An Experiment on Deception, Reputation and Trust*

David Ettinger† and Philippe Jehiel‡

September 2, 2016

Abstract

We report results from an experiment on a repeated sender/receiver game with twenty periods in which one period has higher weight. The sender communicates about the realized state in each period, the receiver takes an action matching his 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 three quarters of subjects reason coarsely when making inferences and forming expectations about others' behaviors.

*We thank Maxim Frolov for assistance on the experimental design, Guillaume Frechette, Guillaume Hollard, Frederic Koessler, Dov Samet, Jean Marc Tallon, and the participants of the Neuroeconomic Workshop, the Extensive form Games in the Lab Workshop, LSE-UCL workshop, LSE behavioral economics seminar, The first Socrates workshop, the ASFEE conference, the IHP, Dauphine, Cerge-EI, HEC-Polytechnique, Paris 1, Technion seminars' participants for helpful comments. Jehiel thanks the European Research Council for funding and Ettinger the Governance and Regulation Chair for its support.

†Université Paris-Dauphine, PSL Research University, LEDa and CEREMADE, 75016 Paris, France and CIRANO ; david.ettinger@dauphine.fr

‡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. A major concern of this network was about maintaining a secret communication channel with Moscow while preserving the security of its members. The chosen solution was to organize the Red Orchestra consisting of several radio transmission cells. 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. Thanks to tricks and torture, they captured the members of the cells connected to these radio transmitters. They even convinced some members of the cells 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 these to send information to Moscow. 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 services was to maintain a high level of trust in the mind of the Russian services regarding 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, who had been arrested and pretended to cooperate with the Gestapo services managed to send a message to the Russian 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).

The case of the Red Orchestra is a vivid example of a repeated information transmission game in which the organization sending the information and the organization receiving the information may possibly have conflicting interests, and some pieces of information may be attached to higher stakes. We believe that there are many environments with similar characteristics. To suggest a very different context, consider the everyday life of politicians: they intervene frequently in the media and elsewhere (the communication aspect); they sometimes care more about being reelected than just telling the truth about the state of the economy (the potential difference of objective between the sender and the receiver), and as reelection time approaches, they become more anxious to convey the belief that they are highly competent and trustworthy (the high stake event).

The key strategic considerations in such multi-stage information transmission environments are: 1) How does the receiver use the information he gets to assess the likelihood of the current state of affairs but also to assess the type of sender he is facing (which may be useful to interpret subsequent messages)? and 2) How does the sender understand and make use of the receiver's inference process? On the receiver side, it requires understanding how trust or credibility evolves. On the sender 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, reputation, 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 (but knows the sender is). The sender sends a message regarding the state of the world, then the receiver chooses an action. The receiver takes an action so as to match his 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 she is malevolent in which case her objective is to induce actions of the receiver as far as possible from the states of the world. The receiver ignores the type of the sender he is matched with, but he discovers at the end of each period whether the message received during this period was truthful or not (in the Red Orchestra case, the Russian services could test reasonably quickly the accuracy of the information sent).

Furthermore, 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 five times as large as the other periods and the initial share of benevolent senders (implemented as machines in the lab) represents 20 % of all senders. In the baseline treatment, we also chose not to inform the receiver that human senders had preferences opposed to them so as to be closer to real life incarnations of this game (in which economic agents would typically have to infer preferences from observed behaviors), and we forced human senders to lie roughly half of the time while leaving them free to choose when to lie so as to focus on the strategic timing of lies rather than on whether subjects are reluctant to lie (receivers were informed of the 50:50 lie rate of human senders).

We considered several variants of the baseline treatment either increasing the weight of the key period to 10 or reducing the share of benevolent senders to 10% or letting free the number of lies or letting the receivers know about the preferences of human senders or stopping the interaction right after the key period (in which case the number of lies was left free).

We obtained the following results :

- In the baseline treatment, a large share of senders (roughly 27%, i.e. 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). 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.

- The behaviors are roughly the same whether or not senders are constrained in their

number of lies and whether or not receivers are informed of senders' preferences (for the latter, this is true when subjects play the game for the first two times, while some learning effect is observed after more plays of the game).

- Senders' behaviors are very different when the interaction stops right after the key period in which case the share of deceptive tactics drops to 5%.

The observation that behaviors are not much affected whether or not receivers are informed of senders' preferences suggests either that this information is not much used (at least in the initial plays of the game) or that, in the absence of it, receivers consider the worst case scenario (and accordingly behave as if they were informed that human senders had opposed preferences). The remaining observations present a challenge for interpretative purposes. Assuming subjects behave as in the sequential equilibrium (SE) of the corresponding game with opposed preferences does not provide a good account of the observations for several reasons:

(a) Senders follow the deceptive tactic too often.[2]

(b) The deceptive tactic is successful in the sense that, in our data, deceptive senders obtain higher payoffs than non-deceptive senders while sequential equilibrium would predict that all employed strategies should be equally good.

(c) 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.

(d) The share of deceptive tactic should be the same whether or not the interaction stops right after the key period.

Faced with these challenging observations, we suggest interpreting our findings by considering that at least some share of our subjects followed an inference process that is less

[2]Indeed, it cannot be part of a sequential equilibrium that the share of deceivers exceeds 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.

sophisticated than the one involved in SE. Specifically, given that receivers knew that human senders lie overall half of the time, a simple inference process for coarse receivers consists in believing that human senders lie half of the time in every period independently of the history and update their beliefs accordingly. If a human sender knew she was facing such a receiver and given her preferences, she would pick the deceptive tactic. Indeed, by telling the truth up to the period just before the key period, she would increase considerably the belief in the coarse receiver's mind that she is a benevolent machine, which she could exploit at the key period by lying.

The kind of reasoning just proposed involving naive inference on the receiver side and deception on the sender side -while at odds with SE- can be captured in the framework of 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 and Jehiel and Samuelson (2012) for some recent application of ABSE). It should be noted that the equilibrium approach of ABSE only requires for coarse receivers that they would be knowledgeable of the aggregate lie rate of the two types of senders, which fits well with our choice in most treatments to provide subjects with that information (and we believe fits also well with real life incarnations of such interactions insofar as lying attitudes are generally known accurately at an aggregate level but not at a detailed one).

We observe within the ABSE framework that allowing subjects -senders or receivers- to be either coarse[3] with probability 3/4 or rational with probability 1/4 provides a good account of the (qualitative) observations made above with the exception of the 5 period-treatment in which the interaction stops right after the key period (in which case the share of coarse senders should be assumed to be much larger to account for the very small proportion of deceptive tactics observed in that treatment). Maybe the most notable feature of our finding is that the same share of coarse subjects works well for most of the variants we considered. From this perspective the main observation that is left unexplained is why the share of deceptive tactics drops so dramatically when the interaction stops right after the key period (which

[3]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. When rational, a subject would play optimally as in standard equilibrium approaches.

may possibly be attributed to the idea that when the key period is final, subjects are less likely to consider that others use the naive inference process as described above).

The rest of the paper is devoted to making the analysis of the game, its SE and its ABSE as well as the statistical analysis more precise. We will also discuss alternative competing approaches as well as whether some learning effect is observed in the various treatments.

Our study is mainly related to the experimental 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. A key difference with that literature is our focus on repeated expert-agent communication games in which there is no value for a malevolent sender to being permanently confounded with a machine always telling the truth, but only a value to being temporarily confounded so as to take advantage of it at the key period.[4] Interestingly, previous studies on reputation games have suggested that the sequential equilibrium may be a powerful tool to organize the data,[5] which contrasts with our finding that theories beyond the sequential equilibrium are needed to give a reasonable account of the data in our experiment.

Our study is also related to a lesser extent to the experimental literature on non-repeated strategic information transmission games à la Crawford and Sobel (1982) that was initiated by Dickhaut et al (1995) and Blume et al (1998) (see also Blume et al (2001), Cai and Wang (2006), Kawagoe and Takizawa (2009) or Wang et al (2010)). That literature has noted that senders have a tendency to transmit more information than theory predicts suggesting that (at least some) senders may be averse to lying.[6] It has also suggested that receivers may be more credulous than theory predicts. Our study is complementary to that strand of literature to the extent that our main interest is focused on the timing of the lies and the dynamic inference process which cannot be studied in non-repeated communication games. In relation to that literature, it may be mentioned that in the variant in which subjects were free in their number of lies, we observed very few senders who never lied. That is, we did not

[4] In the chain-store game, the monopolist would like to be considered to be always fighting in case of entry.

[5] Sometimes, references to homemade (i.e subjective) beliefs were required (Camerer and Weigelt (1988)), or some departures from the theoretical predictions were observed close to the end of the interaction in the mixed strategy phase (Jung et al. (1994)).

[6] Gneezy (2005) also suggests an aversion to lying in his experiment. But, the aversion to lying has been questioned in recent papers which have considered related but different environments (see Embrey et al (2015) or Vespa and Wilson (2016)).

observe the bias toward truth-telling that this previous literature had identified (this may be the consequence of the zero-sum nature of the preferences considered in our experiment).

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 k, the sender (but not the receiver) is informed of the state of the world s_k prevailing in this period. The receiver discovers the state of the world of period 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 $\frac{1}{2}$ for each realization.

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 $a_k \in [0, 1]$ after which he is informed of s_k.

The receiver's payoff in period k is equal to $\delta_k(1 - (a_k - s_k)^2)$ where δ_k is the weight of period k. The overall payoff of the receiver is $\sum_{k=1}^{20} \delta_k(1-(a_k-s_k)^2)$. The choice of a quadratic scoring rule ensures that if the receiver only considers the current period's payoff, he will pick the action that corresponds to what he expects to be the expected value of s_k given the message he 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 α (in the experiment, α will be either $\frac{1}{10}$ or $\frac{1}{5}$), the sender is *benevolent*. This means that she strictly prefers sending a truthful message about the state of the world in all periods.[7] With probability $1 - \alpha$, the sender is

[7] Observe that we do not define benevolent senders as having the same preferences as receivers but being otherwise free to send whatever message they like. We force benevolent senders to transmit truthful messages (which is implemented in the lab by using machines always telling the truth, see below).

malevolent. A malevolent sender's payoff in period k is equal to $\delta_k(a_k - s_k)^2$ and her overall payoff is $\sum_{k=1}^{20} \delta_k(a_k - s_k)^2$. Hence a malevolent sender's objective is to minimize the receiver's payoff. Observe that the weights of the periods are the same for the sender and the receiver.

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 sender is said to employ a *deceptive* tactic if $m_k = s_k$ 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.[8]

2.1 Sequential equilibrium analysis

The strategy of the benevolent sender being fixed by the very definition of her 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. Since a benevolent sender never sends false messages, by sending a false message, a malevolent sender fully reveals her type. It follows by backward induction, that, in any sequential equilibrium, in all periods following this *revelation*, the malevolent sender sends a truthful message with probability $\frac{1}{2}$ and the receiver chooses action $\frac{1}{2}$.[9] 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* $(p_1, p_2, ..., p_{20})$*:*[10]

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

[9] 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.

[10] 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.

- *Conditional on not having sent a false message before, a malevolent sender sends a false message 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.*

- *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 $d_k = \frac{1}{2}$.*

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

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

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

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

Proof: See the Appendix.

The conditions of the Proposition posit that a malevolent sender should be indifferent as to when to make her first lie over periods in which she 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 specific parameter values (δ_{k^*}, α) is essentially a numerical exercise that deals with the indifference conditions. For the sake of illustration (but also for the purpose of interpreting our experimental data), we provide an approximate value of the equilibrium p vector when (δ_{k^*}, α) is equal to $(5, \frac{1}{5})$: (0.477, 0.447, 0.368, 0.177, 1, 1,..., 1). Such a p vector implies that a malevolent sender chooses a deceptive tactic with probability $\prod_{i=1}^{4}(1-p_i)p_5$, which is close to 0.15.

Roughly, the strategic considerations of this game can de understood as follows. Considering the higher value of δ_{k^*} as compared to δ_k, $k \neq k^*$, a malevolent sender would like to

persuade the receiver that she is benevolent by sending truthful messages during the $k^* - 1$ initial periods if it allowed her to obtain a high payoff in period k^* (a deceptive tactic). However, choosing this tactic with too high a probability for a malevolent sender cannot be part of an equilibrium even for very high values of δ_{k^*}. To see this, observe that if the malevolent sender were choosing a deceptive tactic with probability 1, she would obtain $\alpha^2\delta_{k^*}$ adding up payoffs from period 1 to k^* (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 she could obtain $1 + \frac{k^*-2+\delta_{k^*}}{4}$ over the same range of periods if she were deviating, sending a false message in the first period. More intuitively, if a malevolent sender were following a deceptive tactic she would not be much trusted at the key period, which in turn would make the deceptive tactic suboptimal.

The above observation holds true even for large values of δ_k. As δ_k becomes larger, even though malevolent senders 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 $\frac{1}{2}$.The proportion of malevolent senders choosing a deceptive tactic never exceeds the proportion of benevolent senders (as otherwise it would be counterproductive 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 δ_{k^*} but to a limited extent. For instance, if $(\delta_{k^*}, \alpha) = (10, \frac{1}{5})$, a malevolent sender chooses a deceptive tactic with a probability close to 0.18. Lowering the probability α of being matched with a benevolent sender reduces the frequency of deceptive tactic. For instance, if $(\delta_{k^*}, \alpha) = (5, \frac{1}{10})$, a malevolent sender chooses a deceptive tactic with a probability close to 0.075.

Let us also observe that for $(\delta_{k^*}, \alpha) = (5, \frac{1}{5})$, the malevolent senders' behavior is the same in the last $20 - k^*$ periods of the game. They send a false message with probability $\frac{1}{2}$. In equilibrium, the sender's type is always revealed at the end of period k^*. This also implies that the last $20 - k^*$ periods of the game do not affect equilibrium behaviors in the first k^* periods of the game. In particular, 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

To analyze the data, we consider the analogy-based sequential equilibrium (ABSE) as defined in Ettinger and Jehiel (2010), which proves a useful alternative to the sequential equilibrium. Without going into the details of this equilibrium concept, we consider the following cognitive environment. Both malevolent senders and receivers may be of two different cognitive types. With probability $\beta \in [0,1]$, they are standard rational players and, with probability $1-\beta$, they are coarse players not distinguishing their opponent's behavior as a function of history. Types are private information, they are distributed independently across players, and the share β is assumed to be known by rational players.

Let us be more precise about the description of coarse players. On the sender side, coarse players put all the decision nodes of the receivers in the same analogy class and therefore do not perceive the effects of the messages they send on receivers' decisions. As a result, they may choose optimally to randomize 50:50 between telling the truth and lying independently of history. On the receiver side, coarse players are aware that there are benevolent (machine) senders who always tell the truth and human (non-machine) senders and they know the proportion of the two. Regarding human senders, coarse receivers are assumed to know only the aggregate lie rate over the 20 periods, but not how their behaviors depend on the history of play.[11] Moreover, coarse receivers are assumed to reason as if human senders were behaving in a stationary way as given by their aggregate lie rate (this assumption is meant to formalize the idea that coarse receivers consider the simplest theory that is consistent with their knowledge).

In equilibrium, rational players play a best-response to other players' strategies and coarse players play a best-response to their perceptions of other players' strategies, using Bayes' rule to revise their beliefs about the type of player they are matched with.

In order to show how ABSE works, we describe an equilibrium with $\beta = \frac{1}{4}$, assuming that $\alpha = \frac{1}{5}$ so as to match the conditions of the baseline treatment. We choose this specific value of β because we will see later that the corresponding ABSE provides a good approximation of the observed experimental data (we briefly discuss later the effects of changing the value

[11] In our experiment (and we believe in a number of applications too), subjects had explicitly access to these aggregate statistics.

of the parameter β).

Proposition 2 *There exists an analogy-based sequential equilibrium of the game just defined with* $\beta = \frac{1}{4}$ *and* $\delta^* = 5$*, satisfying the following properties:*

- *A coarse malevolent sender uniformly randomizes between sending false and truthful messages during the 20 periods of the game. She sends, on average, 10 false and 10 truthful messages during the 20 periods of the game.*

- *A rational malevolent sender always sends truthful messages during the first 4 periods, sends a false message in period* 5 *and randomizes between truthful and false messages during the last 15 periods of the game. She 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, a coarse receiver chooses* $d_k = \frac{(1-\alpha)(\frac{1}{2})^k}{\alpha+(1-\alpha)(\frac{1}{2})^{k-1}}$*. In any period* k *such that the sender has already sent a false message at least once in a former period, he chooses* $d_k = \frac{1}{2}$*.*

- *In any period* k *such that the sender has already sent a false message at least once in a former period, a rational receiver chooses* $d_k = \frac{1}{2}$*. During the first 4 periods of the game, a rational receiver mimics the behavior of coarse receivers. Conditional on not having observed a false message during the first 4 periods, a rational receiver chooses* $d_5 = \frac{(1-\alpha)(\beta+(1-\beta)(\frac{1}{2})^5)}{(1-\alpha)(\beta+(1-\beta)(\frac{1}{2})^4)+\alpha}$ *and, for* $k > 5$*, if he did not observe a false message in any prior period, he chooses* $d_k = \frac{(1-\alpha)(1-\beta)(\frac{1}{2})^k}{\alpha+(1-\alpha)(1-\beta)(\frac{1}{2})^{k-1}}$*.*

Proof: See the Appendix.

The intuition for Proposition 2 works as follows. As already mentioned, coarse senders find it optimal to send a false message with probability $\frac{1}{2}$ in all periods and after any history because they fail to see any link between the messages they send and receivers' decisions.

Malevolent senders, independently of their cognitive types, send, on average, 10 false messages and 10 truthful messages. Therefore, coarse receivers have the perception that there are two types of senders with the following characteristics: With probability α, senders

are *honest* and always send truthful messages and with probability $1-\alpha$, senders are non-trustworthy -we refer to such senders as *liars*- and send truthful and false messages with probability $\frac{1}{2}$ in each period.

When observing at least one false message, a coarse receiver perceives that he is matched with a liar and chooses $d = \frac{1}{2}$ from then on. In period k, conditional on having observed only truthful messages in past periods, a coarse receiver believes that he is matched with a liar with probability $\frac{(1-\alpha)(\frac{1}{2})^{k-1}}{(1-\alpha)(\frac{1}{2})^{k-1}+\alpha}$ since he believes that a liar sends a false message with probability $\frac{1}{2}$ in all periods of the game. Therefore, he chooses $d_k = \frac{(1-\alpha)(\frac{1}{2})^{k}}{(1-\alpha)(\frac{1}{2})^{k-1}+\alpha}$ which coincides with his overall perceived probability that the sender sends a false message in the current period given that only malevolent senders lie and they are perceived to lie with probability $\frac{1}{2}$. Conditional on only observing truthful messages, d_k is strictly decreasing in k including in period 5 where $d_5 = \frac{1}{10}$.

Coarse receivers perceive that human senders send false messages with probability $\frac{1}{2}$ in all periods, and, as a result, they perceive that a false message in period 5 after 4 truthful messages is quite unlikely (since past behaviors most likely come from a machine). This belief is exploited by rational senders who follow a deceptive strategy with probability 1. To make the deceptive tactic part of a best-response for rational senders, the probability $1-\beta$ of being matched with a coarse receiver should not be too small.

In order to understand numerically the optimality of rational senders' strategy, let us first imagine that they are only matched with coarse receivers. In that case, choosing a deceptive strategy is costly during the first 4 periods of the game since coarse receivers choose $d_1 = \frac{2}{5}$, $d_2 = \frac{1}{3}$, $d_3 = \frac{1}{4}$ and $d_4 = \frac{1}{6}$ so that a rational sender obtains $(\frac{2}{5})^2 + (\frac{1}{3})^2 + (\frac{1}{4})^2 + (\frac{1}{6})^2$ during these 4 periods, as opposed to the larger payoff $(\frac{3}{5})^2 + 3(\frac{1}{2})^2$ she would obtain in the corresponding periods if she were to send a false message in period 1. However, this cost is more than compensated by the extra profit she makes in period 5 in which she obtains $5(\frac{9}{10})^2$, a much higher payoff than the $5(\frac{1}{2})^2$ she would have obtained if she had sent a false message in a prior period. Now, with probability β, a rational sender is matched with a rational receiver who mimics coarse receivers during the first 4 periods but does not choose $d_5 = \frac{1}{10}$ conditional on having observed only truthful messages during the first 4 periods. In this case, the deceptive strategy is not profitable and a rational sender would have been better off sending her first false message before period 5. However, whenever β is not too high, the

extra profit that a rational sender makes with coarse receivers is sufficient to compensate for the loss she makes with rational receivers.

Consider next rational receivers. Clearly, after having observed a first false message, the best response for the remaining periods is $d = \frac{1}{2}$. By mimicking the behavior of coarse receivers up to period 4, a rational receiver maintains rational senders in the ignorance of his type, which the rational receiver takes advantage of in the key period 5. This is better than choosing a myopic best-response in period 1, 2, 3 or 4 because the chance of being matched with a rational sender, i.e. $(1-\alpha)\beta$, is not too small.

Let us also explain why the aggregate lie rate of malevolent senders must be 50:50 in equilibrium (to simplify the exposition of the intuition, we will assume that all receivers are coarse as the presence of rational receivers does not change the intuition). Suppose that malevolent senders send, on average, more false messages than truthful messages. Then, once a sender is identified as malevolent (which must happen at the latest at the key period), 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. Best-response of rational senders then requires to send only truthful messages after the first false message. But this is not consistent with malevolent senders sending in aggregate more false messages than truthful messages given the large number of periods following the key period. 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 in aggregate over the 20 periods of the game, thereby explaining the 50:50 lie rate of malevolent senders in equilibrium.

For expositional purposes and in order to be consistent with experimental observations, Proposition 2 is stated for $\beta = \frac{1}{4}$. However, a similar ABSE arises for a broad range of values of β.[12] For the existence of such an equilibrium, β should not be too high as otherwise receivers are too often rational and the profit made by rational senders with coarse receivers would not compensate for the loss they make with rational receivers when following a deceptive tactic. Similarly, β should not be too low either because otherwise it is more profitable for a rational receiver to reveal himself as rational in the first period (instead of mimicking the behavior of coarse receivers until period 4).[13]

[12]This is so for any β in including $[0.2, 0.57]$. Computations needed to find out the values of the bounds of this interval are not particularly interesting. They are available upon request.

[13]For low values of β, in equilibrium, all types of players behave as in Proposition 2 except for rational

During the first 5 periods of the game, behaviors differ significantly in the sequential equilibrium and in the ABSE. Moreover, contrary to what we obtained in the sequential equilibrium, in the ABSE, the presence of the last 15 periods of the game do matter for the first 5 periods as they have an impact on the aggregate lie rate of malevolent senders. As a matter of fact, with the ABSE, if there were only 5 periods, we would not have the same equilibrium behaviors as with 20 periods with a truncation of the 15 last periods of the game. These last periods are necessary in order to establish the $\frac{1}{2}$ average frequency of false messages of malevolent senders. With only 5 periods, if rational senders were to follow the deceptive strategy, the aggregate lie rate of non-machine senders would fall below $\frac{1}{2}$ making the deceptive strategy less rewarding than in the case considered in Proposition 2.[14]

The comparative static is much simpler than in the sequential equilibrium case. For instance, equilibrium strategies remain unchanged if we lower α from $\frac{1}{5}$ to $\frac{1}{10}$ or increase δ_{k^*} from 5 to 10.[15]

It should also 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, in an attempt to better understand the mechanics of deception, we were interested in understanding the strategic choice of the timing of lies rather than whether subjects are averse to lies. Second, such a constraint has limited (if any) effect on the 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).[16] For analogy-based sequential equilibrium

receivers who reveal themselves before period 5, when they only receive truthful messages. For large values of β, rational senders no longer use a deceptive tactic as they are matched too often with rational receivers.

[14] As it turns out, following a deceptive strategy for rational senders remains part of an ABSE in the 5 period version of the game when the share of rational senders is $\beta = \frac{1}{4}$, but it would no longer be part of equilibrium if the share of rational senders on the sender side were increased (unlike in the case in which there are 15 periods after the key period).

[15] We mention these comparative statics because they are the ones relevant for the study of some variants in the experimental treatments.

[16] To illustrate this remark, we can provide an approximate value of the equilibrium p vector if we add this constraint: $(0.488, 0.483, 0.418, 0.176, 1...1)$ resulting in a deceptive pattern with a probability close to 0.13. This is quite close to what we obtain without this constraint.

(ABSE), it does not affect at all the construction of the equilibrium, since the unconstrained equilibrium satisfies this extra constraint.[17]

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 into Euros using a pre-specified exchange rate. Earnings ranged from 8 Euros to 27.80 Euros with a variance of 9.20 Euros and an average of 15.45 Euros. We arranged standard sessions (6 sessions) with $\delta_{k^*} = 5$ and $\alpha = \frac{1}{5}$ and considered several variants to be described next.

In the baseline treatment, the game was played 5 times (5 rounds), 10 subjects were assigned to the role of receivers and 8 subjects were assigned to the role of senders with a malevolent sender's utility function as described above. 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 her remaining capital of false and truthful messages. Whenever her 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

[17] One may argue that imposing this constraint makes the learning that motivates ABSE simpler (the 10 truthful messages/10 false messages statistics is more salient when it is mentioned in the instructions), and from this perspective our main motivation for this is that we wanted to save on the time spent by subjects in the lab.

toward the end of the game on the receiver side).[18]

Senders and receivers' instructions contained a complete description of the game except that receivers were not told senders' utility functions. 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.[19]

Variants were also considered.

- 10% sessions (3 sessions i.e. 150 rounds). 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 i.e. 150 rounds). In this treatment, the weight of the key period k^* was increased to $\delta_{k^*} = 10$.

- 5 period sessions (3 sessions i.e. 300 rounds[20]). 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.

These first three variants were designed in order to study some comparative statics. SE predicts that in 10% sessions (resp: weight 10 sessions) the rate of deceptive tactic should be smaller (resp: larger) than in the baseline sessions whereas ABSE as shown in Proposition 2 predicts that these rates should remain the same (see the discussion after Proposition 2). Regarding 5 period sessions, our main interest lies in comparing these sessions to the results

[18] It seems that the unpredictability worked well since we did not observe that receivers derived significantly different payoffs in the last period compared to the previous ones ($p > 0.85$).

[19] As already mentioned in Introduction, 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 other players' payoffs are rarely given from the start and must be inferred from behaviors. Besides, in a number of contexts, past behaviors are often framed in the shape of summary statistics somewhat similar to the aggregate lie rate that we consider in our experiment.

[20] Since the game was shorter, participants played it 10 times rather 5 times.

found in the first 5 periods of the baseline sessions which should coincide according to SE.[21]

The next two variants to be described now were designed to test the effects of two features of our baseline treatment: the constraint on the number of false and truthful messages (for free sessions) and the fact that receivers did not know senders' payoff (for RISP).

- Free sessions (4 sessions i.e. 200 rounds). This treatment was identical to the baseline treatment except that 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 the start of a round (except the first one), receivers were informed of the percentage of false and truthful messages sent by human senders in former rounds. Besides, there were 6 rounds so that the data from the last 5 rounds could more easily be compared to the data from our baseline treatment.

- RISP sessions (4 sessions i.e. 200 rounds). This treatment was the same as our baseline treatment except that Receivers were Informed of human Senders' Payoff (RISP) function.

The last variant we considered was designed in order to obtain more information on receivers' beliefs.

- Belief sessions (4 sessions i.e. 200 rounds). In order to obtain extra insights about the mode of reasoning of receivers, we implemented a treatment in which, on the top of the 20-period interaction already described, receivers were asked[22] in each period to report their belief regarding the probability with which 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

[21] The ABSE approach predicts a priori a difference between the 5 period version and the baseline version, but this difference is predicted to be small when $\beta = \frac{1}{4}$ (if players optimize exactly, there is only a difference on the receiver side for coarse subjects).

[22] Conditional on not having received a false message in the previous periods of the session.

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 messages of the current round. Instructions appear in the online appendix.

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 in standard sessions

We first describe some salient observations (out of the 300 rounds).

The receiver 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. Note that we did not identify any significant effect according to whether the signal sent was 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 periods is equal to 0.4 taking 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. As it turns out, the distribution of ds is heavily affected by a very simple statistic: did the receiver already observe a false message during the game or not?

Conditional on no false message being 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 oscillations around 0.05. Even after 15 or 18 truthful messages, the average d_k never falls below 0.05.[23]

If at least one false message has been observed during the game, the average d_k is equal to 0.485. It does not vary much with the period k. Besides, 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. More generally, we did not identify any specific lie pattern during the last 15 periods of the game.[24] These observations are gathered in figure 1.

[23]There is also a small peak in period 11. Some receivers seem to expect that human senders may send 10 truthful messages in the first 10 periods and then 10 false messages.

[24]For instance, once a false message has already been sent, the likelihood of sending a false message is not

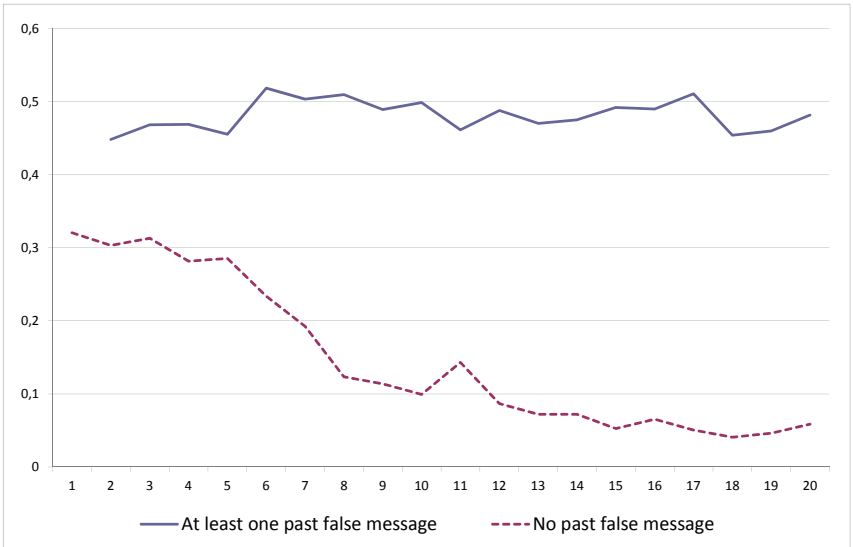

Figure 1: Average d in standard sessions

The contrast between the distribution of ds according to whether or not a lie was previously observed is very much in line with some basic Bayesian understanding of the problem to the extent that a single lie perfectly reveals that the sender cannot be a machine consistently sending truthful messages. The distribution of d after a lie is also broadly consistent with the theories presented above ($d = 0.5$) even if the data are noisier than according to the theories.

The downward sloping pattern of d_k including at the key period k^* when no lie is observed is not consistent with the sequential equilibrium prediction. Conditional on no false message being observed during the game, receivers tend to choose values of d_k higher than the ones that would maximize their payoff given the actual behavior of senders. d_k decreases too slowly across periods. However, there is one major exception: period 5, the key period. Conditional on having observed only truthful messages during the 4 first periods, receivers should choose a d_5 much above d_4 (again considering both actual senders' behaviors and the

significantly affected by the truthfulness of the message of the previous period. 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.

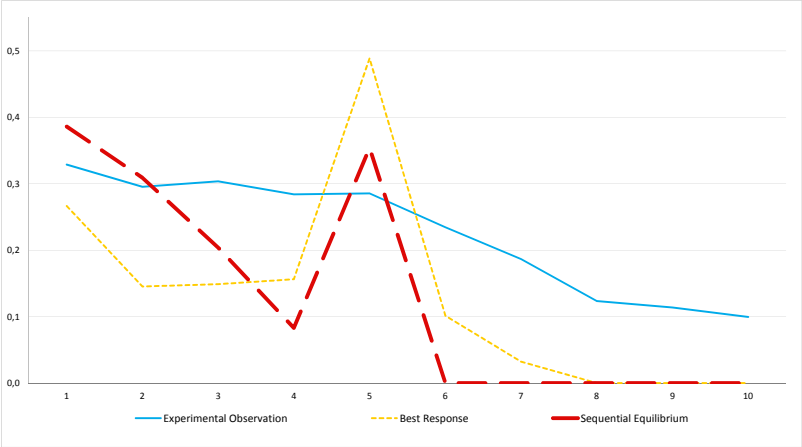

Figure 2: Average d (no false message observed) in standard sessions

sequential equilibrium). However, this is not the case. The average d_5 is very close to the average d_4. All these observations are represented in figure 2 together with the corresponding evolution of d_k according to the Sequential Equilibrium (SE).

The sender side

The more salient observation on the sender side concerns the deceptive tactic which is chosen with a frequency 0.275 by human senders as compared with the 0.15 frequency of SE[25] -or the 0.13 if we consider the variant of the game in which senders must send on average 10 false messages during the 20 periods of the game. We note that choosing such a deceptive tactic is much more profitable as compared with the other used strategies (aggregating over the latter) during the 5 first periods of the game. A sender who sends her first false message during one of the 4 first periods of the game obtains on average 297 during the 5 first periods. When she follows a deceptive tactic, she obtains, on average, 362.[26] This difference is highly

[25]The frequency 0.275 is significantly different from 0.15 with a $p < 10^{-4}$.

[26]In order to obtain more data, we bundled for this statistic, the payoffs obtained by human senders following a deceptive tactic and the payoffs 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 in the two groups is negligible and not significant.

significant ($p < 0.003$).

We did not identify any specific lie pattern during the last 15 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 sent in the previous period.

It is also worth noting that we did not observe any major change over the five repetitions (rounds) of the 20-period game regarding the observations reported above. There is no clear evidence of learning at this aggregate level (we will elaborate on learning effects later on).

4.2 First interpretations

The sender 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 in our experimental data 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). Moreover, a significant difference in the average revenue obtained with different tactics chosen with positive probability by senders is hard to reconcile with an interpretation in terms of rational senders and receivers playing a sequential equilibrium with mixed strategies. In a sequential equilibrium, the tactics chosen with strictly positive probability are supposed to provide the same expected payoff.

As already suggested, we intend to rationalize our data based on the ABSE concept. Of course, allowing ourself to vary the share β of rational players in ABSE gives one more degree of freedom in ABSE as compared with SE, and it is thus not surprising that ABSE with well chosen β can explain data better than SE. But, our main challenge will be to suggest that such an ABSE with the same share β of rational players both on the sender and the receiver sides explains the qualitative features of the complex strategies of the senders and the receivers in the baseline treatment and in a number of variants. Coming back to the observed data, given the proportion 0.275 of observed deceptive tactic, the required proportion β of rational subjects should satisfy $\beta + \frac{1-\beta}{32} = 0.275$.[27] That is, $\beta \approx 0.25$, hence the choice of β

[27] This is because for a broad range of β rational senders would pick the deceptive tactic and among the

in Proposition 2.

For periods 1 to 4 we compare the proportions of lies conditional on not having observed any lie previously in the data with the ABSE ($\beta = 0.25$) and SE theoretical benchmarks. These are depicted in Figure 3 where we observe a good match between the observed data and ABSE with $\beta = 0.25$.

Apart from the good match of lie rate between observations and ABSE, it should also be mentioned that the extra profitability of the deceptive tactic observed in the data agrees with the ABSE prediction.[28]

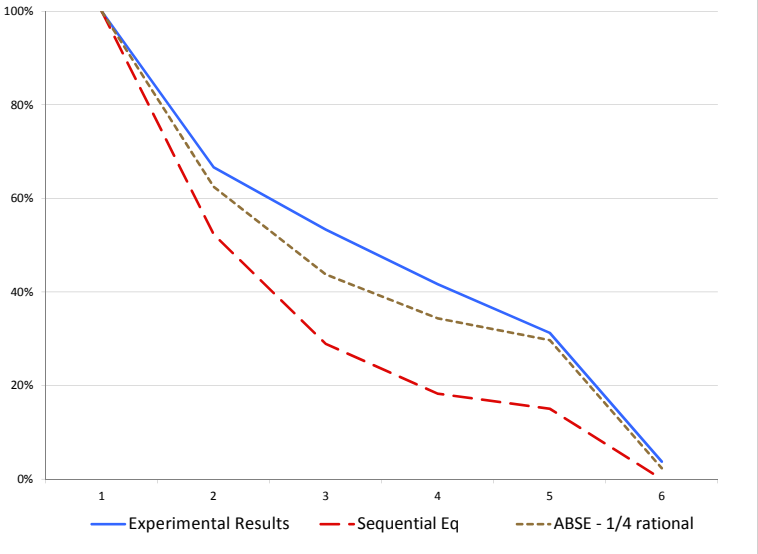

Figure 3: percentage of malevolent senders having sent only truthful messages at the beginning of the period - Standard sessions

The receiver side

On the receiver side, we wish to explore whether the observed data fit the ABSE with

$1-\beta$ Coarse senders, $\frac{1}{2^5} = \frac{1}{32}$ of them would also behave (by chance) according to the deceptive tactic.

[28] Proposition 2 predicts that the deceptive tactic provides a payoff of 371, close to the 362 we observe, and that the revenue if the first false message appears before period 5 is 235. The 297 payoff we observe can be explained by the high variance of the ds after a false message. While best-response would lead receivers to choose $d = 0.5$ in such events, we observe more variations, which may be attributed to the desire of receivers knowing they face malevolent senders to guess what the right state is as kids would do in rock-paper-scissor games. Such a deviation from best-response on the receiver side is beneficial to senders.

$\beta = 1/4$ described in Proposition 2.

For the categorization of receivers into cognitive types, we employ a methodology that retains a salient feature that differentiates the strategies of rational and coarse receivers.

Specifically, coarse receivers as considered above believe that human senders are equally likely to send a false message in each period (independently of the round history). As a result, coarse receivers get more and more convinced that they are facing a machine as they observe no lie in the past with nothing special happening at the key period. Thus, the pattern of d_k for coarse receivers is such that d_k declines up to and including at the key period, as long as no lie is observed resulting in a \-shape for d_k.

As far as rational receivers are concerned, they are ones who anticipate that the lie rate may be quite high at the key period if no lie has been observed so far (because human senders who have not yet lied are expected to lie at the key period). For rational receivers, as long as no lie has been observed, their d_k declines up to period $k^* - 1$ and goes up at k^* resulting in a V-shape for d_k.

Accordingly, we categorize receivers who have observed no lie from period 1 to 4 into two subpopulations:[29] \-receivers and V-receivers. A receiver is a \-receiver (identified as a coarse receiver) if, conditional on having only observed truthful messages in the past, he follows more and more the sender's recommendation up to and including at the key period, or, in symbols, for any $k < 5$, $d_{k+1} \leq d_k$. A receiver is a V-receiver (identified as a rational receiver) if, conditional on having only observed truthful messages in the past, he 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 $d_5 > d_4$.[30]

We observe that most of the receivers who have observed no lie up to period 4 belong to one of these two categories. 59% of the receivers are \-receivers and 24% are V-receivers (out of the 125 observations). Retaining a share $\beta = 0.25$ of rational subjects as in Proposition 2, figure 4 reveals that the average behaviors of these two populations are quite well approximated by identifying \-receivers with coarse receivers playing the analogy-based sequential equilibrium and V-receivers with rational receivers playing the ABSE rational receiver's strategy. The

[29] For other histories, there is no difference in the behaviors of rational and coarse receivers in the ABSE shown in Proposition 2.

[30] 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 $k < 4$.

observed coefficient of the slope slightly differs from the equilibrium predictions but this may be the result of receivers' difficulties in applying an exact version of Bayes' law.

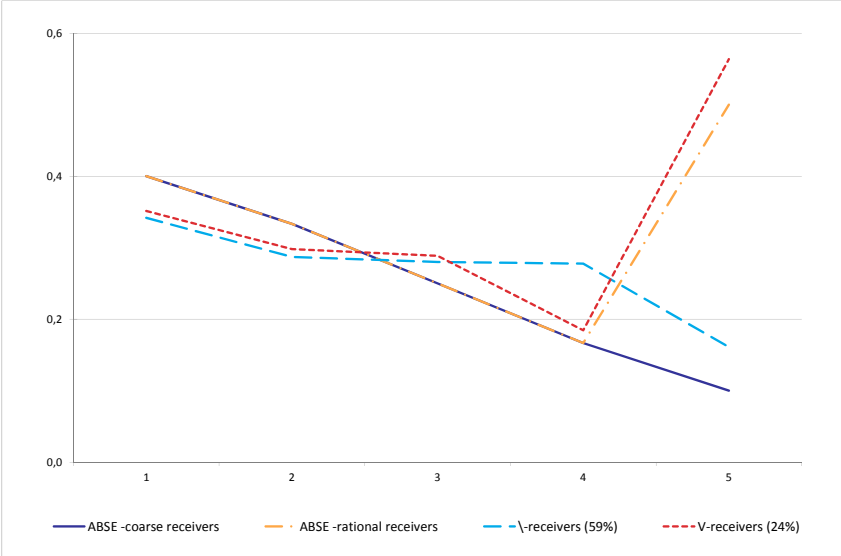

Figure 4: Average d - No past false message - Standard sessions

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. The average gain over the first five periods is 627 for \-receivers and 701 for V-receivers, thereby resulting in a difference of 74 in expected payoff (the prediction of the ABSE is a difference of 80) which is significantly different from 0 ($p < 0.05$).[31] We note also that receivers who belong to none of these two groups got an expected gain comparable to those of \-receivers. These receivers would be closer to the coarse receivers as previously defined, with a slightly different form of bounded rationality and more trials and error (which would explain the non-monotonicity of d in the first periods). These considerations give **some** support to our identification of the share of rational receivers with the share of V-receivers.

As just reported, our analysis of the baseline treatment suggests that the experimental

[31] 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 easily secure.

data are well organized by the ABSE shown in Proposition 2 with a $\beta = \frac{1}{4}$ share of rational subjects and $\frac{3}{4}$ share of coarse subjects both on the sender and the receiver sides. In order to improve the fit, one could allow subjects to use noisy best-responses as in Quantal Response Equilibrium models, but the computation of the corresponding ABSE is quite complicated, which has led us not to follow this route. In an attempt to allow for noisy behavior, in the appendix, we develop a statistical method for the categorization of receivers into rational vs coarse types, explicitly allowing for mistakes and focusing on types who would either if rational take the actual aggregate lie behavior (conditional on no lie being observed so far) as their belief or else they would consider that the lie rate is uniformly 50:50 for human senders exactly as coarse receivers in ABSE. The method assigns each individual to one or the other type according to the likelihood for each type to generate their observed behavior. The results obtained with this alternative method are qualitatively very close to those obtained with the \-receivers and V-receivers categorization. They are reported in the Appendix.

4.3 Variants

We now discuss the experimental findings in the various variants we considered.

10% automata - Weight 10 - Free sessions

First note that in the 10% automata case ($\alpha = \frac{1}{10}$), with β unchanged, the ABSE is the same as in Proposition 2 (with values of d_k adjusted to the changes of α). The SE has the same properties as when $\alpha = \frac{1}{5}$ with a lower frequency of deceptive tactic (≈ 0.075), since a smaller α makes it less profitable for a malevolent sender to be confounded with a machine.

Experimental observations with 10% automata are almost identical to the ones we obtained in the baseline treatment: A ratio 0.25 of deceptive behaviors and of V-receivers. This comparative static is consistent with the ABSE and a share $\beta = \frac{1}{4}$ of rational subjects (see the discussion after Proposition 2), much less with the sequential equilibrium.

If we increase the weight δ_{k^*} of the key period from 5 to 10, this increases the frequency of deceptive behavior in the sequential equilibrium and, ceteris paribus, does not affect the ABSE with $\beta = \frac{1}{4}$ shown in Proposition 2.

In the data of the weight 10 sessions, the frequency of deceptive behavior is slightly lower than in the baseline treatment (0.19) and the ratio of V-receivers slightly higher (30%). The

relative stability of these frequencies is more in line with our interpretation in terms of ABSE with a share $\beta = \frac{1}{4}$ of rational subjects than an interpretation in terms of subjects playing SE. Maybe the slight difference with the baseline case can be interpreted along the following lines. As the weight of the key period increases, it becomes more salient, thereby leading receivers to pay more attention to it, which may result in a higher ratio of V-receivers. Anticipating this effect, rational senders are less eager to follow a deceptive strategy which is less likely to be successful.

For Free sessions, we observe that the actual average ratio of false and truthful messages communicated to receivers by human senders is equal to 0.46, close to the 0.5 of the standard sessions. Thus, in contrast to Cai and Wang (2006) or Wang et al. (2010), we do not observe a strong bias toward truth-telling when senders are free in their number of lies. This is probably due to the zero-sum nature of the interaction. The observed frequency of deceptive behavior is 0.28 and the observed ratio of V-receivers is 24%. Both frequencies are extremely close to those obtained in the standard treatment. We could not identify any major effect of the constraint on the ratio of false and truthful messages on players' behaviors.

RISP sessions

In these sessions, receivers were informed that human senders had preferences opposite to their own. The frequency of deceptive tactic is slightly lower at 0.21. On average, senders obtain a higher payoff when they follow a deceptive tactic as compared with any other tactic (317 as compared to 290) but this difference is not very significant ($p = 0.125$).

However, contrary to what we observe in the other variants of the game, it is interesting here to disentangle what we observe in the first 2 rounds from what we observe in the last 3 rounds. In the first 2 rounds, with less experienced receivers, the deceptive behavior gives a higher average payment than non-deceptive behaviors: 386 against 278 ($p < 0.01$). There is no statistically significant difference between these observations and what is obtained in standard sessions ($p > 0.7$). But if we consider the last 3 rounds, we observe a clear difference. In the baseline treatment (and the other variants), there is no statistically significant difference between rounds. In RISP sessions, in the last 3 rounds, the average payment obtained with a deceptive tactic is 285, statistically different from what is obtained with this same tactic in the baseline treatment or in the first 2 RISP rounds ($p < 0.01$).

This shows that, contrary to what we observe in other variants, there is a learning process

at work in RISP sessions (we will elaborate on the learning dimension on the receiver side later on). From a different perspective, the similarity of the first two rounds of RISP with the five rounds of the baseline sessions suggests that ceteris paribus the heuristics of the sort used by coarse receivers in ABSE may describe better the behaviors of less experienced subjects when other players' preferences are known than when they are not.

5 period sessions

The predictions of the sequential equilibrium in this variant are the same as in the standard treatments. However, in this variant, we observe very different behaviors. On the sender side, the frequency of deceptive tactic is equal to 0.05 (out of 240 observations), much lower than in any other variant and much lower than predicted by the sequential equilibrium or the ABSE with $\beta = \frac{1}{4}$

On the receiver 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). The shares of \-receivers and V-receivers are 58% and 18%, respectively.

In general, behaviors are explained neither by SE nor by ABSE with $\beta = \frac{1}{4}$, even accounting in ABSE for the fact that the key period is final. On the sender side, behaviors seem to be better explained assuming that in every period, human senders would randomize 50:50 between telling the truth and lying, which would require 100% share of coarse senders.[32] On the receiver side, even rational subjects should behave like \-receivers if they were rightly perceiving the high share of coarse subjects on the sender side. So rationalizing V-behaviors would require allowing rational receivers to erroneously think that the share of rational subjects on the sender side is higher than it really is.

Based on the above, it is fair to acknowledge that we do not have a good explanation for the data in the 5 period sessions, and more work is required to rationalize these. We believe though that the contrast between the finding in the baseline case and the 5 period case should be of interest to researchers concerned with reputation and deception.[33]

Learning Process on the receiver side?

[32]This gives a distribution function of the first false message close to the one described in the Sequential Equilibrium but with a higher probability to send a first false message in period 4 than in period 5, contrary to what we observe in the sequential equilibrium in which the first false message is never sent in period 4.

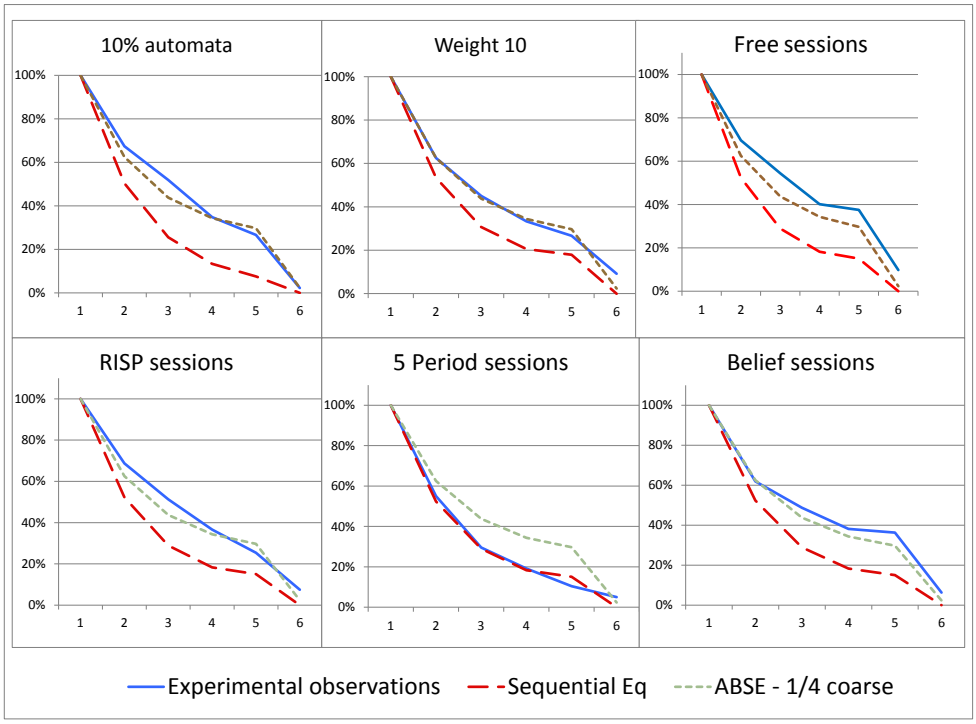

Figure 5: Percentage of human senders having sent only truthful messages at the beginning of the period.

In this part, we ask ourselves whether there is some learning trend on the receiver side (no clear learning trend is apparent on the sender side as reported above). Specifically, we analyze whether there is a change across rounds in terms of the share of \-receivers and V-receivers.

Consider first the baseline treatment. We observe that the percentage of V-receivers is almost identical in rounds 1 and 2, 61%, and in round 5, 59%. This suggests that there is no clear learning trend. We next turn to whether receivers' behaviors depend on their specific history. More precisely, we consider separately four different subsets of histories. H1: The receiver has never been matched in any prior round neither with an automaton nor with a deceiving sender. H2: The receiver has already been matched in a prior round at least once with an automaton and he has never been matched with a deceiving sender. H3: The

This can be seen in Figure 5.

[33]Maybe the 5th period becomes more salient when it is also the last one. The complexity of the game also differs. This may explain why players use heuristics other than the analogical reasoning. Although this may explain why observations do not coincide with ABSE, it is left for future research to find out the type of heuristics followed by subjects in this case.

receiver has already been matched at least once with a deceiving sender. H4: The receiver has already been matched at least once with a deceiving sender and never been matched with an automaton (subset of H3).

We do not observe any significant difference in the share of \-receivers and V-receivers in H1 or H2 (the share of V-receivers is 22% in H1 and 12% in H2 but the size of the sample is not sufficient to make this difference significant). However, there exists a statistically significant difference ($p \approx 0.04$) between the percentage of V-receivers in H1 or H2, 18%, on the one hand and the percentage in H3, 36% on the other. The difference is even more pronounced if we compare H2, 12% of V-receivers, with H4, 39% with $p \approx 0.014$. The difference is less significant for \-receivers, 65% in H2 and 45% in H4 with $p \approx 0.1$ (the size of the two samples is close to 30). To sum up, we find that receivers are more likely to be V-receivers (rational) if they have been previously matched with a deceiving sender, thereby suggesting that with sufficiently many repetitions, a larger share β of rational subjects would be required to explain observed behaviors with ABSE.

It is worth noting that the learning process seems to go faster in RISP sessions (other sessions except the 5 period sessions give similar patterns to those observed in the baseline sessions). In RISP, the percentage of V-receivers (resp: \-receivers) is particularly low (resp: high) in H3, equal to 17% (resp: 56%) and significantly different from the H1 percentage: 46% (resp: 14%) with $p \approx 0.02$ (resp: 0.02). The learning process goes so fast in RISP that the percentage of \-receivers in round 5, 23% (resp: 36%) is statistically lower than the percentage in rounds 1 and 2, 53%, (resp: 16%) with $p \approx 0.03$ (resp: 0.01) unlike in the baseline sessions. These results about receivers in RISP are consistent with the observations we made on the evolution of the profitability of the deceptive tactic.

Overall, our analysis reveals some learning effect, after being exposed to a deceptive tactic. This effect is more pronounced in RISP sessions, presumably because the knowledge of the other party's preference allows to better make sense of the observation of a deceptive tactic in this case.

While such a learning trend may, in the long run, lead experienced subjects to behave more rationally, we believe that the phenomenon of deception as prevailing in our experimental data is of relevance for the real world insofar as in most practical situations there is always a flow of less experienced subjects. Inexperienced subjects may be expected to base their

Treatment	\-receivers	V-receivers	Profit \-receivers	Profit V-receivers	T-test
Standard	59%	24%	361	297	0.002
Belief	54%	24%	363	283	$< 10^{-3}$
10% automata	52%	25%	355	297	0.037
Weight 10	52%	30%	675	403	$< 10^{-11}$
Free, rounds 5, 6	56%	26%	336	280	0.012
5 period	56%	18%	353	316	0.16
RISP, rounds 1, 2	53%	17%	386	278	$< 10^{-2}$
RISP, rounds 3, 4, 5	32%	31%	285	298	0.44

Table 1: Frequency of the two types of receivers and their profits during the 5 first periods of the game.

decisions on coarse reasoning as suggested in this paper essentially because such subjects are unlikely to have been exposed to a deceptive tactic.

Summary

The main findings are summarized in Figure 5 and Table 1 (where for completeness we include also the belief sessions to be described next). Except for the 5 period sessions, results are well organized by ABSE assuming the share of rational subjects is $\beta = \frac{1}{4}$ both on the sender and the receiver sides. By contrast, in the 5 period sessions, neither ABSE nor SE organize the data well. Finally, we observe some learning in terms of a reduction in the share of coarse receivers when receivers have been exposed to a deceptive tactic.

4.4 A specific analysis of belief sessions

We ran 4 belief sessions. The purpose of these sessions was to get separate information about how receivers perceive the probability with which they face a machine and how receivers think human (non-machine) senders behave. A belief session is equivalent to a standard session except that receivers were asked to report the probability q_k (in period k) they assigned to being matched with an automaton after having made their decision and before being informed of the true realization of the signal.[34]

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

We observe that this extra query did not affect the aggregate behaviors of the players in terms of the lying strategy or the sequence of d_k.

As a natural next step, we analyze the extent to which the two populations of receivers, V-receivers and \-receivers, also differ in their belief regarding whether they are matched with a machine or a human sender. As it turns out, we observe major differences as can be seen in figure 6.

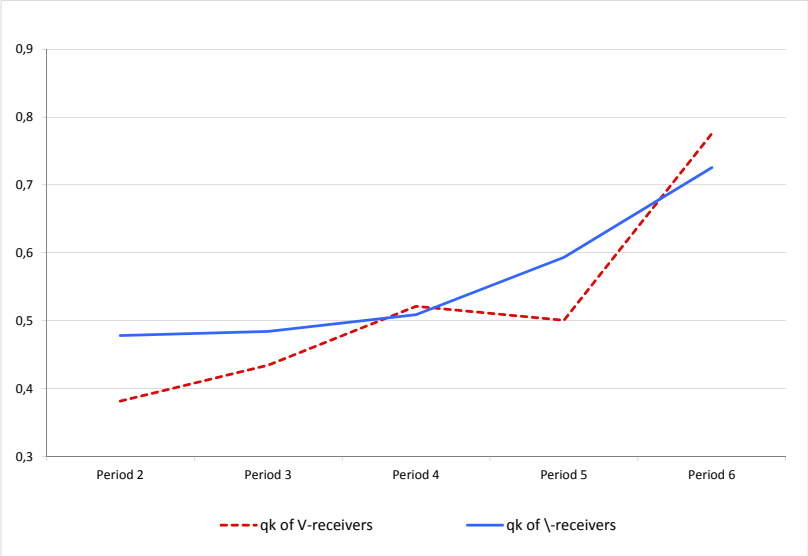

Figure 6: q_k for the two populations of receivers

\-receivers choose a high average q_2, close to 0.5 and during the next two periods, the average q_k slowly increases, it increases more steeply from period 4 onwards. It is particularly striking to observe that \-receivers do not pay attention specifically to period 5. The slope of q_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 both in the experimental data and in the ABSE with $\beta = \frac{1}{4}$, malevolent senders seldom send a first false message in period 4). The slope is almost the same between period 4 and period 5 and between period 5 and period 6. \-receivers do not perceive as more informative a truthful message sent in period 5 than a truthful message sent in period 4 (conditional on having received only truthful messages so far).

These observations are consistent with our interpretation of \-receivers as reasoning in

terms of coarse 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.[35]

V-receivers begin with a belief slightly below 0.4 in period 2 closer to the actual share of automata in period 2 after a truthful message in period 1 (close to 0.3). The average q_k increases from period 2 to period 4, it slowly decreases from period 4 to period 5 and increases very heavily from period 5 to period 6. These observations are in agreement with the interpretation that V-receivers do not expect human senders to send their first false message in period 4 but perceive that a human sender if she has not sent her 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 rational receiver in the ABSE.[36]

Hence, our observations in the belief-sessions corroborate the general conclusion that the reasonings of \-receivers and V-receivers are well captured by those of coarse and rational receivers, respectively. V-receivers quite clearly distinguish period 5 in the pattern of their inference process (their higher d at the key period is not solely the result of cautiousness that would be attached to a higher stake decision) while \-receivers do not.

4.5 Alternative approaches

Alternative popular approaches to study experimental data include the Quantal Response Equilibrium (see McKelvey and Palfrey (1995)) and the 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 observations. 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. Yet, the share of deceptive tactics

[35] The process of belief revision is smooth except in period 2 in which Coarse receivers choose an average a value slightly below 0.5 after a unique truthful message. Receivers may have in mind both the actual proportion of automaton, 0.2 and a probability close to 0.5 that they associate to this uncertain environment.

[36] The small decrease from period 4 to period 5 may be due to the fact that V-receivers re-consider more cautiously the probability that they may be matched with a deceiving sender at the beginning of period 5.

was not bigger in this treatment (it was in fact slightly smaller). These observations are not suggestive that QRE provides a good account of observed data.

Regarding level-k, we note that existing theories are not well suited 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.[37] To bypass these difficulties, we can consider the level-k approach of our game viewed in normal 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 trust what they are told, thereby choosing $d_k = 0$ in all periods k. [38]

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 deceptive tactic appears as a possible focal behavior), it makes no prediction as to whether the deceptive tactic should be profitable. Moreover, the other (focal) behaviors emerging from the approach do not show up in our data (for example, we see almost no human sender telling the truth up to and including the key period nor do we see receivers choosing $d_5 \approx 0.8$ after 4 truthful messages, which would be the best-response to the belief that senders follow the deceptive tactic). Moreover, 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.

In the rest of this section, we briefly consider additional approaches one can think of

[37] 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 to senders.

[38] Some scholars applying the level-k model to cheap talk games have considered that level-0 senders would always tell the truth (Ellingsen and Ostling (2010)) or that level-0 senders would randomize 50:50 between telling the truth and lying. Given our constraint in standard sessions that senders should send between 9 and 11 false messages, our specification for level-0 senders sounds more natural.

(these are not so natural given the way the instructions were provided to subjects, but they are natural to consider from a conceptual viewpoint). First, senders and receivers may be allowed to entertain subjective beliefs regarding the share of benevolent senders. We note that varying these subjective beliefs would not allow to explain the large presence of \-receivers (SE with different shares of benevolent senders would predict V-patterns for receivers).[39] Second, senders and receivers may be allowed to be inattentive to the fact that period 5 has a higher weight. We note that such a variation would have a hard time explaining the observed share of deceptive tactic which is already much beyond the one predicted by SE with the correct weight on the key period (and a fortiori even further away from the one associated with SE and a smaller weight on the key period).

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 deceptive tactics (i.e. sending truthful messages until the key period and a false message at the key period) with a much higher frequency than what the sequential equilibrium of the game would predict. Moreover, deceptive tactics provide 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). Observations favor the idea that both senders and receivers are heterogenous in their cognitive abilities, some share (roughly $\frac{3}{4}$) employing a coarse reasoning with a smaller share (a quarter) employing 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) but not 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).

[39] Of course, one may combine the subjective belief approach with the coarse reasoning approach in ABSE. Such a combination may allow to rationalize ex post the results found in the 5 period treatment (see above) but we have no ex ante rationale for the required choice of subjective beliefs.

Our experimental findings suggest that we should see more deceptive tactic when the interaction is not stopped right after the time at which stakes are higher which may fit better in contexts in which communication stages occur in pre-arranged ways. Moreover, putting aside the findings in the 5 period sessions (for which theories beyond those considered here are needed), our study suggests that solution concepts allowing for coarse reasoning may fruitfully be used to shed light on deception, where a closer look at our data reveals that coarse reasoning is more widespread 1) when subjects are less experienced given that an exposure to a deceptive tactic was shown to shift behavior toward that of rational types,[40] and 2) when subjects are exposed in a more prominent way (in the context of our experiment, at the stage of the instruction) to aggregate statistics about the play of others (deception was a bit less pronounced in Free sessions). Even if coarse reasoning becomes less prominent with experience, deception of the type highlighted here is of practical importance given that experienced agents keep being replaced by less experienced agents in the real world (and yet the novice agents are still exposed to past aggregate statistics, thereby making the equilibrium approach of ABSE compelling for this case).

Finally, let us mention that the kind of deceptive tactic highlighted in our experiment shares a number of similarities with classic influence manoeuvres as reported in Cialdini (2006)'s best-seller book on influence. In particular, at the end of the chapter on authority, Cialdini reports the story of a waiter named Vincent who was particularly successful at increasing the bill and the tip that goes with it in large party dinners. After letting the first customer in the party make her choice of starter, Vincent would invariably suggest that the pick was not the best on that evening and would redirect the customer on a cheaper starter. By gaining the trust of the customers that the waiter was on their side given the redirection to a cheaper starter -this is the analogue of telling the truth up to the key period in our experiment-, Vincent was able to put forward his recommendation of the most expensive wine, which would increase considerably the bill -this is the analogue of lying at the key period.[41] From this perspective and beyond the commonality of theme, our experimental

[40] Our guess in the waiter story is that if too many waiters used the tactic of Vincent, it would become less effective after a while.

[41] Vincent's tactic also required persuading the customers that he was knowledgeable of the quality of the food, which our experiment assumed away by letting receivers know from the start that senders were informed of the state of the world.

work can be viewed as an attempt at quantifying how many people may think of using a deceptive tactic and how many people may be the victims of deception in a simple repeated expert/agent relationship.

6 Appendix

6.1 Proof of Proposition 1

By definition, benevolent senders send truthful messages. Moreover, it is clear (by backward induction) that once a lie has been observed (so that it is common knowledge that the sender is malevolent), 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: $\frac{(1-\alpha)\prod_{i=1}^{k-1}(1-p_i)}{(1-\alpha)\prod_{i=1}^{k-1}(1-p_i)+\alpha}$ and since, by definition, the probability that a malevolent sender will choose a false message is p_k, the receiver's best response is $d_k = \frac{(1-\alpha)\prod_{i=1}^{k-1}(1-p_i)p_k}{(1-\alpha)\prod_{i=1}^{k-1}(1-p_i)+\alpha}$. When $k = 1$, the best response is $(1-\alpha)p_k$.

The last equilibrium condition is only a standard mixed equilibrium indifference condition.

Now, 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.[42] 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 $q_{k_q^r} = 1$ (resp: $p_{k_p^r} = 1$) 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 her 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

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

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 who 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, the receiver chooses a $d_{\widetilde{k},p}$ equal to $\frac{(1-\alpha)\prod_{i=1}^{k-1}(1-p_i)p_{\widetilde{k}}}{(1-\alpha)\prod_{i=1}^{k-1}(1-p_i)+\alpha}$ and in a q-equilibrium, the receiver chooses a $d_{\widetilde{k},q}$ equal to $\frac{(1-\alpha)\prod_{i=1}^{k-1}(1-q_i)q_{\widetilde{k}}}{(1-\alpha)\prod_{i=1}^{k-1}(1-q_i)+\alpha}$ and since $\frac{(1-\alpha)\prod_{i=1}^{k-1}(1-p_i)}{(1-\alpha)\prod_{i=1}^{k-1}(1-p_i)+\alpha} = \frac{(1-\alpha)\prod_{i=1}^{k-1}(1-q_i)}{(1-\alpha)\prod_{i=1}^{k-1}(1-q_i)+\alpha}$, $d_{\widetilde{k},p} < d_{\widetilde{k},q}$ so that the payoff obtained by a malevolent sender sending her 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.[43]

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

Now, we need to show that $p_{\bar{k}+1} > 0$ and $q_{\bar{k}+1} = 0$ is impossible. If this were the case, the payoff of a malevolent sender in the periods between $\widetilde{k}$ and $\bar{k}+1$, in a q-equilibrium, if she deviates and sends her first false message in period $\bar{k}+1$ would be $\sum_{j=k}^{\bar{k}} \delta_j d_{j,q}^2 + \delta_{\bar{k}+1}$. Because of the arguments we have just mentioned $\sum_{j=k}^{\bar{k}} \delta_j d_{j,q}^2 + \delta_{\bar{k}+1} > \sum_{j=k}^{\bar{k}} \delta_j d_{j,p}^2 + \delta_{\bar{k}+1}(1-d_{\bar{k}+1,p})^2$. Therefore, a malevolent sender deviating in a q-equilibrium, sending her first false message in period $\bar{k}+1$ obtains more than a malevolent sender in a p-equilibrium. This cannot be

[43] 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.

possible since we showed that a malevolent sender always obtains a strictly lower payoff in a q-equilibrium than in a p-equilibrium. Hence $p_{\bar{k}+1} > 0$ and $q_{\bar{k}+1} = 0$ is impossible. Hence, for any $i \in [\widetilde{k}, \bar{k}+1]$, $p_i = q_i = 0$ or $p_i < q_i$ and $d_{i,p} < d_{i,q}$. End of the induction proof.

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 her payoff in future period. [44] 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 $p_k = \frac{1}{2}\frac{(1-\alpha)\prod_{i=1}^{k-1}(1-p_i)+\alpha}{(1-\alpha)\prod_{i=1}^{k-1}(1-p_i)}$ exceeds 1 at some point). The same arguments apply to reject $k_q^r = 21$.

Let us consider $k_q^r < k_p^r < 21$ (this is always the case because of the result we proved by induction) 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 her 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 she sends her 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 her first false message in period $\widehat{k}$ obtains strictly more than a malevolent sender in a p-equilibrium sending her 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.

[44] Besides, if sending a false message gives a strictly higher payoff, she 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, she 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

First, we need to describe more completely the strategies that we only partially introduced in proposition 2 for rational senders.

In case she 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, she plays as in the sequential equilibrium of a variant of the game beginning in period $k+1$, with a fraction $\frac{2^k}{3+2^{k+1}}$ of benevolent senders, a fraction $\frac{2^k}{3+2^{k+1}}$ of rational malevolent senders and a fraction $\frac{3}{3+2^{k+1}}$ 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, sends a false message with probability $\frac{9-\frac{15}{2}\beta}{1-\beta}$ during the last 15 periods of the game. If she observes a different vector of d_ks during the first 5 periods of the game, she sends a false message with probability $\frac{1}{2}$ in the 15 last 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, she does not perceive the link between the message she sends and the decision of the receivers she is matched with. Sending a truthful and a false message with probability $\frac{1}{2}$ in all the periods is a best response with this belief.

Considering senders' strategies, a rational receiver cannot raise his payoff choosing a $d \neq \frac{1}{2}$ conditional on having observed at least one false message. Therefore, we can focus on his behavior conditional on not having received any false message. A rational receiver must decide in that case whether he mimics coarse receivers or he reveals his type choosing a different d. If he reveals her type in period k, a coarse sender will continue sending a false message with probability $\frac{1}{2}$ 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 a fraction $\frac{2^k}{3+2^{k+1}}$ of benevolent senders, a fraction $\frac{2^k}{3+2^{k+1}}$ of rational malevolent senders and a fraction $\frac{3}{3+2^{k+1}}$ of mechanical senders sending truthful messages with probability $\frac{1}{2}$ 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 $\frac{1}{2}$ in each period. Now, a rational receiver must choose the period k in which he reveals his type and d_k. Since the value of d_k does not affect the payoff in the following periods

as long as $d_k \neq \frac{4(1/2)^k}{1+4(1/2)^{k-1}}$, his best choice is a d_k which maximizes his 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 $k = 5$.

Rational senders. Again, the key element is the period of the first false message. After this first false message, in all the remaining periods, $d_k = \frac{1}{2}$, therefore any choice is a best response and she 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 she believes that she is matched with a coarse receiver with probability $\frac{3}{4}$, she obtains a higher payoff sending her first false message in period 5 (her 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.

Q.E.D.

6.3 An alternative statistical analysis

We developed a second methodology based on a statistical model in order to characterize receivers' behaviors. Given the seemingly noisy character of decisions, we allow subjects to play noisy best-responses using a logit specification which is commonly used in econometrics or in experimental work. That is, if in period k, the belief that the sender is lying is b, the receiver will choose action d with a probability proportional to[45]

$$\exp \lambda V_k(b, d)$$

where

$$V_k(b, d) = \delta_k[1 - b(1 - d)^2 - (1 - b)d^2]$$

denotes the expected period k utility of playing d given the belief b and λ denotes a noise parameter to be estimated (the higher λ the less noisy the best response).

[45] In the experiment, there were finitely many possible d (because it could move by increments of 0.01). The ratio of the probability that d vs d' is played is $\exp \lambda V_k(d, b) / \exp \lambda V_k(d', b)$.

The two types Coarse and Rational of receivers correspond to different specifications of b. The belief of a Coarse receiver in period k when no lie was observed is given by:

$$b_k^c = \frac{(1-\alpha)(1/2)^k}{\alpha + (1-\alpha)(1/2)^{k-1}}$$

given that Coarse receivers expect human senders to lie with probability $\frac{1}{2}$ in every period (and they know that honest senders always tell the truth).

Regarding non-coarse receivers, we considered several variants for b^r,[46] but picking the one that gave the best fit, we considered that their belief in period k after no lie was observed was given by the empirical proportion of lie in period k obtained from the overall population of senders -humans and machines- when no lie was observed up to period $k-1$.

Since the beliefs are almost identical for Rational and Coarse receivers after the first observed false message, we focus on receivers' choices conditional on no lie being made so far.[47] [48]

We first test whether the data are best explained assuming all receivers are Coarse or assuming all receivers are Rational. We obtain that data are best explained when all Receivers are Coarse (the log-likelihood ratio is higher than 60 and extremely significant).

We next move to estimating a mixed model in which we allow receivers to be either Coarse or Rational and we estimate the proportion of the two types (as well as the noise parameter λ). Formally, let

$$LR = \prod_{i=1}^{n} \prod_{s \in S_{d,a}^i} \prod_{k=1}^{5} [x_{is}^r \exp \lambda V_k(b_{ks}^r, d_{ks}^i) + (1 - x_{is}^r) \exp \lambda V_k(b_{ks}^c, d_{ks}^i)]$$

where $S_{d,a}^i$ is the set of sessions in which receiver i is matched either with a deceiving sender or an automaton, x_{is}^r is a binary variable equal to 1 if the receiver is categorized as Rational and equal to 1 if he is categorized as Coarse in session s, and d_{ks}^i is the decision of receiver i in period k of session s.

[46] Another variant that we considered was the b_k that derives from SE.

[47] More precisely, in order to obtain a coherent data set and because many first lies appear in the key period, we focus on receivers' decisions during the first 5 periods of a round conditional on not having observed a false message during the first 4 periods of the game.

[48] We also chose to restrict the dataset this way because we did not want to add any extra uninformative noise by considering periods posterior to the first false message.

Maximizing LR with respect to x_{is} yields $x_{is} = 1$ if the vector d^{is} is better explained (higher likelihood ratio) by referring to rational belief, b^r, than coarse belief, b^c, and 0 otherwise.

In standard sessions, maximizing LR with respect to to (x_{is}) and λ, this two population model gives a much higher likelihood ratio and a much higher λ than models with only one (homogeneous) population. Our estimation gave $\lambda = 3.5$ and a 58% share of coarse receivers.

We also considered adding a third type of receiver referred as Skeptical who would consistently use $b^s = 0.5$. In this case, we found that the share of Coarse was 55%, the share of Rational was 30% and the share of Skeptical was 15% with $\lambda = 4.1$. These results are quite in line with our categorization in terms of \-receivers and V-receivers. Besides, trying to connect the two methods, we found that 77% of \-receivers were categorized as Coarse receivers $(x_{is} = 0)$ and 90% of V-receivers were categorized as Rational $(x_{is} = 1)$ according to the statistical method, thereby giving some extra support as to why our heuristic categorization captures an essential element that differentiates the behaviors of Coarse receivers from those of Rational receivers.

We obtain qualitatively equivalent results with the other variants of the game (except 5 period sessions), with a fraction of coarse receivers varying between 51% and 58% in the main ones.[49]

In the 5 period sessions, in accordance with our difficulties in identifying receivers' behaviors, we obtain a low value for λ, 0.63.

[49] For Free sessions and the 5 period sessions, in order to evaluate the ABSE behaviors, we took into account the actual average ratio of false and truthful messages communicated to receivers by human senders. This average lie rate was equal to 0.46.

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. and Weigelt, K. (1988): 'Experimental Tests of Sequential Equilibrium Reputation Model', *Econometrica*, **56**, 1-36.

[5] Cialdini, R. B. (2006): *Influence: The Psychology of Persuasion*, HarperBusiness.

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

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

[8] Ellingsen, T. and Ostling, R. (2010): 'When does communication improve coordination?', *American Economic Review*, **100**, 1695-1724

[9] Embrey, M., Frechette,G. R. and Lehrer, S. F. (2015): 'Bargaining and Reputation: Experimental Evidence on Bargaining in the Presence of Irrational Types', *The Review of Economic Studies*, **82**, 608-631.

[10] Ettinger, D. and Jehiel, P. (2010): 'A Theory of Deception', *American Economic Journal: Microeconomics*, **2**, 1-20.

[11] Gneezy, U. (2005): 'Deception: The Role of Consequences', *American Economic Review*, **95**, 384-394.

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

[13] Jehiel P. and L. Samuelson (2012): 'Reputation with analogical reasoning', *Quarterly Journal of Economics*, **127**, 1927-1969.

[14] 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.

[15] Kawagoe, T., and Takizawa, H. (2009): 'Equilibrium refinement vs. level-k analysis: An experimental study of cheap-talk games with private information', *Games and Economic Behavior*, **66**, 238-255.

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

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

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

[19] 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.

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

[21] Stahl, D. and Wilson, P. (1994): 'Experimental Evidence on Players Models of Other Players', *Journal of Economic Behavior and Organization*, **25**, 309-327.

[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] Vespa, E. and Wilson, A. (2016): 'Communication With Multiple Senders: An Experiment', *Quantitative Economics*, **7**, 1-36.

[25] Wang, J. T. , Spezio, M. and Camerer, C. F. (2010): 'Pinocchio's Pupil: Using Eyetracking and Pupil Dilation To Understand Truth-telling and Deception in Games', *American Economic Review*, **100**, 984-1007.