# An Experiment on Deception, Credibility and Trust*

David Ettinger† and Philippe Jehiel‡

November 15, 2014

**Abstract**

We report results from an experiment on a repeated sender/receiver game with twenty periods in which one of the periods has higher weight, the sender communicates about the realized state in each period, the receiver takes an action matching her belief about the state, and then learns whether the sender lied. Receivers are matched either with malevolent (human) senders who prefer the agents to take *wrong* decisions or with benevolent (machine) senders who always tell the truth. Our findings do not support the predictions of the Sequential Equilibrium. The deceptive tactic in which malevolent senders tell the truth up to the key period and then lie at the key period is used much more often than it should and it brings higher expected payoff. We suggest that our data are well organized by the analogy-based sequential equilibrium (ABSE) in which (some subjects) may reason coarsely when making inferences and forming expectations about others' types and behaviors.

*We thank Maxim Frolov for assistance on the experimental design, Guillaume Hollard, Frederic Koessler, Jean Marc Tallon, and the participants to the Neuroeconomic Workshop, the Extensive form Games in the Lab Workshop, the IHP, Dauphine, Cerge-EI, HEC-Polytechnique, LSE-UCL workshop seminars' participants for helpful comments. Jehiel thanks the European Research Council for funding.

†LEDa and CEREMADE, Université Paris-Dauphine ; david.ettinger.fr@gmail.com

‡PSE, 48 boulevard Jourdan, 75014 Paris, France and University College London ; jehiel@enpc.fr

*Any false matter in which they do not say a bit of truth at its beginning does not hold up at its end.*

RASHI, *Comments on Numbers*, XIII, 27.

*Adapted from Talmud Bavli Tractatus Sotah 35a.*

# 1 Introduction

During World War II, the Red Orchestra was the most important spying network of the Soviet Union in Western Europe. One of the major concerns of such a network was about maintaining a secret communication channel with Moscow while preserving the security of its members. The solution chosen was to organize the Red Orchestra with several almost independent cells. Each cell had its own radio transmitter which was almost its only way to communicate with Moscow. After the discovery of the importance of the Red Orchestra network and the quality of the data sent to Moscow, the German counter-spying services decided to attack it through its weakest point: the communication system. With *high-tech* goniometric instruments and some luck, the German counter-spying services managed to detect several radio transmitters. Besides, thanks to tricks and torture, they captured the members of the cells connected to these radio transmitters. They even convinced some of them to work for them.

Then, began a new strategy for the German counter-spying services: the *Funkspiel*. Rather than interrupting the information transmission from the radio transmitters that they had identified, they kept on using them to send information to Moscow. Besides, not only did the German counter-spying services sent information but they even sent accurate and important pieces of information.

One can guess that the idea of the German spies was to maintain a high level of trust from the Russian services in the quality of the Red Orchestra information (because Moscow also knew that radio transmitters could be detected) and to use this communication network to intoxicate the Russian services at a key moment.[1]

[1]The Funkspiel did not quite succeed. Leopold Trepper, the leader of the Red Orchestra, which had been imprisoned and pretended to cooperate with the Gestapo services managed to send a message to the Russian

The case of the Red Orchestra is a vivid example of a repeated information transmission game in which the organization sending the information may have objectives other than the organization receiving the information. We believe that there are many environments with similar characteristics. The key strategic considerations in such environments are: 1) How does the receiver use the information she gets to assess the likelihood of the current state of affairs but also to assess the type of sender she is facing (which may be useful for future interactions)? and 2) How does the sender understand and make use of the receiver's inference process? On the receiver's side, it requires understanding how trust or credibility evolves and, on the sender's side, it requires understanding the extent to which deception or other manipulative tactics are effective.

This paper proposes an experimental approach to shed light on deception, credibility, and trust. Specifically, we summarize below results found in experiments on a repeated information transmission game (à la Sobel (1985)). Senders and receivers are randomly matched at the start of the interaction. Each pair of sender/receiver plays the same stage game during twenty periods. In each period, a new state is drawn at random, the sender is perfectly informed of the state of the world while the receiver is not. The sender sends a message regarding the state of the world, then the receiver must choose an action. The receiver takes an action that matches her belief about the state of the world (using a standard quadratic scoring rule objective). The sender is either benevolent and always sends a truthful message or he is malevolent in which case his objective is to induce actions of the receiver as far as possible from the states of the world. A receiver ignores the type of the sender she is matched with, but she discovers at the end of each period whether the message received during this period was truthful or not (in the Red Orchestra case, Moscow services could test reasonably quickly the accuracy of the information sent). Besides, one of the periods, the 5th period, has a much higher weight than the other periods both for the sender and the receiver (in the Red Orchestra case, this would coincide, for instance, with an information affecting a key offensive). In our baseline treatment, the key period has a weight 5 times as much as the other periods and the initial share of benevolent senders (implemented as machines in the lab) respresents 20 % of all senders. We considered several variants either increasing the

Headquarter explaining that the Gestapo services controlled the radio transmitters. He even managed to escape later on. For more details on the Red Orchestra, see Trepper (1975) or Perrault (1967).

weight of the key period to 10 or reducing the share of benevolent senders to 10%.[2]

We observed the following results :

- A large share of senders (typically larger than the share of benevolent senders) chooses the following deceptive tactic: they send truthful messages up to the period just before the key period (as if they were benevolent senders) and then send a false message in the key period. The share of deceptive tactics followed by malevolent senders is roughly the same whether the initial proportion of benevolent senders is 10% or 20% and whether the weight of the key period is 5 or 10.

- Receivers are (in aggregate) deceived by this strategy. In the key period, they trust too much a sender who has only sent truthful messages until the key period (i.e., they choose an action which is too close to the message sent **as compared to what would be optimal to do DAVID J'AI REALISE QUE L ON N' AVAIT PLUS DE STATISTIQUE A CE SUJET NI DE GRAPHE. JE PROPOSE PLUS LOIN D AJOUTER QQUECHOSE.** ). In the data we have collected, Receivers get on average a lower payoff against malevolent senders who follow the deceptive tactic than against the other malevolent senders who do not. The deceptive tactic is successful.

These observations are not consistent with the predictions of the sequential equilibrium of this game for several reasons. First, senders follow the deceptive tactic too often.[3] Second, the deceptive tactic is successful in the sense that in our data deceptive senders obtain higher pay-off than non deceptive senders while sequential equilibrium would predict that all employed strategies should be equally good. Third, while the sequential equilibrium would predict that the share of senders following the deceptive tactic should increase if the weight of the key period increases and/or if the initial proportion of benevolent senders increase, we see no such comparative statics in our data.

[2]We also imposed in most treatments that malevolent senders should send a total number of lies equal to either 9, 10, 11 so that we can focus on the timing of lies rather than on whether some subjects are less willing to lie than others. We also ran a control treatment in which the number of lies were free and did not observe very different findings (except that they were a little noisier).

[3]Indeed, it cannot be part of a sequential equilibrium that the share of deceivers exceed the share of truthful senders as otherwise if receivers had a correct understanding of the strategy employed by senders (as the sequential equilibrium requires), malevolent senders would be strictly better off telling the truth at the key period.

As an alternative to the sequential equilibrium, we suggest referring to the analogy-based sequential equilibrium (ABSE) developed in Ettinger and Jehiel (2010) (see Jehiel 2005 for the exposition of the analogy-based expectation equilibrium in complete information settings on which EJ build). In that approach, agents may differ in how finely they understand the strategy of their opponents. In a simple version, receivers would only understand the aggregate lie rate of the two types of senders (benevolent or malevolent) and would reason (make inference as to which type they are facing and what the recommendation implies about the current state of the world) as if the two types of senders were playing in a stationary way over the twenty periods (so as to reflect that receivers have no clue how the lie rate depends on history). If senders were all rational, they would optimally exploit the mistaken inference process of receivers. As a result, malevolent (rational) senders would follow a deceptive tactic and such a tactic would result in a higher payoff than any non-deceptive strategy. Note that the conclusion would hold true for a wide range of weights of the key periods and a wide range of initial probabilities of being matched with benevolent or malevolent senders. The reason for this theoretical finding is simple. In aggregate, it turns out that malevolent senders must be lying with probability 50%. Based on this aggregate lie rate and given the assumed coarse reasoning of receivers, if the truth is told up to the key periods, a receiver thinks that she is facing a benevolent sender with a very large probability, which can then be exploited by a malevolent sender in the key period.

Our data do not fit with the prediction of the above ABSE in particular because we do not see all malevolent senders employing a deceptive tactic. Yet, amending the above cognitive scenario to allow for a mixed share of rational and coarse senders (when coarse, a sender would randomize between telling a lie and telling the truth in every period because she would fail to see the impact of her current lie on future behavior) allows us to get a good fit. In particular, the lack of response of the share of deceptive senders to the weight of the key period and to the initial share of the two types of senders is perfectly in line with the prediction of this model as is the observation that the deceptive tactic is rewarding.

The rest of the paper is devoted to making the analysis of the game, the SE and ABSE of it as well as the experimental data more precise. We will also consider additional variants of the main game, not discussed so far.

Our study is mainly related to two strands of experimental literature. First, a literature

which considers sender/receiver games à la Crawford and Sobel (1982) initiated by Dickhaut et al (1995) and Blume et al (1998) and also illustrated by Blume et al (2001), Cai and Wang (2001) or Wang et al (2006). That literature has noted that senders somehow transmit more information than theory predicts suggesting that (at least some) senders may be averse to lying. In our baseline treatment, senders have a fixed number of lies to make, which allows us to focus on the timing of the lies rather than the aversion to lie that this previous literature has identified.[4]

Second, a strand of literature on reputation games initiated by Camerer and Weigelt (1988), Neral and Ochs (1992) and Jung et al (1994) which considers reputation games such as the chain-store game or the borrower-lender game. That literature has suggested that the sequential equilibrium may be a powerful tool to organize the data, which contrasts drastically with our finding that theories beyond the sequential equilibrium are needed to give a reasonable account of our experiment on deception, credibility and trust.

## 2 The game and some theoretical benchmarks

We consider a game played by an informed sender and an uninformed receiver which shares a number of features with that studied by Sobel (1985). The game consists of twenty periods. At the beginning of each period, the sender (but not the receiver) is informed of the state of the world prevailing in this period. The receiver discovers the state of the world of period $k$, $s_k$, at the end of period $k$. States of the world may take two values, 0 and 1. The states of the world in the different periods are independently drawn with a probability 1/2 for each value.

In each period $k$, the sender sends a message $m_k$ which can be equal to 0 or 1: $m_k$ is supposed to be representing the current state of the world. The sender can choose a truthful ($m_k = s_k$) or a false ($m_k = 1 - s_k$) message about the state of the world. The receiver observes the message, $m_k$ but does not observe whether the message is truthful or false, the receiver is aware that the sender may choose strategically to send a false message. Then, the receiver makes a decision $a_k \in [0, 1]$.

[4] In our treatment in which senders are free to choose the number of lies they wish, we do not find the bias toward truthtelling that the previous literature has identified. This may be due to the zero-sum nature of the interaction.

The receiver's payoff in period $k$ is equal to $\delta_k(1-(a_k-s_k)^2)$ where $\delta_k$ is the weight of period $k$. The overall payoff of the receiver is $\sum_{k=1,\dots,20}\delta_k(1-(a_k-s_k)^2)$. The choice of a quadratic scoring rules ensures that if the receiver only considers the current period's payoff, she will pick the action that corresponds to what she expects to be the mean value of $s_k$ given the message she received and the history of interaction (i.e. the sequence of messages sent up to the current period).

All periods have the same weight, 1, except one, the *key* period, period $k^*$ (we will assume that $k^*=5$), which has weight $\delta_{k^*}>1$ (we will assume that $\delta_{k^*}\in\{5,10\}$).

There are two types of senders. With probability $\alpha$ (in the experiment, $\alpha$ will be either 1/10 or 1/5), the sender is *benevolent*. This means that he strictly prefers sending a truthful message about the state of the world in all periods. With probability $1-\alpha$, the sender is *malevolent*. A malevolent sender's payoff in period $k$ is equal to $\delta_k(a_k-s_k)^2$ and his overall payoff is $\sum_{k=1,\dots,20}\delta_k(a_k-s_k)^2$. Hence a malevolent sender's objective is to minimize the receiver's payoff.

For expositional purposes, we define $d_k=|m_k-a_k|$, the distance between the signal sent by the sender and the decision made by the receiver. We also introduce the following definition. A tactic of a sender is *deceptive* if it is such that $m_k=s_k$ for $k<k^*$ and $m_{k^*}=1-s_{k^*}$. In a deceptive tactic, a sender sends truthful messages before the key period and a false message at the key period.[5]

## 2.1 Sequential equilibrium analysis

The strategy of the benevolent sender being fixed by the very definition of his type (i.e. sending truthful messages in all periods), a sequential equilibrium of the game is characterized by the strategies of the malevolent sender and the receiver. Besides, since a benevolent sender never sends false messages, by sending a false message, a malevolent sender fully reveals his type. Then, we can show, by backward induction, that, in any sequential equilibrium, in all periods following this *revelation*, the malevolent sender will send a truthful message with

[5] We refer to such patterns of behavior as deceptive tactic as we believe they capture common sense (outside game theory) of deception in so far that they contain a good looking phase (up to the key period) followed by an exploitation phase (at the key period).

probability $\frac{1}{2}$ and the receiver will choose action $\frac{1}{2}$.[6] Hence, to characterize a sequential equilibrium, it remains only to determine the strategies for histories that do no include a past false message.

We introduce the notation $p_i$, the probability for a malevolent sender to send a false message in period $i$ conditional on not having sent a false message before. A sequential equilibrium is characterized by a vector $(p_1, p_2, ..., p_{20})$ and the receiver's best response. We show that there is a unique sequential equilibrium.

**Proposition 1** *There is a unique sequential equilibrium. Such an equilibrium satisfies the following conditions for a uniquely defined* $(p_1, p_2, ..., p_{20})$*:*[7]

- *Conditional on not having sent a false message before, a malevolent sender sends a false messages in periods $k$ with probability $p_k$. A malevolent sender sends a false message with probability $\frac{1}{2}$ conditional on having sent a false message before.*

- *For any $\{i, j\} \in \{1, 2, .., k^*\}^2$ such that $p_i > 0$, $p_j > 0$ and that $\forall l < \max\{i, j\}$, $p_l < 1$ ,*

$$\sum_{l=1}^{i-1} \delta_l d_l^2 + \delta_i (1 - d_i)^2 + \sum_{l=i+1}^{k^*} \delta_l / 4 = \sum_{l=1}^{j-1} \delta_l d_l^2 + \delta_j (1 - d_j)^2 + \sum_{l=j+1}^{k^*} \delta_l / 4$$

  *and for any $\{i, j\} \in \{1, 2, .., k^*\}^2$ such that $p_i > 0$, $p_j = 0$ and that $\forall l < \max\{i, j\}$, $p_l < 1$ ,*

$$\sum_{l=1}^{i-1} \delta_l d_l^2 + \delta_i (1 - d_i)^2 + \sum_{l=i+1}^{k^*} \delta_l / 4 \geq \sum_{l=1}^{j-1} \delta_l d_l^2 + \delta_j (1 - d_j)^2 + \sum_{l=j+1}^{k^*} \delta_l / 4.$$

- *In any period $k$ such that the sender has never sent a false message in any earlier period, a receiver chooses $d_k = \frac{(1-\alpha)\prod_{i=1}^{k-1}(1-p_i)p_k}{(1-\alpha)\prod_{i=1}^{k-1}(1-p_i)+\alpha}$ when $k \neq 1$ and $d_1 = (1 - \alpha)p_1$. In any period $k$ such that the sender has already sent a false message at least once in a former period, a receiver chooses $d_k = \frac{1}{2}$.*

- *A benevolent sender sends truthful messages during all the periods.*

[6] This is the unique Nash equilibrium of the constant-sum game played in one period when it is common knowledge that the sender is malevolent.

[7] The uniqueness of the vector is defined up to the first period $k$in which $p_k = 1$, since behaviors are unaffected by the following values of $p$.

The conditions of the Proposition posit that a malevolent sender should be indifferent as to when to make his first lie over periods in which he may consider lying for the first time and receivers make rational inferences using Bayes' law when no lie has yet been made. As already mentioned, the behaviors after a lie are dictated by the equilibrium of the stage (zero-sum) game.

Solving the sequential equilibrium for a specific value of $(\delta_{k^*}, \alpha)$ is essentially a numerical exercise that deals with the indifference conditions. As an example, we provide approximate values of the equilibrium $p$ vector when $(\delta_{k^*}, \alpha)$ is equal to (5,1/5): $(0.477, 0.447, 0.368, 0.177, 1, 1, ..., 1)$. Such a vector $p$ implies that a malevolent sender chooses a deceptive tactic with a probability close to 0.15.

We observe standard properties of reputation games. Considering the high value of $\delta_{k^*}$, a malevolent sender would like to persuade the receiver that he is benevolent by sending truthful messages during the $k^* - 1$ initial periods if it allowed him to obtain a high payoff in period $k^*$ (a deceptive tactic). However, choosing this tactic with probability 1 for a malevolent sender cannot be part of an equilibrium even for very high values of $\delta_{k^*}$. If the malevolent sender were choosing a deceptive tactic with probability 1, he would obtain $\alpha^2\delta_{k^*}$ during the first $k^*$ periods (in this case, the receiver would choose a $d$ equal to 0 during the first $k^* - 1$ periods of the game and a $d_{k^*}$ equal to $1 - \alpha$.) while he could obtain $1 + \frac{k^*-2+\delta_{k^*}}{4}$ during these same periods if he were deviating, sending a false message in the first period. More intuitively, if a malevolent sender were always following a deceptive tactic he would not be much trusted at the key period, which in turn would make the deceptive tactic suboptimal.

Whatever the chosen value of $\delta_{k^*}$, conditional on not having sent a false message during any prior period, a malevolent sender always sends a false message at the key period with probability 1. But he is not much trusted. This is even more pronounced for higher values of $\delta_{k^*}$. As a matter of fact, when $\delta_{k^*}$ becomes higher, the qualitative properties of the malevolent senders' strategy remains unchanged. Even though they follow a deceptive tactic with a slightly higher probability, this probability is always below $\frac{\alpha}{1-\alpha}$ so that conditional on not having observed any prior false message, $d_{k^*}$ is always below 1/2. The proportion of malevolent senders choosing a deceptive tactic never exceeds the proportion of benevolent senders (as otherwise it would be counter-productive for a malevolent sender to send a lie at the key period after having always sent truthful messages, thereby undermining the equilibrium

construction). Hence, the frequency of deceptive tactic increases with $\delta_{k^*}$ but to a limited extent. For instance, if $(\delta_{k^*}, \alpha) = (10, 1/5)$, a malevolent sender chooses a deceptive tactic with a probability close to 0.18. Lowering the probability $\alpha$ of being matched with a benevolent sender reduces the frequency of deceptive tactic. For instance, if $(\delta_{k^*}, \alpha) = (5, 1/10)$, a malevolent sender chooses a deceptive tactic with a probability close to 0.075.

Let us also observe that the malevolent senders' behavior is always the same in the $20 - k^*$ last periods of the game. They always send a false message with probability 0.5. At the equilibrium, the sender's type is always revealed at the end of period $k^*$. This also means that the last $20 - k^*$ periods of the game do not affect equilibrium behaviors in the first $k^*$ periods of the game. For instance, if we were to consider a variant of the game with only the first $k^*$ periods of the game, the sequential equilibrium would be exactly the same except that we would truncate the equilibrium strategies of the last $20 - k^*$ periods of the game.

### 2.2 A setup with cognitive limitations

As an alternative to the sequential equilibrium, consider the analogy-based sequential equilibrium as defined in Ettinger and Jehiel (2010).[8] Without going into the details of this equilibrium concept, we consider the following cognitive environment. Receivers are aware that there exist two types of senders and they also know the frequency of the two types of senders. Receivers are assumed to be knowledgeable of the aggregate lie rate of the two types of senders over the twenty periods, but they are assumed not to be knowledgeable of how the behaviors of Senders depend on the history of play. Moreoever, we assume that Receivers reason as if Senders of a given type were behaving in a stationary way as given by the aggregate lie rate of the corresponding type of Senders (this asumption is meant to formalize the idea that Receivers consider the simplest theory that is consistent with their -coarse- knowledge). Senders are standard fully rational agents.

At the equilibrium, senders play a best-response to receivers' strategy and receivers play a best-response to their perception of senders' strategies, using Bayes' rule to revise their beliefs about the type of sender they are matched with. As already mentioned, we also require that the average frequencies of false messages perceived by the receivers, for each type of senders,

[8] For a full description of the equilibrium concept, see Ettinger and Jehiel (2010).

to be equal to the actual frequencies of false messages sent by the two types of senders.[9]

**Proposition 2** *Any analogy-based sequential equilibrium of the defined game must satisfy the following properties:*

- *A benevolent sender sends truthful messages during all periods.*
- *A malevolent sender always sends truthful messages in periods $k$ such that $k < k^*$, sends a false message in period $k^*$ and sends, on average, 9 false messages during the $20 - k^*$ last periods so that he sends, on average, 10 false and 10 truthful messages during the 20 periods of the game.*
- *In any period $k$ such that the sender has never sent a false message in any former period, the receiver chooses $d_k = \frac{(1-\alpha)(1/2)^k}{\alpha+(1-\alpha)(1/2)^{k-1}}$. In any period $k$ such that the sender has already sent a false message at least once in a former period, a receiver chooses $d_k = \frac{1}{2}$.*

The equilibrium works as follows. A malevolent sender sends, on average, 10 false messages and 10 truthful messages. Therefore, receivers have the following perception regarding senders. There exist two types of senders. With probability $\alpha$, senders are *honest* and always send truthful messages and with probability $1 - \alpha$, senders are non-trustworthy referred to as *liars* and send truthful and false messages with probability $1/2$ in each period.

During the first $k^* - 1$ periods, receivers observe truthful messages. Therefore, at the end of period 1, a receiver perceives that the sender is *honest* with probability $\frac{\alpha}{\alpha+(1/2)(1-\alpha)} = \frac{2\alpha}{\alpha+1}$ (since she believes that a liar would send a false message with a probability $1/2$ in this period) and she chooses $d_2 = \frac{2\alpha}{\alpha+1}(0) + (1 - \frac{2\alpha}{\alpha+1})\frac{1}{2} = \frac{1-\alpha}{2\alpha+2}$. The same belief revision works in the following periods so that, at the end of period $k^* - 1$, if she has only observed truthful messages, a receiver perceives that she is facing an *honest* sender with probability $\frac{\alpha}{\alpha+(1/2)^{k^*-1}(1-\alpha)}$. For $\alpha = 1/5$, this is equal to $4/5$. Even though both types of senders send truthful messages during the first $k^* - 1$ periods, because of the analogical reasoning and the aggregate lie rate of malevolent senders, a receiver perceives that she is much more likely to

[9] While the motivation for this assumption is based on learning, in our experiment this consistency was imposed by design.

be matched with a benevolent sender after $k^* - 1$ truthful messages. This belief is exploited in the key period by malevolent senders who follow a deceptive strategy with probability 1.

If a sender sends a false message before period $k^*$, he obtains $\delta_{k^*}/4$ in period $k^*$. If he follows a deceptive tactic, he obtains $\delta_{k^*}(1 - \frac{(1-\alpha)(1/2)^k}{\alpha+(1-\alpha)(1/2)^{k-1}})^2$ in period $k^*$ (which is equal to 0.8). Therefore, because $\delta_{k^*}$ is sufficiently high and $\alpha$ is not too small, even though it is costly not to send a false message in the first $k^* - 1$ periods of the game, malevolent senders prefer postponing their first false message because of the extra profit that they can derive in the key period by persuading the receiver that they are much likely to be benevolent. Hence, if we assume that receivers have these cognitive limits, a deceptive tactic becomes profitable.

Let us also explain why the aggregate lie rate of malevolent senders must be 50:50 in equilibrium. Suppose that malevolent senders send, on average, more false messages than truthful messages. Then, once a sender is identified as malevolent, the receiver always chooses $d_k > 0.5$ since the receiver perceives that malevolent senders are more likely, in any period of the game, to send false messages. Then, the sender should always choose to send truthful messages after the first false message. But this is not consistent with malevolent senders sending more false messages than truthful messages. We can apply the same reasoning in the other direction in order to reject the possibility that malevolent senders send more truthful messages than false messages during the 20 periods of the game.

During the first $k^*$ periods of the game, malevolent sender's behaviors differ significantly in the sequential equilibrium and in the ABSE. Besides, contrary to what we observe in the sequential equilibrium, in the ABSE, the presence of the $20 - k^*$ last periods of the game do matter for the first $k^*$ periods. As a matter of fact, with the ABSE, if there were only $k^*$ periods, we would not have the same equilibrium behaviors as with 20 periods with a truncation of the $20 - k^*$ last periods of the game. These last periods are necessary in order to establish the 0.5 average frequency of false messages of malevolent senders. With only $k^*$ periods, malevolent senders would choose a mixed strategy in which the deceptive tactic would be played with a probability strictly lower than 1 (which would maintain the average frequency of false messages at a sufficiently high level for the deceptive tactic to be worth playing).

The comparative static is much simpler than in the sequential equilibrium case. On the senders' side, the equilibrium strategy remains exactly the same if we modify the values of

$\alpha$ and $\delta_{k^*}$ provided that $\delta_{k^*} > 2$. Whether the fraction of benevolent sender is high or low, as long as the weight of the key period is sufficiently high, it is worth giving up on the gain during the first periods of the game so as to obtain an extra profit in the key period by deceiving the receiver[10]. On the receiver's side, the equilibrium strategy requires that for $k \leq k^*$ $d_k = \frac{(1-\alpha)(1/2)^k}{\alpha+(1-\alpha)(1/2)^{k-1}}$, which decreases in the ex ante share $\alpha$ of benevolent senders but is not affected by changes in $\delta_{k^*}$ (as long as $\delta_{k^*} \geq 2$).

It should be mentioned at this stage that in our benchmark experimental setting, we imposed the total number of lies of malevolent Senders to be approximately 10 (see below for a detailed presentation of how this was implemented). There are several reasons for this choice: First, we were interested in testing the strategic choice of the timing of lies rather than testing the propensity of subjects to make lies. Second, such a constraint has limited if any effect on the main theories presented above. For sequential equilibrium, this extra constraint affects in a negligible way the computations of the equilibrium mixed lying strategies (and of course not the property that all employed strategies should be equally good for payoffs). For Analogy-Based sequential equilibrium (ABSE), it does not affect at all the construction of the equilibrium, since the unconstrained equilibrium satisfies this extra constraint. One may argue that imposing this constraint somehow simplifies the learning that motivate ABSE, and from this perspective our main motivation for this is that we wanted to save somehow on the time spent by subjects in the lab.

## 3 Experimental Design

The experiment was conducted in the Laboratoire d'Economie Experimentale de Paris, located in the Maison des Sciences Economiques with the software REGATE. Sessions lasted from 1.4 to 1.7 hours and subjects (18 or 19 per session) were predominantly Paris 1 undergraduate students, 40% of them majoring in economics. During the experiments, subjects interacted with each other only through computer terminals. There was no show-up fee, subjects only obtained what they earned from playing the game. Their point payoffs were converted in Euros using a pre-specified exchange rate. Earning ranged from 8.60 Euros to 21.60 Euros with a variance of 5.48 Euros and an average of 14.40 Euros.

[10]The extra profit decreases in $\alpha$ but the extra cost in sending truthful messages during the first period also decreases in $\alpha$.

We organized standard sessions (6 sessions) with $\delta_{k^*} = 5$ and $\alpha = 1/5$ and considered several variants to be described next.

In standard sessions, the game was played 5 times (5 rounds), 10 subjects were assigned to the role of receivers and 8 subjects were assigned the role of senders with a malevolent sender's utility function. Two computerized machines played the role of benevolent senders.

At the beginning of each round, a sender was assigned a capital of false and truthful messages summing to 20. During the game, this capital evolved depending on the number of false and truthful messages sent earlier. During a round, a sender was constantly informed of his remaining capital of false and truthful messages. Whenever his capital of one of the two types of messages was equal to zero, the computer system forced the sender to send the other type of messages until the end of the current round. At the start of an interaction (round), a sender's capital of false messages was randomly drawn. It could be equal to 9, 10 or 11 with an equal probability for all these draws (so as to introduce an element of unpredictability toward the end of the game on the receiver' side).[11]

Senders and receivers' instructions contained a complete description of the game except that receivers were not told senders' utility functions.[12] Receivers were informed that with probability $\frac{4}{5}$ they would be paired with human senders and, with probability $\frac{1}{5}$, with an automaton that always sends truthful messages. They knew that human senders' strategies were such that they send, on average, 10 false messages and 10 truthful messages across the 20 periods of the baseline treatment. Variants were also considered. These are described as follows.

- Free-sessions (4 sessions). Human senders were not constrained to send a specified number of truthful or false messages during a 20-period interaction (round). Consequently, receivers were not informed that human senders' strategies were such that they send, on average, 10 false messages and 10 truthful messages across the 20 periods of the game. However, before beginning a round (except the first one), receivers were informed of the percentage of false and truthful messages sent in the former rounds by human senders.

[11] It seems the unpredictability worked since we did not observe that receivers derived significantly different payoffs in the last period compared to the previous ones.

[12] We believe that not letting receivers know the payoffs of the senders is in better agreement with real life incarnations of the above information transmission game in which payoffs are rarely given from the start and must be inferred from behaviors.

Besides, there were 6 rounds so that the data from the last 5 rounds can more easily be compared to the data from our baseline treatment.

- 10%-sessions (3 sessions). In this treatment, the chance of meeting a truthful machine was reduced from 20% to 10%. This was implemented by having 9 malevolent senders and only one benevolent automaton sender.

- Weight 10-sessions (3 sessions). In this treatment, the share of truthful automata was kept at 20% as in our baseline treatment, but the weight of the key period $k^*$ was increased to $\delta_{k^*} = 10$.

- 5 periods free-sessions (3 sessions). In this treatment, the interaction stopped right at the end of the key period $k^*$. There was no constraint on the number of false messages. After the first round, receivers were informed of the past aggregate lie rate of human senders exactly as in Free-sessions.

- Belief-sessions (4 sessions). Finally, in order to obtain extra insights about the mode of reasoning of receivers, we implemented a treatment in which receivers were asked in each period to report their belief as to whether they were facing a machine or a human sender.

During all sessions, subjects had at their disposal a written version of the instructions and a pencil as well as a piece of paper. Before the beginning of a session, we presented to the subjects the screens that they would have to face during the game. In all sessions, subjects, during a round, could see on the lower part of the screen the history of false and truthful past messages.

In order to facilitate the computations, the payoffs of the participants of the game were multiplied by one hundred as compared with the game introduced in the previous section.

# 4 Results

## 4.1 First observations, the standard sessions

We first describe some salient observations.

**The receiver's side**

We focus on the variable $d_k$ rather than $a_k$ since what really matters is the distance between the message sent and the action of the receiver. Besides, we did not identify any significant effect according to whether the signal sent is equal to 0 or 1 on the value of $d_k$. A message 0 is neither more nor less trusted than a message 1.

The average value of $d$ over all the periods is equal to 0.4 but this takes into account both receivers matched with benevolent and malevolent senders. If we only consider receivers matched with malevolent senders, this statistic is equal to 0.46, slightly less than 0.5. Now, the distribution of $d$s is much affected by a very simple statistic: did the receiver already observe a false message during the game or not?

If no false message has been observed during the game, the average $d_k$ slowly decreases from period 1 to 5 (from 0.32 to slightly more than 0.28), decreases faster from period 6 to 9 and reaches 0.1, then again slowly decreases with some movements around 0.05. Even after 15 or 18 truthful messages, the average $d_k$ never falls below 0.05. There is also a small peak in period 11. Some receivers seem to expect that human receivers may send 10 truthful messages in the first 10 periods and then 10 false messages. These observations are represented in figure 1.

MAYBE ADD A FIGURE THAT WOULD DEPICT THE BEST RESPONSE AGGREGATING ALL THE DATA AFTER k truths? OR ELSE REPORT THIS FOR THE KEY PERIOD.

If at least one false message has been observed during the game, the average $d_k$ is equal to 0.485, slightly less than 0.5[13] as represented in figure 1. Neither the number of observed false messages, as long as it is strictly positive, nor the truthfulness of the message sent at period $k-1$ affect $d_k$. The only statistic that really matters is whether at least one false message has been observed during the game. This is perfectly in line with the Bayesian feature that the type of the sender is fully revealed after the sender has sent one false message[14].

[13] It is slightly lower than 0.5 in the first periods of the game when no false messages has been already observed.

[14] Observe that the reported finding is inconsistent with interpretations in terms of law of small numbers or extrapolation. With extrapolation in mind, a receiver should choose a higher $d_k$ if the ratio of false messages is high in the past periods of the game (one may relate extrapolation to the so called *hot hand fallacy*). With the law of small numbers in mind (which may be motivated here on the ground that we explicitly told the receivers that the average lie rate was 50%), a receiver should choose a lower $d_k$ if the ratio of false messages is high in the past periods of the game. We did not observe any of these two biases.

**The sender's side**

The more salient observation on the sender's side concerns the deceptive tactic which is chosen with a frequency 0.275 by human senders. Besides, choosing such a tactic is much more profitable than the other used strategies (aggregating over the latter) during the 5 first periods of the game. A sender who sends his first false message during one of the 4 first periods of the game obtains on average 297.2 during the 5 first periods. When he follows a deceptive tactic, he obtains, on average, 361.5[15]. This difference is highly significant ($t < 0.003$).

Besides, if we consider the average payoff obtained by the sender whenever the first lie took place in period $k$, we observe that a sender obtains a higher payoff by sending his first false message in period 5 than in period 1, 2, 3, 4 or 6. This difference is highly significant (with $t$ always lower than 0.03) except for period 1 for which $t = 0.095$. We also observe that sending a first false message in period 1 provides a higher payoff than sending it in period 2, 3, 4 or 6 (with $t$ respectively equal to 0.06, 0.13, 0.12 and $10^{-7}$).

We did not identify any specific lie pattern during the 15 last periods of the game. For instance, once a false message has already been sent, the likelihood of sending a false message is not significantly affected by the truthfulness of the message of the previous period.

## 4.2 First interpretations

**The sender's side**

Neither the high frequency of observed deceptive tactic nor the extra profitability of this tactic is consistent with the predictions of the sequential equilibrium. We note here that a receiver has a higher chance of facing a human sender who employs a deceptive tactic than of facing a machine, which, even without getting into the details of the sequential equilibrium, is at odds with the predictions of the rational model (see the discussion surrounding the description of the sequential equilibrium in Section 2). Besides, a significant difference in the average obtained revenue between different tactics chosen with positive probability by senders is hard to reconcile with an interpretation in terms of rational senders and receivers playing a

[15] In order to obtain more data, we gather for this statistic, the payoff obtained by human senders following a deceptive tactic and the payoff that the automata senders would have obtained if they had sent a false message in period 5, supposing that $d_5$ would have remained the same in that case. Let us also mention that the difference between the average payoffs of the the two groups is negligible.

sequential equilibrium with mixed strategies. In a sequential equilibrium, the tactics chosen with a strictly positive probability are supposed to provide the same expected payoff.

By contrast, the extra profitability of the deceptive tactic is a prediction of the ABSE. It also predicts a high frequency of deceptive tactics. In the ABSE we introduced, the deceptive tactic is chosen with probability 1. This probability 1 is due to our assumption that all human senders are fully rational. However, we may reconsider this strong assumption we made on cognitive abilities. It may be more realistic to assume that only a share of senders is fully rational, thereby being able to perceive the game in its full complexity.

Suppose instead that a share $\beta$ of senders is fully rational and a share $1-\beta$ is coarse in the sense that coarse senders only perceive the aggregate distribution of actions of receivers in all periods of the game, but not how these distributions depend on history. To the extent that receivers' behaviors are perceived to be independent of senders' messages, it is plausible (and can indeed be sustained in equilibrium) that such coarse senders would in each period of the game send a false message with probability 0.5 as if it were only a one period game. In that case, the distribution of first false messages would be as follows : $\frac{1-\beta}{2}$ in period 1, $\frac{1-\beta}{4}$ in period 2, $\frac{1-\beta}{8}$ in period 3, $\frac{1-\beta}{16}$ in period 4, $\beta + \frac{1-\beta}{32}$ in period 4 and $\frac{1-\beta}{32}$ later. This coincides quite well with the observations if we choose $\beta = 1/4$ as can be seen in figure 2. The distribution of first false messages coincides much better with the one obtained with this scenario than with the one described in the sequential equilibrium. We will see that this observation is still verified, with the coefficient $\beta$ equal to $1/4$, in several variants of the game. We will come back later to the derivation of the ABSE when all subjects (whether senders or receivers) are assumed to be coarse with probability $3/4$ and rational with probability $1/4$.

**The receiver's side**

We suggest that receivers may be better understood by disaggregating the whole population of subjects assigned to the role of receivers. More precisely, we will focus on receivers' behaviors during the 5 first periods of the game conditional on not having observed any prior false messages. We will consider two populations of receivers: \-*receivers* and *V-receivers*. A receiver is a \-receiver if, conditional on having only observed truthful messages in the past, the receiver follows more and more the sender's recommendation up to and including the key period, or, in symbols, for any $k < 5$, $d_{k+1} \leq d_k$. (We adopt the notation \-receiver to represent the declining trend of $d_k$ till the key period.) A receiver is a V-receiver if, conditional on

having only observed truthful messages in the past, the receiver follows more and more the recommendation before the key period but becomes cautious at the key period, or in symbols, for any $k < 4$, $d_{k+1} \leq d_k$ and $d_5 > d_4$[16]. (The notation V-receiver should be interpreted graphically.) Note that according to these definitions, the receivers in the above ABSE are \-receivers and those in SE are V-receivers. We propose that in the game under study, this difference in attitude at the key period distinguishes fundamentally coarse reasoning (as in ABSE) from more sophisticated reasoning (as in SE).

We observe that most of the receivers belong to one of these two categories. 59% of the receivers are \-receivers and 24% are V-receivers.

\-receivers can be related to the coarse receivers of the ABSE we described. They do not perceive the specific status of the key period and put all the decision nodes of the senders in the same analogy class. V-receivers would be more rational receivers, distinguishing the key period and understanding its specific status. Their behaviors could be compared to the one described by the sequential equilibrium. As figures 6 and 7 indicate it, the behaviors of these two populations are quite well approximated by associating receivers with coarse receivers playing the analogy-based sequential equilibrium and V-receivers with rational receivers playing the sequential equilibrium. The observed coefficient of the slope is lower than the equilibrium predictions but this may be explained by receivers' difficulties in applying Bayes' law. As already mentioned, we will come back toward the end of the paper to the formal description of ABSE with 75% share of coarse reasoners (whether on the sender or the receiver' side).

Given our suggestion that V-receivers can be thought of as being more sophisticated than \-receivers, it is of interest to compare how these two groups performed in terms of their expected gains. We obtain that the average gain over the first five periods is 627 for \-receivers and 701 for V-receivers, thereby suggesting a difference in expected payoff which is significant ($t < 0.05$).[17] We note also that receivers who belong to none of these two groups

[16] In fact, because receivers' decisions are somehow noisy, we allow $d_{k+1}$ to be higher than $d_k$ by at most 0.1, not more than once and for $k < 4$.

[17] It should also be mentioned that \-receivers and V-receivers were matched with machines with the same frequency in our data, thereby reinforcing the idea that the difference in performance can be attributed to difference in cognitive sophistication (rather than luck). It may also be reminded that if receivers always choose $d = 0.5$ they ensure a gain of 675, thereby illustrating that \-receivers get less than a payoff they could

got an expected gain comparable to those of \-receivers.

### 4.3 Variants

**10% automata**

The sequential equilibrium of the game with 10% of benevolent senders has the same qualitative properties as with 20% benevolent senders except that malevolent senders send more often their first false message during earlier periods of the game so that the deceptive tactic is only chosen with a probability 0.075 (a reduction of 50%) and the ABSE is not modified.

In the data, we do not observe any statistically significant difference in the frequency of deceptive tactic between the 20% automata case (0.275) and the 10% automata case (0.25). Thus, we do not see the comparative statics suggested by the sequential equilibrium. Again, this high frequency of deceptive tactics goes with extra profits (355 compared to 297 during the 5 first periods of the game and this difference is significant ($t < 0.03$)). 52% of the receivers are \-receivers and 25% are V-receivers. It is quite striking to observe that on both sides (senders and receivers), behaviors are almost the same as in the Baseline treatment. Such a finding is in line with our general interpretation in terms of a fixed share (75%) of subjects relying on coarse reasoning.

**Weight 10 sessions**

The prediction of the sequential equilibrium would be that compared to standard sessions, the share of deceptive tactics should increase from 15% to 18% and as usual the deceptive tactic should be as good/bad as other (employed) strategies for senders. We do not get this in the data. On the sender's side, the ratio of deceptive tactic slightly decreases. Since receivers' behaviors are not modified, choosing a deceptive tactic is even more profitable than in the standard sessions. Here the deceptive tactic gives an average 675 during the 5 first periods of the game as compared to 403 for other tactics. This difference is highly significant ($t < 10^{-11}$). On the receiver's side, 52% of the receivers are \-receivers and 30% are V-receivers, close to the observations in standard sessions. On the receiver's side, the results are very close to the one we observed in the other sessions. Again, these findings are in line with our proposed interpretation of having a share (75%) of coarse reasoners.

easily secure.

**Free sessions**

In these sessions, human senders were not constrained regarding the number of false and truthful messages sent. Therefore, in the receivers' instructions, the average ratio of false message by human senders was not mentioned. However, when they played again the game (from round 2 on), receivers were informed of the average ratio of false messages sent by human senders in the past games played in the session.

On the receiver's side, the main difference with standard sessions lies in the evolution of the average $d_k$ conditional on having only observed truthful messages. While in the standard sessions, the average $d_k$ was (slightly) decreasing up to and including the key period, in the Free sessions, $d_k$ slightly decreases with $k$ between periods 1 and 4, and then increases between periods 4 and 5 (from 0.262 to 0.313) and then again slowly decreases. We observe this difference with standard sessions even though we only consider rounds 2 to 6 in which receivers are informed that, in the past rounds of the session, the frequency of false messages sent by human senders on average lies between 0.43 and 0.49 (depending on the sessions and the round) with an average of 0.46, close to the 0.5 of the standard sessions[18]. Besides, 51% of the receivers are \-receivers and 24% are V-receivers which is quite close to what we obtain in standard sessions.

On the sender's side, the frequency of deceptive tactic is almost the same as in the standard session, 0.25. Senders still obtain more, on average, during the 5 first periods of the game when they follow a deceptive tactic than when they follow any other tactic (325 rather than 295). However, this difference is no longer significant precisely because of the change in receivers' behaviors previously mentioned. V-receivers do not choose a higher average $d_5$ than in standard sessions and the share of V-receivers does not increase. The main difference lies in the average $d_5$ chosen by \-receivers which is higher than in standard sessions (0.19 rather than 0.16) and the standard deviation of $d_5$ chosen by \-receivers which is lower (0.17 compared to 0.24 in standard sessions). These two elements have a negative effect on senders' payoff, which is magnified by the convexity of the payment function.

[18]This shows that, contrary to what has been observed in other experiments (such as Cai and Wang (2006) or Wang et al (2006)), here, when senders can freely choose the percentage of false messages, they do not send more informative messages than what is predicted by the equilibrium, sequential equilibrium or ABSE. This may be due to the structure of the game which clearly appears as a constant-sum game to senders.

**5 Periods Free sessions**

The prediction of the sequential equilibrium in this variant should be the same as in the standard treatments. However, in this variant, we observe very different behaviors. On the sender's side, the frequency of deceptive tactic is equal to 0.05, much lower than in any other variant and in fact much lower than predicted by the sequential equilibrium. The share of deceptive tactics is better accounted for by referring to *random* behavior in which in each period, independently of the history of the game, malevolent senders would send a false message with probability 1/2. There are three major differences between behaviors in the sequential equilibrium and *random* behaviors. In the sequential equilibrium, the deceptive tactic is chosen with probability 0.15 (0.03 in the *random* case). In the sequential equilibrium, conditional on having sent truthful messages during the first three periods, the probability of sending a false message in the fourth period is low, 0.17, while it is equal to 0.5 in the *random* case. Finally, in the sequential equilibrium, malevolent senders never send 5 truthful messages while they do so with probability 0.03 in the *random* case. The observed frequency of deceptive behavior is 0.05, the frequency of false message in the fourth period after 3 truthful messages is equal to 0.4 and the series of 5 truthful messages appear with a frequency 0.05. These three elements favor the interpretation in terms of random behavior.

On the receiver's side, we observe higher $ds$. Except between periods 1 and 2, the average $d_k$ conditional on having observed only truthful messages is decreasing in $k$ but the values are higher than in all the other variants, between 0.44 and 0.38 (in period 5). However, the shares of \-receivers and V-receivers are not much different from what we observe in other sessions: 56% and 18%, respectively. Hence, except that $d_1$ is higher than in standard sessions, on the receivers' side, the general trend is the same as in other sessions. Finally, the deceptive tactic provides a higher profit than other tactics (358 rather than 316), but the difference is not significant ($t = 0.16$) [19]. In sum, the deceptive tactic is not much less profitable but it is much less often chosen.

**Summary**

With the exception of the 5periods treatment, three major elements are almost always verified (with more noise in the Free sessions). When receivers do not observe false message

[19] We have less observations since the frequency of deceptive is much lower than in other sessions.

in a game, $d_k$ tends to slowly decrease without any strong inversion in period 5. Following a deceptive tactic during the 5 first periods of the game provides a higher payoff to a sender than any other tactic. Human senders' behaviors during the 5 first periods coincide quite well with the behaviors of a population of 25% of sophisticated senders playing the ABSE and 75% of coarse senders who would play in each period of the game as if it were only a one period game (this general observation appears clearly on figures 2, 3 and 4).

On the receiver's side, the ratios of \-receivers and V-receivers are also quite stable across treatments, between 51% and 59% for \-receivers and between 24% and 30% for V-receivers. Receivers do not tend to pay more attention to the key period or to fear more deceptive tactics when we lower the ratio of benevolent senders or we raise the weight of the key period. The lack of adjustment to these aspects is inconsistent with the sequential equilibrium.

Let us also mention that we do not observe any clear-cut learning effect across the 5 repetitions of the game. There is no trend in the frequency of deceptive tactic (it is slightly increasing but this is not significant) and on the receiver's side, there is no clear-cut learning process either.[20]

Regarding the 5 period session, we see a big difference with the other sessions mostly on the sender's side (the receiver's side is very similar to what is observed in other sessions). The difference between the 5 periods session and the baseline treatment is hard to reconcile with the sequential equilibrium. The reduction in the share of deceptive tactic in the 5 periods session is in line with the ABSE, but the reduction seems to go much further than what this theory would predict, suggesting that human senders are anxious (even though wrongly so) that a deceptive tactic would be too easy to detect in this case.

### 4.4 A specific analysis of belief sessions

We ran 4 belief-sessions. A belief session is equivalent to a standard session except that receivers were asked to report the probability $p_k$ (in period $k$) they assigned to being matched with an automaton after having made their decision and before being informed of the true

[20]The frequency of deceptive tactic may be too low to observe a learning process with 5 iterations. We ran sessions in which all the senders were automata : 20% benevolent always sending truthful messages and 80% following a deceptive tactic and sending 9 false messages in the 15 last periods of the game. In this case, even with 5 periods, receivers do learn. The average $d_5$ is increasing in the number of games already played, from 0.17 the first time the game is played to 0.45 the 5th time.

value of the signal[21]. Of course, receivers were only asked in the case they did not observe any false message during the game, and they were not asked at the start of the game.

First, we observe that this extra querry did not affect the aggregate behaviors of the players in terms of the lying strategy or the sequence of $d_k$. The ratio of deceptive tactic is 30% and receivers' behaviors are almost identical to what we observe in standard sessions. Being asked about the likelihood of facing a human sender does not affect behaviors. Second, we observe many violations of Bayes' law and inconsistencies. This is not specific to this experimental design. In general, we observe that $p_k$ is less adjusted than required if receivers were applying Bayes' law (in agreement with their $d_k$ choices). They apply a *smoothened version* of Bayes law.

Now, we intended to use this variant in order to test our hypothesis regarding the two populations of receivers. Again, in these four belief-sessions, we distinguish two populations of receivers: 54% of \-receivers and 24% of V-receivers. During the 5 first periods of the game, when they do not observe any false message, the average behavior of \-receivers is close to the behavior of coarse receivers of the ABSE and the average behavior of V-receivers is close to the receivers' behaviors in the sequential equilibrium.

If we consider the evolution of $p_k$ for these two populations (see figure 8), we observe major differences[22].

\-receivers choose a high average $p_2$, close to 0.5 and during the next two periods, the average $p_k$ is very stable, it increases in period 5 and slowly increases after period 6 (it is still below 0.8 in period 8). The belief revisions is slow. \-receivers do not pay attention specifically to period 5. As a matter of fact, the slope of $p_k$ increases between period 4 and period 5 which indicates that for \-receivers, a truthful message in period 4 is a good indicator that the sender is less likely to be a human sender (although in the sequential equilibrium, malevolent senders seldom send a false message in period 4). This could be considered as the behavior of receivers thinking in terms of analogy classes, thereby failing to make any distinction regarding period 5, and a process of belief revision consisting in a *smooth* version of Bayes' law.

[21]For each period in which the receiver chooses a $p_k$, she receives a payment equal to $(p_k - \beta)^2$ with $\beta$ equal to 0 if the sender is an automaton and 1 otherwise.

[22]From period 2 to period 5, we use the average $p_k$ when matched with a *deceiving* human sender or an automaton and for period 6 onwards, the average $p_k$ when matched with an automaton.

V-receivers begin with a belief in period 2 closer to the actual share of automata. The average $p_k$ increases from period 2 to period 4, remains stable from period 4 to period 5 (which is in agreement with the low frequency of false message in period 4 at the sequential equilibrium) and increases very strongly from period 5 to period 6. The latter observation is also in agreement with the fact that V-receivers expect that a human senders if he has not sent his first false message prior to period 5 is very likely to do so in period 5. These elements are qualitatively reminiscent of the equilibrium strategy of the receiver in the sequential equilibrium.

These results tend to corroborate the proposed distinction of two populations of receivers and the interpretations that we suggested for the behaviors of these two populations.

### 4.5 A richer cognitive framework

Considering the results of the experiments, it is natural to suggest a richer setting regarding cognitive types than the ones introduced in section 2.2. More precisely, the observations of the experiments suggest that both on the sender's side and the receiver's side, a distribution with 25% of rational players and 75% of coarse players putting all the decision nodes of their opponents of a given type (machine or human subject) in a unique analogy class could organize the data well. In the next result, we report an ABSE that corresponds to such a cognitive environment. The proof of the claim appears in Appendix.

**Claim 1.** There exists an analogy-based sequential equilibrium of the game with $\alpha = 1/5$, $k^* = 5$, 25% of rational senders, 75% of coarse senders putting all the nodes of the receivers in a unique analogy class, 25% of rational receivers and 75% of coarse receivers putting all the decision nodes of the (human) senders in a unique analogy class in which:

- A coarse malevolent sender sends truthful and false messages with probability 1/2 during all the periods of the game.
- A coarse receiver chooses $d_k = \frac{4(1/2)^k}{1+4(1/2)^{k-1}}$ conditional on not having observed any false message in any past period and 1/2 otherwise.
- A rational receiver chooses 1/2, if she has observed a false message in any past period of the game. As long as no lie has been observed, she mimics the behaviors of coarse

senders during the first 3 periods. Then she chooses $d_4 = 3/38$ (lower than $\frac{4(1/2)^4}{1+4(1/2)^{4-1}}$), $d_5 \approx 0.43$, $d_6 \approx 0.16$ and $d_k < 0.03$ for any $k > 6$.

- As long as he observes behaviors consistent with those of coarse receivers, a rational sender follows a deceptive tactic (he sends truthful messages in the first four periods and lies in period 5) then he randomizes between lying and telling the truth. If he observes a $d_k$ other than $\frac{4(1/2)^k}{1+4(1/2)^{k-1}}$ in periods $k = 1, ..4$ he infers that he is not facing a coarse receiver and behaves accordingly. Specifically, on the equilibrium path, he observes $d_4 = 3/38$ when facing a rational receiver; he sends a false message with probability higher than 0.88 in period 5 and with a probability 1 in period 6 if he did not send it in period 5. After the first false message, the rational sender randomizes uniformly over all periods between telling the truth and lying.

This equilibrium works as follows. Coarse receivers reason as in the simple cognitive scenario case. As long as they observe truthful messages, they increase their belief that they are facing a benevolent sender and since they believe that malevolent senders send a false message with probability 1/2 in every period of the game, they do not give a specific status to period 5. Besides, because they perceive that malevolent senders lie with probability 1/2 in all periods, they perceive that the sender they face is unlikely to be malevolent after a few truthful messages.

Coarse malevolent senders, as an effect of their coarseness, do not perceive the effects of creating a reputation of being a benevolent sender and therefore send a false message with probability 1/2 in all periods.

Rational malevolent senders exploit coarse receivers by following a deceptive behavior. They also identify rational receivers in period 4 (because rational receivers, conditional on not having observed any false message during the 4 first periods, choose a $d_4$ that differs from the one chosen by coarse receivers). Then, rational senders play a *standard* mixed strategy in periods 5 and 6 (taking into account that in the corresponding subgame, they may still be confounded with machines that always tell the truth or with coarse senders who randomize 50:50 in every period).

Regarding rational receivers, they face the following trade-off. If they do not play like

coarse receivers, they reveal themselves as rational senders, which is not good. If they play like coarse senders up to period $k$, they can exploit their fine understanding of rational senders' behavior in period $k+1$. This would suggest that it is best for rational receivers to wait until the key period before they reveal themselves as not being coarse. This would indeed be the best strategy if senders were most likely to be rational. Yet, because the share of rational senders is smaller (in fact close to the share of truthful machines), exploiting the knowledge of rational senders' behaviors is more profitable at period 4 than at period 5, thereby explaining their strategy.

Two elements of this equilibrium differ from our experimental observations. In the experiments, conditional on not having observed any false message, $d$ decreases more slowly and with more noise than in the proposed equilibrium. Besides, because of this noise that we may explain by the difficulties of applying Bayes' Law, a malevolent rational sender is unlikely to perfectly infer whether he faces a rational receiver by observing a choice of $d_k$ in period 4 other than $\frac{4(\frac{1}{2})^k}{1+4(\frac{1}{2})^{k-1}}$.

However, except for these two points, the proposed equilibrium shares many qualitative properties with our experimental observations if we identify the V-receivers in our observations with rational receivers, the \-receivers with coarse receivers and the human senders following a deceptive tactics with rational malevolent senders. To make the comparisons more accessible, we report the following.

- The frequency of deceptive tactics is 27% in the proposed equilibrium and between 25 and 30 % in the observations.

- The deceptive tactic induces a higher payoff than the other employed strategies. Considering the payoffs obtained over the 5 first periods, in the proposed equilibrium, a sender choosing a deceptive tactic obtains, on average, 374 and 267 if he does not. In the observations, he obtains respectively 362 and 297 with no statistically significant differences between the distribution of payoffs obtained by senders following a deceptive tactic and 374, and the distribution of payoffs obtained by senders following a different tactic and 267.

- In both the proposed equilibrium and the data, conditional on sending the first false message before period 5, the best period to do so is period 1 (another observation that

is inconsistent with the predictions of the sequential equilibrium).

- During the 5 first periods of the game, rational receivers in the proposed equilibrium and V-receivers in the experiments obtain a higher expected payoff than respectively coarse receivers and \-receivers, 666 rather than 644 and 701 rather than 627. The difference is even more important in the experimental observations because V-receivers also revise better their beliefs and their $d$s in the 4 first periods of the game and, because of the noise, they manage to not reveal their type in period 4.

### 4.6 Alternative approaches

Alternative popular approaches to study experimental data include the Quantal Response Equilibrium (see McKelvey and Palfrey (1995)) and the so called level-k approach (following Stahl and Wilson (1994, 1995) and Nagel (1995)).

While it is cumbersome/complex to characterize the QRE in the context of our game with a continuum of actions (on the receiver' side), we make the following comments. First, we conjecture that the share of deceptive tactics would not be higher in the QRE than in the sequential equilibrium (think of the extreme version of QRE in which strategies would not be responsive to payoffs in which case the deceptive tactic would appear with probability 3%). Second, considering the data, we observed that the deceptive tactic was more rewarding when the weight of the key period was 10 instead of 5 or in the 5 period version in which the interaction stopped right after the key period. Yet, the share of deceptive tactics was not bigger in those two treatments (it was in fact much smaller in 5-period sessions). These observations are not suggestive that QRE provides a good explanation of the observations here.

Regarding level k, we note that existing theories are not well adapted to deal with multi-stage games insofar as they do not provide a theory of how level-k beliefs should be revised once an inconsistent behavior is observed.[23] **DAVID/ Finalement on ne mentionne meme pas la version ou les receivers etaient informes de ces payoffs?**To avoid these difficulties, we have briefly considered the level-k approach of our game viewed in normal

[23]**It should also be mentioned that level k theories are less adapted to deal with situations in which players would not know the payoff structure of their opponent, which applies to receivers but not senders.**

form.

There are several possible choices for level 0. Let us assume that level 0 (human) senders randomize between telling the truth and lying with probability 50:50 in every period, and level 0 receivers always choose $d_k = 0$ in all periods $k$.

With this specification, level-1 senders would use a deceptive tactic, level-1 receivers would behave as our coarse receivers, level-2 senders would again use a deceptive tactic, level-2 receivers would choose $d_k = 0$ for $k = 1, ..4$ and $d_5 = 0.8$ (anticipating a deceptive tactic on the sender' side), level-3 senders would tell the truth up to and including the key period (anticipating that the deceptive tactic is what receivers expect), level-3 receivers would behave like level-2 receivers, and the behaviors of senders and receivers would cycle for higher levels.

While such an approach would provide some account of our observations (the deception tactic appears as a possible focal behavior), it makes no prediction as to whether the deceptive tactic should be profitable and the other (focal) behaviors emerging from the approach do not show up in our data. Besides, like the sequential equilibrium, the level-k approach predicts that there should be no difference between the 5 period sessions and our main treatments, which is not so in our data.

Finally, given that receivers did not know the payoff structure of senders, it may be argued that they faced an ambiguous environment. Based on the idea that subjects may be ambiguous averse **(Ellsberg ???)**, one might have expected that this would induce more cautiousness on receivers' type, thereby leading receivers to choose decisions closer to $d = 0.5$ in every period. There is no clear evidence of this in our data.

# 5 Conclusion

We have reported results from experiments on multi-period sender-receiver games in which one period has a significantly higher weight. We have observed that players' behaviors are not well captured by the sequential equilibrium of the game. More precisely, senders tend to follow a deceptive tactic (i.e. sending truthful messages until the key period and a false message at the key period) with a much higher frequency than what is described in the sequential equilibrium of the game. Besides, this tactic provides a higher payoff than other

chosen tactics (averaging over those).

We suggest that the high frequency of the deceptive tactic as well as its success can be explained by a different equilibrium concept, the analogy-based sequential equilibrium (ABSE). More than half of the receivers' behavior qualitatively coincide with the behaviors described in the ABSE while the behavior of one fourth of the receivers is close to the sequential equilibrium predictions. This favors the idea that receivers are heterogenous in their cognitive abilities, some share (more than half) employing a coarse reasoning to make their inferences whereas a smaller share (a quarter) would employ a more sophisticated mode of reasoning. Our observations are robust to the introduction of several modifications of the game (notably a change in the share of non-human senders or a change in the weight of the key period) except in the variant in which the game ends at the key period (in which senders seem to be excessively afraid of using the deceptive tactic and instead seem to be playing randomly).

# 6 Appendix

## 6.1 Proof of Proposition 1

By definition, benevolent senders have a unique strictly dominant strategy: always to send truthful messages. Moreover, it is clear (by backward induction) that once a lie has been observed (so that it is common knowledge that both the sender and the receiver are rational), the sender and the receiver play as in the unique Nash equilibrium of the stage game (i.e., the Sender randomizes 50:50 between telling the truth and lying, and the Receiver chooses $d = 0.5$).

Now, the value of $d_k$ conditional on not having observed any false message during the game. In this case, when $k \neq 1$, the conditional probability that the sender is malevolent is: $\frac{4\prod_{i=1}^{k-1}(1-p_j)}{4\prod_{i=1}^{k-1}(1-p_j)+1}$ and since, by definition, the probability that a malevolent sender will choose a false message is $p_k$, the receiver's best response is $d_k = \frac{4\prod_{i=1}^{k-1}(1-p_i)p_k}{4\prod_{i=1}^{k-1}(1-p_i)+1}$. When $k = 1$, the best response is $\frac{4p_k}{5}$.

It remains to prove the uniqueness of the vector $(p_1, ..., p_{20})$ satisfying the conditions of Proposition 1. Suppose that $(p_1, ..., p_{20})$ and $(q_1, ..., q_{20})$ are two different equilibrium vectors[24]. We define $\widetilde{k}$ such that $p_{\widetilde{k}} \neq q_{\widetilde{k}}$ and $\forall\ k$ such that $k < \widetilde{k}$, $p_k = q_k$. We also assume, without loss of generality that $q_{\widetilde{k}} > p_{\widetilde{k}}$.

We introduce $k_q^r$ (resp: $k_p^r$) , the revelation period, defined as follows. for any integer $i < k_q^r$ (resp : $i < k_p^r$), $q_i < 1$ (resp: $p_i < 1$) and $q_i = 1$ (resp: $p_i = 1$) or $k_q^r = 21$ (resp: $k_p^r = 21$). We also denote $d_{k,q}$ (resp: $d_{k,p}$), the equilibrium $d$ chosen by receivers in an equilibrium with a $q$-vector (resp: $p$-vector) in period $k$ conditional on not having received a false message in any prior period.

Let us compare the equilibrium payoff of a malevolent sender sending his first false message in period $\widetilde{k}$ with both types of equilibrium, a $p$-equilibrium and a $q$-equilibrium. In any period $k$ before $\widetilde{k}$, since $p_k = q_k$, the best response of the receiver is the same and the payoff is the same for a malevolent sender sending a truthful message either in a $p$-equilibrium or in $q$-equilibrium. In any period $k$ after a false message in period $\widetilde{k}$, the receiver chooses $d_k = 1/2$ so that the payoff is the same for the malevolent sender which has sent a false message in period $\widetilde{k}$ either in a $p$-equilibrium or in $q$-equilibrium. Now, in period $\widetilde{k}$, in a $p$-equilibrium,

[24] At least one coordinate of these two vectors which is not posterior to a "1" differs in these two vectors.

the receiver chooses a $d_{\widetilde{k},p}$ equal to $\frac{4\prod_{i=1}^{k-1}(1-p_i)p_{\widetilde{k}}}{4\prod_{i=1}^{k-1}(1-p_i)+1}$ and in a $q$-equilibrium, the receiver chooses a $d_{\widetilde{k},q}$ equal to $\frac{4\prod_{i=1}^{k-1}(1-q_i)q_{\widetilde{k}}}{4\prod_{i=1}^{k-1}(1-q_i)+1}$ and since $\frac{4\prod_{i=1}^{k-1}(1-p_i)}{4\prod_{i=1}^{k-1}(1-p_i)+1} = \frac{4\prod_{i=1}^{k-1}(1-q_i)}{4\prod_{i=1}^{k-1}(1-q_i)+1}$, $d_{\widetilde{k},p} < d_{\widetilde{k},q}$ so that the payoff obtained by a malevolent sender sending his first false message in period $\widetilde{k}$ is strictly higher in a $p$-equilibrium than in a $q$-equilibrium. Then, because of the properties of the mixed equilibrium, a malevolent sender always obtains a strictly lower payoff in a $q$-equilibrium than in a $p$ -equilibrium.[25]

We intend to show, by induction, that for any $i \in [\widetilde{k}, k_q^r]$, $p_i = q_i = 0$ or $p_i < q_i$ and $d_{i,p} < d_{i,q}$.

First, we observe that this property is verified for $i = \widetilde{k}$. Now, suppose that for any $i \in [\widetilde{k}, k^*]$ with $k^*$ such that $\widetilde{k} \leq k^* < k_q^r$, $p_i = q_i = 0$ or $p_i < q_i$ and $d_{i,p} < d_{i,q}$. Let us first observe that, since for any $i \in [\widetilde{k}, k^*]$, $p_i = q_i = 0$ or $p_i < q_i$, $k^* < k_p^r$. Now, suppose that $p_{k^*+1}, q_{k^*+1} > 0$ and let us consider a malevolent sender sending his first false message in period $k^* + 1$. He obtains the same payoff in all the periods whether he plays a $p$-equilibrium or a $q$-equilibrium except in periods from $\widetilde{k}$ to $k^* + 1$. In these periods, in a $p$-equilibrium, he obtains $\sum_{j=k}^{k^*} \delta_j d_{j,p}^2 + \delta_{k^*+1}(1 - d_{k^*+1,p})^2$ and in a $q$-equilibrium, he obtains $\sum_{j=k}^{k^*} \delta_j d_{j,q}^2 + \delta_{k^*+1}(1 - d_{k^*+1,q})^2$. Besides, for any $j \in [\widetilde{k}, k^*]$, $d_{j,p}^2 \leq d_{j,q}^2$, this inequality being strict at least for $j = \widetilde{k}$. Because of the indifference in mixed strategies $\sum_{j=k}^{k^*} \delta_j d_{j,p}^2 + \delta_{k^*+1}(1 - d_{k^*+1,p})^2 > \sum_{j=k}^{k^*} \delta_j d_{j,q}^2 + \delta_{k^*+1}(1 - d_{k^*+1,q})^2$. Therefore, $(1 - d_{k^*+1,p})^2 > (1 - d_{k^*+1,q})^2$ which also implies $d_{k^*+1,p} < d_{k^*+1,q}$ and $p_{k^*+1} < q_{k^*+1}$.

Now, we need to show that $p_{k^*+1} > 0$ and $q_{k^*+1} = 0$ is impossible. If this were the case, the payoff of a malevolent sender in the periods between $\widetilde{k}$ and $k^* + 1$, in a $q$-equilibrium, if he deviates and sends his first false message in period $k^* + 1$ would be $\sum_{j=k}^{k^*} \delta_j d_{j,q}^2 + \delta_{k^*+1}$. Because of the arguments we have just mentioned $\sum_{j=k}^{k^*} \delta_j d_{j,q}^2 + \delta_{k^*+1} > \sum_{j=k}^{k^*} \delta_j d_{j,p}^2 + \delta_{k^*+1}(1 - d_{k^*+1,p})^2$. Therefore, a malevolent sender deviating in a $q$-equilibrium, sending his first false message in period $k^* + 1$ obtains more than a malevolent sender in a $p$-equilibrium. This cannot be possible since we showed that a malevolent sender always obtains a strictly lower payoff in a $q$-equilibrium than in a $p$-equilibrium. Hence $p_{k^*+1} > 0$ and $q_{k^*+1} = 0$ is impossible. Hence, for any $i \in [\widetilde{k}, k^* + 1]$, $p_i = q_i = 0$ or $p_i < q_i$ and $d_{i,p} < d_{i,q}$. End of the induction proof.

[25] This implies that $p_i = 0$ for $i < \widetilde{k}$ as otherwise by lying in those periods, the sender would get the same expected payoff both in the $p$ and the $q$-equilibrium, which is not possible as just proven.

Now, we use the result we proved with the induction proof.

First, we show that $k_q^r = 21$ or $k_p^r = 21$ is not possible. Suppose that $k_p^r = 21$. In a $p$-equilibrium, a malevolent sender who did not have sent a false message in any prior period must be indifferent between sending a false and truthful message in period 20 since the choice of a truthful or a false message does not affect his payoff in future period[26].This means that $d_{20,p} = 1/2$. By backward induction, we also obtain that $d_{19,p} = 1/2$, $d_{18,p} = 1/2$ . But, this is not possible at the equilibrium (because the sequence $p_k = \frac{1}{2}\frac{4\prod_{i=1}^{k-1}(1-p_i)+1}{4\prod_{i=1}^{k-1}(1-p_i)}$ exceeds 1 at some point). The same arguments apply to reject $k_q^r = 21$.

Let us consider the $k_q^r < k_p^r < 21$ case and define $\widehat{k}$ as follows: $k_q^r < \widehat{k}$, $p_{\widehat{k}} > 0$ and $\forall i$ such that $k_q^r < i < \widehat{k}$, $p_i = 0$ ($\widehat{k}$ is the first period posterior to $k_q^r$ in which a malevolent sender sends his first false message with a strictly positive probability in a $p$-equilibrium) . We consider the following deviation in a $q$-equilibrium: send the first false message in period $\widehat{k}$. In all the periods before $\widetilde{k}$, after $\widehat{k}$ and between $k_q^r$ and $\widehat{k}$, the payment is the same with this strategy as what a malevolent sender obtains in a $p$-equilibrium if he sends his first false message in period $\widehat{k}$. In a $q$-equilibrium, the receiver does not expect any false message in period $\widehat{k}$ conditional on not having observed any prior false message so that sending a false message, the malevolent sender obtains $\delta_{\widehat{k}}$ in this period, the highest possible payoff. Besides in any period $i$ from $\widetilde{k}$ to $k_q^r$ (including these periods), $d_{i,p} < d_{i,q}$ or $d_{i,p} = d_{i,q} = 0$ (but this cannot be the case in all the periods) so that a malevolent sender deviating in a $q$-equilibrium sending his first false message in period $\widehat{k}$ obtains strictly more than a malevolent sender in a $p$-equilibrium sending his first false message in period $\widehat{k}$. But we found that a malevolent sender always obtain a strictly lower payoff in a $q$-equilibrium than in a $p$-equilibrium. Hence, the deviation is strictly profitable, the $q$-equilibrium is not valid and we can reject the possibility of multiple equilibria. Q.E.D.

[26]Besides, if sending a false message gives a strictly higher payoff, he will send it with probability 1 and $k_p^r = 21$ will not be verified. If sending a false message gives a strictly lower payoff, he will send it with probability 0. Then, the best response will be $d_{20,p} = 0$ but in that case sending a false message gives a strictly higher payoff than sending a truthful payoff.

## 6.2 Proof of Proposition 2

Again, benevolent senders have a unique strictly dominant strategy: always to send truthful messages.

Now, let us consider malevolent senders.

Suppose that, at the equilibrium, on average, they send strictly more false messages than truthful messages, with an average frequency, $q > 1/2$. Then, once a receiver identifies that she is matched with a malevolent sender, she will choose a $d$ equal to $q > 1/2$ during all the following periods of the game since she puts in a unique analogy class all the decision nodes of the sender. Then, the malevolent sender has a unique best response: always to send truthful messages after having being identified as a malevolent sender. But, in this case, it is not possible that, on average, malevolent senders send more false messages than truthful messages.

Suppose that, on average, malevolent senders send strictly more truthful messages than false messages, with an average frequency, $q > 1/2$. Then, once a receiver identifies that she is matched with a malevolent sender (after the first period during which a false message is sent), she will choose a $d$ equal to $1 - q < 1/2$ during all the following periods of the game since she puts in a unique analogy class all the decision nodes of the malevolent sender. Then, the malevolent sender has a unique best response: always to send false messages after having being identified as a malevolent sender. Besides, at the equilibrium, in any period $k$ such that the sender never sent a false message in any past period of the game, a Coarse receiver will choose a $d_k = \frac{4q^{k-1}(1-q)}{4q^{k-1}+1}$ (and $d_1 = 4(1-q)/5$). Since, on average, malevolent senders send strictly more truthful messages than false messages and after the first false message, senders only send false messages, it must be the case that with a strictly positive probability, a malevolent sender sends his first false message after period 11.

We intend to show that a malevolent sender can obtain a higher payoff always sending false messages than sending his first false message in any period $\widehat{k} > 11$ and then only false messages until the end of the game. This would show that sending, on average, strictly more truthful messages than false messages cannot be part of an equilibrium for a malevolent sender.

Suppose that a malevolent sender always sends false messages. He obtains $(1-(4/5)(1-q))^2 > q^2$ in the first period and $q^2$ in all the other period except $k^*$ in which he obtains $5q^2$ so

that $24q^2$ is a lower bound on his payoff. If a malevolent sender sends his first false message in any period $\widehat{k} > 11$ and then only false messages until the end of the game, we obtain the following upper bound on his pay-off: $(3+\widehat{k})(\frac{4(1-q)}{5})^2+1+(20-\widehat{k})q^2 \leq 15(\frac{4(1-q)}{5})^2+1+8q^2$.

$24q^2-15(\frac{4(1-q)}{5})^2-1-8q^2 = (-53+96q+32q^2)/5$ which is strictly positive for $q \in (1/2, 1]$. Hence, a malevolent sender can obtain a strictly higher payoff always sending false messages than sending his first false message in any period $\widehat{k} > 11$ and then only false messages until the end of the game.

Therefore, at the equilibrium, it must be the case that a malevolent sender sends, on average 10 false messages and 10 truthful messages. This implies that once the receiver observed that the sender has sent her a false message, she chooses a $d$ equal to $1/2$ until the end of the game (since the receiver believes that she is matched with a malevolent sender with probability 1 and she perceives that a malevolent sender sends false messages with probability $1/2$ in all his decision nodes). Besides, if in period $k$, a receiver never received any false message in a previous period, he perceives that the sender is malevolent with probability $\frac{4(1/2)^{k-1}}{4(1/2)^{k-1}+1}$ and she will choose $d_k = \frac{4(1/2)^k}{4(1/2)^{k-1}+1}$.

Now, let us consider a malevolent sender. At the equilibrium, he sends, on average, 10 false messages. Besides, we have already identified the equilibrium strategy of the coarse receiver. After having sent his first false message, he is indifferent between sending false or truthful messages since the receiver always chooses a $d$ equal to $1/2$. Therefore, we only need to identify when the receiver sends his first false message and the strategy after this false message must be such that he send on average 10 false messages (the choice of the periods where he sends this 9 other false messages does not affect his payoff since in all these periods $d$ is equal to $1/2$). If we compare the payoff that the sender obtains depending on the first period he sends a false message, easy computations show that he maximizes this payoff choosing $k^*$. Q.E.D.

### 6.3 Proof of Claim 1

First, we need to describe more completely the strategies that we only partially introduced in Claim 1 for rational players.

The rational receiver's strategy is almost fully described except that we need to specify that after period 4, if she has only received truthful messages, she plays as in the sequential

equilibrium of a variant of the game beginning in period 5, with $\frac{16}{35}$ of benevolent senders, $\frac{16}{35}$ of rational malevolent senders and $\frac{3}{35}$ of mechanical senders sending truthful messages with probability 1/2 in each period[27].

For rational senders, we need to specify that in case he observes a $d_k$ different from $\frac{4(1/2)^k}{1+4(1/2)^{k-1}}$ in period $k = 1$, 2, 3 or 4, he plays as in the sequential equilibrium of a variant of the game beginning in period $k+1$, with $\frac{2^k}{3+2^{k+1}}$ of benevolent senders, $\frac{2^k}{3+2^{k+1}}$ of rational malevolent senders and $\frac{3}{3+2^{k+1}}$ of mechanical senders sending truthful messages with probability 1/2 in each period. Let us also mention that conditional on having observed $d_k = \frac{4(1/2)^k}{1+4(1/2)^{k-1}}$ in the 5 first periods of the game, a rational sender sends, on average, 9 false messages during the last 15 periods of the game except if he observes a $d_j \neq 1/2$ for $j > 5$. In such a case, he sends a false message with probability 1/2 in the remaining periods of the game.

Now let us check that these strategies are constitutive of an ABSE.

A coarse malevolent sender puts all the decision nodes of the receivers in a unique analogy class. Therefore, he does not perceive the link between the message he sends and the decision of the receivers he is matched with. Sending a truthful and a false message with probability 1/2 in all the periods is a best response with this belief.

The strategy of a coarse sender is the same as in Proposition 2 since her perception of the game is the same: Benevolent senders always send truthful messages and malevolent senders send false and truthful messages with a probability 1/2 in all their decision nodes of the game.

Considering senders' strategies, a rational receiver cannot raise raise his payoff choosing a $d \neq 1/2$ conditional on having observed at least one false message. Therefore, we can focus on her behavior conditional on not having received any false message. A rational receiver must decide in that case whether she mimics coarse receivers or she reveals her type choosing a different $d$. If she reveals her type in period $k$, a coarse sender will continue sending a false message with probability 1/2 in all the periods and a rational sender will play as in the sequential equilibrium of a variant of the game beginning in period $k+1$, with $\frac{2^k}{3+2^{k+1}}$ benevolent senders, $\frac{2^k}{3+2^{k+1}}$ rational malevolent senders and $\frac{3}{3+2^{k+1}}$ mechanical senders

[27]This sequential equilibrium is uniquely defined. This can be proved using the same elements as in the proof of Proposition 1

sending truthful messages with probability 1/2 in each period. Therefore, the best response for a rational receiver will be also to play as in the sequential equilibrium of a variant of the game beginning in period $k+1$, with $\frac{2^k}{3+2^{k+1}}$ benevolent senders, $\frac{2^k}{3+2^{k+1}}$ rational malevolent senders and $\frac{3}{3+2^{k+1}}$ mechanical senders sending truthful messages with probability 1/2 in each period. Now, a rational receiver must choose the period $k$ in which she reveals her type and $d_k$. Since the value of $d_k$ does not affect the payoff in the following periods as long as $d_k \neq \frac{4(1/2)^k}{1+4(1/2)^{k-1}}$, her best choice is a $d_k$ which maximizes her period expected payoff i.e. if $k<5$, $d_k = \frac{3(1/2)^k}{2+3(1/2)^{k-1}}$, $d_5 = \frac{1}{2}$ and if $k>5$, $d_k = \frac{3(1/2)^k}{1+3(1/2)^{k-1}}$. Finding the $k^*$ that maximizes the rational receiver expected payoff is only a matter of computations (requiring to compute expected payoff in the sequential equilibria of all the considered variants of the game). The solution is $k=4$

Rational senders. Again, the key element is the period of the first false message. After this first false message, in all the remaining periods, $d_k = 1/2$, therefore any choice is a best response and he obtains $\frac{\delta^k}{4}$ in period $k$. Then, considering the strategies of the different types of receivers, it is only a computation issue to find the best choice for a rational sender. As long as he believes that he is matched with a coarse receiver with probability 3/4, he obtains a higher payoff sending his first false message in period 5 (his expected payoffs conditional on sending a first false message in period 1, 2; 3, 4 or 5 are respectively and approximatively 2.36, 2.3544, 2.3336, 2.2781 and 3.711), following a deceptive tactic. When he discovers, in period 4, that he is matched with a rational sender, he plays a best response that coincides with his strategy in a sequential equilibrium of a variant of the game beginning in period 5, with $\frac{2^4}{3+2^5}$ benevolent senders, $\frac{2^4}{3+2^5}$ rational malevolent senders and $\frac{3}{3+2^5}$ mechanical senders sending truthful messages with probability 1/2 in each period.

Now, the $d_4 = 3/38$ chosen by rational receiver is precisely equal to $\frac{3(1/2)^4}{2+3(1/2)^3}$ and the other approximated numerical values that we mention in the claim derive all from the values of the sequential equilibrium of the variant of the game beginning in period 5, with $\frac{2^5}{3+2^6}$ benevolent senders, $\frac{2^5}{3+2^6}$ rational malevolent senders and $\frac{3}{3+2^6}$ mechanical senders sending truthful messages with probability 1/2 in each period.

Q.E.D.

## References


[1] Blume, A., DeJong, D. and Sprinkle, G. (1998): 'Experimental Evidence on the Evolution of Meaning of Messages in Sender-Receiver Games', *American Economic Review*, **88**, 1323-1340.

[2] Blume, A., DeJong, D., Kim, Y.-G. and Sprinkle, G. (2001): 'Evolution of Communication with Partial Common Interest', *Games and Economic Behavior*, **37**, 79-120.

[3] Cai, H. and Wang, J. T. (2006): 'Overcommunication in Strategic Information Transmission Games', *Games and Economic Behavior*, **56**, 7-36.

[4] Camerer, C. F. (2003): Behavioral Game Theory: Experiments on Strategic Interaction, Princeton University Press.

[5] Camerer, C. F. and Weigelt, K. (1988): 'Experimental Tests of Sequential Equilibrium Reputation Model', *Econometrica*, **56**, 1-36.

[6] Crawford, V. P. (2003): 'Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentations of Intentions', *American Economic Review*, **93**, 1193-1235.

[7] Crawford, V. P. and Sobel, J. (1982): 'Strategic Information Transmission', *Econometrica*, **50**, 1431-1451.

[8] Dickhaut, J., McCabe, K. and Mukherji, A. (1995): 'An Experimental Study of Strategic Information Transmission', *Economic Theory*, **6**, 389-403.

[9] Ettinger, D. and Jehiel, P. (2008): 'A Theory of Deception', UCL Working paper.

[10] Jehiel, P. (2005): 'Analogy-based Expectation Equilibrium', *Journal of Economic Theory*, **123**, 81-104.

[11] Jehiel, P. and Koessler, F. (2007): 'Revisiting Games of Incomplete Information with Analogy-based Expectations', *Games and Economic Behavior*, **62**, 533-557.

[12] Jung, Y. J., Kagel, J. H. and Levin, D. (1994): 'On the Existence of Predatory Pricing: An Experimental Study of Reputation and Entry Deterrrence in the Chain-Store game', *Rand Journal of Economics*, **25**, 72-93.

[13] McKelvey, R. D. and Palfrey, T. R. (1995): 'Quantal response equilibria for normal form games', *Games and economic behavior*, **10**, 6-38.

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

[15] Kreps, D. and Wilson, R. (1982): 'Reputation and Imperfect Information', *Journal of Economic Theory*, **27**, 253-279.

[16] Nagel, R. (1995): 'Unraveling in Guessing Games: An Experimental Study', *American Economic Review*, **85**, 1313-1326

[17] Neral, J. and Ochs, J. (1992): 'The sequential equilibrium Theory of Reputation: A Further Test', *Econometrica*, **60**, 1151-1169.

[18] Perrault, G. (1967): *L'orchestre rouge*, Fayard translated as: *The Red Orchestra: Anatomy of the most Successful Spy Ring in WWII* (1967), Simon and Schuster.

[19] Sobel, J. (1985): 'A theory of Credibility', *Review of Economic Studies*, **52**, 557-573.

[20] Spence, A. M. (1973): 'Job Market Signaling', *Quarterly Journal of Economics*, **87**, 357-374.

[21] Stahl, D Wilson, P. (1994): 'Experimental Evidence on Players Models of Other Players', *Journal of Economic Behavior and Organization*, **25**, 30927.

[22] Stahl, D. and Wilson, P. (1995): 'On Player's Modals of other Players: Theory and Experimental Evidence', *Games and Economic Behavior*, **10**, 218-254.

[23] Trepper, L. (1975): *Le Grand Jeu, Memoires du Chef de l'Orchestre Rouge*, Albin Michel translated as: *The Great Game: Memoirs of the Spy Hitler couldn't Silence* (1977), McGraw Hill.

[24] Wang, J. T. , Spezio, M. and Camerer, C. F. (2006): 'Pinocchio's Pupil: Using Eyetracking and Pupil Dilation To Understand Truth-telling and Deception in Games', working paper California Institute of Technology.

[25] Von Neuman J. and Morgenstern, O. (1944): *Theory of Games and Economic Behavior*, Princeton University Press.

evolution.pdf

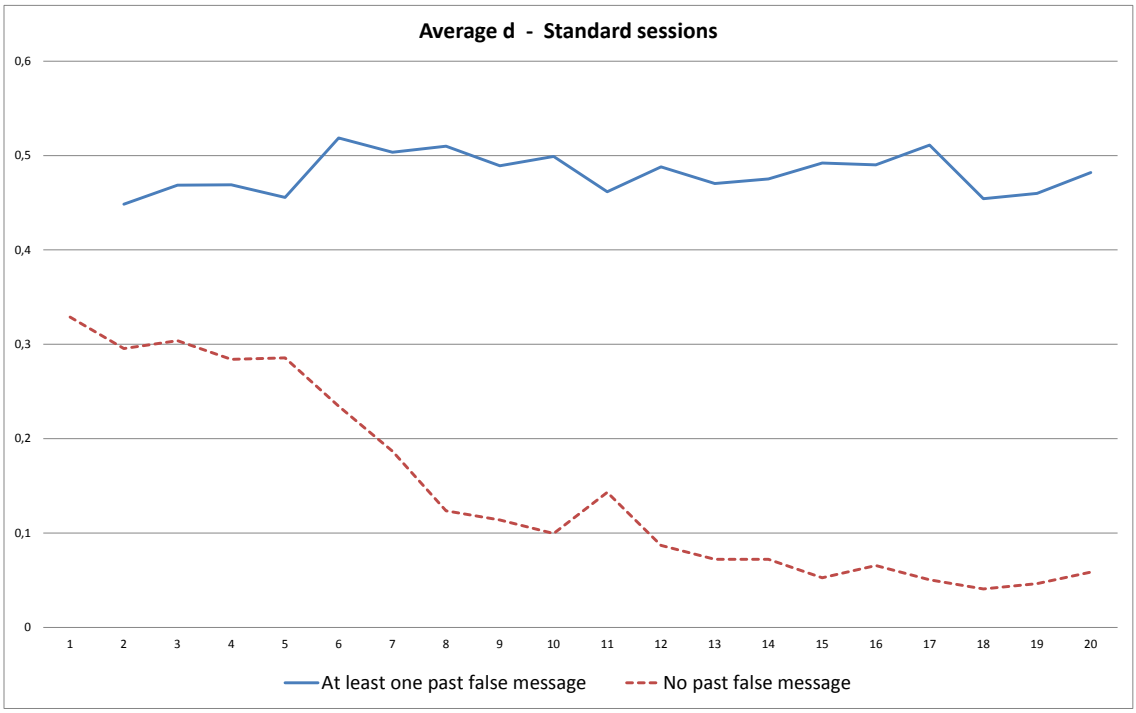


Figure 1:

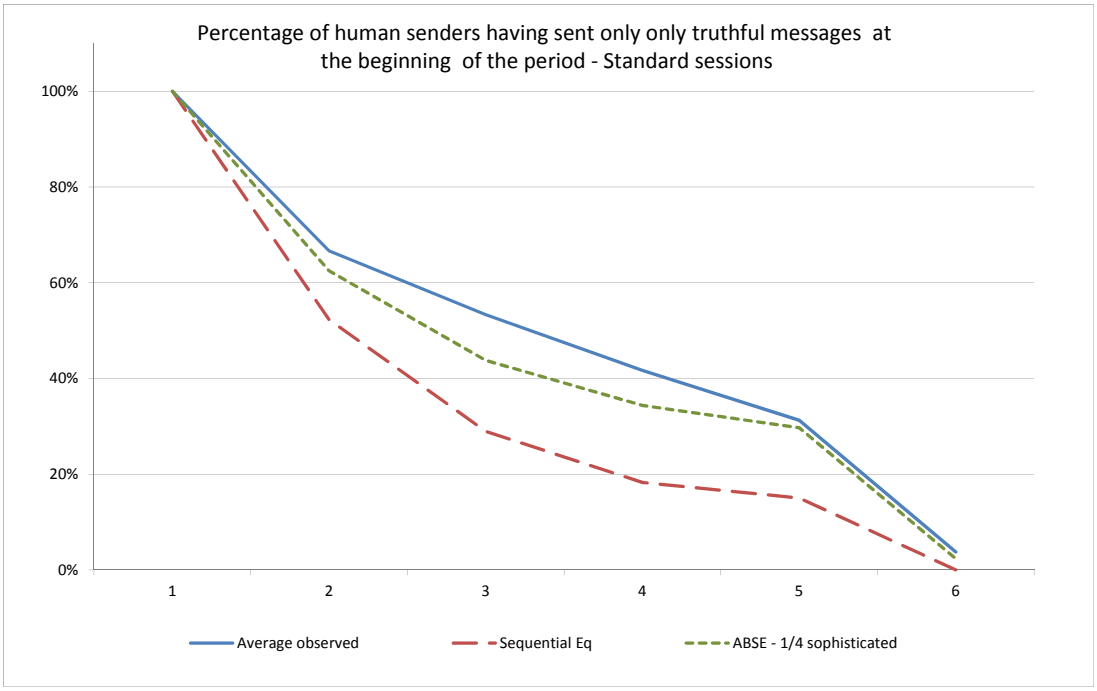
Percentage of human senders having sent only only truthful messages at the beginning of the period - Standard sessions
100%
80%
60%
40%
20%
0%
1
2
3
4
5
6
Average observed
Sequential Eq
ABSE - 1/4 sophisticated

Figure 2:

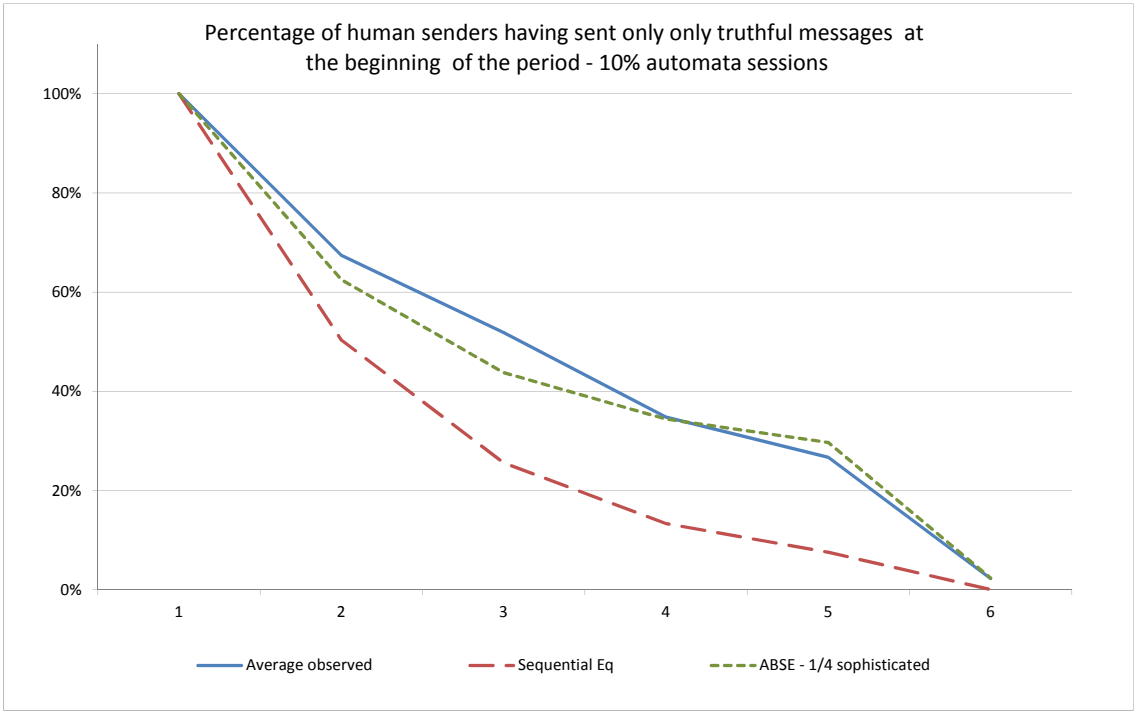
Percentage of human senders having sent only only truthful messages at the beginning of the period - 10% automata sessions
100%
80%
60%
40%
20%
0%
1
2
3
4
5
6
Average observed
Sequential Eq
ABSE - 1/4 sophisticated

Figure 3:

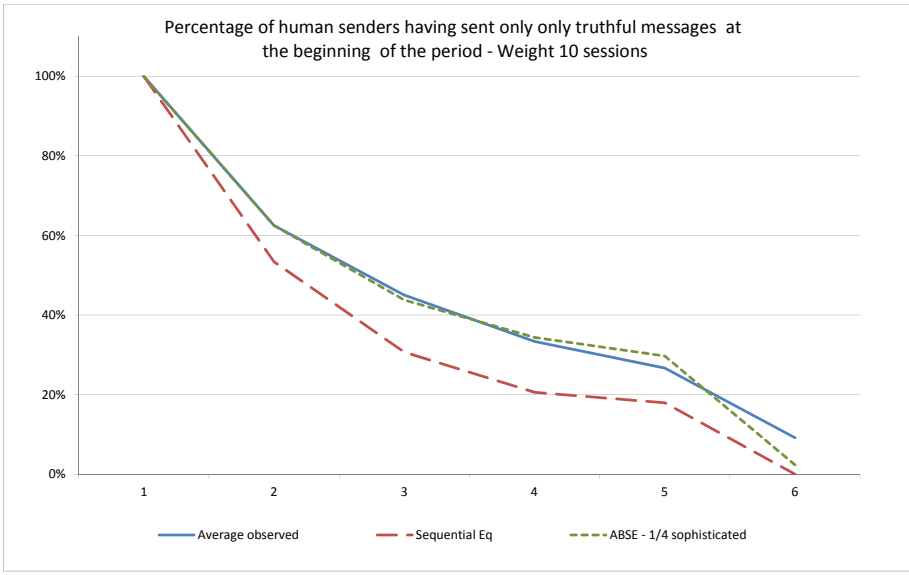
Percentage of human senders having sent only only truthful messages at the beginning of the period - Weight 10 sessions
100%
80%
60%
40%
20%
0%
1
2
3
4
5
6
Average observed
Sequential Eq
ABSE - 1/4 sophisticated

Figure 4:

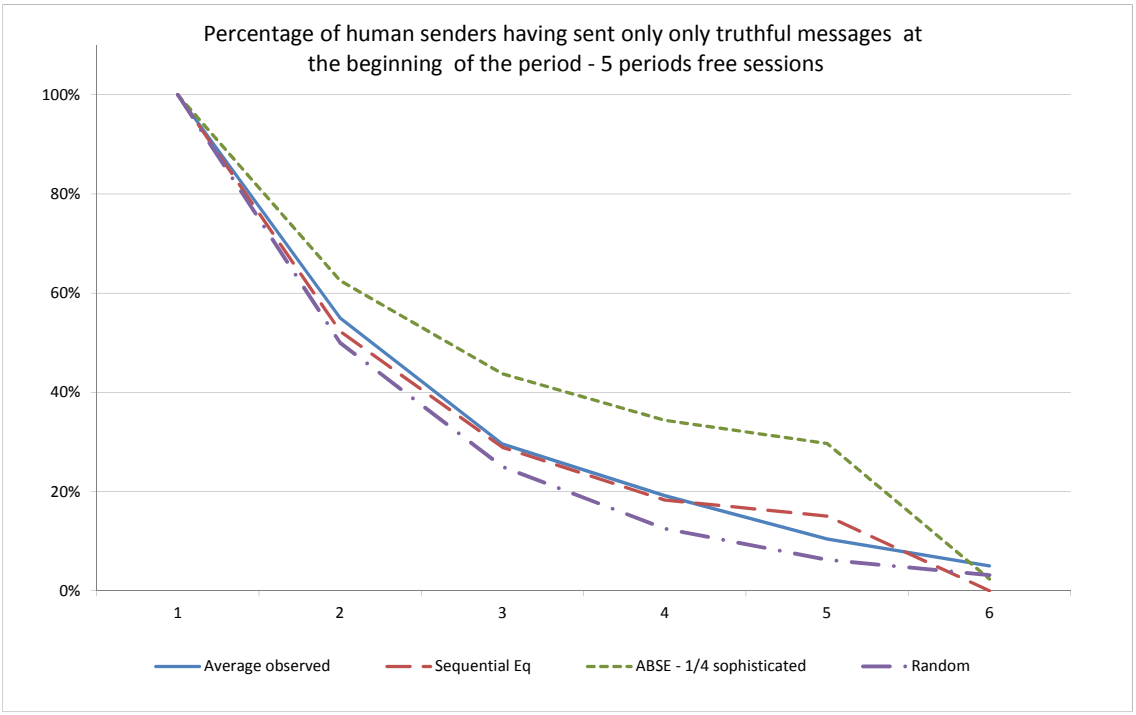
Percentage of human senders having sent only only truthful messages at the beginning of the period - 5 periods free sessions
100%
80%
60%
40%
20%
0%
1
2
3
4
5
6
Average observed
Sequential Eq
ABSE - 1/4 sophisticated
Random

Figure 5:

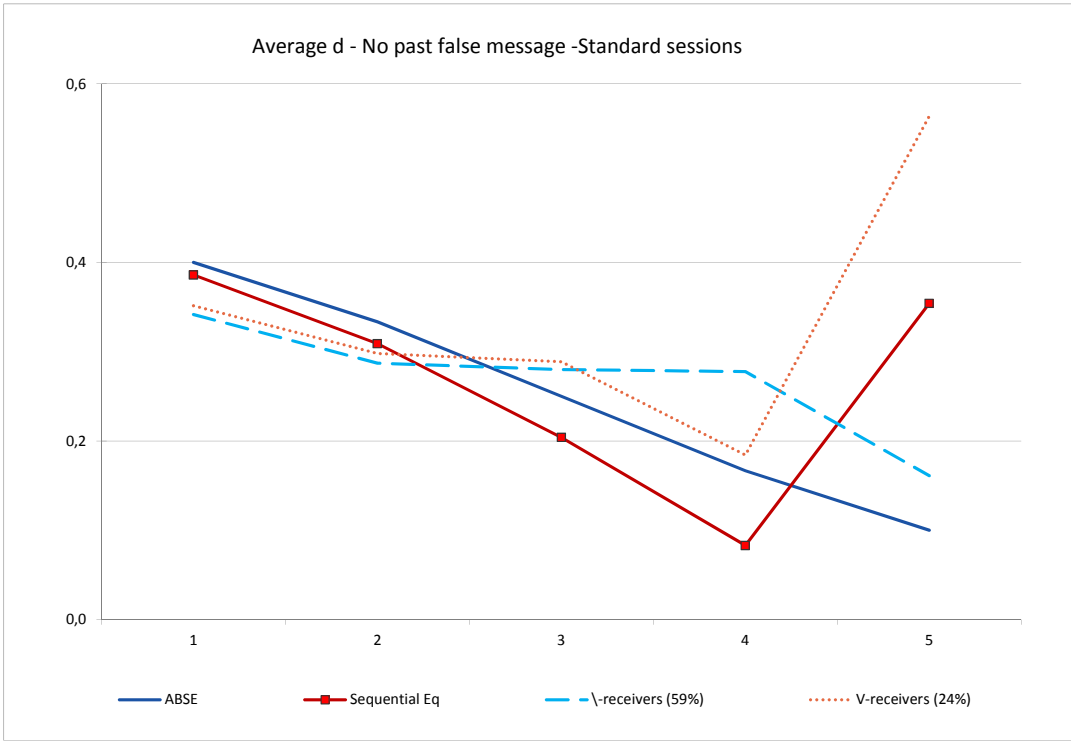
Average d - No past false message -Standard sessions
0,6
0,4
0,2
0,0
1
2
3
4
5
ABSE
Sequential Eq
\-receivers (59%)
V-receivers (24%)

Figure 6:

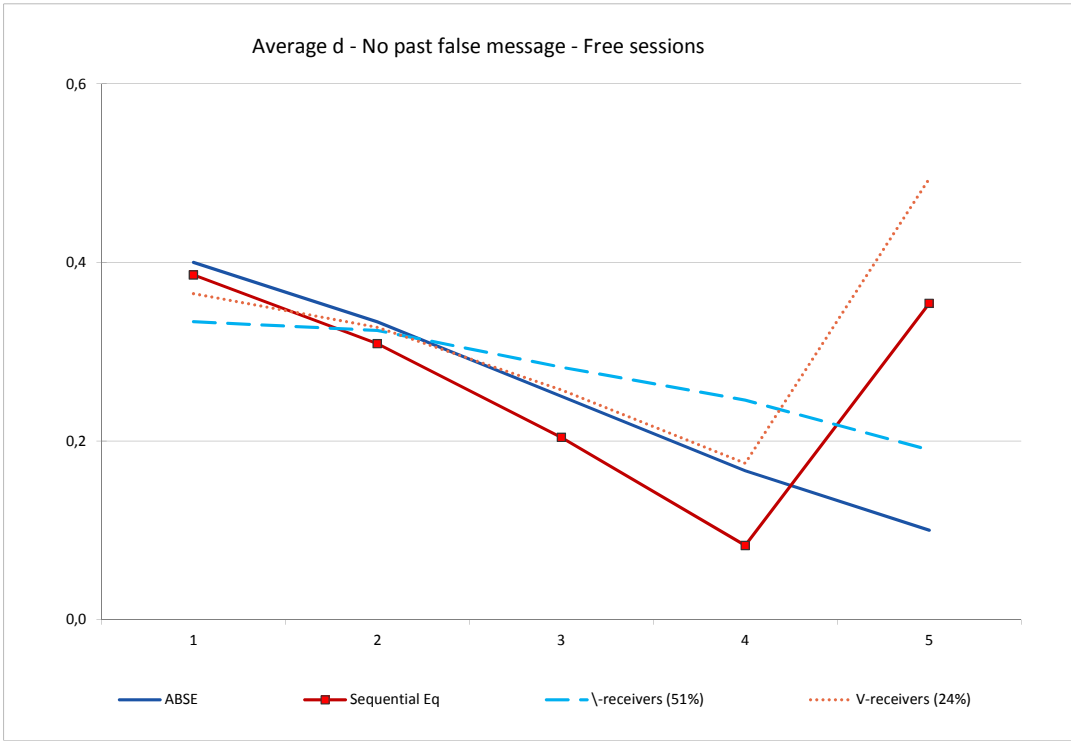
Average d - No past false message - Free sessions
0,6
0,4
0,2
0,0
1
2
3
4
5
ABSE
Sequential Eq
\-receivers (51%)
V-receivers (24%)

Figure 7:

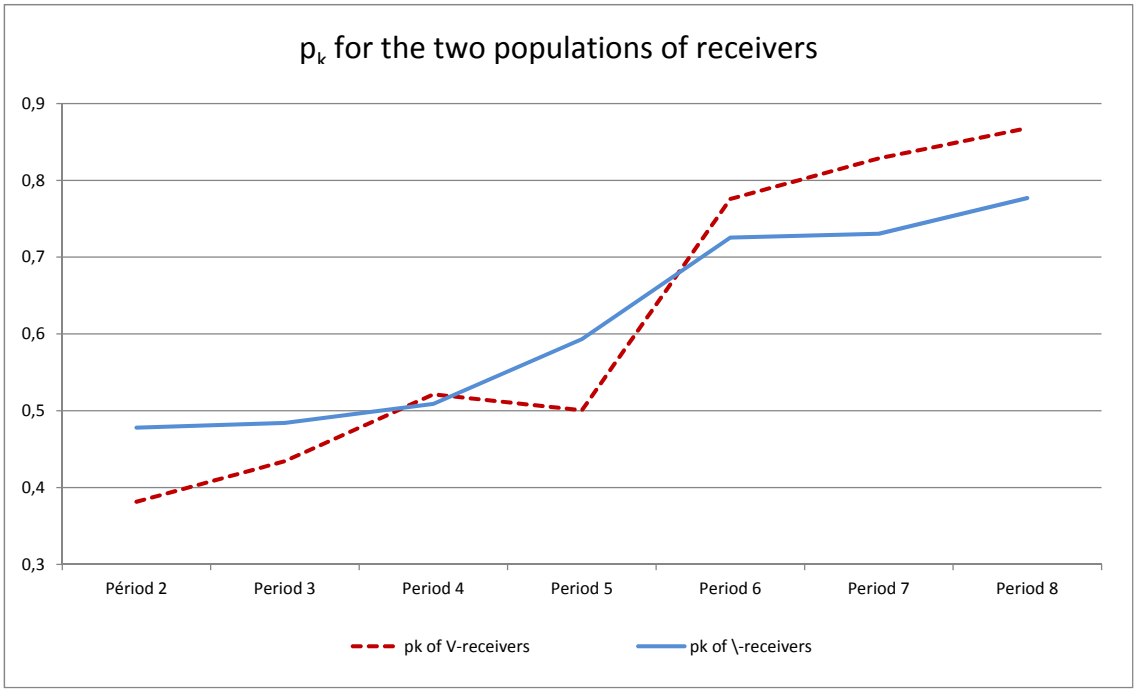
$p_k$ for the two populations of receivers
0,9
0,8
0,7
0,6
0,5
0,4
0,3
Période 2
Period 3
Period 4
Period 5
Period 6
Period 7
Period 8
pk of V-receivers
pk of \-receivers

Figure 8: