Experimental Cognitive Psychology
Are we good lying detectors?

Raffaele Canale

Introduction

In society, it is important to be able to make a quick and reliable judgment upon the trustworthiness of others. As soon as we face someone, we use every cue at our disposal (such as the body posture, voice and facial expression) to determine if we can rely on him, or instead, if he is lying. But how good are we at detecting lies? Do we naturally possess skills at liars detection or is it something that we can learn?

Some theories suggest that lying detection is a skill, and thus can be practiced up to become reliable enough in real time (Ekman et al, 1999). In fact, in his study, Ekman challenged different groups of people to detect lies from video tapes of people lying or telling the truth. The results showed that people that had a proper training in that field (in this case, law-enforcement or psychologists) had better results than the others. This would lead to believe that lying detection is an active cognitive process, thus can be learnt and improved by experience.

On the other hand, other studies support that our lying detection skills (good or bad) are a modular process. That is an innate faculty that we all possess and have no control upon. More specifically, that means we are able to perform judgments about liars through a very fast and unconscious process that is encapsulated and thus, adding a cognitive load does not influence our ability (Verplaetse, 2007). Bonnefon tries to show in his experiment (J-F Bonnefon, 2012) that the human mind developed a module for cheater detection and that such a module is unrelated to intelligence or any cognitive processing. In his experiment, Bonnefon created groups according to their level of intelligence score[1] and challenged them in a trust game. The result of this study show that the intelligence score had no impact on the results because participants were all equally capable of trust judgments. However, Bonnefon's experience is based on trustworthiness and the images used in his experiment where taken from a previous study (Centorrino, 2011). These images where taken under the assumption that the perceived trustworthiness is function of the genuineness of the smile and do not entirely fit our purpose.

Bonnefon supports that our liar detection process is modular while Ekman shows that experience and practice are also effective. These results seem to contradict. A possible explanation can be brought by Arminjon's hypothesis that the type of the process (modular or central) is not predefined but depends mainly on the complexity of the subject to analyze. That means, if someone manifests obvious signs of a liar, we would automatically detect him through a modular process. But if that person sends mixed signals, ambiguous cues, then we would need to switch to a central process to issue a judgment. Through a first experiment, Arminjon defines facial cues that can be categorized as either lying or non-lying cues. The facial cues (eg. mouth overbite, smile closed, etc...) where selected from another study (DePaulo et al, 2003). Then, they used a computer software to generate a neutral virtual face and added each cue uniquely to this face (each face only had one cue). Finally, they presented the faces to the participants, asking them to judge whether that face was lying or not. From the results of that study, Arminjon was able to categorize facial cues as lying or non-lying cues. Then, in his further experiment, he selected 4 lying cues and 4 non-lying cues, combined them in all possible combinations and finally submitted the resulting faces to the participants that would perform the same task as before. He additionally captured the response time of the participants in order to try and determine if the process is modular or not. Results showed that faces with high discrepancy and intermediate amount of cues where the ones that caused the most randomness in the participants' responses. In such case, there was no clear tendency in the participants' answers to whether that face was lying or not. In other words, if a liar would exhibit such cues, it would be hard to detect him. Also, the response times for this case were significantly higher than usual. This would confirm the hypothesis that when the analyzed face is too ambiguous, our modular liar detector process would fail to ensue a proper judgment and we would switch to a controlled process. Such a switch would be the cause for a slower and a more random, unsure response.
However, a limitation of this study is that it has been conducted using computer generated faces that have no context and no relation to an actual lie. They are not necessarily a good representation of a real lying face.

For our experiment, we want to further test Arminjon's results. But instead of taking computer generated faces that are unrelated to real lies, we will use real images of liars. Our first step was to capture images of people during real lying conditions. We follow the assumption that the discrepancy of the two type of cues (as described in Arminjon's study) are mainly due to the fact that the liar tries to hide his lie by covering it with controlled non-lying cues. Thus, for our study we need the liar to act on a stake in order to fully reproduce a

[1] using Raven's advanced progressive matrices (Bors and Stokes, 1998)

real lie. To capture the images, we organize several tournaments based on the game ”Two truths, one lie”. In this game, participants are required to make three declarations, two of which must be true and the other a lie. If the player’s lie is not discovered, he wins else the opponent wins. To create a stake, this game is organized in a tournament form such that the winner receives a prize. Every statement is captured on video from which we extract images of the same people either lying or telling the truth. These images are then used in the same experiment as Arminjon. That is, we select an equal number of pictures with lying or non lying faces and then we record participants’ answer and response time when asked if the presented images depict a liar or not. Our hypothesis is that lying faces should cause a higher randomness in the participants’ answers. In fact, we assume that liars try to hide their lies by forcefully expressing non-lying cues, but at the same time, they cannot entirely repress lying cues. Thus, this is a very similar situation as the high discrepancy, intermediate number of cues discussed in Arminjon’s study and we should have the same type of results. As specified, we have images of the same people either lying or telling the truth. Then, if we compare the difference of response times for theses cases, we would expect to have a longer response time for the images of liars because they are more ambiguous. If that is the case, then it would confirm that our lying detection process is either a modular or central process depending on the complexity of the analysis. Else, different results could be explained either by the fact that our assumption is wrong and real liars do not express ambiguous signs (or we failed to properly capture them), or that our cheater detection process is actually not affected by the subject complexity (that would contradict Arminjon’s results).

Method

For our experiment, we want to confront the participants to subjects in real condition lies. That is, showing them subjects that are lying while trying to hide it. Also, we need to show the same subjects sometimes telling the truth so that we can build our experiment. The best way we found to capture such images is to invite subjects to play the game *“Two truths, one lie”*.

Two truths, one lie

In this simple game, two opponents face each other. One is the speaker, the other the observer. The speaker then proceeds to make three declarations (of any kind), where two must be true, and one a lie. The goal of the observer is to guess which of the three is the lie. If he succeeds, the he is awarded a game point, else the speaker receives the point.

In our case, we separated the opponents in two different rooms so that the speaker is not given any possible distraction. The speaker is placed in front of a uniform white wall and facing our camera. We also placed a softbox light above the camera for a better control and consistency of lightning throughout our takes. A second camera is also present and directly retransmits to the observer in an other room. The observer can view and hear the speaker from his room, however the inverse is not true. In order to create a stake, this game was organized in a tournament form. We organized two tournaments of four people that is presented in the following form:

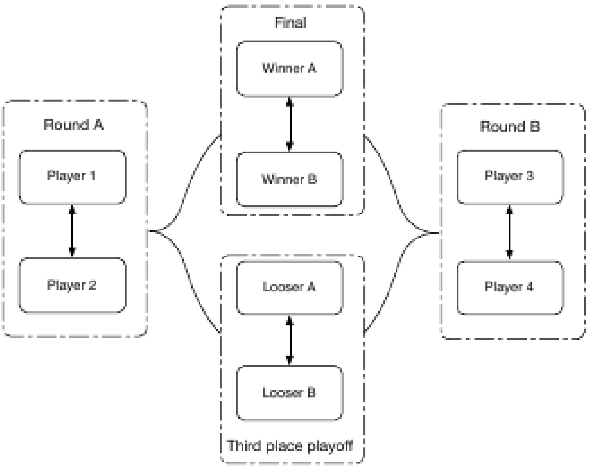

Figure 1: 4 people tournament form

For each game, the first roles (speaker or observer) are chosen arbitrarily. Then, the speaker announces his three declarations and a point is attributed to the speaker or observer as explained above. After that, the roles are inversed, and so on. Doing so, every player gets a chance to be the observer and the speaker. The role inversions go on until one of the player reaches three points, making him the winner of the current game.

The winner of the final and the third place playoff where attributed a price. Obviously, the price for the winner was bigger than the one for third place. Thus, for any game of the tournament, each player had a real motivation to win. This was to ensure that the speakers would really try to hide their lies.

In total, we organized two tournaments of this form, gathering 8 different subjects (5 men, 3 women, all around 21 to 23 years old).

Online survey

In our experiment, we wanted to confront the participants to the videos we captured during the game presented above. In order to organize that, we decided to present our experiment in a online survey form where participants could take part by themselves by following the instructions on our website.

A prior step was to process the videos we captured. Firstly, we separated each declarations (truths or lies) in single videos. From there, we filtered all videos that where too short, where the subject did not face the camera as instructed or any other reason that would make the video unusable. We removed the sound of the videos because that is out of the scope of this experiment. In fact, we only want to test the ability to recognize lying facial cues, so any distraction must be removed. Following that idea, we also set the video in black-and-white and cropped the frame to fit only the subjects face. In fact, external features such as hairstyle, skin color, clothing impair our judgment ability as demonstrated by Bonnefon (J-F Bonnefon, 2012).

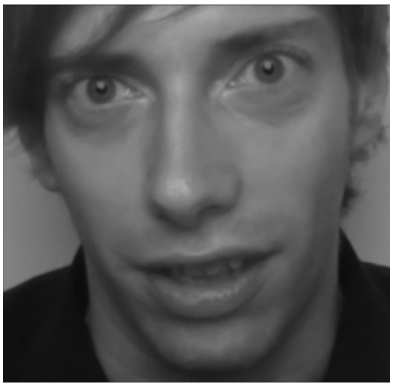

Figure 2: Example of the video format

Then, we decided to only use short videos to keep the experiment feasible in reasonable time. Also, we wanted a standardized way to select which portion of the video to keep so that we have a full consistency between the videos. But how to decide when to cut the video? Unfortunately, there isn't much literature about the precise moment when the visual cues of a lying face appear. The may appear at the beginning of the liar's declaration, or maybe at end, during the short pause after his declaration. That is unclear. Thus, we made the arbitrary choice to set the beginning the video 1.2 seconds *before* the end of the subject's statement and stop it 0.3 seconds *after* it. Thus, all our videos have a definite length of 1.5 seconds.

We made a selection of 36 videos (containing 18 truths and 18 lies) that we used for the survey. In this survey, we show the participants the videos discussed above. Then, they answer (by pressing the keyboard key `F` or `J`) whether they think the subject of the video is lying or not. The participant is instructed to answer as spontaneously as possible. Also, the videos automatically disappear when they reach the end and the participant hasn't any control upon them, thus he only has a very limited time to analyze them.

Each participants is first shown four videos as training so that they learn to use the interface. Then, they're shown each of our 36 videos four times (so that we can calculate the consistency), totaling 144 videos. All the videos are displayed in a random order, but still ensuring that no repetitions occur. For each video, we register the participant's answer as well as his response time (counting from the start of the video).

Results

Filtering

From the online survey, we had 198 participations (104 men, 94 women). We then applied several filtering criterias.

Firstly, there were some participants who did not properly finish the experiment, that is, they quit before the end. We decided to exclude them. We also excluded participants who answered with the same response for more than 85% of the time. The reason is to filter out any participant that is always pressing the same key, and thus, not seriously taking part to the experiment. When excluding a participant, we excluded *all* the answers associated with those participants.

That leaves us with 101 participants (51 men, 50 women).

Then, from the remaining participants, we excluded some unfitting answers. More specifically, we excluded all answers that where issued too fast (in less than 0.3 seconds) because it is impossible that the participant could normally process an input and answer so quickly. Such a quick answer would most likely be due to a participant accidentally pressing a key, thus we filtered it out. We also excluded the answers with a response time higher than seven seconds. We decided to exclude these answers because such case is likely to be due either to a loss of attention from the participant, a distraction that made him loose time, or because he was over thinking his answer (which contradicts our instruction to answer spontaneously). The specific value of seven was chosen arbitrarily.

Finally, we also filtered all results with a low consistency. As mentioned before, every video was shown four times to each participant. We then calculated a level of consistency as follows:

If we consider a video in particular, then, since each video is displayed four times, a participant has four answers for that video. If the participant submitted the same answer to all four of them, then his consistency is **1**. If he submitted the same answer to three of them, and the other answer for the last one, then his consistency is **0.75**. If he answered the same answer to two of them, then his consistency is **0.5**. Note that the consistency is independent of the fact that the participant answered right or wrong, but is a measure that shows if the participant was answering in a consistent manner.

A consistency of 0.5 represents actually no consistency at all, that is what we would get if we would answer at random. This can be explained by the fact that, for that specific video, the participant is very insecure of his answer. Therefore, when asked several times, his answer may vary.

For our results, we discarded all answers with consistency of 0.5. Also, we aggregated the participants answers to only one per video. That means that for a given video, instead of having four answers for each participant, we only took one, that is the one most answered by the participant. As for the response time, we consider the average time of the four answers.

Accuracy

The main measure we use to compare groups is the accuracy. We define the accuracy as being the level of correctness of the participants answers. For instance, an accuracy of 100% would indicate that the participants made every guess right.

In our case, if we take all participants, we reach a general accuracy of **55.65%**. That value is close to 50%, meaning that the participants answers are almost random. However, as we can see from this plot, participants accuracy is slightly above random chance:

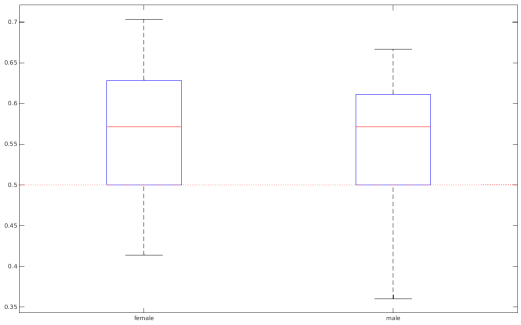

Figure 3: Accuracy by gender

Furthermore, we can see that participants are more effective at detecting truths rather than lies. This can be shown with following table:

	yes	no
lie	45.14%	54.86%
truth	33.61%	66.39%

Figure 4: Accuracy for lies/truths

This tables show the accuracy conditioned by the fact that the subject is a liar or not. For instance, if we only consider the answers to the lying subjects, then, participants show an accuracy of 45.14%. Inversely, for subjects telling the truth, participants show an accuracy of 66.39%.

Accuracy relation to age

Our result show no correlation between accuracy and age. However, we do not have well distributed samples of ages thus, we cannot draw any significant conclusion.

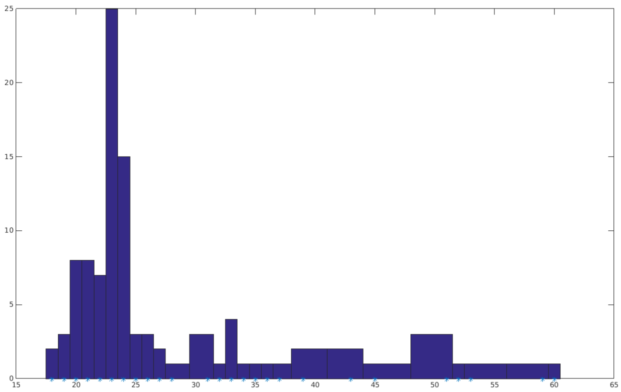

Figure 5: Distribution of participants age

Accuracy relation to gender

The distribution between men and women is almost 50%. However, their accuracy show no significant difference.

Accuracy relation to degree

We have a fair distribution of degrees in our population (except for others, which can thus be ignored).

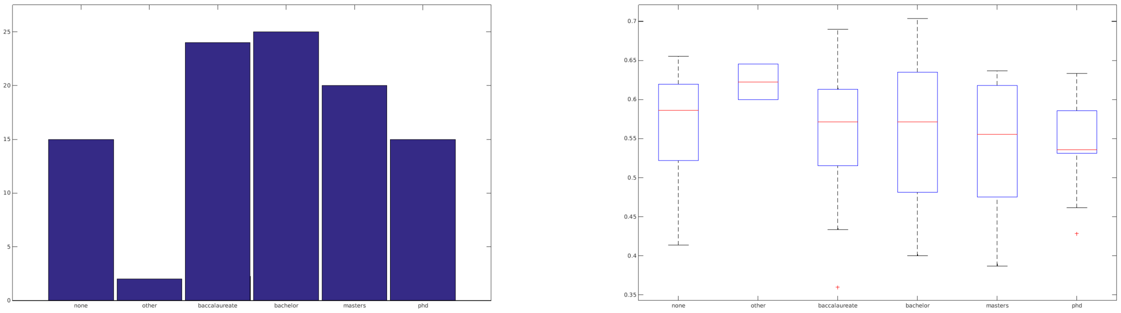

Figure 6: Distribution (left) and accuracy (right) by degree

If we observe the mean of the accuracy, it seems descending with higher degree. However, the variance is too large to be able to make such an observation. Indeed, the only valuable observation that can be made is that there is none to very little difference of the accuracy for these groups.

Accuracy relation to response time

We now compare the accuracy with respect to the response time. First, we calculated the mean response time for each participant and observed that they follow a normal distribution of mean **2432.46 ms** and standard deviation of **452.67**.

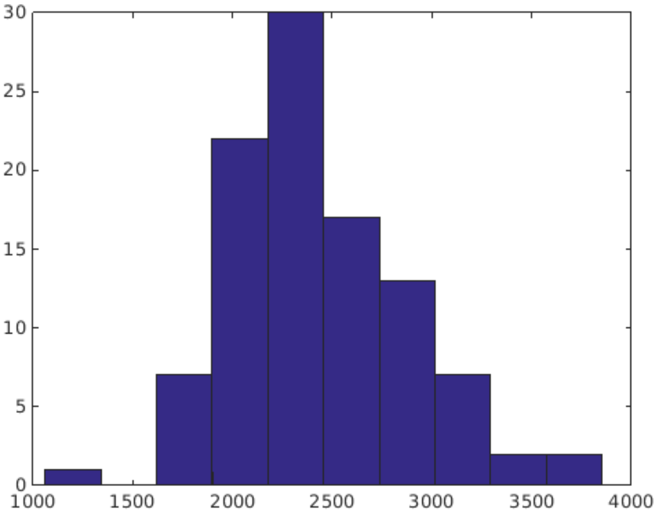

Figure 7: Response time distribution

We then proceeded to search any correlation between the response time and accuracy of the participants. To bring out a relation, we used a linear model to estimate the general tendency of our samples.

The following plot shows, for each participant, his accuracy over his average response time in milliseconds. The red line represents the linear regression of theses samples.

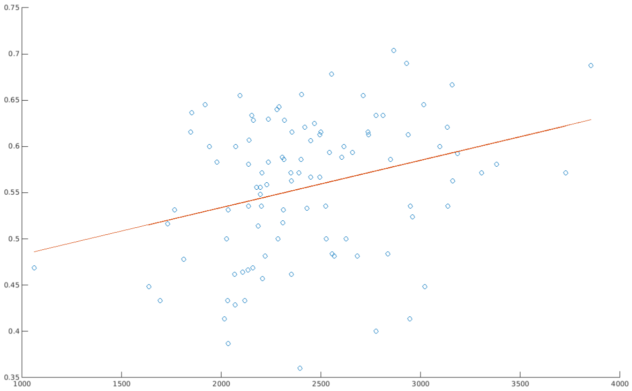

Figure 8: Accuracy in relation to response time
$t(97) = 3.169, p = .002$

The linear regression is growing, showing that higher response time generally results in higher accuracy. In other words, participants that used about one second more, generally achieved a better result that others.

Discussion

The results we obtained first show that our general accuracy at detecting lies is weak, close to randomness. This aligns to previous studies suggesting that we are generally bad liar detectors. We also see that trustworthiness detection seems easier than liar detection. In fact, participants reached significantly higher accuracy when considering their results for non-liars only than when considering liars only. These results go in accordance with Bonnefon's results (J-F Bonnefon, 2012) showing that we are able (to some extent) to detect trustworthiness. However, the accuracy results are, even if very slightly, above the random level. That would indicate that despite the fact that we are bad liar detectors, we are still able to make a better prediction than random, showing that we naturally possess some minimal skills. Even if theses skills are absolutely not reliable, it still indicates that we are not entirely helpless faced with this task. This can be very likely explained with Verplaetse's results (Verplaetse, 2007) explaining that the ability to hide behind a lie, and its converse, the ability to assess someone's trustworthiness are in fact subject to an evolutionary predator-prey arms race. Following this reasoning, we possess an ability to detect trustworthiness, but at the same time, liars also possess abilities to hide themselves making it very difficult in general to spot them.

Our result show no significant impact of factors such as the age, the gender or the degree of the participants over their general performance. That would be indicating that experience (by age or gender) does not actually help us to better detect liars. This would seem to be in contradiction with Ekman's results (Ekman et al, 1999) which state that experienced individual can reach higher accuracy than others. However, Ekman defines an experienced individual as someone belonging to some group highly involved with lying detection (such as law-enforcers, psychologist, ...) as in our case, we can only base our groups on age and degree. In fact, if a participant degree has no relation with deception, then it might be irrelevant in the consideration of his experience. Since we do not have any information about the nature of the participants degrees, we cannot properly make a measure of experience in deception for our participants.

A difference is however noticeable amongst participants who answer slower. Indeed, participants with longer response times tend to have higher accuracy. When related to the question of modularity, this would indicate that the liar detection process might not be modular. In fact, if it was, the participants would be able to answer quickly and yet reach a reasonably good accuracy. However, participants that spent more time reached better results. This shows that using a few more seconds to think, which would indicate that the participant does not answer spontaneously/instinctively but involves some amount of active reflexion, actually helps to improve the accuracy. However, to ensure such statement, we would need to conduct more studies on the point in time of alternation from modular process to central. Indeed, we cannot properly affirm that, in our experiment, people who answered quickly (and generally less effectively) performed a modular treatment and the others (with better results) used a central process.

For further studies, it could be interesting to have a more distributed age range amongst the participants so that analysis can be performed and try to detect variations according to age. Also, a major improvement would be to categorize the subjects videos. In their actual state, the only information we have per video is either the subject is lying or not. However, if we could analyze facial cues of our subject and categorize them, we might be able to observe variations amongst the different categorizes. For instance, it could be possible that subjects show the most discrepancy in their facial cues are the ones that cause the longest response times if we follow Arminjon's hypotheses (M Arminjon, to be released).

References

[1] Paul Ekman, Maureen O'Sullivan and Mark G. Frank (1999). A few catch a liar Psychological science.

[2] Jan Verplaetse, Sven Vanneste and Johan Braeckman (2007). You can judge a book by its cover: the sequel. A kernel of truth in predictive cheating detection.

[3] Jean-François Bonnefon, Astrid Hopfensitz, Wim De Neys (2013). The modular nature of trustworthiness detection.

[4] Mathieu Arminjon, Amer Chamseddine, Vladimir Kopta, Aleksandar Paunović, Christine Mohr, François Ansermet, Pierre Magistretti (to be released). Are we modular lying cues detectors? The answer is "maybe sometimes"

[5] S. Centorrino, E. Djemai, A. Hopfensitz, M. Milinski and P. Seabright (2011). Smiling is a costly signal of cooperation opportunities: Experimental evidence from a trust game.

[6] D. A. Bors and T. L. Stokes (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form.

[7] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, H. Cooper (2003). Cues to deception.

[8] W. Hippel and R. Trivers (2011). The evolution and psychology of self-deception