Bayesian Persuasion with Fact-Checking

Zeeshan Samad
Utah State University

Lucas Rentschler
Utah State University

Abstract

This paper experimentally investigates the impact of a fact-checking device that probabilistically flags false messages in a Bayesian persuasion framework.

In theory, such a device should not reduce the effectiveness of persuasion because the sender can simply compensate for increases in fact-checking by lying more frequently. However, our experimental data contradicts this prediction.

We find that senders do not lie any more frequently in the presence of fact-checking than in its absence, a behavior consistent with lying aversion. By contrast, receivers' actions are monotonic in their induced posterior, a behavior consistent with Bayes rationality.

Introduction

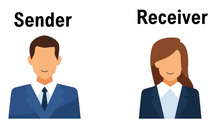
Bayesian persuasion is a framework to study any situation where one person is trying to persuade another to do something. Originally developed by Kamenica and Gentzkow (2011), this model has been well studied since then—with over 2000 extensions of this framework!

An important extension is to incorporate fact-checking in this framework. This is because all real-life applications of Bayesian persuasion include some sort of fact-checking. For example, allegations made by prosecutors are often challenged by a witness presented by the defense (so the witness essentially *fact-checks* the prosecutor's claims). Similarly, in the context of lobbying, a policymaker's staff fact-checks information provided by lobbyists. Likewise, when people share some disinformation on social media, it can get flagged as misleading.

However, very little research effort has been spent on exploring the effect of fact-checking in a Bayesian persuasion framework. Despite the vast literature in this topic, there is only one theoretical paper (Ederer and Min, 2022) and no experimental paper that studies this.

Game

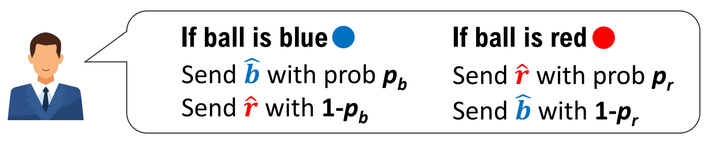
Two players: Sender and Receiver



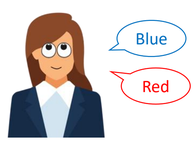
An urn contains 2 balls: 1 red, 2 blue

Two possible messages: $\hat{b}$ = "ball is blue"
$\hat{r}$ = "ball is red"

Stage 1:
One ball is drawn randomly.
Sender sees the ball. Receiver does not.
Sender commits to a messaging strategy $p = (p_b, p_r)$ where: $p_b = \Pr(m = \hat{b} \mid \bullet)$, $p_r = \Pr(m = \hat{r} \mid \bullet)$



Stage 2:
Message gets realized according to p_b and p_r
If message is true, it is not flagged.
If message is false, gets flagged with prob $\boldsymbol{q}$ and not flagged with $\mathbf{1} - \boldsymbol{q}$.
Receiver sees 3 things:
(1) Sender's messaging strategy
(2) Realized message
(3) Whether it is flagged or not
Then Receiver guesses the ball's color

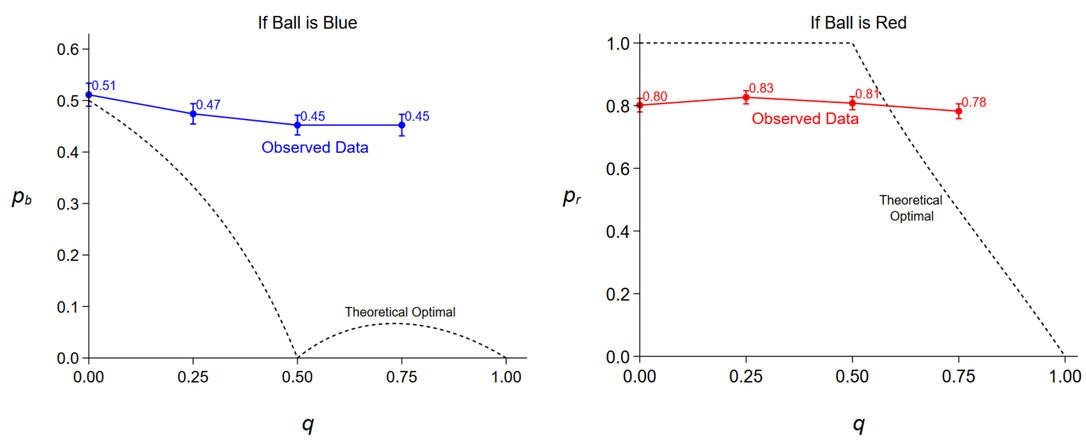


Payoffs:
If Receiver's guess is correct, Receiver earns $2
If Receiver's guess is Red, Sender earns $2

Experiment Design

Treatment variable = q (probability of fact-checking)
Four treatments: $q = 0\%, 25\%, 50\%, 75\%$
Between-subject design
60 subjects in each treatment (240 subjects in all)
All subjects were USU students
Everyone played 20 paid rounds (after 2 practice rounds)
Roles (sender/receiver) remained fixed for all rounds
Random rematching between rounds
We used the strategy method

Result 1

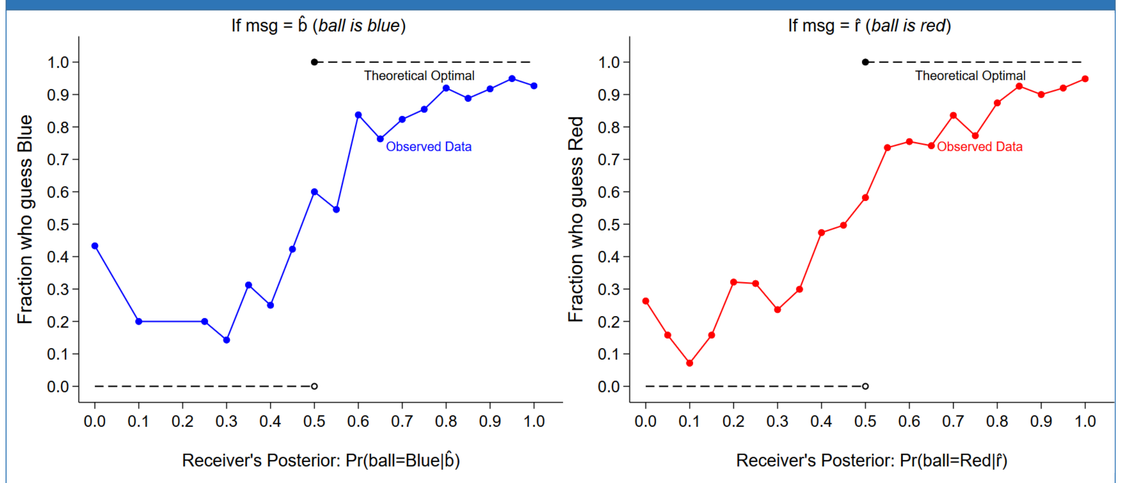


In the graphs above, y-axis corresponds to probability of sending true message.

Theory predicts that senders will lie more as probability of fact-checking increases. But this is not what we observe in our experiment. In fact, we don't observe any meaningful increase or decrease in lying as q increases.

This means that fact-checking is effective after all! Real-life senders are either not as sophisticated or they don't expect receivers to be Bayesian.

Result 2



The graphs above are pooled for all treatments. This is because the receiver's posterior accounts for the probability of fact-checking.

Meaning of upward sloping blue and red lines: As the posterior probability about the ball being blue (red) increases, more receivers guess blue (red). This means that most receivers behave how a Bayesian Receiver would.

Conclusion

Theory:
Fact checking won't help because senders will just compensate by lying more.

Experimental evidence:
Actually senders don't lie more. So fact checking helps.

Contact

Zeeshan Samad
Utah State University
Email: zeeshan.samad@usu.edu
Website: ZeeshanSamad.com
Phone: +1-301-503-8232