node n_2. In this equilibrium, the coarse non-sophisticated player 2 would perform better than the the coarse sophisticated player 2 (the former gets 29 while the latter gets only $\frac{43}{3}$), thus illustrating that more sophistication need not be beneficial.

5 The monitoring game

5.1 Presentation of the game

We consider a monitoring game with an employer and an employee. In three consecutive periods, the employee has to choose between working, W, or shirking, S. In the first period, the employee has entire discretion (he is not controlled). Then, the employer decides whether or not to control the employee in the last two periods. At $t = 0$, the employee makes his first decision concerning period 1. If the employee works in period 1, he gets 0 and the employer gets 1. If the employee shirks in period 1, he gets 1 and the employer gets 0. At date $t = 1$, the employer observes the decision made by the employee in period 1. The employer can choose either to control (C) the employee or to delegate (D) the decision power to the employee for the last two periods. At periods $t = 2$ and $t = 3$, the employee chooses between working or shirking.

If, at $t = 1$, the employer chooses C, he is sure to obtain a payoff of 2 from the employee's actions in the last two periods.[27] If the employee tries to shirk in period 2 and 3, it will be costly for him but not for the employer. The employee pay-off is strictly decreasing in the number of times he shirks from period 2 to 3. He gets 2 if he works twice, 1 if he works once and shirks once and 0 if he shirks twice. An interpretation of the control technology is

[26] By definition of his type he would not make any updating about the type of his opponent. He would believe that player 1 concedes on average with probability 2/3 and $29 > \frac{86}{3}$.

[27] This holds whatever the employee does in these two periods and independently from what he obtained in the first period.

that it is such that the employee always fulfills his task. If he shirks, he is punished and eventually does what he should do.

If, at $t = 1$, the employer chooses D, his payoff is strictly increasing in the number of times the employee works in the last two periods. If the employee shirks twice, the employer gets 0, if he shirks once and works once, the employer gets 2 and if he works twice, the employer gets 3. The pay-off of the employee is strictly decreasing in the number of times he works in periods 2 and 3: he gets 1 if he never shirks, 2, if he shirks once and 4 if he shirks twice. The game is represented in figure 2.

Observe that except if the employee decides to shirk when he is controlled[28], the sum of players' payoffs is equal to 5. With fully rational agents, this game has a unique subgame perfect Nash equilibrium. In this equilibrium, the employee shirks in period 1, the employer chooses to control him and the employee works in the last two periods. The employee gets 3 and the employer gets 2.

5.2 The game played by boundedly rational agents

We consider two types of employees. *Coarse* employees who put in the same analogy class all the decision nodes in which the employer has to make a decision and *Rational* employees who use two analogy classes, one for each decision node of the employer. The employee is Coarse with probability 2/3 and Rational with probability 1/3.[29]

The employer is a *Sophisticated Coarse* type. That is, he has a unique analogy class that contains all the decision nodes of the employee. Besides, $\delta_2 = 1$, i.e. he distinguishes between the behaviors of the two different types of employees.

We first explain why the strategy profile of the Subgame Perfect Nash equilibrium is not an equilibrium in this setup. By contradiction, assume it is an equilibrium. Then the belief of the employer should be that the employee works with probability 2/3. But, with such a belief, the employer would choose to Delegate and not to Control after the employee's decision to Shirk at $t = 0$ (he would choose D and not C because $(\frac{2}{3})^2 \times 3 + (\frac{1}{3})^2 \times 0 + 2(\frac{1}{3})(\frac{2}{3}) \times 2 > 2$), violating the prescription of the Subgame Perfect Nash equilibrium. As it turns out there is a unique equilibrium in pure strategies:

Proposition 4 *The game has a unique equilibrium in pure strategies. At $t = 0$, the employee shirks when Coarse and works when Rational. In the*

[28] In this case, we consider that the employee is punished which destroys a fraction of the created added value.

[29] Whether $\delta_1 = 0$ or 1 is irrelevant, since the employer can be of only one type.

last two periods, the employee, whatever his type and his behavior at $t = 0$, shirks if the employer chose D and works if the employer chose C at $t = 1$. The employer chooses D if he observes that the employee worked in period 1 and C if he observes that the employee shirked in period 1.

At the equilibrium, if the employee is Rational, he works in the first period, the employer chooses D and the employee shirks in periods 2 and 3. If the employee is Coarse, he shirks in the first period, the employer chooses C and the employee works in periods 2 and 3. A Rational employee gets 4, a Coarse employee gets 3 and the expected payoff of the employer is $5/3$. Compared to what happens when all agents are rational, a Rational employee obtains a higher revenue and the Coarse employer gets a lower expected payoff.[30] This difference is due to the employer's decision to delegate to the Rational employee. This strategy is costly for the employer. In the rest of this Section we elaborate on why the employer chooses this strategy.

First, observe that, in equilibrium, a Coarse employee works twice and shirks once while a Rational employee works once and shirks twice. Since the employer puts in the same analogy class all the decision nodes of the employee, he perceives a Coarse employee to be a (relatively) *working employee* and a Rational employee to be a (relatively) *shirking employee*. When he chooses between D and C, the employer cares about the type of the employee insofar as it is indicative of whether the employee is perceived to be working or shirking.

The most interesting aspect of this example is about the updated belief of the employer after observing the employee's action at date $t = 0$. When the employer observes that the employee works in the first period, he puts more weight on the probability that the employee is the *working type* (while he is, with probability 1, a Rational employee, i.e. a *shirking type*). The employer chooses to delegate, i.e. D, because he is sufficiently confident that the employee will work next with a high probability. Symmetrically, when the employee shirks in the first period, the employer puts more weight on the probability that the employee is a *shirking type* (while he is, with probability 1, a Coarse employee of the *working type*). He chooses to control.

Interestingly, we see that a Rational employee, in the first period, in order to have the employer believe that he is Coarse -that is, to be a relatively *working employee* as perceived by the employer- behaves differently from a Coarse employee. He mimics the most frequent behavior of a Coarse employee which differs, in this node, from the actual behavior of a Coarse employee.

It remains to explain the behavior of the employee at date $t = 0$. A Coarse employee puts the two decision nodes of the employer in the same

[30] The Coarse employee gets the same payoff as in the rational paradigm!

analogy class. He does not perceive that the employer choice may depend on his own decision at $t = 0$. That is why he decides to shirk in period 1. A Rational employee perceives that by working in the first period, he will deceive the employer who will believe that he is more likely to be of the *working type*. Besides, it is worth losing 1 in the first period to obtain that the employer chooses to delegate next, thereby making an extra profit of 2 in the last two periods (2 is equal to the difference between what he gets if the employer chooses D and he shirks twice and what he gets if the employer chooses C and he works twice). This explains why a Rational employee does work in the first period.

We believe that this example captures important features of the deception process. In equilibrium, the rational employee manages to impose a false belief about his type. The employer is deceived by the Rational employee who behaves in the first period in a way that the Coarse employer associates with the Coarse employee (he works). The employer subsequently chooses to delegate the decision power and gets a lower outcome than what he would get by controlling the employee. What is striking here is that a Coarse employee does not even behave that way in the first period, he shirks. A Rational agent in order to be identified with a Coarse agent follows the most frequent behavior of a Coarse agent even though at this specific decision node a Coarse agent would not behave that way. We recognize here standard swindlers' stratagems. The swindler tries initially to build a confidence relationship with his prey. To do so, in the first interactions, he follows an excessively *honest* behavior (even a standard *honest* agent would not behave that way). The coarse prey infers from this behavior that the agent he is facing is *honest*. He drops his guard and the swindler takes advantage of it in the following periods. The swindler's strategy relies on the coarseness of his prey who wrongly interprets his initial extreme honesty.[31] A rational prey would rightly interpret this excessively honest behavior of the swindler in the initial periods, "too good to be true" or "too nice to be honest", and would not believe in the *honesty* of the swindler. But, a boundedly rational prey as modelled in this paper is deceived. The example illustrates how incentives may be profoundly affected in the presence of boundedly rational agents.

[31]This phenomenon is well exposed in many movies such as "The House of Games" (1987) by David Mamet, "The Sting" (1973) by George Roy Hill or "The Color of Money" (1986) by Martin Scorsese (*honest* is replace here by *bad pool player*).

6 A wage negotiation game

6.1 Description of the game

We consider the following wage negotiation between a professor and the dean of his department. At $t = 1$, the professor chooses between accepting the status quo (SQ) and developing contacts with another university (D) in view of an alternative faculty position in another department. Establishing these contacts costs him $\gamma > 0$. If he develops contacts, the professor asks for a pay rise Δ (> 0) to the dean. At $t = 2$, the dean decides either to refuse (R) or to accept (A) the pay rise. If the dean accepts, the professor stays in the department and the negotiation process is over, the professor ends up with a higher wage and stays in his original position. If the dean refuses, at $t = 3$, the professor chooses again between accepting the status quo (SQ), staying in his department with his initial salary or developing further contacts (D) with the other department at cost γ, getting from it an alternative offer. If the professor chooses the second option, he goes back to the dean, exhibits his alternative offer and asks for a pay rise Δ' (> 0).[32] At $t = 4$, the dean decides whether he accepts (A) the pay rise Δ' or refuses it (R). The professor stays in the department if the dean accepts the pay rise or leaves it and goes to the other department if the dean refuses it. If the professor accepts the offer of the other department, the original department incurs a cost $-X$ and the professor gets $U - 2\gamma$.[33]

We normalize payoffs so that in the original situation both the department and the professor have a pay-off 0. We further assume that $X < \Delta'$ and $U < \gamma$, and to fix ideas, we let $\Delta = 3$, $\Delta' = 4$, $\gamma = 1$, $X = \frac{7}{2}$ and $U = \frac{1}{2}$. In a perfect rationality world, $X < \Delta'$ implies that at $t = 4$, the dean prefers to let the professor go rather than accept the pay rise. Given that there is no pay rise at $t = 4$, $U < \gamma$ implies that at $t = 3$ the professor does not find it useful to generate an outside offer of U for an extra cost γ. Anticipating that no further search effort will be made by the professor, the dean at $t = 2$, finds it optimal not to accept a pay raise. Finally, at $t = 1$, the Professor does not develop contacts because he anticipates no pay rise will be accepted. The game is represented in figure 3.

In the standard rationality paradigm, even though the Professor has the possibility to go for another job, the outside option is perceived as non-credible and there is no pay rise.

[32] We have in mind that $\Delta' > \Delta$ so that the new pay rise compensates at least partially for the extra search cost.

[33] U is equal to the value of the alternative offer minus the costs the professor incurs leaving his university.

6.2 The game played by boundedly rational agents

We wish to illustrate now why the limited cognitive abilities of the Professor and the Dean may explain the possibility of search activity and pay rise in this model in which the outside option of the Professor would traditionally be thought of as non-credible.

Specifically, the distribution of players' cognitive types is the following. With probability 1/2, the dean is *Sophisticated Coarse*, $An = \{\{n_1, n_3\}\}$, he puts in the same analogy class both decision nodes of the professor and $\delta = 1$, he distinguishes between the various types of the professor. With probability 1/2, the dean is *Rational*, $An = \{\{n_1\}, \{n_2\}\}$ and $\delta = 1$.

With probability 1/2, the professor is *Coarse*, $An = \{\{n_2, n_4\}\}$, he puts in the same analogy class both decision nodes of the dean and $\delta = 0$, he does not distinguish between the various types of the dean. With probability 1/2, the professor is *Rational*, $An = \{\{n_2\}, \{n_4\}\}$ and $\delta = 1$.

The following proposition illustrates the possibility of search activity and pay rise in equilibrium:

Proposition 5 *A Sophisticated Coarse dean accepts the pay rise in n_2 and refuses to give a pay rise in n_4. A Rational dean always refuses to give a pay rise. A Coarse professor establishes contacts with an alternative department whenever he has the opportunity to and a Rational professor establishes contacts in n_1 and accepts the status quo in n_3.*

Remark : The Subgame Perfect Nash equilibrium strategy profile (described earlier) is also an equilibrium. Together with the strategy displayed in Proposition 5 these are the only equilibria in pure strategies (see the Appendix).

In the equilibrium of Proposition 5, the professor always makes an effort to get an alternative offer at $t = 1$. With probability $\frac{1}{2}$, the dean concedes a pay rise, with probability $\frac{1}{4}$, the professor leaves his department and, with probability $\frac{1}{4}$, he stays in his department with his initial salary.

The logic of the equilibrium is as follows. First, it is readily verified that the behaviors of the rational professor and the rational dean are optimal (given the behaviors of other players). So let us focus on the Coarse Professor and the Sophisticated Coarse dean.

A Coarse professor puts n_2 and n_4 in the same analogy class. His perception of the dean's behavior in n_4 is *contaminated* by the dean behavior in n_2. Because the dean accepts a pay rise with a positive probability at $t = 2$, a Coarse professor perceives that he accepts it with a probability equal to $\frac{2}{5}$ $(= \frac{1/2}{1+1/4})$ at $t = 4$. For such a high probability of acceptance, it is worthwhile developing further contacts at $t = 3$ since $\frac{2}{5}(2) + \frac{3}{5}(\frac{-3}{2}) = \frac{-1}{10} > -1$. It is also readily verified that a Coarse Professor finds it optimal to develop outside contacts at $t = 1$.

Now, let us consider the Sophisticated Coarse dean. He perceives that a Coarse professor always chooses D both in n_1 and in n_3 and a Rational professor chooses D with a probability $\frac{2}{3}$ in these two nodes[34]. At $t = 2$, he observes that the professor chose D in n_1. After having observed the professor's behavior in n_1, the Sophisticated Coarse dean's updated belief is that the Professor is Coarse with probability $\frac{(1/2)1}{(1/2)1+(1/2)(2/3)} = \frac{3}{5}$. Combining the perceived behaviors and beliefs, the Sophisticated Coarse dean expects the Professor to choose D in n_3 with probability $(\frac{3}{5})1 + (\frac{2}{5})\frac{2}{3} = \frac{13}{15}$. According to this expectation, the dean prefers to accept the pay rise in n_2.[35]

It should be noted that the dean accepts the pay rise here because he puts a sufficiently high probability on the Professor being coarse due to this erroneous belief updating. If the dean had kept the prior belief that the Professor is Coarse with probability $\frac{1}{2}$ (which is actually correct at node n_2), he would have chosen not to accept the pay rise at n_2.[36]

For rational agents, the presence of boundedly rational agents has two main consequences. First, at the start of the interaction, Rational professors mimic Coarse professors and develop efforts toward the outside university. That way, they deceive Coarse deans, and make them believe that they are facing a Coarse professor with a high probability. Coarse deans accept the pay rise at $t = 2$ because they are sufficiently afraid that the professor will otherwise leave (after making extra search effort). Second, Rational deans fail to be identified as Rational deans by Coarse professors at $t = 3$. Coarse professors do not perceive that there exist two types of deans. Thus, even though a Rational dean behaves differently from a Coarse dean at $t = 2$, a Coarse professor keeps on believing, at $t = 3$, that the dean will concede with probability $\frac{2}{5}$ at $t = 4$. In this case, Rational deans would prefer being identified as what they are : Rational deans who never accept pay rises. Coarse professors would then choose the status quo at $t = 2$ and the rational dean would get 0 rather than $\frac{-7}{4}$ (expected payoff in the second equilibrium with boundedly rational agents). This illustrates that it may be costly for

[34]In the first node, reached with probability 1, a Rational professor chooses D and in n_3, reached with probability $\frac{1}{2}$, a Rational professor chooses SQ. Therefore, a Sophisticated Coarse dean perceives that a Rational Professor chooses D with a probability $\frac{1+0}{1+1/2} = \frac{2}{3}$ in the analogy class gathering n_1 and n_3.

[35]This is because $-3 > \frac{13}{15}(\frac{-7}{2}) + \frac{2}{15}(0)$.

[36]He would have perceived that his payoff obtained by not accepting the pay rise is $\frac{-35}{12}$ $(= (\frac{1}{2} \times 1 + \frac{1}{2} \times \frac{2}{3})(\frac{-7}{2}) + (1 - (\frac{1}{2} \times 1 + \frac{1}{2} \times \frac{2}{3}))(0))$ and he would not have accepted the pay rise (since $\frac{-35}{12} > -3$).

a rational agent not to be distinguished from other types due the cognitive limitations of other players.

7 Appendix

7.1 Proof of Proposition 3

Proof :

First, we prove that no other strategies than the one we mention in Proposition 3 can be part of an equilibrium.

It is a strictly dominant strategy for player 2 to take in n_4, therefore, at the equilibrium, player 2 always take in n_4. By backward induction, we can infer that at the equilibrium, if player 1 is Rational, he always takes in n_3.

Suppose that player 2 takes in n_2. If player 1 is Rational, he plays the unique best response, take in n_1. Besides, since player 2 takes in n_2, he always takes when he is in the analogy class $\{n_2, n_4\}$. Then either player 1 always takes in n_1 or player 1, if he is Sophisticated Coarse, perceives that player 2 always takes when he is either in n_2 or in n_4. Taking in n_1 and n_3 is also his unique best-response to his belief.[37] If player 2 takes in n_2, player 1, whatever his type is, has a unique best response, take in n_1 and in n_3.

Now, suppose that player 2 passes in n_2. A Rational player 1 has a unique best response, pass in n_1. Suppose that a Sophisticated Coarse player 1 takes in n_1. Player 2 passes in n_2 and never has the opportunity to play in n_4. Therefore, whenever nodes n_2 or n_4 are reached, player 2 passes. A Sophisticated Coarse player expects that player 2 will pass with probability 1 in n_2 and n_4. Taking in n_1 is not a best response to such a belief and cannot be an equilibrium strategy for a Sophisticated Coarse player 1. Now, suppose that a Sophisticated Coarse player 1 passes in n_1 and takes in n_3. Again, n_4 is never reached and player 2 passes in n_2. Since he puts n_2 and n_4 in the same analogy class, a Sophisticated Coarse player 1 expects that player 2 will pass with probability 1 in n_4. Taking in n_3 is not a best response to such a belief.

There only remains two pairs of strategies that can be part of an equilibrium:

- Players take whenever they have the opportunity to.

- Player 1 passes in n_1 whatever his type is. He takes in n_3, if he is Rational and passes, if he is Sophisticated Coarse. Player 2 passes in n_2 and

[37] We should also notice that there exists no belief about the average behavior of player 2 in $\{n_2, n_4\}$ such that taking in n_1 and passing in n_3 would be a best response to this belief for a Sophisticated Coarse player 1.

takes in n_4.

Suppose that players choose to take whenever they have the opportunity to. This is an equilibrium with player 1 believing that player 2 takes with probability 1 in the analogy class $\{n_2, n_4\}$[38] and player 2 believing that player 1 passes with probability 1 in the analogy class $\{n_1, n_3\}$ whatever his type is. Strategies are best responses to beliefs and beliefs are consistent with strategies. This is the first type of equilibrium that we mention.

Let us consider the second type of equilibrium. First, if player 1 is Rational and player 2 follows the defined strategy, it is clearly a best-response for him to pass in n_1 and to take in n_3. Now, if player 1 is Sophisticated Coarse and players follows the defined strategies, he perceives that player 2 passes with a probability $\frac{1}{1+1/3} = \frac{3}{4}$ in the analogy class $\{n_2, n_4\}$. Therefore, in n_3, he perceives that his expected payoff, if he passes, is equal to $(\frac{3}{4})50$ which is strictly higher than 36, his payoff if he takes. Hence, passing is a best-response to his belief. In n_1, he also perceives that player 2 passes with a probability $\frac{3}{4}$ in n_2 and n_4. His perceived expected payoff if he passes -knowing that he will pass if n_3 is reached- is $(\frac{9}{16})50$ which is strictly higher than 22, his payoff if he takes. The strategy of a Sophisticated Coarse player 1 is a best response to the unique analogy based expectation strategy of player 2 that is consistent with its observed behavior.

Now, let us consider player 2. First, taking in n_4 is a strictly dominant strategy. Furthermore, if players follows the defined equilibrium strategies, player 2 perceives that a Rational player 1 passes with a probability $\frac{1}{2}$ in the analogy class $\{n_1, n_3\}$ and a Sophisticated Coarse player 1 passes with probability 1 in this same analogy class. In n_2, conditional on having observed that player 1 passed in n_1, he believes that player 1 is Sophisticated Coarse with probability $\frac{1\times(1/3)}{1\times(1/3)+(1/2)\times(2/3)} = \frac{1}{2}$ (application of the Bayes' rule with the analogy based expectation strategies of player 1 perceived by player 2). Then, in n_2, player 2 perceives that player 1 will pass in n_3 with a probability $(\frac{1}{2})1 + (1)\frac{1}{2} = \frac{3}{4}$. His perceived expected payoff, if he passes is $(\frac{3}{4})43$ which is strictly higher than 29, his payoff if he takes.

Q.E.D.

7.2 Proof of Proposition 4

First, a Coarse employee puts in the same analogy class the two nodes in which the employer has to make a decision. He perceives that his decision in the first node does not affect the decision made by the employer. Therefore, at the equilibrium, he always shirks in the first period.

[38] Or in n_2 and n_4 if player 1 is Rational.

Second, if the employer chooses C, there is a unique best-response for the employee whatever his type and belief are: To work in both final periods. Conversely, if the employer chooses F, there is a unique best-response for the employee whatever his type and belief are: To shirk in both final periods.

Therefore, to find an equilibrium, we only need to determine what will be the decision of a Rational employee at $t = 0$ and the decision of the employer.

Suppose that the employer chooses D when he observes that the employee shirks in the first period. A Rational employee has a unique best-response: Always to shirk. In that case, the employee shirks three times whatever his type is. Therefore, the employer perceives that the employee always shirks and choosing D is not a best response to such a belief. Hence, choosing D when the employee shirks cannot be part of an equilibrium strategy.

Suppose now that the employer chooses C both if the employee works and if the employee shirks in the first period. The employee, whatever his type is, has a unique best response: to shirk in the first period and to work in the last two periods. The employer perceives that an employee, whatever his type is, when he has to make a decision, shirks with a probability 1/3 and works with a probability 2/3. C is not a best response to such a belief. Hence, choosing C in both cases cannot be part of an equilibrium either.

The only remaining possibility for the employer is to choose C when he observes that the employee shirks in the first period and to choose D when he observes that the employee works in the first period. A Rational employee has a unique best response to such a behavior, to work in the first period (we already found the other elements of the employee's strategy).

Now, to check if this is an equilibrium, we need to establish if it exists a belief consistent with these behaviors such that the employer behavior is a best response to this belief.

If players follow the described behaviors, a Coarse employee shirks once and works twice and a Rational employee shirks twice and works once. Since the employer puts in the same analogy class all the nodes in which the employee has to make a decision, he considers that a Coarse employee chooses to work with a probability 2/3 and a Rational employee chooses to work with a probability 1/3 when they have to make a decision. Therefore having observed that an employee works in the first period, he revises his belief and considers that he is Coarse with probability: $\frac{(2/3)(2/3)}{(2/3)(2/3)+(1/3)(1/3)} = \frac{4}{5}$. If the employer observes that the employee shirks in the first period, he believes that he is Coarse with probability $\frac{(1/3)(2/3)}{(1/3)(2/3)+(2/3)(1/3)} = \frac{1}{2}$. Besides, since the employer puts in the same analogy class all the employee's nodes, he thinks that, in the last two periods, a Coarse (resp: Rational) employee works twice with a probability 4/9 (resp: 1/9), shirks once and works once with a probability 4/9 (resp: 4/9) and shirks twice with a probability 1/9 (resp: 4/9).

If we cross beliefs regarding the type of the employee and the average behavior of employee of each type, we obtain the following. After having observed that the employee shirks (resp: works) in the first period, the employer believes that he will shirk twice with a probability $\frac{25}{90}$ (resp: $\frac{16}{90}$), work once and shirk once with a probability $\frac{40}{90}$ (resp: $\frac{40}{90}$) and work twice with a probability $\frac{25}{90}$ (resp: $\frac{34}{90}$). With such beliefs, if the employer observes that the employee works in the first period, he prefers choosing D and if he observes that the employee shirks in the first period, he prefers choosing C.

Q.E.D.

7.3 Proof of Proposition 5

We first prove that these strategies are constitutive of equilibria.

The first equilibrium. Suppose that the dean always refuses to give a pay rise to the professor. Then whatever his type is, it is a best response for the professor to always accept the status quo since $-\frac{3}{2} < -1 < 0$. Whether the professor gathers in the same analogy class n_2 and n_4 does not matter since the dean behaves the same way in these two nodes if they are reached. Now, if the professor never develops contacts with another department, it is a best response for the dean never to concede a pay rise since $-3 < 0$ and $-4 < -\frac{7}{2}$. Again, whether the dean gathers in the same analogy class n_1 and n_3 does not matter since the professor behaves the same way in these two nodes when they are reached.

The second equilibrium. Suppose that a Coarse professor chooses D in n_1 and n_3, a Rational professor chooses D in n_1 and SQ in n_3, a Sophisticated Coarse dean chooses C in n_2 and R in n_4 and a Rational professor chooses R in n_2 and n_4. First, let us remark that the strategies of the Rational professor and dean are clearly best response to the strategy of the other agents. Now, a Coarse professor perceives that the dean concedes with a probability $1/3$ in the analogy class $\{n_2, n_4\}$ and a best response to this belief is to choose D in n_3 and n_1 since $-1 < -1 - 1 + (1/3)4 + (2/3)(1/2)$ and $0 < -1 + (1/3)3 + (2/3)[-1 + (1/3)4 + (2/3)(1/2)]$. Besides, a Sophisticated Coarse dean chooses R in n_4 since it is a dominant strategy. Now, if agents plays according to the defined strategy, a Sophisticated Coarse dean believes that, in the analogy class $\{n_1, n_3\}$, a Coarse professor always chooses D and a Rational professor chooses D with a probability $2/3$. When n_2 is reached, a Sophisticated Coarse dean perceives that the professor is Coarse with a probability $\frac{1\times 1/2}{1\times 1/2+2/3\times 1/2} = \frac{3}{5}$. Then, he perceives that the professor will choose D in n_3 with a probability $\frac{3}{5} \times 1 + \frac{2}{5} \times \frac{2}{3} = \frac{13}{15}$. If the dean concedes in n_2, he gets -3, if he refuses, his perceived expected pay-off is $-\frac{13}{15} \times \frac{7}{2} = -\frac{91}{30} < -3$ so that his perceived best response is to concede in n_2.

We now prove that no other equilibrium can exist.

First, whatever his type is, it is a strictly dominant strategy for the dean to choose R in n_4. Therefore, choosing C in n_4 cannot be part of an equilibrium. Now, by backward induction, we can infer that at the equilibrium, if the professor is Rational, choosing D in n_3 cannot be part of an equilibrium either.

Suppose that a Coarse professor chooses D in n_1 and SQ in n_3. A Rational dean has a unique best response, to choose R in n_2. Besides, a Rational professor always chooses SQ in n_3. Then, a Sophisticated Coarse dean, whatever his strategy is, perceives that a professor, whatever his type is, chooses SQ with a probability at least 1/3 in the analogy class $\{n_1, n_3\}$. The Sophisticated Coarse dean has a unique best response to this belief, to choose R in n_2. Now, the behavior of the Coarse professor is not a best response to his belief since he perceives that the dean always passes in the analogy class $\{n_2, n_4\}$. Therefore, there cannot exist an equilibrium in which the Coarse professor chooses D in n_1 and SQ in n_3.

Suppose that a Coarse professor chooses SQ in n_1. A Rational dean has a unique best response, to choose R in n_2. Besides, a Rational professor always chooses SQ in n_3. Then, whatever his behavior is, a Sophisticated Coarse dean perceives that a professor, whatever his type is, chooses SQ with a probability at least 1/3 in the analogy class $\{n_1, n_3\}$. The Sophisticated Coarse dean has a unique best response to this belief, to choose R in n_2. The behavior of the Coarse professor is a best response to his belief since he perceives that the dean always passes in the analogy class $\{n_2, n_4\}$ and he will also choose SQ in n_3. The Rational professor has unique best-response, to choose SQ in n_1 and the Rational dean's best response is to choose R in n_2. This is the first type of equilibrium that we mention.

Suppose that a Coarse professor chooses D in n_1 and D in n_3 and the Rational professor chooses SQ in n_1. The best response of the dean, Sophisticated Coarse or Rational, is C in n_2. Then, choosing SQ in n_1 cannot be an equilibrium strategy for a Rational professor. Hence, there cannot exist n equilibrium in which a Coarse professor chooses D in n_1 and D in n_3 and the Rational professor chooses SQ in n_1.

Suppose that a Coarse professor chooses D in n_1 and D in n_3 and the Rational professor chooses D in n_1. The best response of the Rational dean is R in n_2 since a Rational professor chooses SQ with probability 1 in n_3. Now, what will a Sophisticated Coarse dean do in n_2? If he chooses R in n_2, a Rational professor is better off choosing SQ in n_1 so that it cannot be an equilibrium. If a Sophisticated Coarse dean concedes in n_2, then he has the following belief. In the analogy class $\{n_1, n_3\}$, a Coarse professor chooses D with probability 1 and a Rational professor chooses D with probability 2/3. Now, when n_2 is reached, a Sophisticated Coarse dean perceives that the professor is Coarse with a probability $\frac{1\times1/2}{1\times1/2+2/3\times1/2} = \frac{3}{5}$. He perceives that the professor will choose D in n_3 with a probability $\frac{3}{5}\times 1 + \frac{2}{5}\times\frac{2}{3} = \frac{13}{15}$. If the dean concedes in n_2, he gets -3, if he refuses, his perceived expected pay-off is $-\frac{13}{15}\times\frac{7}{2} = -\frac{91}{30} < -3$ so that his perceived best response is to concede in n_2.

Q.E.D.

References

[1] Crawford, V.P. (2003): 'Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions', *American Economic Review*, **93**, 133-149.

[2] Fudenberg D and J. Tirole (1991): *Game theory*, MIT Press.

[3] Gigerenzer, G. and Selten, R. (2002): *Bounded Rationality: The Adaptive Toolbox*, MIT Press.

[4] Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. and Krüger, L. (1989): *The Empire of Chance. How Probability Changed Science and Everyday Life*, Cambridge University Press.

[5] Gneezy, U. (2004): 'Deception : The rolle of consequences', Working Papers Chicago Graduate School of Business.

[6] Jehiel, P. (2003): 'Analogy-Based Expectation Equilibrium', Working Paper Ceras.

[7] Kahnemann, D., Slovic, P. and Tversky, A., eds (1982): *Judgement Under Uncertainty, Heuristic and Biases*, Cambridge University Press.

[8] Kreps, D. , P. Milgrom, J. Roberts and R. Wilson (1982): 'Rational cooperation in the finitely repeated prisoners' dilemma', *Journal of Economic Theory*, **27**: 245-252.

[9] Kreps, D. and R. Wilson (1982): 'Sequential Equilibria', *Econometrica*, **50**: 863-894.

[10] Mullainathan, S. (2002): 'A Memory-Based Model of Bounded Rationality', *Quaterly Journal of Economics*, **117**: 735-774.

[11] Rabin, M. (2002): 'Inference by Believers in the Law of Small Numbers', *Quaterly Journal of Economics*, **117**: 775-816.

[12] Reny, P. (1993): 'Common belief and the theory of games with perfect information', *Journal of Economic Theory*, **59**: 257-274.

[13] Simon, H.A. (1956):'Rational Choice and the Structure of Environment', *Psychological Review*, **63**: 129-138.

[14] Thaler, R.H., (1991): *Quasi Rational Economics*, Russell Sage.

[15] Vrij, A. (2001): *Detecting Lies and Decit*, Wiley.

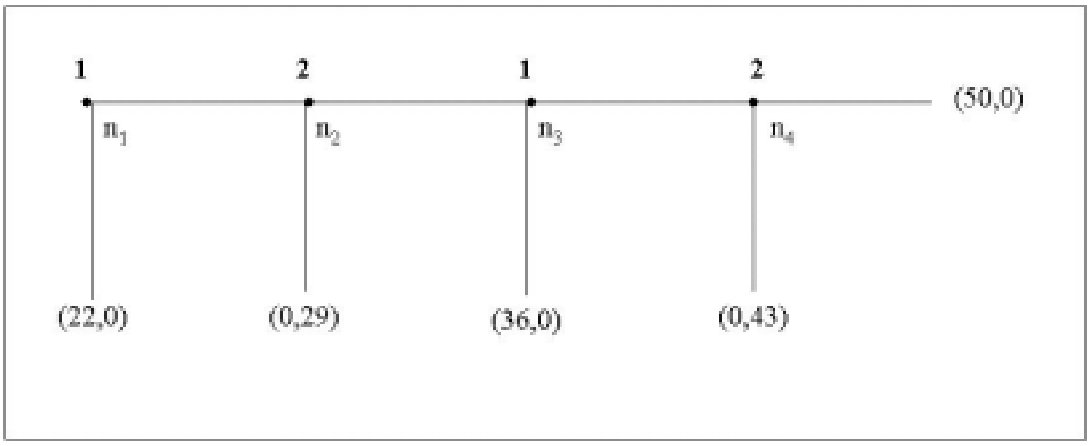

Figure 1: The concession game.

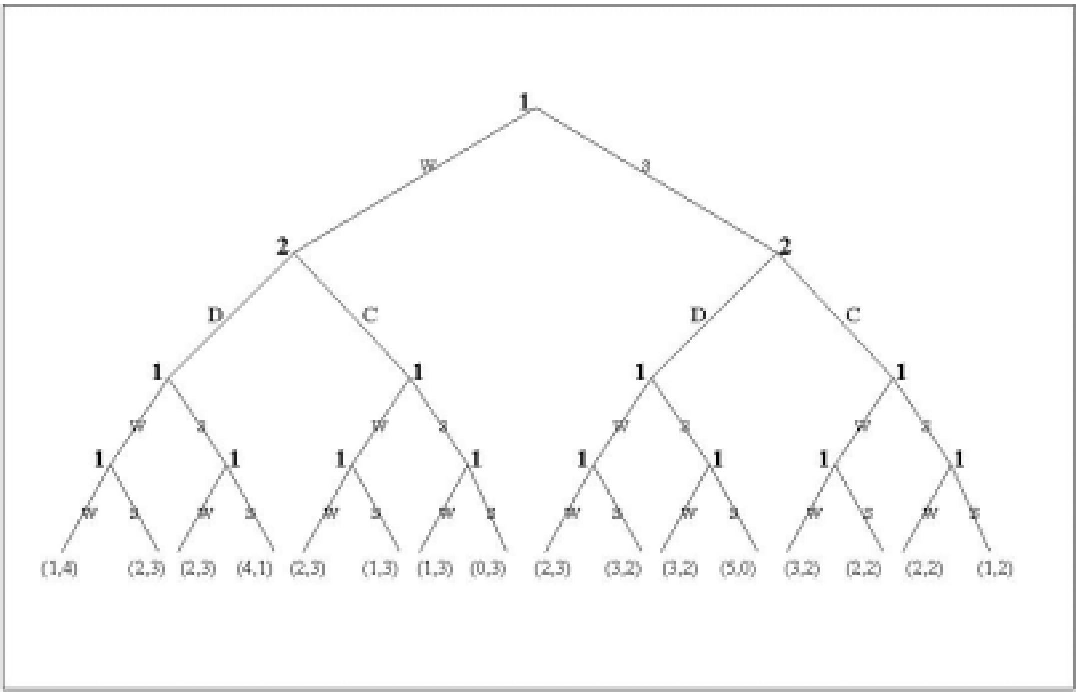

Figure 2: The monitoring game.

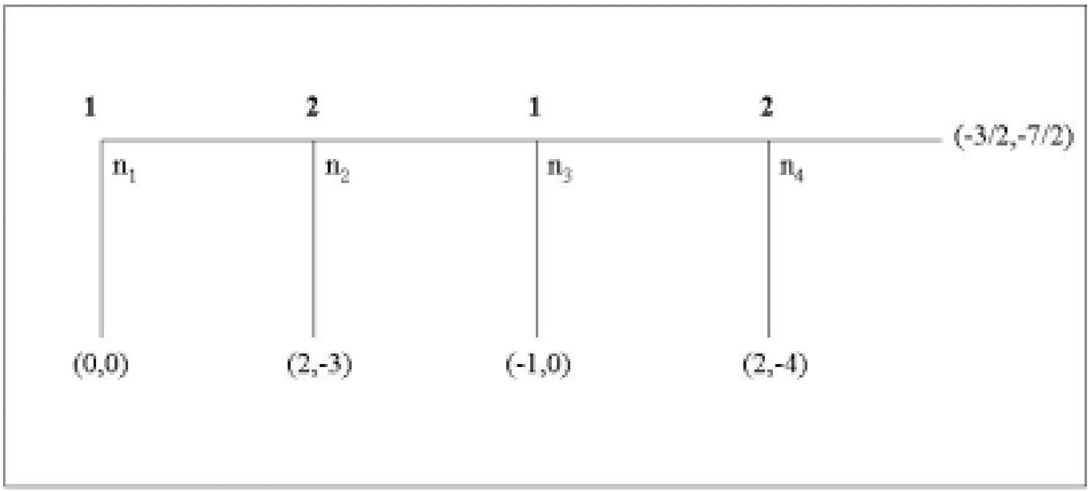

Figure 3: The wage negotiation game.