Preregistration

How to sway voters? Part 1

Ralf H. J. M. Kurvers[1], Uri Hertz[2], Bertrand Jayles[1], Bahador Bahrami[1,3]

[1] Max Planck Institute for Human Development (Berlin), Center for Adaptive Rationality

[2] University of Haifa, Department of Cognitive Sciences

[3] Ludwig-Maximilians Universität München, Faculty of Psychology and Educational Siences, General and Experimental Psychology

27. February 2019

Study Information

Research questions

RQ1: Can a dishonest advisor succesfully draw the attention of a client?

RQ2: How is the ability of a dishonest advisor to draw the attention of a client affected by the level of uncertainty in the environment?

Hypotheses

H1 (RQ1): We expect that a dishonest advisor is better able to draw the attention of a client than an honest advisor.

H2 (RQ2): We expect that a dishonest advisor is better able to draw the attention of a client when there is high uncertainty (i.e., when the outcome of a trial is difficult to predict).

Design Plan

Existing data

Registration prior to creation of data. As of the date of submission of this research plan for preregistration, the data have not yet been collected, created or realized.

Study design

This study will use an experimental paradigm similar to the one in Hertz et al. (2017). In brief, there will be three agents in the experiment: two advisors and one client. On each trial, the client (human) has to make a decision which of the two advisors (algorithms) to follow. The advisors observe part of the avialable evidence, and then decide how to communicate this evidence to the client. The client does not observe any evidence and thus has to rely on the advisors.

Each trial consists of the following steps:

- A rack of 100 balls is filled with a mix of white and black balls.
- Both advisors observe the same (randomly sampled) 75 balls from the rack. (The value of 75 adds a small amount of noise to the signal; that is, not all evidence is available).
- The client selects one of the two advisors to follow.
- Both advisors communicate their advice to the client. The advice consists of (i) colour (black or white), and (ii) level of confidence (1-5 scale).
- One ball from the rack is randomly drawn. If the colour advice of the selected adviser matches (/does not match) the colour of the drawn ball, the client wins (/loses).

In this version of the experiment, the two advisors will be algorihtms and play two predefined strategies:

- Honest Advisor (HA) The HA reports the colour and confidence honestly: if the majority of observed balls are white (/black), the HA reports white (/black) to the client. The HA also reports confidence honestly, using the following linear mapping: majority of one colour: 50-60%: Confidence (CF) = 1. 60-70%: CF = 2. 70-80%: CF = 3. 80-90%: CF = 4. 90-100%: CF = 5. Additionally, we add a small amount of noise: in 20% of cases the CF level was increased with one unit; and in 20% of cases lowered with one unit (provided this was possible).

- Dishonest Advisor (DA) When selected, the DA performs the same strategy as the HA, thus reporting the evidence honestly. When not selected DA's strategy depends on the strenght of the evidence:

- if the evidence is strong (> 75% of observed balls are of the same colour), DA performs the same stategy as the HA.
- if the evidence is weak (<= 75% of observed balls are of the same colour), DA reports the colour of the **minority** of the balls to the client with a confidence level 2, 3 or 4 (all values equally likely).

The **client** is a human player and will at the start of each trial select one of the two advisors to follow. The experiment runs for 20 trials.

We plan four treatments, varying the uncertainty in the environment. Specifically, we manipulate the ratio of black vs. white balls between treatments:

1. 25% of trials: 90 balls of one colour. 75% of trials: 50 balls of each colour.
2. 25% of trials: 90 balls of one colour. 75% of trials: 60 balls of one colour.
3. 25% of trials: 90 balls of one colour. 75% of trials: 70 balls of one colour.
4. 25% of trials: 90 balls of one colour. 75% of trials: 80 balls of one colour.

From treatment 1 to 4 the level of uncertainty decreases. That is, the trials become increasingly easier. We have specific predictions based on a pilot experiment and simulations.

Pilot experiment

In a pilot, we tested 29 participants (recruited via MTurk) playing for 20 trials using the settings of treatment 1. Figure 1 shows the results of this pilot. Fig. 1A shows that participants had an overall preference for the dishonest advisor. Fig. 1B shows (part of) the behavioural mechanism underlying this result: participants were more likely to switch advisors between trials when (i) they lost, (ii) the non-selected advisor gave advice that was opposing the selected advisor, and (iii) when it did so with high confidence.

Simulations

To predict the behaviour of the client across different environments we used simulations, closely replicating the experimental settings. The simulations (implemented in R) consisted of the following steps:

- A rack of 100 balls is filled with a mix of white and black balls.
- Two advisors (one HA and one DA, as described above) observe 75 balls.
- In trial 1, the client randomly selects either of the two advisors. For all next trials, this choice depends on trial outcome and advice (see below).
- Both advisors communicate their advice to the client (colour + confidence) following their respective strategies.
- One ball is randomly drawn from the rack. If the advice of the selected adviser matches (/does not match) the drawn ball, the client wins (/loses).
- The client decides whether to switch advisor or not. This switching likelihood is based on the observed values in the pilot (Fig. 1b). To illustrate, if the client lost and the non-selected advisor gave the opposing advice with confidence level 4, the client had a likelihood of 0.9 to switch advisor (most left bar in Fig. 1B) .

We used the four different parameter settings of the four treatments as described above (i.e, systematically varied the level of uncertainty). On each setting, we ran 10,000 simulations, each consisting of 20 trials. We report averaged values over all runs. Fig. 2 shows the results of the simulations, which are also **our main predictions** of this preregistration. In the treatment with the highest uncertainty (50 balls of each colour), we replicate our empirical finding: a strong preference for the DA (see Fig. 1A). This is not surprising, since the simulations are based on the observed behaviours, but this suggests that the simulations capture the key variables of interest. Increasing the easiness of trials, results in a gradual shift in the direction of the HA. Hence our second main prediction: the DA is better able to draw the attention of a client when there is high uncertainty. Note that nevertheless, the DA never performs worse than the HA. At the lowest level of uncertainty (80 balls of one colour), the DA effectively turns into an HA (since the evidence is always strong). This thus serves as a control condition.

Randomization In our planned study, the **HA** and **DA** will appear either on the left or right side of the screen and this will be counterbalanced between participants. That is, approximately half of the participants will experience the HA on the left side, and the other half will experience the HA on the right side of the screen.

Blinding	In our planned study, participants in all four treatments will receive the same instructions. Participants are unaware of the exact treatment they are in, and they are unaware that there are different treatments.
Data collection procedures	The data will be collected using an online study implemented in *javascript*. Participants will be recuited over *Amazon Mechinal Turk* (https://www.mturk.com/). The study will take approximately 10 minutes and participants who complete the study receive 3 dollar compensation.
Sample size and stopping rule	In each of the four treatments we plan to collect data for 30 participants. We will stop data collection within a treatment as soon as 30 participants succesfully completed the treatment. A successfull completion means that a participant started and finished the experiment. Participants who quit the experiment prior to completion will be excluded from the analysis.
Measured variables	The key variable of interest is the *choice* of the clients. Additionally, the following demographic variables will be elicited at the beginning of the study: 1. Age. 2. Gender: female, male, other, do not want to report. 3. Education: basic, high school, college, posgraduate.
Data exclusion	Participants who did not complete the entire experiment will be exlcuded from the analysis. Likewise, participants who never switched between advisors will be excluded.

Reference

Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C. D., & Bahrami, B. (2017). Neural computations underpinning the strategic management of influence in advice giving. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-02314-5

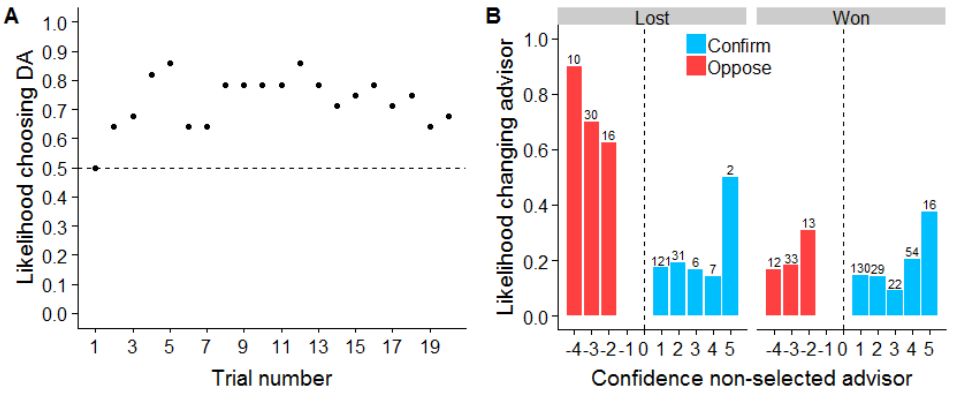

Figure 1: Results of the pilot experiment. (A) The likelihood that the client selected the dishonest advisor (DA) over the 20 trials. Horizontal dashed line indicates chance level. (B) The likelihood that the client switched advisor after a trial as a function of whether the client lost or won, whether the non-selected reported the same (Confirm) or the different (Oppose) color, and the confidence level of the non-selected advisor. Numbers above the bars indicate the number of occurences.

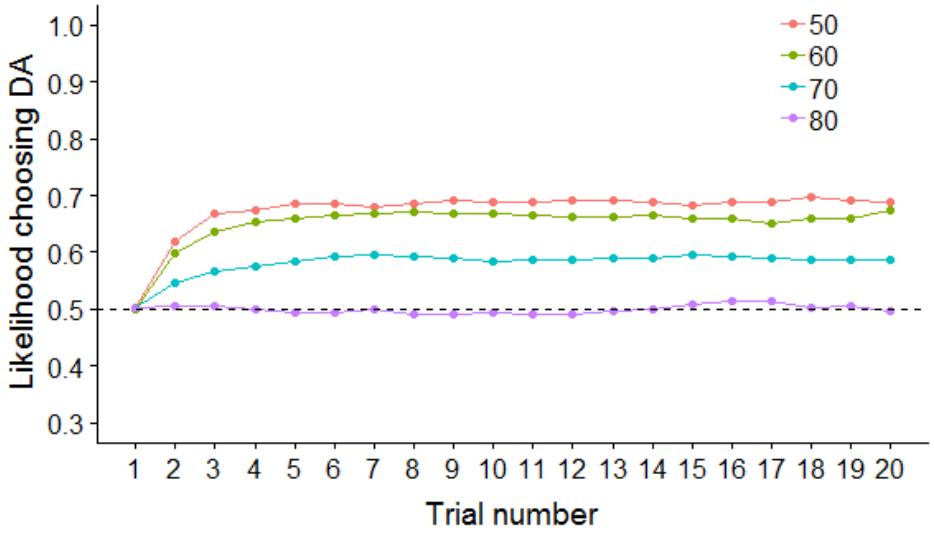

Figure 2: Results of simulations. The likelihood that the client selects the dishonest advisor (DA) over the 20 trials. Horizontal dashed line indicates chance level. Numbers in legend indicate the proportion of balls of one color in the hard trials. The higher this number the lower the uncertainty.