## Title

Mistakes, Overconfidence, and the Effect of Sharing on Detecting Lies

Permalink
https://escholarship.org/uc/item/3rr210r7

Journal
American Economic Review, 111(10)

ISSN
0002-8282

Authors
Serra-Garcia, Marta
Gneezy, Uri

Publication Date
2021-10-01

DOI
10.1257/aer.20191295

# Mistakes, Overconfidence, and the Effect of Sharing on Detecting Lies

Marta Serra-Garcia and Uri Gneezy*

May 2021

**Abstract**

Mistakes and overconfidence in detecting lies could help lies spread. Participants in our experiments observe videos in which senders either tell the truth or lie, and are incentivized to distinguish between them. We find that participants fail to detect lies, but are overconfident about their ability to do so. We use these findings to study the determinants of sharing and its effect on lie detection, finding that even when incentivized to share truthful videos, participants are more likely to share lies. Moreover, the receivers are more likely to believe shared videos. Combined, the tendency to believe lies increases with sharing.

**JEL Classification:** D83, D91, C72, C91.

**Keywords:** Detecting lies, overconfidence, sharing behavior, fake news.

*Serra-Garcia: Rady School of Management, UC San Diego and CESifo (email: mserragarcia@ucsd.edu.); Gneezy: Rady School of Management, UC San Diego (email: ugneezy@ucsd.edu). This research was conducted under IRB #161827. We are grateful to Johannes Abeler, Michele Belot, Tim Lohse, Jeroen van de Ven, and several seminar and conference participants for helpful comments.

# 1 Introduction

“Fake news” came into the spotlight during the 2016 US presidential election, referring to fabricated news stories with the intent to deceive (Lazer et al., 2018). An important way in which fake news spread and affect behavior across various domains, including voting, healthcare decisions, and political violence, is by sharing on social media. For fake news to be effective in influencing behavior, its target audience needs to see them, and then believe the lies that are inherent in them.

Much of the discussion in the literature on fake news focuses on motivated beliefs (e.g., Kunda, 1990; Bénabou and Tirole, 2011; Bénabou, 2013): Individuals wish to believe news that is in line with their ideology. This motivation prevents people from being skeptical and could make them more likely to spread lies, which increases their momentum and impact even further (e.g., Nisbet, Cooper, and Garrett, 2015; Kahan, 2016a, b; Allcott and Gentzkow, 2017; Pennycook and Rand, 2019 and 2020).[1] In this paper, we take a step back and show that individuals make mistakes and are overconfident in their ability to detect lies even without the motivation to believe them. We then connect the findings to sharing behavior and show that our participants are more likely to share lies, and that receivers believe that videos are more likely to be true when they are shared, relative to when they are not. Combining these effects, we find that sharing can increase the belief in lies.

To study the ability to detect lies, we developed a new experimental paradigm and conducted experiments with 3,508 receivers. We first created 30-second recorded videos with senders, who were incentivized to convince the receivers, who would later watch the videos, that they were seeing and describing a true news event. In one case, senders saw a real headline and associated picture from the front cover of the international edition of the *New York Times* (NYT). In the other case, they saw a blank screen. Each sender was recorded twice: after seeing the news event and after seeing the blank screen. We then showed these videos to receivers and asked them

[1] See also, De keersmaecker and Roets (2017), Lazer et al. (2018), Pennycook et al. (2018), Tappin et al. (2018), Grinberg et al. (2019), Bronstein et al. (2019) and Thaler (2019).

to decide, for each video, whether the sender was telling the truth (i.e., describing a real headline and picture) or lying. Receivers were paid based on accuracy.

We found that receivers displayed a limited ability to detect lies. The distribution of receivers' ability was close to chance in distinguishing between lies and truthful videos, and, on average, their accuracy was between 50% and 53% (compared with a 50% chance of a random answer). Receivers in our experiment made both type I errors (believing a false video) and type II errors (not believing a true one). These results are broadly in line with the existing literature on lie detection. In a meta-analysis of studies on lie detection, Bond and DePaulo (2006) found that 54% of statements were correctly classified (see also Bond and DePaulo, 2008). Konrad, Lohse, and Qari (2014) and Dwenger and Lohse (2019) showed that receivers were slightly worse than chance at detecting false taxable income reports. In contrast, Belot and van de Ven (2017) found that receivers in the role of buyers were better than chance at detecting the lies of sellers. Similarly, trustors in a trust game tend to exhibit a limited ability to detect trustworthiness based on trustees' facial pictures (e.g., Bonnefon, Hopfensitz and De Neys, 2013, 2017), and trustors also easily form first impressions from faces, though such impressions are unrelated to stable personality traits (e.g., Todorov et al., 2015).

We then asked what the source of receivers' low ability is: simple noise or receivers putting wrong weights on predictor variables. We explored whether senders' speech and emotions contain cues that may make a video more or less believable and whether receivers pick up some of those cues. To that end, we used facial-expression-recognition software to measure emotions and facial movements of senders during their videos, and collected other measures of their speech (e.g., word count). A number of emotions, facial movements, and speech characteristics of senders affected receivers' beliefs. However, the data show that receivers put the wrong weight on senders' cues: the effects were often in the wrong direction or concentrated on indicators that did not predict a video's truthfulness. For example, more words said in a video were associated with a higher likelihood that the video was false. Yet, receivers tended to believe videos with a higher word count more.

The wrong use of cues by receivers may explain why they were overconfident about their ability. We measured both absolute overconfidence (beliefs about correctly detecting lies) and relative overconfidence (quartile of the distribution of the ability to detect lies). Overconfidence can be strategically valuable as shown by Charness, Rustichini and van de Ven (2018), who gave participants incentives to convince others that their ability on an IQ test was high. Participants often convinced themselves that they performed better than they actually did, and by that achieved advantage. Most previous studies in lie detection have investigated self-perceived competence without objective benchmarks.[2] To our best knowledge, we are the first to measure overestimation and overplacement in lie detection based on incentivized measures.

The limited ability of receivers to detect lies, combined with their overconfidence, lays the foundation for studying sharing behavior. Fake news often spread through sharing on social media. The literature on why people share content on social media, and what its impact is, is small. Data in these papers are based on descriptive studies about what people share on Facebook and Twitter (e.g., Guess, Nagler, and Tucker, 2019; Allcott, Gentzkow, and Yu, 2019), or hypothetical choices based on news headlines with political motivations, from true and fake news sources (e.g., Pennycook and Rand, 2018a). Little is known about which kind of news is shared more and what the effect of sharing on receivers is. If lies are shared more often, and shared news tends to be trusted more, this association could fuel the spread of lies.

We use our experimental paradigm to examine the determinants of sharing and test the effect of sharing on the rate at which lies are believed. To that end, we test

[2] For example, Frank and Ekman (1997) asked participants, “In this video, how well do you think you did in telling who was lying?” The answers were rated on a 5-point scale (1 = very poor and 5 = very good). To measure accuracy of confidence, answers need to have an objective measure to compare them to. DePaulo et al. (1997) identified six studies that had an objective benchmark, and all reported overconfidence in absolute terms and without incentives. Later studies found mixed results regarding overconfidence (e.g., Mann et al., 2004; Swann et al., 1995). For a general discussion on overconfidence, see Moore and Healy (2008) and Benoit, Dubra and Moore (2015).

two main questions. First, would lies be shared more often than truthful videos? Second, does knowing that a video is shared affect the receivers' beliefs? If the answer to these two questions is positive, then sharing would increase the likelihood that receivers believe a lie. This effect is important because believing a lie is the type of consequence of sharing that contributes to fake news being effective and potentially concerning.

To that end, we extended our design to include two receivers. The first receiver, to whom we refer as Receiver 1 (or R1), watched eight videos and was incentivized to guess for each if it was true or false. The second receiver, to whom we refer as Receiver 2 (or R2), was shown the pictures and titles from the eight videos and was asked to choose four of them to watch, knowing that she/he would also be incentivized to guess the truthfulness of each chosen video. The new element of the design is that Receiver 1 was also asked to choose one of the eight videos to share with Receiver 2, who was informed about the video that was shared with her/him before choosing which videos to watch. To control for different motivations for sharing, in the first treatment, Receiver 1 was paid a bonus if Receiver 2 picked the shared video to watch and that video was true. In the second treatment, we replaced the incentive criterion from true to believable; that is, Receiver 1 was paid if Receiver 2 chose to watch the shared video and believed it. In these main treatments, Receiver 2 knew about Receiver 1's incentive. In two control treatments, either no sharing information was provided, or Receiver 2's choices had no effect on Receiver 1.

This design allows us to answer both questions: Whether lies are shared more often, even when receivers are incentivized to share true or believable videos, and how knowing a video was shared affects receivers' beliefs. Our results show that despite their incentives, Receiver 1s were more likely to share lies than truthful videos (between 58% and 62% of the videos shared were lies). Sharing significantly increased the chance that a video was believed when Receiver 1s were incentivized to share a truthful video. This result can help us understand how fake news spread even without motivated beliefs. By contrast, if videos were shared with the intention of finding content that persuades others, beliefs did not exhibit significant changes in response to sharing. This finding suggests that perceived intentions behind sharing could affect how lies spread, and, perhaps strikingly, what may fuel the spread of lies could be the intention to share truthful videos. The combined effect of sharing was that it increased the likelihood that a lie was believed (between 33% and 45% of the time, compared to 26% to 29% when it was not shared), even though a shared video should be believed *less* often given the fact that it is more likely to be false.

# 2 Experimental Design

We start by describing the common features of our experiments, and then explain the differences between them.

## 2.1 Senders

We recruited individuals to participate as senders. Each sender generated two 30-second videos: one "false" and one "true." We chose to use videos to allow receivers to use language characteristics and facial expressions to detect the truthfulness of a video. Upon arriving to the laboratory, the experimenter gave senders the instructions (see Online Appendix A for all instructions used). We told senders they would either see a headline and picture from the front cover of the NYT or a blank slide. When senders observed a headline and picture from the NYT, we asked them to describe it. When senders observed a blank slide, they could make up a news event or report that the screen was blank. We incentivized senders to convince receivers they were seeing a news event in both cases: for each of the two videos they recorded, senders received $10 if they convinced a receiver that they were seeing a true news event, and $0 if they did not. Senders were informed about the task of receivers. A single receiver was recruited afterwards specifically to determine the senders' payments.

Senders were given time to read the instructions and ask questions. When they stated that they were ready to start, they were shown the first slide and given 30 seconds to describe it. At the end of the 30 seconds, senders saw the second slide and were given another 30 seconds to describe it. News and blank slides were presented in random order.

For Experiments 1 and 3, we recruited 10 senders, five women and five men, all of whom were research assistants or graduate students at UC San Diego. Because each sender was recorded in two videos, a total of 20 videos were recorded. For Experiment 2, we recruited an additional 42 new senders—20 male and 22 female undergraduate students at UC San Diego—leading to a total of 84 videos. This larger sample of senders allowed us to use facial-expression-recognition software and speech characteristics to explore the predictors of lying and telling the truth, and compare them with receivers' beliefs. We followed the same procedure and used the same instructions as with the first group of 10 senders, with the only difference being that we explicitly instructed senders that when they saw a blank screen, if they chose to make up a news event, it should not be a real event and they should not describe a news event that they recalled from a different date or source. A research assistant checked the video transcripts, and an overwhelming majority of participants described made-up news events. To avoid bias, we chose not to exclude any videos based on this. All transcripts are provided in Online Appendix A.[3] We refer to videos in which the sender saw and described a NYT headline and picture as "truthful" videos, and videos in which the sender saw a blank screen and described a made-up news event as lies.

[3] Precisely, an RA classified the 42 videos in which the sender saw a blank screen in Experiment 2. The RA classified 41 out of 42 into 3 categories: (a) made-up story (21 videos), (b) made-up story that was similar to news stories (17 videos), and (c) true story (3 videos). For one video the RA was unsure. According to this classification, a majority of receivers made up news stories when seeing a blank screen, some of which had elements related to familiar news stories (category (b)). Only a small minority did not follow the instructions and told a true story.

## 2.2 News Events

We systematically chose pictures and headlines from the front page of the international edition of the NYT. The actual headlines are reproduced in Online Appendix A. Most news events reported on the front cover deal with international conflicts (e.g., the meeting of South and North Korean leaders), international catastrophes (e.g., an airplane crash), and news from the arts (e.g., a new movie by a renowned actor).

To avoid experimenter selection bias in sampling news events, we decided (before we saw the headlines) to select front-page headlines and corresponding photos from the NYT dating exactly six months prior to when the first videos were scheduled to be recorded. We used all events that were presented with a photo in the main section of the front page of the NYT, without exclusions.

## 2.3 Receivers

Table 1 presents an overview of the experiments we conducted. The receivers in our experiments were participants on Amazon Mechanical Turk (AMT). In Experiment 1, each receiver watched the same 20 videos and was asked to decide after watching each video whether the sender was telling the truth or lying. Receivers were informed that one of the 20 videos they evaluated would be randomly selected, and they would be paid $5 if they guessed correctly whether the sender in the selected video was telling the truth or lying. Receivers had to stay on the video question for at least the duration of the video (30 seconds) but could remain longer if they wished to. We included a question at the beginning of the experiment that required sound recognition to ensure receivers could hear the videos in the experiment. Participants could not continue with the experiment if they failed this test. We also had two control questions that participants answered before starting the experiment. As pre-registered, we excluded subjects who failed to answer either one of the control questions correctly.

Receivers watched the 20 videos in blocks of five, presented in random order.

After each block of five videos, we asked receivers to state their confidence in their absolute ability to detect lies: “How many of the 5 video guesses you just made do you believe are correct?” Receivers earned $1 each time their belief was correct. After watching all 20 videos, we elicited receivers’ relative confidence in their ability, using the question, “Compared with previous participants in this experiment, how well do you think you did?” They were asked to choose a quartile, and earned $1 if their answer was correct.

To test whether receivers knew the answer to some questions more accurately than others, we also elicited “safe bets” with the question “For 3 videos you saw, you can earn an additional bonus of $0.25 if your guess is correct. Which 3 videos would you choose to receive an additional $0.25 bonus if your guess was correct?”[4]

Experiment 1 had two treatments, which varied whether receivers knew the fraction of lies and truthful videos out of the 20. In the No Prior treatment, receivers did not know how many videos out of the 20 were true. To test whether knowing this information (the “prior”) would help receivers detect lies, we conducted the Prior treatment, in which the only difference was that receivers were informed about the 50-50 prior in advance.

In Experiment 2, we extended the set of videos that receivers could watch to a new group of 84 videos. Each receiver watched eight videos. One video was randomly chosen for payment. Because this experiment was shorter in duration, the receiver was paid $1 if he or she correctly assessed the truthfulness of the selected video. We also elicited the receiver’s absolute confidence twice–first after watching the first four videos and then after watching the second four videos. The receiver earned $0.25 each time her/his belief about the number of correct assessments was correct. We did not elicit receivers’ relative confidence in this experiment, because the potential score varied only from one to eight correct video assessments, and due to the number of ties, we would have observed little to no difference in ability across

[4] We also elicited sharing decisions through the use of a coordination game, by which receivers were incentivized to choose a video to share that was also chosen by other receivers. The design and results of the coordination game are presented in Online Appendix B.

quartiles. We also asked the same “safe bet” question as in Experiment 1, but asked the receivers to select only one video for it.

In Experiment 3, we examined whether videos containing lies are shared more often and their effects on beliefs. The first two treatments of Experiment 3 had the same structure as Experiment 2. First, receivers in the role of Receiver 1 watched eight videos, and one video was randomly chosen for payment. Absolute confidence was elicited twice, and a “sure” bet was elicited in the end, as in Experiment 2. For this experiment, we used 16 out of the original 20 videos used in Experiment 1, making two sets of eight videos. Each set contained eight different senders and was balanced on gender and the share of truthful videos. That is, each set contained two women and two men in a truthful video and the same numbers for false ones. The other set of eight videos included the other videos of the same senders. The goal of this balancing was to prevent effects due to some people being more believable than others (e.g., due to their attractiveness, gender, age, etc.). Our design counterbalances these effects (which is why we included 16 and not the entire 20 videos).

After Receiver 1s watched and assessed the videos, they were presented with the titles and screenshots of each of the eight videos in random order (details on the titles and screenshots are provided in Online Appendix A). Receiver 1s were asked to pick one video to share with another participant. Motivated by the fact that on social media people often choose which videos to watch or stories to read more about based on what others share with them, Receiver 1s were informed that the other participant (Receiver 2) would see the same titles and screenshots, and would be asked to choose four out of the eight videos to watch and assess whether the sender was truthful or lying in these four videos. Videos are shared for many reasons. For example, one may choose to share a video that shows an absurd behavior because she/he finds it funny and entertaining. To control for the motivation of sharing a video by Receiver 1s, we varied the incentives regarding which kinds of videos to share across two treatments. In the Shared-True treatment, Receiver 1s were asked to pick a video they found "interesting and true," and were paid an additional $0.50 if the other participant chose to watch the video they shared and the video was true. In the Shared-Believable treatment, the incentives were the same, except that instead of being true, the video had to be believed by the other participant for Receiver 1 to earn the additional $0.50 payment.

After Receiver 1s made their sharing decision, they were asked about their belief regarding whether Receiver 2 would assess the video they shared as true. They were asked to pick one of 10 intervals, from 0-10, to 91-100, and were paid an additional $0.25 if the interval they selected was correct. This measure allows us to assess whether Receiver 1 thought Receiver 2 would believe her/his shared video, and whether she/he anticipated differential effects depending on the treatment.

Table 1: Overview of Experiments

| Experiment | Treatments | Videos seen per receiver | Number of receivers |
|---|---|---|---|
| 1 | No Prior | 20 | 380 |
|  | Prior | 20 | 192 |
| 2 | No Prior-8 | 8 | 1056 |
| 3 | Shared-True | 8 if R1, 4 if R2 | 384 R1 and 384 R2 |
|  | Shared-Believed | 8 if R1, 4 if R2 | 371 R1 and 371 R2 |
|  | Shared-No-Incentive | 4 | 198 |
|  | No-Sharing-Information | 4 | 185 |

*Notes:* In Experiments 1 and 3, videos were drawn from the set of 20 videos initially collected. For Experiment 3, two groups of eight videos were created, with an equal number of male and female senders and an equal number of lies and truthful videos. In Experiment 2, videos were drawn from the set of 84 additional videos.

In the Shared-True and Shared-Believed treatments, we matched each Receiver 1 with a Receiver 2. Each Receiver 2 saw the same screenshots and titles as Receiver 1 for eight videos at the beginning of the experiment. Receiver 2 was also informed about which video Receiver 1 shared. In both treatments, Receiver 2 was informed about the incentives of Receiver 1. Receiver 2 chose four videos to watch, and then assessed the truthfulness of each video. When watching the shared video, she/he was reminded of the fact that the video was shared. At the end of the experiment, Receiver 2 also reported her/his absolute confidence and received $0.25 for providing a correct answer.

We ran two additional treatments in Experiment 3. First, we included a control treatment, the No-Sharing-Information treatment, in which Receiver 2 saw no sharing decisions by any other participant and was only asked to choose four out of eight videos at the beginning of the experiment. Second, we added the Shared-No-Incentive treatment, in which we removed the link between Receiver 2's decision of which videos to watch and Receiver 1's payment. This treatment mimics the Shared-True treatment in that Receiver 2 was informed that another participant shared a video she/he thought was interesting and true. But Receiver 2's decision had no effect on Receiver 1, and no further information on Receiver 1's incentives was thus provided. This treatment allows us to test whether the effects of sharing in Shared-True are similar, even if Receiver 2's decisions had no impact on anyone else.

The incentive for Receiver 1 in Shared-True (and Shared-Believed) was the same as the participant's fixed fee for completing in the experiment ($0.50). While not large in absolute terms, it could double Receiver 1s' total payment, and Receiver 1 shared videos they believed to be true in the Shared-True treatment (in over 90% of the cases). However, if results in the Shared-No-Incentive treatment are similar to those in Shared-True, one reason could be the limited size of incentives.

## 2.4 Procedures

We pre-registered our analyses on Aspredicted.org for all three experiments.[5] We conducted Experiment 1 in two waves (pre-registrations #16666 and #18131). We initially recruited 300 receivers, 287 of whom correctly answered two control questions about the instructions. As pre-registered, we excluded the 13 receivers who failed the control questions. We then added the Prior treatment and included a

[5] Replication data are available in Serra-Garcia and Gneezy (2021).

small replication of the first wave ($N = 93$) to confirm that results remained similar (as shown in Online Appendix B). We thus pool all observations in Experiment 1 ($N = 380$ in the No Prior treatment, and 192 in the Prior treatment). In Experiment 1, receivers were paid a $2 fixed fee for participation and earned an additional $3.75 bonus on average.

In Experiment 2, we aimed to recruit 1,100 receivers, such that each video would be viewed 100 times. Excluding receivers who failed the control question provided us with 1,056 receivers, as pre-registered (pre-registration #19319). Receivers were paid a $0.50 fixed fee for participation and earned an additional $0.84 bonus on average.

In Experiment 3, we aimed to recruit 400 subjects in the role of Receiver 1 and 400 in the role of Receiver 2, in the Shared-True and the Shared-Believed treatments. In the Shared-No-Incentive and No-Sharing-Information treatments, we aimed to collect 200 subjects.[6] In total, 1893 participants answered the control questions correctly. In Experiment 3, receivers were paid a $0.50 fixed fee for participation and earned an additional $0.61 bonus on average.

To be eligible to participate in the experiments, participants had to have a United States IP address and previously completed at least 100 tasks on AMT with a 95% approval rating. We used CloudResearch (Litman et al., 2016) for recruitment, which additionally blocks duplicate IP addresses, suspicious geocode locations, and verifies country location. Across the three experiments, 50% of participants were female and were 39 years old on average, with ages ranging from 18 to 84 (25% of participants were 61 and older). Among them, 7% reported reading the NYT on a daily basis. Receivers in Experiment 1 were somewhat less likely to be female than receivers in Experiments 2 and 3 (45% vs. 50% and 52%, respectively) and were slightly younger (37.0 vs. 38.8 and 40.5 years old, respectively). Receivers in Experiment 3 were more likely to report reading the NYT on a daily basis (9% vs.

[6] All treatments were conducted simultaneously, for each receiver role (R1s first, and then the matched R2s) and pre-registered in #41933, except for the Shared-No-Incentive, which was pre-registered and conducted less than a week later (#42342).

5% and 4%, respectively). We controlled for experiment fixed effects throughout. Further details on the sample are presented in Online Appendix B, and the preregistration materials are in Online Appendix C.[7]

## 2.5 Hypotheses

All the hypotheses are based on our preregistration. Our first hypothesis concerns the ability of receivers to correctly detect whether a video is true or false. As we mentioned in the Introduction, the literature in psychology has mostly found a limited ability in such detection tasks (e.g., Bond and DePaulo, 2006; Bond and DePaulo, 2008). The few experiments in the economics literature, in which receivers are incentivized to detect fake materials correctly, have found more mixed results: Belot and van de Ven (2017) found a better-than-chance ability, whereas receivers in Konrad, Lohse, and Qari (2014) and Dwenger and Lohse (2019) were slightly worse than chance at such a task.

**Hypothesis 1–Ability to classify videos:** Receivers are not better than chance at classifying videos containing lies and true statements.

Our second hypothesis concerns confidence. In general, people are frequently overconfident in their ability, although sometimes they are underconfident (see Moore and Healy, 2008; Benoit, Dubra, and Moore, 2015). We hypothesized that men are more overconfident than women. Men are typically more overconfident in knowledge-based tasks (e.g., Mondak and Anderson, 2004; Coffman, 2014; Bordalo et al., 2019), though women may be better at detecting others' emotions, which could be important for detecting lies.

**Hypothesis 2–Overconfidence:**

**Hypothesis 2a:** Receivers are overconfident in their absolute ability to detect lies.

**Hypothesis 2b:** Receivers are overconfident in their relative ability to detect lies.

[7] We conducted a pilot with undergraduate students at UC San Diego (N=100). The ability to detect lies of students in the laboratory does not differ significantly from that on AMT.

**Hypothesis 2c:** Men are more overconfident than women.

In addition to measuring ability and confidence in assessing lies and truthful videos, we pre-registered that we would explore whether receivers' assessments are related to their gender, age, and readership of the NYT, as well as the characteristics of senders' speech, facial expressions, and movements. We collected these data for exploratory analyses, and we did not have hypotheses regarding them.

Next, we turn to sharing behavior. We hypothesized that lies are more interesting than truthful videos. In general, making up a story allows senders to make it more interesting than when they need to stick to the facts. Berger and Milkman (2011) found that NYT articles that go viral tend to generate higher arousal among readers. When asked to pick the most interesting video out of 8, independent raters (blind to the truthfulness of videos) indeed picked lies 61% of the time.[8] If Receiver 1 also evaluates videos containing lies as more interesting, they will be more likely to be shared.

**Hypothesis 3–Sharing rates of lies:** Receivers are more likely to share videos that contain a lie than ones that contain a true statement, both when they are incentivized to share a truthful video and when they are incentivized to share a video that they think another receiver will believe.

The final hypothesis concerns the effects of sharing. Receivers 2s were informed that Receiver 1s were encouraged to choose an interesting video to share. Because we expect Receiver 2s to prefer to watch interesting videos, we expect them to be more likely to choose to watch a video when they know Receiver 1s shared it. Our prediction was that, both when Receiver 1s were incentivized to share a video that was true and a video that would be believed, knowing that a video was shared by Receiver 1s would increase its believability for Receiver 2s.

[8] A sample of 200 independent raters was asked to watch 8 videos, like Receiver 1, and pick the most interesting one at the end. To incentivize their decisions, they received a $0.50 bonus payment if they chose the video that is considered as the most interesting by other participants. Receiver 1s were uninformed about the truthfulness of the videos.

**Hypothesis 4–Effect of sharing:** Receivers are more likely to watch a video and believe it when they learn another receiver chose to share it with them.

# 3 Results

## 3.1 Ability to Detect Lies

In Experiment 1, receivers correctly evaluated the truthfulness of videos in 50.4% of cases in the No-Prior treatment, and in 51.2% of the cases in the Prior treatment. These frequencies are not significantly different from chance ($t-$test, $p = 0.4956$ and 0.1555, respectively; all tests in the paper are two-tailed), and the detection rates between treatments did not differ significantly ($t-$test, $p = 0.3730$). In Experiment 2, receivers correctly detected the veracity of 53.2% of the videos. This rate is significantly higher than chance ($t-$test, $p < 0.001$), but only 3.2 percentage points better. It is significantly better than in Experiment 1 ($t-$test, $p = 0.0039$).[9,10]

**Result 1: Ability to classify videos**
In line with Hypothesis 1, receivers demonstrated almost no ability to correctly classify lies and truthful videos.

Receivers could make two types of errors: type I was failing to detect when a sender was lying, and type II was failing to detect when a sender was telling the truth. In the No-Prior treatment, receivers made type I errors 46.7% of the time and type II errors 52.6% of the time. Overall, they believed videos to be true 47% of the time, reflecting that receivers were slightly skeptical and believed videos less than 50% of the time ($t-$test, $p < 0.001$). In the Prior treatment, however, this skepticism disappeared: they believed 50.1% of videos were true, a frequency that

[9] Separating by treatment, ability in Experiment 2 is higher than in the No-Prior in Experiment 1 ($t-$test, $p = 0.007$), but not significantly better than in the Prior treatment ($t-$test, $p = 0.1713$) treatment. Detailed results on the distribution of ability are shown in Online Appendix B.

[10] We also find a very limited "wisdom of the crowd" effect (e.g., Surowiecki, 2005; Lee and Lee, 2017), as detailed in Online Appendix B. At the video level, the difference between the share of receivers who believed a truthful video and the share who believed a lie was less than 7 percentage points on average.

is not significantly different from 50% ($t-$test, $p = 0.8996$).

In Experiment 2, with a different set of videos, receivers believed lies in 48.2% of the cases, and did not believe truthful videos in 45.4% of the cases. In this experiment, receivers were slightly gullible: they guessed that 51.4% of the videos were true, which was more than the true rate of 50% ($t-$test, $p = 0.0093$).

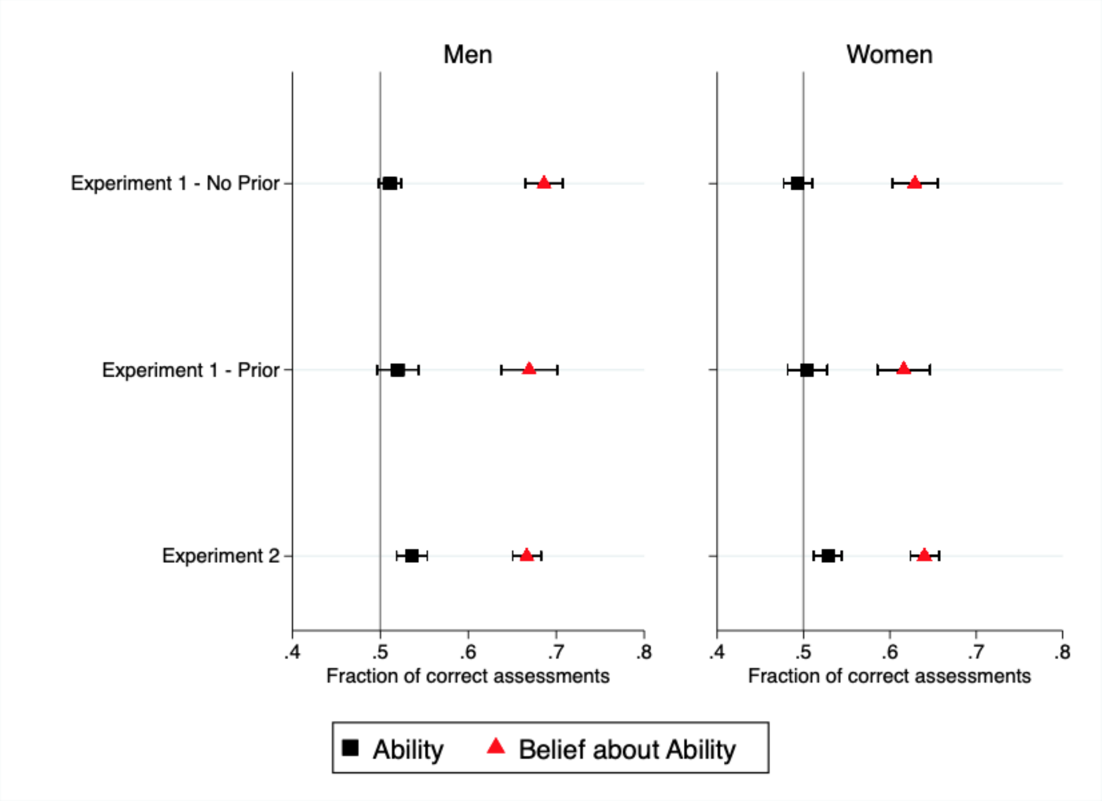
When we asked receivers to choose three videos for which they would obtain a bonus payment if they guessed correctly in Experiment 1 (and to choose one video in Experiment 2), their guesses were correct in 48.1% of the cases in the No-Prior treatment, 48.8% of the cases in the Prior treatment, and 51.2% in Experiment 2. These guessing rates were not better than chance ($t-$test, $p > 0.1$ in all cases). Hence, on average, their assessments were not better than chance, even in videos that they chose themselves. This finding opens up the question of whether their beliefs about their ability to detect lies were accurate, which we study by measuring overconfidence in absolute and relative terms.

## 3.2 Overconfidence in the Ability to Detect Lies

*Absolute overconfidence:* We start with receivers' confidence regarding their ability to correctly assess the truthfulness of videos. Although the ability of men was not statistically different from that of women (52.8% vs. 51.8%), men were more confident than women ($p < 0.001$). Considering Experiments 1 and 2 together, men ($N = 837$) believed they correctly assessed 67.2% of all videos, whereas women ($N = 791$) believed they correctly assessed 63.5% of all videos, as shown in Figure 1, separately by treatment and experiment. Comparing ability to beliefs at the individual level reveals that over 65.4% of men and 61.2% of women believe they performed better than they actually did.

*Relative overconfidence*: In Experiment 1, a majority of receivers (61.9%) believed their ability was in the second quartile of the distribution. Less than 1.6% believed their ability was in the bottom quartile, and 17.0% believed it was in the third quartile. Comparing receivers' beliefs to their actual placements, we find that only

Figure 1: Absolute Overconfidence, by Experiment and Gender

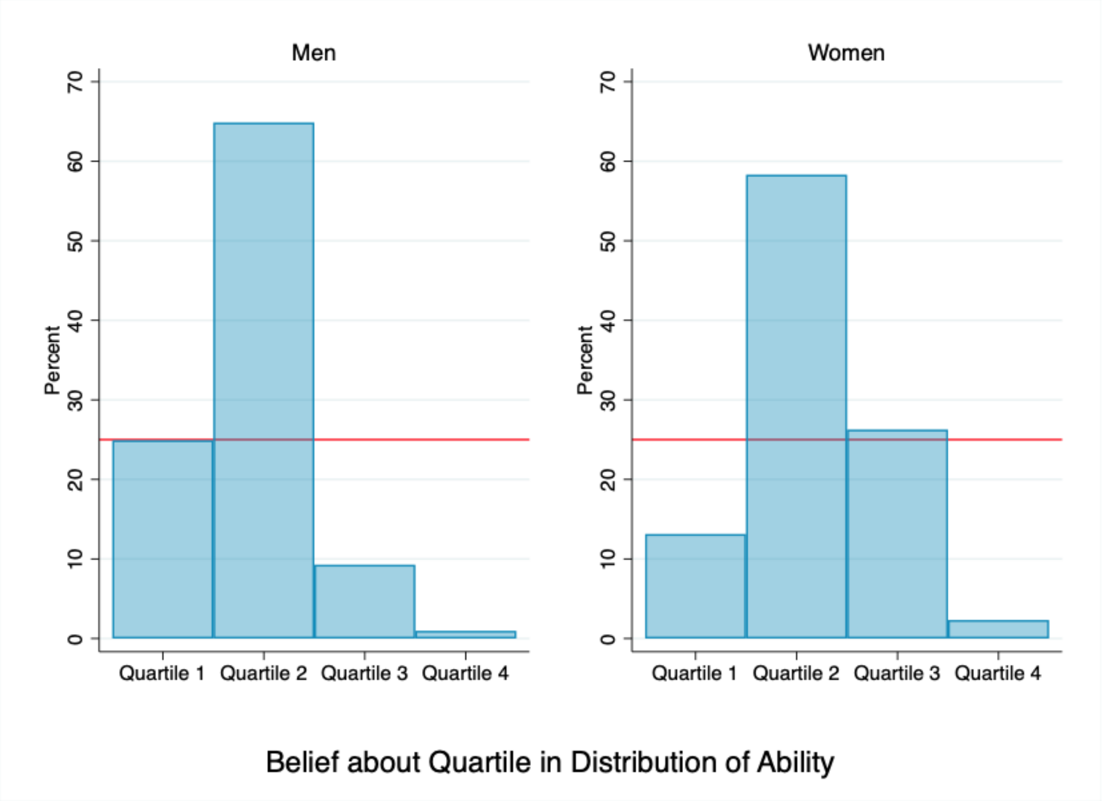


*Notes:* This figure presents the average share of correct assessments (ability) and average confidence (belief about ability) for men and women in Experiments 1 and 2. Error bars indicate 95% confidence intervals.

13.4% of receivers in the third quartile of ability place themselves in the third quartile, while 84.8% of these receivers place themselves in the first and second quartile of the distribution. Similarly, only 2.3% of receivers in the bottom quartile of ability placed themselves in the correct quartile. These findings violate the condition that the largest group of subjects placing themselves in a given quartile must belong to that quartile, consistent with overconfidence (Benoit and Dubra, 2011; Burks et al., 2013; Benoit et al., 2015). Men placed themselves at higher quartiles than women ($\chi^2$-test, $p < 0.001$), as shown in Figure 2.

Overall, actual ability was not related to beliefs about ability (detailed results are provided in Online Appendix B), suggesting that receivers were systematically overconfident, independent of their performance.

Figure 2: Relative Overconfidence in Experiment 1, by Gender



**Result 2: Overconfidence**

In line with Hypothesis 2, receivers were overconfident in their absolute and in their relative ability to correctly classify lies and truthful videos. Men were more overconfident than women.

## 3.3 Beliefs and Truth: The Role of Speech, Emotions, and Facial Expressions

The disconnect between receivers' actual ability and their beliefs about ability is large and concerning. One possible driver of this disconnect is that receivers base their beliefs on cues about the sender and sender's behavior that are wrong. Prior research suggests that individuals may use different speech, facial expressions, and movements when lying (e.g., Ekman, 1970), in bargaining environments (e.g., van Leeuwen et al., 2018; Chen et al., 2019) and when signaling their trustworthiness (e.g., Centorrino et al., 2015 a and b). We test whether certain facial expressions or speech features are associated with lying in our setting, and whether receivers' beliefs responded to such features in a similar manner.

To measure emotions and facial movement, we ran the videos through FaceReader, a facial-expression-recognition software (Bijlstra and Dotsch, 2011).[11] The emotions measured by the software are happy, sad, angry, surprised, scared, disgusted, and neutral. The software analyzed 10 frames per second of each video and measured the intensity of each emotion (from 0 to 1). We computed the average intensity of each emotion in a video and included each emotion separately (excluding neutral) as a predictor. The software also provided a measure of valence, which measures whether the emotional status of the sender was positive or negative, and arousal, which indicates the level of the sender's facial activity. In additional robustness checks shown in Online Appendix B, we include only these summary measures (valence and arousal) and find qualitatively similar results.

The software also provided several measures of head and facial movement. It measured head orientation along the x, y, and z axes, in degrees deviating from looking straightforward. It also measured whether the participant gazed forward, left, or right. We computed the average degrees of head orientation along each axis throughout the video, as well as how often the software detected the sender looking left or right, versus forward. Furthermore, the software measured how often the sender opened her/his mouth, and provided a measure of image quality of each video (from 0 to 1). Other measures of eye and eyebrow movement collected by the software can also be included in the model, after which results remain qualitatively similar. The software returned measures of emotions for 102 out of 104 videos.[12]

To measure speech, we collected the number of words used and the Google sen-

[11] We used the main version of the software. An additional module is available to analyze action units within the face, and to conduct, among others, analyses of the role of different types of smiling (see, e.g., Ekman and Friesen, 1982; Centorrino et al. 2015b).

[12] The complete face of one out of 51 senders could not be captured by the software, because the sender lowered it. For three videos, the video coded some facial movements as unknown. We included these videos in the analysis, though conclusions remain similar if these videos are excluded.

timent score of each video. A research assistant manually transcribed the videos, and transcriptions excluded filler words. Sentiment analysis measures the valence of the language used in the videos. It has been used, for example, to examine the language of discrimination (Bohren, Imas, and Rosenberg, 2018). The sentiment score we used in the analysis was obtained from Google's natural language API,[13] and ranges from -1 (clearly negative) to 1 (clearly positive). It allows us to measure whether the language used in truthful videos conveyed a more negative or more positive sentiment than that in videos containing a lie. We standardized all continuous measures. We then examined the effect of a one-standard-deviation increase in each emotion, facial movement, or speech characteristic on the likelihood that a video was actually true, and the likelihood that a video was believed. Descriptive statistics for all variables are provided in Online Appendix B.

Table 2 shows the results from probit regressions on the probability that a video is believed (columns (1)-(3)) and that a video is actually true (columns (4)-(6)). The coefficients displayed are marginal effects of each characteristic of the senders' behavior (speech, emotions, facial expressions).[14] Because we tested the effects of each measure on two outcome variables, we also show adjusted $p-$values for multiple hypotheses testing (Romano and Wolf, 2005) in columns (3) and (6).

To compare the effects of each sender characteristic on receiver beliefs and actual truthfulness, Figure 6 displays the marginal effects of each characteristic on the probability that a video is believed (y-axis), from column (1) of Table 3, and the probability that the video is actually true (x-axis), from column (4) of Table 3. If receivers put the right weight on each cue offered by the sender (speech characteristics, emotions, etc.), the coefficients in Figure 6 should be close to a 45–degree line. Instead, Figure 6 shows that the relationship between senders' cues of truthfulness and their impact on beliefs was negative. For example, receivers were significantly less likely to believe videos in which senders said fewer words or were

[13] https://cloud.google.com/natural-language/, accessed April 8 and 9, 2019.

[14] In Online Appendix B, we also include whether the video is true and receiver characteristics in regression models for beliefs about truthfulness. Results remain qualitatively the same.

Table 2: Determinants of Beliefs and Truth

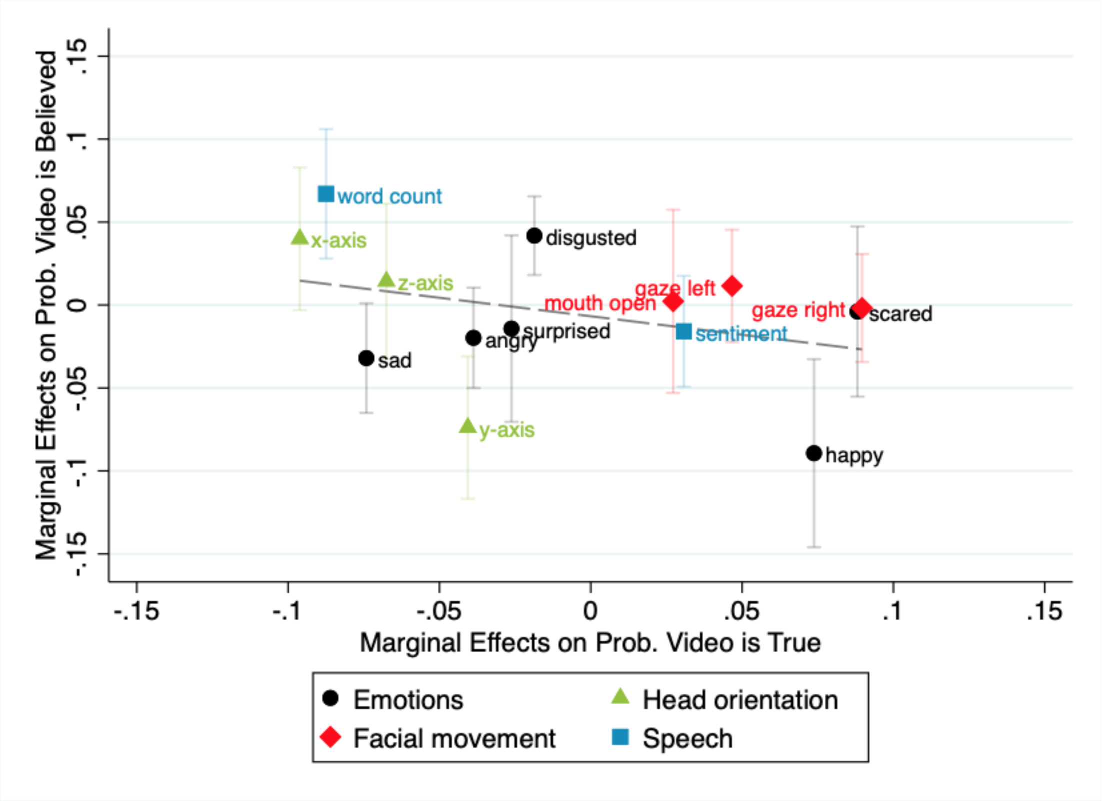
| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Video is believed (=1) | | | Video is true (=1) | | |
| | | | FWER | | | FWER |
| | Coeff | Std.Error | *p*-val. | Coeff | Std.Error | *p*-val. |
| **Sender Gender & Speech** | | | | | | |
| Female sender | 0.150*** | (0.048) | [0.0074] | -0.028 | (0.097) | [0.7342] |
| Word count | 0.067*** | (0.020) | [0.0279] | -0.087** | (0.041) | [0.0420] |
| Sentiment score | -0.016 | (0.017) | [0.5570] | 0.031 | (0.054) | [0.5570] |
| **Emotions & Facial Movement** | | | | | | |
| Happy | -0.089*** | (0.029) | [0.0282] | 0.074 | (0.053) | [0.1183] |
| Sad | -0.032* | (0.017) | [0.0957] | -0.074 | (0.050) | [0.0957] |
| Surprised | -0.014 | (0.029) | [0.8613] | -0.026 | (0.053) | [0.8613] |
| Scared | -0.004 | (0.026) | [0.9038] | 0.088** | (0.041) | [0.1178] |
| Angry | -0.020 | (0.015) | [0.3512] | -0.039 | (0.059) | [0.4303] |
| Disgusted | 0.042*** | (0.012) | [0.0083] | -0.019 | (0.058) | [0.7633] |
| y-axis head orientation | 0.040* | (0.022) | [0.2219] | -0.096** | (0.043) | [0.1499] |
| x-axis head orientation | -0.074*** | (0.022) | [0.0350] | -0.041 | (0.057) | [0.4163] |
| z-axis head orientation | 0.014 | (0.024) | [0.6272] | -0.068** | (0.031) | [0.1201] |
| Eyes gaze left | -0.002 | (0.017) | [0.9139] | 0.090** | (0.043) | [0.0576] |
| Eyes gaze right | 0.012 | (0.017) | [0.5788] | 0.047 | (0.055) | [0.5788] |
| Mouth open | 0.002 | (0.028) | [0.9470] | 0.027 | (0.054) | [0.6923] |
| Image quality | -0.055** | (0.024) | [0.0995] | -0.083* | (0.050) | [0.1369] |
| Observations | | 19,692 | | | 102 | |

*Notes:* This table presents marginal effects of probit regression models on the likelihood that the receiver believes a video (columns 1-3) and that sender tells the truth (columns 4-6) at the means of the covariates. The regression models in columns 1-3 include experiment fixed effects, and robust standard errors for these models, clustered at the sender level, are presented in parentheses. ***, **, and * indicate 1%, 5%, and 10% significance levels respectively for unadjusted p-values, and columns (3) and (6) show FWER p-values adjusted for multiple hypotheses testing (Romano and Wolf, 2005).

happier. But such videos are weakly more likely to be true. A similar pattern is observed for head orientation (shown in green). The senders' eye gaze did not affect beliefs (all effects are close to zero), but when senders gazed more toward the left, they were more likely to be telling the truth.

Receivers also believed female senders more than male ones (see also, Lohse and Qari, 2019). Interestingly, this tendency is in line with findings in the deception literature showing that men are more likely to tell lies, especially when lies help them but may hurt their counterpart (see, e.g., Dreber and Johannesson, 2008; Erat and

Figure 3: Marginal Effects of Emotions, Speech and Facial Expressions on the Probability that a Video Is Believed and the Probability that a Video Is True



*Notes:* This figure presents the marginal effects of sender emotions, speech, and facial expressions on the probability that a video is believed (from column (1) of Table 3) and the probability that the video is believed (from column (4) of Table 3). Error bars around each marginal effect indicate 95% confidence intervals for the marginal effects on the probability that a video is believed.

Gneezy, 2012; Abeler, Nosenzo, and Raymond, 2019), though this gender difference is not always found (e.g., Abeler, Becker, and Falk, 2014). In our study, however, by design, men and women told the truth equally often. We also explored the interaction between the sender's and the receiver's gender and find that, compared to male receivers, female receivers were significantly less likely to believe female senders, but were more likely to believe a video is true overall (detailed results in Online Appendix B).

Overall, we find that receivers' beliefs are explained by a number of emotional expressions, facial movements, and speech characteristics of senders, but the effects are often in the wrong direction, concentrated on indicators that do not predict a sender's truthfulness. Suppose, by contrast, that receivers would have focused on the correct indicators in the right direction. How good would they have been at detecting lies? We predict the likelihood that a video is true based on the emotional expressions, facial movements, and speech characteristics of senders. If receivers would have rated videos as true whenever they believed that the video had a likelihood higher than 0.5 of being true, their detection rate would have been 60%. This finding suggests the videos contained information that receivers could use to detect lies. Future research could test the validity of these indicators out of sample and explore whether receivers can be informed about these mistakes, reducing the overreliance on the wrong predictors or in the wrong direction.

# 4 Sharing

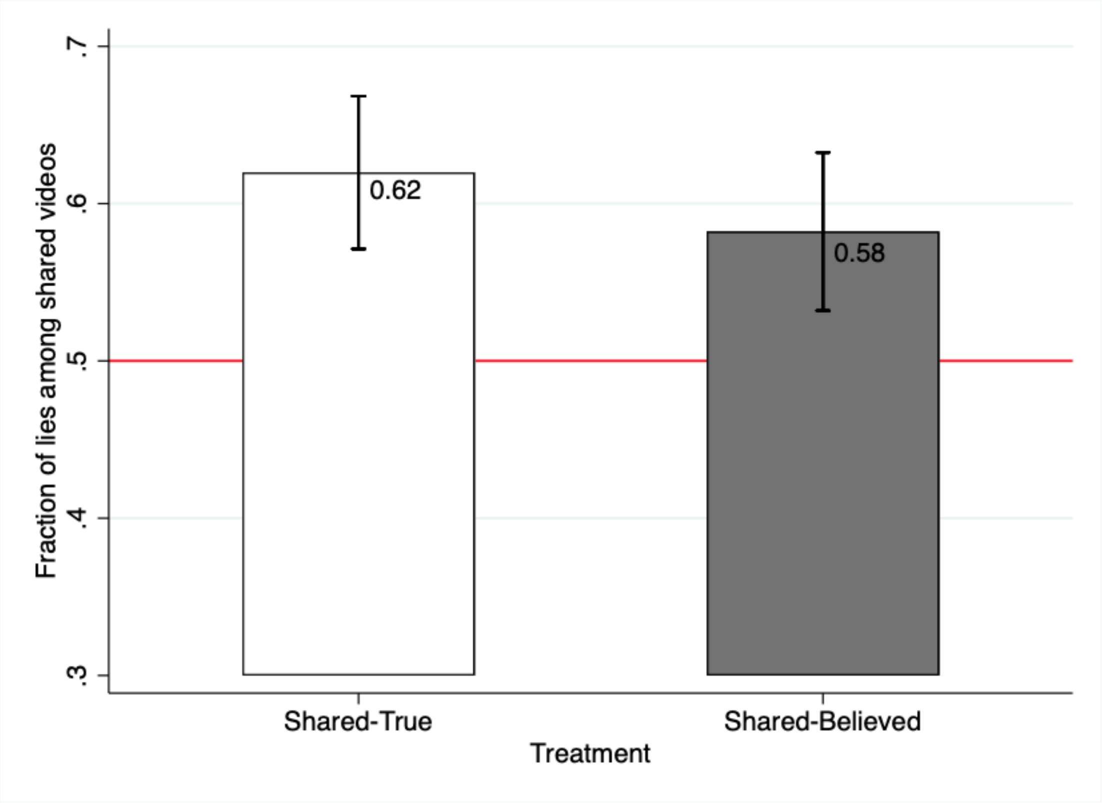
Taken together, the findings thus far document that receivers exhibit a limited ability to detect lies, are overconfident about their ability, and often put the wrong weight on predictor variables. These results set the stage to study sharing behavior and the effects of sharing. Given the limited ability to detect lies, a central concern is that receivers may share videos while believing they are true when they are not, and that others may fall prey to the same mistakes when they see shared videos.

## 4.1 Lies Are Shared More Often

In Experiment 3, receivers in the role of Receiver 1 watched eight videos, assessed whether they were true or false, and provided their estimated accuracy. Their ability was not better than chance (50.1%, $t-$test $p = 0.9315$), and they were significantly overconfident (66.7% believed ability, $t-$test $p < 0.001$). These results confirm the above findings (detailed results are provided in Online Appendix B).

Thereafter, Receiver 1s were asked to share with Receiver 2s one of the eight videos they watched. As Figure 4 shows, in the Shared-True treatment when Receiver 1s were incentivized to share a truthful video, 62.0% of them actually shared a lie, a rate that is significantly higher than 50%, which is the share of lies ($t-$test, $p < 0.001$). In the Shared-Believed treatment, when Receiver 1s were incentivized to share a video that would be believed, they shared lies in 58.2% of the cases, again a rate that is significantly higher than 50% ($t-$test, $p = 0.0015$) but not statistically different than that in Shared-True ($t-$test, $p = 0.2923$). Receiver 1s shared videos they believed to be true in most cases (in 90.4% of the cases in the Shared-True treatment, and 86.8% of the cases in the Shared-Believed treatment).

Figure 4: Sharing Rates of Lies



*Notes:* This figure presents the average share of lies that were shared by Receiver 1 in Experiment 3, by treatment. Error bars indicate 95% confidence intervals.

Table 3 explores which video is chosen out of the 8. If chosen randomly, each video has a chance of 12.5% of being selected. Confirming the finding in Figure 4, lies were 5 percentage points more likely to be shared (15%) than truthful videos (10%). Receiver 1s were also significantly more likely to share a video they believed to be true. This relationship is directionally stronger in the Shared-True treatment, consistent with how Receiver 1s somewhat adjusted their sharing decision based on whether Receiver 2s were supposed to believe the shared video. In column (3) of Table 3, we also explored whether Receiver 1s anticipated Receiver 2s' choices. Receiver 2s would know the gender of the sender when choosing which video to watch, and Receiver 1s could potentially anticipate a tendency towards picking male or female senders. However, we do not find evidence supporting that the senders' gender guided Receiver 1s' sharing decision, suggesting that Receiver 1s mostly focused on sharing videos that they believed to be true. In additional analyses, shown in Online Appendix B, we examined the relationship between Receiver 1s' confidence and sharing lies. We find that overconfident Receiver 1s were more likely to share lies.

Table 3: Determinants of Sharing a Lie

| | (1) | (2) | (3) |
|---|---|---|---|
| | Video is Shared= 1 | | |
| Lie | 0.051*** | 0.045*** | 0.039*** |
| | (0.008) | (0.007) | (0.010) |
| Video was believed | | 0.183*** | 0.168*** |
| | | (0.008) | (0.011) |
| Shared-True Treatment | -0.000 | -0.000 | -0.030 |
| | (0.008) | (0.007) | (0.018) |
| Shared-True Treatment X Lie | | | 0.012 |
| | | | (0.015) |
| Shared-True Treatment X Video was believed | | | 0.029* |
| | | | (0.017) |
| Female sender | | | -0.000 |
| | | | (0.007) |
| Fraction of videos shared (1 of 8) | 0.125 | 0.125 | 0.125 |
| Observations | 6,040 | 6,040 | 6,040 |

*Notes:* This table reports marginal effects, at the mean of covariates, for the likelihood that 1 video out of 8 is chosen to be shared by Receiver 1. Lie is an indicator that takes value 1 if the video is a lie, and 0 otherwise. Video was believed is an indicator that takes value 1 if the receiver guessed that the video was true. Shared-True treatment is an indicator that takes value 1 in that treatment. Female sender is an indicator that takes value 1 if the sender was female (4 out of 8 were) Robust standard errors are reported in parentheses. ***, **, and * indicate 1%, 5%, and 10% significance levels, respectively.

**Result 3: Sharing Rates of Lies**
In line with Hypothesis 3, receivers were significantly more likely to share lies than truthful videos, even when they were incentivized to share truthful videos.

Most Receiver 1s believed that Receiver 2 would believe the video they shared. When asked about the likelihood of Receiver 2 believing the video they shared, Receiver 1s indicated a belief of 69.1% in the Shared-True treatment and of 67.9% in the Shared-Believed treatment. Hence, they expected a small drop in the share of Receiver 2s believing the video they shared, but the difference is not significant ($t-$test, $p = 0.4350$). Next, we address how Receiver 2s reacted to sharing, and whether they reacted differently across treatments, as indicated directionally by Receiver 1's beliefs.

## 4.2 The Effects of Sharing

Knowing that a video was shared increased the probability that Receiver 2s would choose to watch it and the probability that they would believe it. To estimate the effects of sharing on the decision to watch and believe a video, we estimate Heckman selection models, which account for the fact that Receiver 2s only reported beliefs on videos they watched. In estimating the model of the decision to watch, we include one new variable that is orthogonal to Receiver 2s' beliefs: the order with which the eight videos were presented to Receiver 2s on a grid with four columns and two rows, which was randomly determined. Receiver 2s were less likely to select videos presented in the bottom row or last column. We report marginal effects on the probability that a video is watched, and on beliefs conditional on being watched in Table 4. All videos were watched several times when they were shared and when they were not. Only two videos were watched less than 10 times when they were shared, and we also estimated models excluding them as pre-registered and report the results in Online Appendix B.

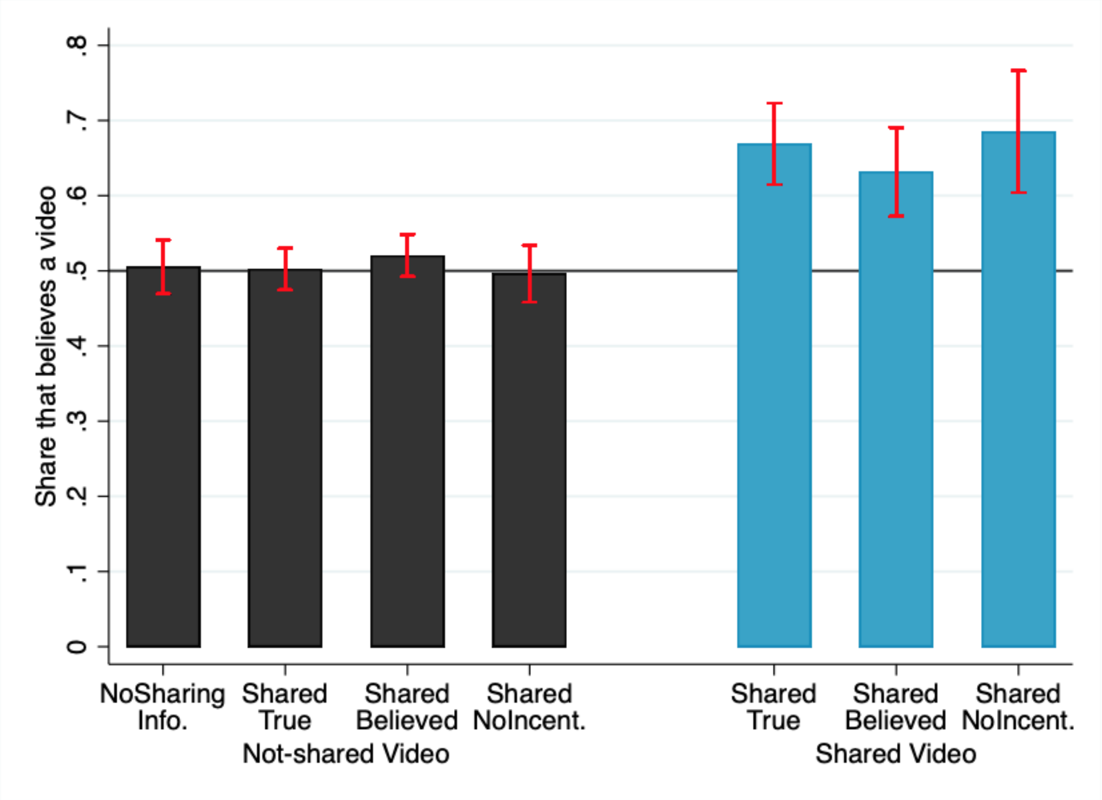
We start by discussing the effects of sharing on the decision of which video to watch. Receiver 2s chose to watch the video shared by Receiver 1 in 75.5%

of the cases in the Shared-True treatment, in 69.5% of the cases in the Shared-Believed treatment, and in 66.1% of the cases in the Shared-No-Incentive treatment. Columns (1)-(3) of Table 4 show that sharing significantly increased the likelihood that a video was watched, though the effects are weaker in the Shared-Believed and Shared-No-Incentive treatments. Receiver 2s were also significantly more likely to choose to watch a video with a female sender. However, Receiver 1s did not appear to anticipate the importance of this feature in choosing which videos to share.

Next, we examined Receiver 2s' beliefs. First, we compare the beliefs of Receiver 2s across treatments, separating beliefs when videos were shared and when they were not shared by Receiver 1. Detailed results at the video level are provided in Online Appendix B. Receiver 2s were more likely to believe shared videos, as shown in Figure 5. In the Shared-True treatment, a shared video was believed 66.9% of the time; in Shared-Believed, 63.2% of the time; and in Shared-No-Incentive, 68.5% of the time. Yet, shared videos were more likely to be false: Receiver 2's optimal response would be to believe videos 38% of the time in the Shared-True treatment and 42% of the time in the Shared-Believed treatment.

The main effect of sharing is to increase the believability of the video shared when Receiver 1s were motivated to share truthful videos. Across all videos, a video was 16 percentage points more likely to be believed in the Shared-True and the Shared-No-Incentive treatments, as shown in column (4) of Table 4. Approximately half of this effect remains with video fixed effects as shown in column (6), indicating that, conditional on watching the same video, Receiver 2 was 8 percentage points more likely to believe it if it was shared. These results are in line with Hypothesis 4. However, sharing did not increase believability in the Shared-Believed treatment, in contrast to Hypothesis 4. This finding reveals a limiting condition for the effect of sharing. When Receiver 1 shared what she/he believed was true, it influenced Receiver 2. But when Receiver 1 shared a video that he or she believed Receiver 2 would believe, Receiver 2 was less inclined to believe it. In both cases, however, the likelihood that a shared video was a lie was above chance (as shown above), and

Figure 5: The Effects of Sharing on Beliefs



*Notes:* This figure presents the average share of Receiver 2s who believed videos, when they were not shared by Receiver 1 (Not-Shared Video) and when they were shared (Shared Video), by treatment in Experiment 3. Error bars indicate 95% confidence intervals.

shared videos should be believed less.[15]

**Result 4: The effect of sharing**

Receiver 2s were significantly more likely to watch and believe a video that was shared if Receiver 1 believed it was true. However, sharing did not affect beliefs if Receiver 1's incentive was to share a video that Receiver 2 would believe.

Receiver 2s also believed more lies (type I error) when they were shared, as shown in column (1) of Table 5. Without sharing information, receivers assessed a video as true when it was a lie 27.3% of the time (not significantly different from 25%, $t-$test, $p = 0.141$). This rate is similar for videos that were not shared in

[15] In exploratory analyses, we tested whether the effect of shared videos was stronger for those receivers who believed they were more accurate (their confidence level). We did not find a significant relationship.

Table 4: Regression Analyses of the Effects of Sharing

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Choose video to watch = 1 | | | Believe video = 1 | | |
| Shared video | 0.312*** | 0.310*** | 0.292*** | 0.163*** | 0.153*** | 0.082*** |
| | (0.029) | (0.034) | (0.030) | (0.031) | (0.038) | (0.030) |
| Shared-True | -0.036** | -0.037** | -0.034** | -0.003 | -0.003 | 0.004 |
| | (0.016) | (0.016) | (0.016) | (0.023) | (0.023) | (0.021) |
| Shared-Believed | -0.028* | -0.028* | -0.026 | 0.015 | 0.015 | 0.026 |
| | (0.016) | (0.016) | (0.016) | (0.023) | (0.023) | (0.022) |
| Shared-No Incentive | -0.022 | -0.022 | -0.020 | -0.009 | -0.010 | -0.011 |
| | (0.019) | (0.019) | (0.019) | (0.027) | (0.027) | (0.025) |
| Shared-Believed X Shared video | -0.080* | -0.083** | -0.079* | -0.054 | -0.053 | -0.061 |
| | (0.041) | (0.041) | (0.042) | (0.046) | (0.046) | (0.044) |
| Shared-No-Incentive X Shared video | -0.122** | -0.131*** | -0.126** | 0.024 | 0.025 | 0.048 |
| | (0.049) | (0.050) | (0.050) | (0.055) | (0.055) | (0.054) |
| Female sender | | 0.068*** | | | -0.023 | |
| | | (0.011) | | | (0.016) | |
| Female sender X Shared video | | 0.014 | | | 0.018 | |
| | | (0.036) | | | (0.040) | |
| Video on last column/row of screen | -0.021* | -0.021* | -0.025** | | | |
| | (0.011) | (0.011) | (0.012) | | | |
| Effect of Shared video: | | | | | | |
| - in Shared-Believed treatment | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.216 |
| - in Shared-NoIncentive treatment | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 |
| Video Fixed Effects | No | No | Yes | No | No | Yes |
| Observations | 9104 | 9104 | 9104 | 9104 | 9104 | 9104 |
| Selected | 4552 | 4552 | 4552 | 4552 | 4552 | 4552 |
| Nonselected | 4552 | 4552 | 4552 | 4552 | 4552 | 4552 |

*Notes*: Estimates from Heckman selection models on the likelihood that a video is chosen to be watched (columns (1)-(3)), selection equation, and believed (columns (4)-(6)) by Receiver 2. Shared video is an indicator variable that takes the value of 1 if the video was shared by Receiver 1. Shared-True, Shared-Believed, and Shared-No-Incentive are indicator variables for each treatment. Female sender is an indicator that takes value 1 if the sender was female (4 out of 8 were), and video on last column/row is an indicator if the video is presented (randomly) in the bottom row or last column of the screen. Columns (3) and (6) include video fixed effects. and controls for the receiver's gender, his/her standardized age, whether the receiver reported reading the NYT daily. The correlations in the error terms between the regression on the choice to watch and believe a video are 0.121, 0.123 and 0.131, respectively, with p-values of 0.021, 0.015, and 0.046, respectively. Robust standard errors are presented in parentheses. ***, **, and * indicate 1%, 5%, and 10% significance levels, respectively.

Shared-True, Shared-Believed, and Shared-No-Incentive. However, receivers (incorrectly) believed shared videos that were lies in 43.1% of the time in the Shared-True treatment. They believed shared lies in 34.1% of the cases in the Shared-Believed treatment, and Receiver 2s made type I errors in 44.9% of the cases in the Shared-No-Incentive treatment. In all cases, they believed lies that were shared significantly more often than in the No-Sharing-Information treatment ($t-$test, $p < 0.001$ in the Shared-True and Shared-No-Incentive treatments, $p = 0.0380$ in the Shared-Believed treatment).

Table 5: Receiver 2 Beliefs for Shared and Not-Shared Videos, by Treatment

| | | (1) | (2) | (3) | (4) | |
|---|---|---|---|---|---|---|
| | | Believe video is true | | Believe video is a lie | | |
| Shared | Treatment | Lie | Truthful video | Lie | Truthful video | N |
| No | No-Sharing-Information | 27.3% | 23.2% | 25.7% | 23.8% | 740 |
| | Shared-True | 25.9% | 24.3% | 26.1% | 23.7% | 1246 |
| | Shared-Believed | 29.2% | 22.8% | 25.6% | 22.3% | 1226 |
| | Shared-No Incentive | 27.2% | 22.4% | 26.3% | 24.1% | 665 |
| Yes | Shared-True | 43.1% | 23.8% | 21.4% | 11.7% | 290 |
| | Shared-Believed | 34.1% | 29.1% | 22.1% | 14.7% | 258 |
| | Shared-No Incentive | 44.9% | 23.6% | 18.9% | 12.6% | 127 |

*Notes*: Notes: This table shows Receiver 2 beliefs by treatment for shared and not-shared videos. Columns (1) to (4) depict four possible cases: a video is believed to be true, but is a lie (type I error); a video is believed to be true, and is true; a video is believed to be false, and is indeed a lie; a video is believed to be false, but is true (type II error).

A first implication of these results is that Receiver 2s were more likely to make mistakes when assessing a shared video that is false. Considering all shared videos, which are false in 58-62% of the cases, Receiver 2s correctly assessed the truthfulness of 48.9% of videos in the No-Sharing-Information treatment, compared to 45.2% of shared videos in the Shared-True treatment, and 42.5% in the Shared-No-Incentive treatment. These changes are not statistically significantly ($t-$test, $p > 0.10$ in both cases). These directional effects of sharing only apply to shared videos: Receiver 2 assessed videos that were not shared correctly at similar rates as in the No-Sharing-

Information treatment: 50.4% of the time in the Shared-True treatment, and 48.7% of the time in the Shared-Believed treatment.

A second implication is that the incentives behind sharing decisions can be important. When Receiver 1s were incentivized to share what Receiver 2 would believe, sharing did not significantly affect beliefs. This result could partly stem from the fact that Receiver 1s were 3-4 percentage points less likely to believe the video when sharing it in Shared-Believed, compared to Shared-True.

The stronger belief in shared videos in the Shared-True treatments is consistent with Receiver 2s (wrongly) believing that Receiver 1s are sufficiently accurate. Receiver 2s may have (correctly) anticipated that sharing is a weaker signal of Receiver 1s' own beliefs in Shared-Believed than in Shared-True. Then, believing shared videos less in the Shared-Believed treatment would follow. Additionally, in the Shared-Believed treatment Receiver 1s were incentivized to make inferences about what others believe to be true. This was not necessary in Shared-True, in which their own beliefs about truthfulness were what mattered, and such belief had already been elicited. As a consequence, Receiver 2s in Shared-Believed may have not inferred as much about Receiver 1s' second-order beliefs.

# 5 Conclusion

The growing influence of fake news has spurred a large number of questions regarding individuals' ability to detect fraudulent news stories. In this paper, we study receivers' ability to detect lies, absent motivated beliefs. We find that even when motivation is removed, receivers' ability to detect lies is limited. Yet, receivers are not aware of their limited ability to detect these lies and exhibit significant overconfidence. Our exploratory analysis reveals that receivers use cues, such as the facial expressions of the sender, as indicators of truthfulness but use them wrongly, making predictable mistakes. This discrepancy between the perception of cues and the ability to recognize them could partially explain why receivers are overconfident. Taken together, our findings inform us of an important reason fake news is likely to be believed–people are not good at detecting lies, and are overconfident in their ability to do so.

We use the above experimental design and results to examine whether sharing behavior may increase the spread of fake news. We find support for this hypothesis on several levels: when receivers are incentivized to share a truthful video, shared videos are more likely to be false, more likely to be watched, and more likely to be believed. As a result, allowing for sharing significantly increases the probability of type I errors (believing a false video).

The combination of overconfidence and being bad at detecting lies, with an overreliance on shared content, may explain why fake news is so prominent and influential. The results could add to the heated discussion regarding flagging false news on social media. There are different types of actors on social media, including bots and groups interested in spreading certain kinds of messages, who may be able to use sophisticated tools to intentionally spread false stories, and potentially predict which false stories are more likely to influence beliefs via sharing. Understanding if, as in our experiments, shared content on social media is more likely to increase the impact of fake news, and whether adding motivation to believe certain news stories (e.g., depending on one's political views) could make the effects of sharing even stronger, is important.

# References

[1] Abeler, J., Becker, A., and A. Falk (2014). Representative Evidence on Lying Costs. *Journal of Public Economics* 113, 96-104.

[2] Abeler, J., Nosenzo, D., and C. Raymond (2019). Preferences for Truth-telling. *Econometrica*, 87 (4), 1115-1153.

[3] Allcott, H. and M. Gentzkow (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31(2), 211-236.

[4] Allcott, H., Gentzkow, M. and C. Yu (2019). Trends in the Diffusion of Misinformation on Social Media. *Research and Politics* April-June 2019, 1-8.

[5] Belot, M. and J. van de Ven (2017). How private is private information? The ability to spot deception in an economic game. *Experimental Economics* 20 (1), 19-43.

[6] Bénabou, R. (2013). Groupthink: Collective Delusions in Organizations and Markets. *Review of Economic Studies* 80, 429-462.

[7] Bénabou, R. and J. Tirole (2011). Identity, Morals and Taboos: Beliefs as Assets. *The Quarterly Journal of Economics*, 126, (2011), 805-855.

[8] Benoit, J.P. and J. Dubra (2011). Apparent Overconfidence. *Econometrica* 79(5), 1591-1625.

[9] Benoit, J.P., Dubra, J., and D.A. Moore (2015). Does the better-than-average effect show that people are overconfident? Two experiments. *Journal of the European Economic Association* 13(2), 293-329.

[10] Berger, J. and K. Milkman (2012). What Makes Online Content Viral? *Journal of Marketing Research* 49 (2), 192-205.

[11] Bijlstra, G., and R. Dotsch (2011). Facereader 4 emotion classification performance on images from the radboud faces database. Unpublished manuscript,

Department of Social and Cultural Psychology, Radboud University Nijmegen, The Netherlands.

[12] Bohren, A., Imas, A., and M. Rosenberg (2018). The Language of Discrimination: Experimental versus Observational Data. *AEA Papers and Proceedings* 108, 169-174.

[13] Bonnefon, J.F., Hopfensitz, A., and W. De Neys (2013). The Modular Nature of Trustworthiness Detection. *Journal of Experimental Psychology: General* 142 (1), 143-150.

[14] Bonnefon, J.F., Hopfensitz, A., and W. De Neys (2017). Can We Detect Cooperators by Looking at Their Face? *Current Directions in Psychological Science* 26 (3), 276-281.

[15] Bond, C. F., and B.M. DePaulo (2008). Individual differences in judging deception: accuracy and bias. *Psychological Bulletin* 134(4), 477-492

[16] Bond, C.F., and B.M. DePaulo (2006). Accuracy of Deception Judgments. *Personality and Social Psychology Review* 10 (3), 214-234.

[17] Bordalo, P., Coffman, K., Gennaioli, N. and A. Shleifer (2019). Beliefs about Gender. *American Economic Review*, 109 (3), 739-773.

[18] Bronstein, M., Pennycook, G., Bear, A., and Rand, D., and T. Cannon (2019) Belief in Fake News Is Associated with Delusionality, Dogmatism, Religious Fundamentalism, and Reduced Analytic Thinking. *Journal of Applied Research in Memory and Cognition* 8 (1), 108-117.

[19] Burks, S.V., Carpenter, J.P., Goette, L., and A. Rustichini (2013). Overconfidence and Social Signalling. *Review of Economic Studies* 80 (3), 949-983.

[20] Charness, G., Rustichini, A., and J. van de Ven (2018). Self-confidence and strategic behavior. *Experimental Economics* 21, 72-98.

[21] Chen, D., A. Hopfenstiz, B. van Leeuwen, and J. van de Ven (2019). The Strategic Display of Emotions. CentER Discussion Paper, 2019-014.

[22] Centorrino, S., Djemai, E., Hopfensitz, A., Milinski, M., and Seabright, P. (2015). Honest Signaling in Trust Interactions: Smiles Rated As Genuine Induce Trust and Signal Higher Earning Opportunities. *Evolution and Human Behavior* 36, 8-16.

[23] Centorrino, S., Djemai, E., Hopfensitz, A., Milinski, M., and Seabright, P. (2015). A Model of Smiling as a Costly Signal of Cooperation Opportunities. *Adaptive Human Behavior and Physiology* 1, 325-340.

[24] Coffman, K. (2014). Evidence of Self-Stereotyping and the Contribution of Ideas. *Quarterly Journal of Economics*, 1625-1660.

[25] De keersmaecker, J., and A. Roets (2017). 'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence* 65, 107-110.

[26] DePaulo, B.M., Charlton, K., Cooper, H., Lindsay, J. J., and L. Muhlenbruck (1997). The Accuracy Confidence Correlation in the Detection of Deception. *Personality and Social Psychology Review* 1(4), 346-357.

[27] Dreber, A., and M. Johannesson (2008). Gender Differences in Deception. *Economics Letters* 99 (1): 197-199.

[28] Dwenger, N., and T. Lohse (2019). Do individuals successfully cover up their lies? Evidence from a compliance experiment. *Journal of Economic Psychology* 71, 74-87.

[29] Ekman, P. (1970). Universal facial expressions of emotion. *California Mental Health Research Digest*, 8, 151-158.

[30] Ekman, P. and W. Friesen (1982). Felt, False, and Miserable Smiles. *Journal of Nonverbal Behavior* 6 (4), 238-252.

[31] Erat, S., and U. Gneezy (2012). White Lies. *Management Science* 58 (4): 723-733.

[32] Eyster, E., Rabin, M. and G. Weizsacker (2018). An Experiment on Social Mislearning. Working paper.

[33] Frank, M. and P. Ekman (1997). The ability to detect deceit generalizes across different types of high-stakes lies. *Journal of Personality and Social Psychology* 72 (6), 1429-1439.

[34] Guess, A., Nagler, J. and J. Tucker (2019). Less Than You Think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* 5 (1), eeau4586.

[35] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and D. Lazer (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 374-378.

[36] Kahan, D. (2016a). The Politically Motivated Reasoning Paradigm, Part 1: What Politically Motivated Reasoning Is and How to Measure It. *Emerging Trends in Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*. Eds. R. A. Scott and M. C. Buchmann.

[37] Kahan, D. (2016b). The Politically Motivated Reasoning Paradigm, Part 2: Unanswered Questions. *Emerging Trends in Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*. Eds. R. A. Scott and M. C. Buchmann.

[38] Konrad, K., Lohse, T., and S. Qari (2014). Deception choice and self-selection: The importance of being earnest. *Journal of Economic Behavior & Organization*, 107, 25-39.

[39] Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin* 108 (4), 480-498.

[40] Lazer, D., Baum, M., Benkler, J., Berinsky, A., Greenhill, K., Metzger, M., and J. Zittrain (2018). The science of fake news. *Science*, 9, 1094-1096.

[41] Lee, M., and M. Lee (2017). The relationship between crowd majority and accuracy for binary decisions. *Judgment and Decision Making* 12 (4), 328-343.

[42] Litman, L., Robinson, J., and Abberbock, T. (2016). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 1-10.

[43] Lohse, T. and S. Qari (2019). Gender differences in face-to-face deceptive behavior. Working paper.

[44] Mann, S., Vrij, A., and R. Bull (2004). Detecting true lies: police officers' ability to detect suspects' ies. *Journal of Applied Psychology* 89 (1), 137-149.

[45] Mondak, J. J., and M. R. Anderson (2004). The knowledge gap: A reexamination of gender-based differences in political knowledge. *The Journal of Politics* 66 (2), 492-512.

[46] Moore, D. A., and P.J. Healy (2008). The trouble with overconfidence. *Psychological Review*, 115(2), 502-517.

[47] Nisbet, E., Cooper, K., and R.K. Garrett (2015). The Partisan Brain: How Dissonant Science Messages Lead Conservatives and Liberals to (Dis)Trust Science. The Annals of the American Academy of Political and Social Science.

[48] Pennycook, G., and D. G. Rand (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39-50.

[49] Pennycook, G., and D. G. Rand (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, 88 (2), 185-200.

[50] Pennycook, G., Cannon, T. D., and D. G. Rand (2018). Prior Exposure Increases Perceived Accuracy of Fake News. *Journal of Experimental Psychology: General,* 147 (12), 1865-1880.

[51] Romano, J.P and M. Wolf (2005). Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica* 73 (4), 1237-1282.

[52] Serra-Garcia, Marta and Uri Gneezy. 2021. Replication data for: Mistakes, Overconfidence and the Effect of Sharing on Detecting Lies" American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. http://doi.org/10.3886/E140961V1.

[53] Surowiecki, J. (2005). *The wisdom of the crowds.* Anchor Books, New York.

[54] Swann, W. B., Jr., Silvera, D. H., and C.U. Proske (1995). On "knowing your partner": Dangerous illusions in the age of AIDS? *Personal Relationships* 2, 173-186.

[55] Tappin, B., Pennycook, G., and D. G. Rand (2018). Rethinking the link between cognitive sophistication and identity-protective bias in political belief formation. Working paper.

[56] Thaler, M. (2019). The "Fake News" Effect: Motivated Reasoning and Trust in News. Working paper.

[57] Todorov, A., Olivola, C.Y., Dotsch, R., and P. Mende-Siedlecki (2015). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology* 66, 519-545.

[58] van Leeuwen, B., C. N. Noussair, T. Offerman, S. Suetens, M. van Veelen, and J. van de Ven (2018). Predictably Angry–Facial Cues Provide a Credible Signal of Destructive Behavior. *Management Science* 64 (7), 3352-3364.