To be submitted for the Special Issue at *Philosophical Transactions of The Royal Society B: Biological Sciences*

Is It Costly to Deceive?
People Are Adept at Detecting Gossipers' Lies but May Not Reward Honesty

Miguel A. Fonseca[1,2] & Kim Peters[1,3]

[1] University of Exeter Business School, UK

[2] University of Minho, Portugal

[3] School of Psychology, University of Queensland

Both authors have contributed equally to this work.

Correspondence should be addressed to Miguel Fonseca (m.a.fonseca@exeter.ac.uk) or Kim Peters (k.o.peters2@exeter.ac.uk)

Abstract

The possibility that gossipers may share dishonest reputational information is a key challenge to claims that gossip can shore up cooperation in social groups. It has been suggested that imposing social costs on dishonest gossipers should increase the honesty of these reputational signals. However, at present, there is little evidence of people's willingness to impose costs on dishonest gossipers; there is also little evidence of their ability to detect gossipers' lies in the first place. This paper aims to shed light on people's abilities to detect dishonest gossip and their treatment of those who share it. To do this, we report the results of two trust game studies using the strategy method (Study 1) and repeated interactions in the lab (Study 2). We show that in an environment where gossipers tell spontaneous lies people are more inclined to believe honest than dishonest gossip. We also show that people are more likely to treat favourably gossipers they believe to be honest, but that this does not always result in more favourable treatment for gossipers who were actually honest. We discuss the implications for the potential utility of social sanctions as a tool for securing honesty.

Word count: 194

Keywords: reputation, cooperation, gossip, lies, honesty, trust

Is It Costly to Deceive?

People Are Adept at Detecting Gossipers' Lies but May Not Reward Honesty

Gossip — a communicative act involving two or more people's exchange of social information about absent third parties — has received attention as a naturally occurring mechanism for the dissemination of reputational information [1]. Reputational information is believed to play an important role in cooperation in large social groups where direct observation is difficult and repeated interactions are rare [2-5]. This is because information about a person's character, past actions and the esteem in which they are held allows others to direct their cooperative behaviours towards those who are likely to deserve and return them, and away from those who are not. This in turn increases the value of a positive reputation and incentivises the hard work that is required to build it [6-7]. If gossip does indeed support these reputational processes, then it may ultimately boost cooperation in many everyday situations.

However, this reasoning rests on an important assumption: gossip needs to be honest, such that good reputations are assigned to good actors and bad reputations to bad ones [8-9]. If this does not hold, such that much of the gossip that people share is dishonest, then its capacity to support cooperation will unravel. In particular, people who act on dishonest gossip may cooperate with those who do not deserve it and withhold cooperation from those who do. This may not only harm those who act on dishonest gossip (making it less likely that they will attend to it in the future), but also those who have worked hard to create a positive reputation (disincentivising further such efforts). This expectation has been borne out in recent empirical work [10-11] that showed that when there was a high likelihood that gossip would be misdelivered, such that the behaviours gossip described were ascribed to the wrong people, levels of cooperation declined.

There are good reasons for doubting the honesty of gossip in everyday life. First, if people wish to share dishonest reputational information, it is not especially hard for them to do so. As Zahavi and Zahavi (1997, p.223) [15] put it, “It is easy to lie with words.” Second, it has been argued that there may be circumstances where gossipers may benefit from lying. For instance, gossipers may wish to advantage allies by assigning them undeservedly positive reputations and to damage enemies by assigning them undeservedly negative ones [12-13]. Gossipers may also wish to gain an advantage over others by providing them with misleading reputational information. In line with this, recent experimental work shows that dishonest gossip may not be rare, and that gossipers believe that it serves a range of social goals, including securing just deserts for those they are talking about or misleading those they were talking to [10, 14]. This work also shows that when there are strategic incentives for lying, for instance when gossipers are competing with one another, rates of dishonesty increase.

Whether this dishonest gossip ultimately undermines cooperation may depend on the extent to which people are able to verify its content before choosing whether or not to act on it. For this reason, dishonest gossip may be less problematic in small groups, including those that were typical of our ancestral environment. In these contexts, group members would have many more opportunities to aggregate gossip about the same individual across multiple sources than is possible in larger, more geographically dispersed societies [16]. Dishonest gossip may also be less problematic when it concerns relatively permanent or visible states of being that can be independently verified (e.g., physical characteristics), as opposed to relatively transient behaviours or hard-to-verify states (e.g., one-off behaviours or claims of witchcraft). It follows, therefore, that the contexts in which dishonest gossip may be most problematic are precisely those in which it is believed to be most helpful: that is, in large mobile social groups and in relation to transient cooperative behaviours that have relatively few witnesses.

However, this is not to say that there is nothing to stop dishonest gossip and ameliorate its harmful effects in these contexts. If people see honest gossip as serving a social function, there should be a demand for it, as well as an incentive to impose social costs on those who lie. Such social costs could, if they exist, offset the short-term selfish benefits of dishonesty and increase the likelihood that gossipers will tell the truth. This rationale is supported by Lachmann, Számadó and Bergstrom (2001) [17], who argue that even a low-cost signal like language can be reliable as long as people readily detect lies and impose costs on those who share them. While the possibility that people may impose social costs on dishonest gossipers has received some theoretical attention [9, 18, 19], it has not been tested empirically in a behavioural setting. However, this expectation is supported by evidence that people generally consider liars to be immoral (at least if the lie is selfishly motivated [20]) and are also strongly motivated to reciprocate others' moral and immoral actions [21, 22].

Our primary aim in this paper is to test whether people are indeed more likely to impose social costs on dishonest gossipers than honest ones. To do this, we consider a sequence of interactions where an individual first acts on a piece of gossip, where this interaction has hypothetical or real payoff consequences, and then has the opportunity to interact directly with the gossiper. We examine whether people who act on gossip that is dishonest treat gossipers less favourably in a subsequent interaction than those who act on gossip that is honest. This expectation is expressed in our first hypothesis:

H1: *Participants will treat gossipers who previously shared dishonest gossip less favourably, on average, than gossipers who previously shared honest gossip.*

Our second aim is to explore a psychological precursor to the above effect: people's ability to detect dishonest gossip. This is important because if people find it difficult to distinguish between honest and dishonest gossip then any social costs that are imposed are likely to be directed towards honest as well as dishonest gossipers. Ultimately, this could be expected to reduce people's willingness to share or act on gossip of any kind, rather than to increase the quality of what is shared. We address this aim in our second study by asking participants to rate the accuracy of the gossip that they have received *both before and after* interacting with the target of the gossip. There is a large body of existing work that shows that people are very poor at detecting lies when their only cues are the verbal and non-verbal cues of liars [23]. This suggests that absent any opportunity to test the honesty of the gossip, people may struggle to distinguish honest from dishonest gossip. However, when people are able to compile more information to help them verify the claims of a gossip item, including by interacting with gossip target themselves, then they are likely to do much better [24, 25, 16]. The expectation that people will be able to discriminate honest from dishonest gossip is expressed in our second hypothesis (we test this expectation before and after participants interact with the target):

H2: *Participants will evaluate dishonest gossip as less accurate, on average, than honest gossip.*

Study Overview

We conducted two studies to test these hypotheses. Study 1 tested H1 in a hypothetical setting, while Study 2 tested H1 and H2 in a behavioural setting. In both studies, participants, in the role of Investor, played repeated one-shot trust games with Agents [26] (for clarity, we call these *Agent-trust* games). Before each Agent-trust game participants received a piece of gossip that described the Agent's trustworthiness in a previous game. Some of these descriptions were honest and some were not. Participants then decided how much to trust this Agent, before being informed of their Agent's response and their respective payoffs. In an environment where Agents behave reasonably consistently over time, the discrepancy between participants' gossip-based expectations and actual Agent behaviour provides information about the gossiper's honesty. To see whether participants would impose social costs on dishonest gossipers (i.e., gossipers who create expectations that are not met), we then asked them to play the one-shot trust game a second time (we call these *gossiper-trust* games). This time, participants were asked to say how much they would trust the gossiper. Because a failure to trust reduces the gossiper's potential payoff it provides evidence of participants' willingness to impose a social cost.

As should be clear from this explanation, there are two different sets of social roles: those related to the trust game (Investor and Agent) and those related to the exchange of gossip (gossiper and audience). Importantly, gossip (about Agents) is only exchanged between Investors. In these studies, we are interested in the behaviour of participants who take the role of Investor. In Study 1, these Investor participants receive gossip from other (hypothetical) Investors, and therefore only take the additional role of audience. In Study 2, these Investor participants exchange gossip with one another, and are therefore also in the roles of gossiper and audience.

There were two other major differences between the two studies. First, while participants in Study 1 completed a gossiper-trust game after every Agent-trust game, those in Study 2 only completed one gossip-trust game after they had completed 24 Agent-trust games. We were able to alternate rounds in Study 1 because in this study gossiper behaviour was pre-determined. This meant that in Study 1 any tendency for participants to impose costs could not eliminate the independent variable (i.e., dishonest gossip). This was not the case for Study 2, which limited us to a single gossiper-trust game. Second, to address H2, in Study 2 we also asked participants to rate their confidence in the honesty of the gossip on first reading it, and again, after their interaction with the target. Materials, data, analysis code and the supplementary information (SI) document are available here: https://osf.io/t2vr9/. Study 2 pre-registration is available here: https://osf.io/8af4s/. The full pre-registered analysis is included in SI.

Study 1

Method

Participants

Participants were 184 UK-based adults who responded to an invitation on Prolific.co to complete a study on social decision making, 96 of whom failed at least one of four comprehension checks, were funnelled out of the study and reimbursed £0.50. The remainder completed the study and were reimbursed £2.50. Of these, six failed a final attention check at the end of the study, and their data were excluded. The final sample includes the remaining 82 participants. The majority were British (74%), female (70%), between 25 and 34 years of age (range 18-64) and had a post-secondary qualification (66%).

Procedure

Participants completed 12 rounds in the experiment. Each round consisted of an Agent-trust game (which manipulated the honesty of the gossip) followed by a gossiper-trust game (which measured participants' treatment of gossipers). The 12 rounds were created by the orthogonal interaction of four types of pre-programmed Agent and three types of gossip content. The four types of Agent consisted of two *trustworthy* Agents who returned either 42 or 50 percent of received tokens (to the nearest whole token), thereby increasing hypothetical participant payoffs, and two *untrustworthy* Agents who returned either 17 or 25 percent of received tokens (to the nearest whole token), thereby reducing hypothetical payoffs. The three types of gossip content consisted of honest gossip that accurately described the Agent's return rate, and two types of dishonest gossip: a positive lie that overstated the Agent's return rate, and a negative lie that understated it. The order in which participants faced each Agent-gossip combination was random.

The experimental instructions told participants that they would take the role of Investor in a series of trust games [26]. They were told that all of their trust game interactions were hypothetical (i.e., none of the other players were real) and for this reason their decisions on these games carried no payoff consequences. It was explained that in each trust game Investors started with an endowment of 10 tokens and decided how many of them to send to their Agent (in whole tokens from 0 to 10). Agents started each round with 0 tokens, and then receive three times the number of tokens sent by their Investor. Agents then chose how many (if any) of these tokens to return to the Investor. After this, participants viewed their own hypothetical payoff (10 *less* tokens sent *plus* tokens returned), as well as the hypothetical payoff to their Agent (tokens sent *times* three *less* tokens returned). After completing comprehension checks, participants played two practice rounds.

Participants were then asked to imagine that they were part of a population that consisted of Investors and Agents who interacted repeatedly. The instructions told participants (using neutral language that avoided terms like 'gossip' or 'gossiper') that they would interact with 12 Agents in turn (these were the Agent-trust games), and that before interacting with each Agent they would receive gossip from another Investor (the gossiper) who had recently played with this Agent. The gossip would describe this Agent's behaviour in a previous interaction. To alert participants to the possibility that this gossip could be dishonest, we told them that there were three kinds of gossipers in the population: honest, who always told the truth about the Agent's previous behaviour; dishonest, who always lied; and mistaken, who sometimes made a mistake when describing the Agent's behaviour. Participants were told that after their interaction with each Agent was complete, they would be asked how many tokens they would send to the gossiper if this person took the Agent role (these were the gossiper-trust games). These games did not play out, so the decision of how many tokens to send to the gossiper was the final act in each round. Participants then answered another two comprehension questions before starting the 12 rounds.

At the start of each round, participants received a piece of gossip that stated "*I sent this Agent 8 tokens; they received 24 tokens, and returned [X] tokens*." The accuracy of the gossip was therefore solely determined by the number of tokens that the Agent was said to have returned. Honest gossip claimed that trustworthy Agents returned 12 or 10 tokens and that untrustworthy Agents returned 6 or 4 tokens. Positive lies claimed that trustworthy Agents returned 15 or 13 tokens and untrustworthy Agents 12 or 10. Negative lies claimed that trustworthy Agents returned 6 or 4 tokens and untrustworthy Agents 3 or 1. After reading the gossip, participants decided how many tokens to send to this Agent. They were then told how many tokens the Agent returned in accordance with the pre-programmed return rate and what their hypothetical payoff was. They were then asked how many tokens they would send to the gossiper if this person was their Agent in this game. After completing twelve rounds of this sequence, participants responded to demographic information questions and a final attention check.

Results

Descriptive Statistics

Participants were moderately willing to trust Agents, sending them an average of *M*=5.55 (*SD*= 2.76) tokens. They were also moderately willing to trust gossipers, sending them an average of *M*=5.29 (*SD*= 2.61) tokens.

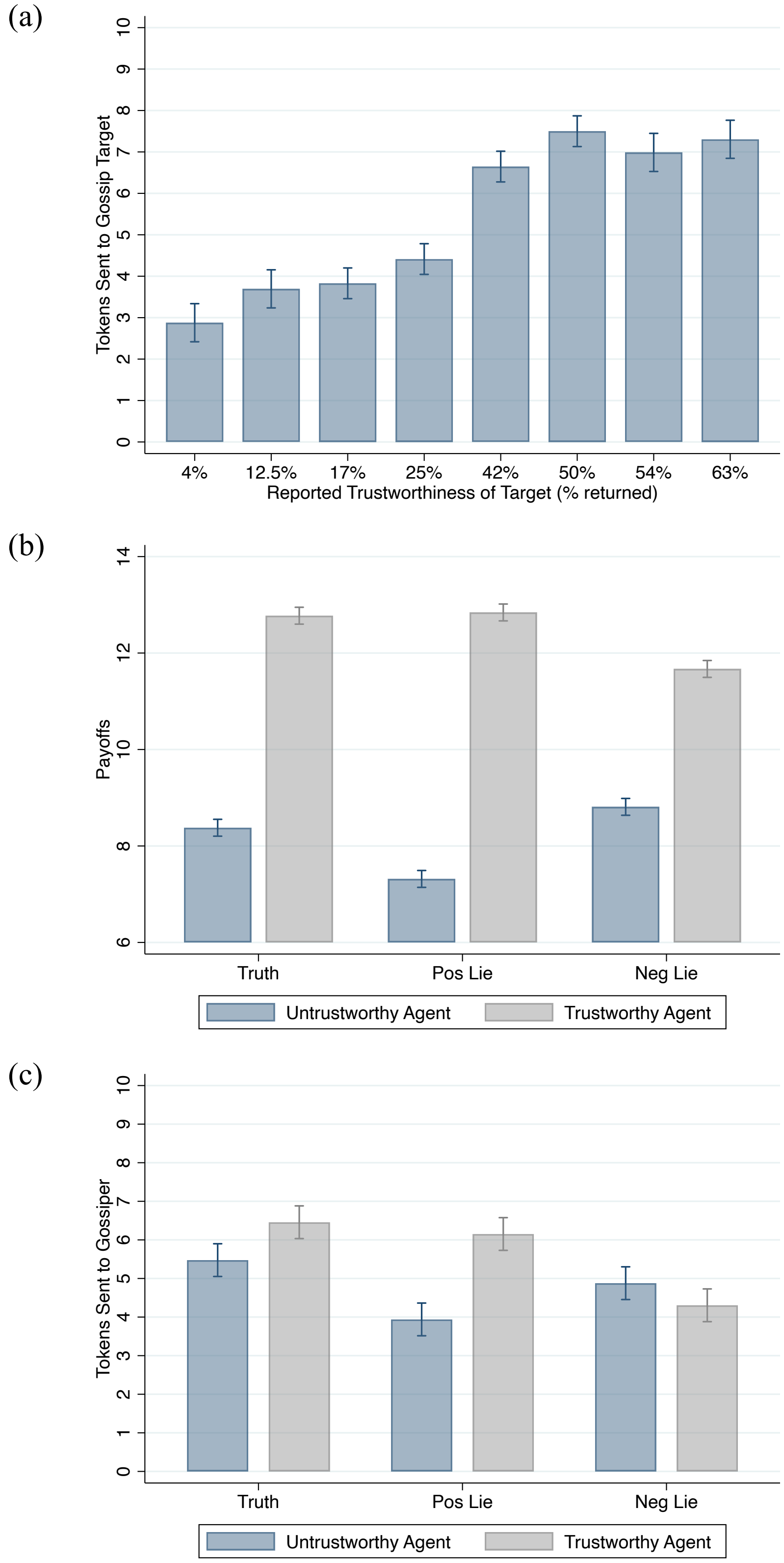
Is Dishonest Gossip Damaging?

The claim that dishonest gossip may be harmful presumes that people (1) act on gossip and (2) that acting on honest gossip is more beneficial than acting on dishonest gossip.

To test the first condition, we regressed the number of tokens participants sent to the Agent onto the gossip about the Agent's trustworthiness (computed as the percentage of tokens returned). In this, and all analyses that follow, we included random effects at the individual level to account for the fact that individuals provided multiple responses; all p-values refer to two-sided tests. In line with this assumption (see Figure 1a), participants said that they would send more tokens if Agents were said to have been more trustworthy (*b*=9.03, χ^2(1)=792.82, *p*<.001).

To test the second condition, we regressed participants' payoffs onto a dummy representing the Agent's previous trustworthiness (1 if Agent had previously returned at least 33% of tokens in the previous round, otherwise 0), a categorical variable representing the honesty of the gossip that claimed to describe this Agent's previous trustworthiness (1 if truth, 2 if positive lie, 3 if negative lie) and the two-way interaction between these variables. This showed that at least some lies decrease payoffs (see Figure 1b). [1] In the case of previously untrustworthy Agents (who were generally disadvantageous to interact with, resulting in payoffs lower than the endowment of 10), participants who received positive lies had significantly lower payoffs than those acting on the truth (b=-1.06, χ^2(1)=70.74, *p*<.001), although those who received negative lies had a small significant increase in payoffs (*b*=0.43, χ^2(1)=11.78, *p*=.001). In the case of previously trustworthy Agents (who were generally advantageous to interact with), participants who acted on positive lies did not differ significantly in their payoffs from those acting on truth (*b*=0.07, χ^2 (1)=0.28, *p*=.595), while those who acted on negative lies had significantly lower payoffs (*b*=-1.10, χ^2 (1)=76.55, *p*<.001).

Figure 1. Trust decisions and payoffs as a function of gossip honesty.



Notes (a) graphs marginal means of trust in Agent as a function of reported trustworthiness; (b) graphs marginal means of payoffs as a function of gossip content and Agent type, where payoffs = 10 - tokens sent + tokens returned; (c) graphs marginal means of trust in gossiper as a function of gossip content and Agent type; Agent trustworthiness is that observed in the previous round and described by the gossiper; error bars denote 95% CIs.

Treatment of Dishonest Gossipers

Our findings are consistent with the possibility that dishonest gossip may erode the ability for gossip to bolster cooperation. To test H1 — the tendency for participants to impose social costs on dishonest gossipers — we regressed the number of tokens that participants sent to their gossiper onto the dummy representing the Agent's previous trustworthiness, the categorical variable representing the honesty of the gossip claiming to describe this trustworthiness and their two-way interaction (codes as described above). The results are depicted in Figure 1c. In the case of previously untrustworthy Agents, participants were less willing to trust gossipers who told positive lies than those who told the truth (*b*=-1.54, χ^2 (1)=46.45, *p*<.001); they were also less willing to trust gossipers who told negative lies (*b*=-0.60, χ^2 (1)=7.03, *p*=.008). In the case of previously trustworthy Agents, participants were less willing to trust gossipers who told negative lies than those who told the truth (*b*=-2.15, χ^2(1)=91.15, *p*<.001); however, they were not significantly less willing to trust those who told positive lies (*b*=-0.30, χ^2(1)=1.83, *p*=.176). Interestingly, participants appeared to trust gossipers who told the truth about previously trustworthy Agents more than gossipers who told the truth about previously untrustworthy Agents, *b*=0.98, χ^2(1)=18.96, *p*<.001.

Discussion

Participants were sensitive to the accuracy of gossip and were more likely to trust honest than dishonest gossipers, supporting H1. This tendency varied with the *type* of lie that gossipers told. In particular, lies that resulted in the worst payoffs for participants — positive lies about untrustworthy Agents and negative lies about trustworthy Agents — led to the lowest levels of gossiper trust. This suggests that the extent to which acting on a lie is harmful is an important factor in people's treatment of gossipers. However, it may not be the only factor. Participants trusted gossipers who shared negative lies about untrustworthy Agents less than gossipers who told the truth about these Agents, even though these lies were payoff-improving. This pattern was not present for positive lies about trustworthy Agents. Participants also trusted gossipers who told the truth about untrustworthy Agents less than gossipers who told the truth about trustworthy Agents. Together these findings suggest that people may be less willing to trust gossipers who share negative gossip, even if this gossip is honest or ultimately helpful. In other words, the valence of gossip may matter for how gossipers are treated.

While this study provides initial evidence for the possibility that there may be social costs to dishonesty, it has limitations. In particular, it is unclear whether the effects that we observe here generalise to a context of naturally occurring lies where decisions have material consequences for participants. In addition, because this study did not assess participants' beliefs about the accuracy of gossip, we could not directly test H2. In order to address these limitations, Study 2 assessed people's ability to detect spontaneous lies and their willingness to trust honest and dishonest gossipers where this decision has payoff consequences.

Study 2

Method

Participants

Participants were 184 university students who had agreed to be contacted about studies run through the FEELE lab at the University of Exeter. A subset of these participants (*N*=24) provided the Agent strategies; the remainder (N=160) took part in the study proper. Participants completed their study on computerised networks in the lab. Our study proper sample size was informed by simulations that showed that this would give us a 90% chance of detecting a 15-point difference in mean ratings of participants' confidence in honest and dishonest gossip (measured on a scale from 0 to 100) across 20 rounds if one-fifth of messages were inaccurate (see SI). About equal numbers of participants were male and female (*N*=87 and *N*=96, respectively; 1 did not state) and their average age was 21.06 (*SD*=3.60). Participants were paid on the basis of their decisions plus a show-up fee. The average payment for those providing Agent strategies was £12.28 (*SD*=5.41) and the average for those in the study proper was £10.61 (*SD*=2.01).

Strategy Elicitation

In order to ensure that Agent behaviour was stable over time, we elicited Agent strategies to use in the study proper. To do this, we introduced an initial group of participants to the trust game [26] with gossip (the form of this game including endowments was identical to that described in Study 1). Participants were told that they would take the role of Agent and play against Investors who would come into the lab in the future. They were asked to provide us with three strategies that specified how many experimental currency units (i.e., ECU) they would return to their Investor for every possible number they could receive (i.e., 3 times each ECU integer between 0 and 10). By asking participants to generate three different mappings of tokens received to returned, i.e., ‘strategies’, we aimed to boost variation in the strategies that were generated. They were told that one of these strategies would be implemented, and that in addition to their £3 completion fee, they would receive a bonus that was based on their payoffs from three randomly selected rounds (where 5 ECU=£2). Other than ensuring that we had one strategy for each Agent and represented the distribution of the types of generated strategies, our selection of Agent strategies for the study proper was random (see SI for further information).

Study Proper Procedure

In the study proper, participants were allocated to eight-person Investor networks. At the beginning of the session, participants sat in individual cubicles, and were told that there were two parts to the experiment and that they would receive information about each part when they reached it. They were told that their payoff for the experiment was denominated in

ECU, where 8 ECU were worth £1. Their payment included their £3 show-up fee, *plus* their earnings from three randomly selected rounds from Part 1, *plus* a 16 ECU (£2) bonus for top performers in Part 1, *plus* their payoff from Part 2 (details of which were provided at the start of Part 2).

Part 1. Part 1 consisted of 24 *Agent-trust games*. Participants were introduced to the trust game with gossip (the form of the trust game was again identical to Study 1). They were told that they would take the role of Investor and play a number of rounds against Agent strategies that had been provided by an earlier group of participants who would also receive payoffs for their performance. To avoid end game effects, we did not tell participants how many rounds they would play in Part 1. To elicit gossip, after each game, participants were told that they would be required to write a message describing their interaction with the Agent. This involved stating the number of ECU they had sent, and the number of ECU their Agent had returned — the actual number sent and returned were summarised on screen for participants to refer to if they wished. There were told that this piece of gossip would be given to the participant who would next interact with this Agent.

It was explained that in each round participants would be matched at random with an Agent strategy, conditional on two restrictions: (1) they would never play a given Agent strategy twice in a row (in fact, they played each of the 24 strategies only once); and (2) they would not receive a piece of gossip from the same Investor in two consecutive rounds. It was further explained that if they were paired with an Agent that no-one had previously interacted with, they would not receive gossip (because of the matching scheme, this occurred in rounds 1, 9, and 17). To introduce an element of competition among participants in Part 1, and thereby increase the possibility of lies, participants were told that the four members of each network who earned the most over the course of Part 1 would receive a 16 ECU bonus payment at the end of the experiment. After reading through these instructions, participants completed three comprehension checks and calculated the payoffs from three hypothetical interactions. They then completed the 24 Agent-trust rounds. On the 21 rounds with gossip, participants were asked to rate its accuracy on first reading it, and again after they were informed about their trust game payoffs. Accuracy ratings were made on 101-point scales (0 = there is no chance the message is accurate, 50 = there is a 50/50 change the message is accurate, 100 = it is certain that the message is accurate).

Part 2. Part 2 consisted of one *gossiper-trust* game. After participants had played the 24th round and been informed of their payoffs, they were told that Part 1 was complete. They were then presented with the instructions for Part 2 and asked to complete five comprehension questions. In Part 2, participants played the one-shot trust game twice more. In the first trust game, they again took the role of Investor. They received the usual endowment of 10 ECU, were presented with a summary of the events in round 24 (i.e., the gossip that they received, their decision and the Agent's response and their respective payoffs) and asked how many ECU they wanted to send to the person who had sent them that gossip. To calculate payoffs, participants were then required to play a second trust game in the role of Agent. They received the usual endowment of 0 ECU, were told how many ECU they had been sent by the Investor they had sent gossip to in round 24, and asked how many they wanted to return. Their payoff for Part 2 was the sum of the payoffs from both trust games.

Post-Study Questionnaire. As a final activity, participants were asked to complete a short questionnaire that, among other things, measured participants' social value orientation (primary items, [27]), narcissism, psychopathy and Machiavellianism (Short Dark Triad, [28]; scale αs ranged from .71 to .80) and basic demographic information.

Results

Descriptive Statistics

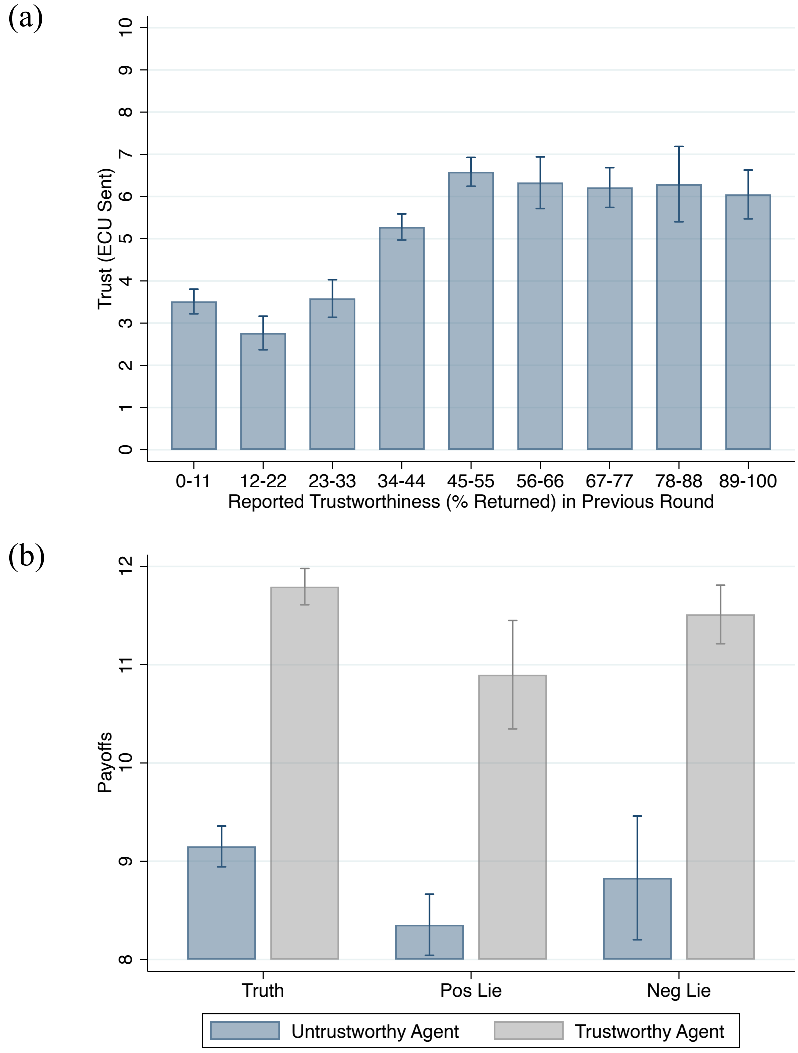
Rates of trust were reasonably high, with Investors sending a positive amount on almost 80 percent of rounds (*N*=3,057; ECU sent *M*=5.78, *SD*=2.67). On just over two-thirds of trust rounds (those in which the Investor sent a positive amount), Agents were trustworthy, returning at least one-third of the ECU they had received (*N*=2,093; *M*=45.91%, *SD*=15.95). On the remaining rounds, Agents were untrustworthy and returned fewer than one-third of the ECU they received (N=964; *M*=10.19%, *SD*=10.85). Investors on average earned *M*=10.37 ECU (*SD*=3.63) per round, and Agents *M*=8.83 ECU (*SD*=7.20).

Just over one-third (36% of *N*=3,680) of messages were lies. Comparing the difference between Agents' actual trustworthiness (computed as percentage returned) and the gossip describing this trustworthiness reveals that, on average, these lies were large. Just under half (47%) of lies were *positive,* claiming that *M*=6.39 ECU had been sent and *M*=11.12 returned, when in fact *M*=3.73 ECU had been sent and *M*=2.51 returned. A very similar proportion (49%) of lies were *negative,* claiming that *M*=6.39 ECU had been sent and *M*=2.60 returned, when in fact *M*=5.54 ECU had been sent and *M*=7.57 returned. The remainder (4%) were ambiguous and omitted from the lie typology. [2]

Is Dishonest Gossip Damaging?

We start by testing the preconditions that people (1) act on gossip and (2) are harmed when this gossip is dishonest. In order to test the first condition, we regressed the number of ECU that participants sent to the Agent as a function of 9 dummies that each corresponded to an interval of the percentage of ECU the Agent was said to have returned ([0%, 11%], [12%, 22%], …, [89%, 100%]). In this regression, as well as those that follow, regressions include random effects at the individual level and session level (where appropriate) and all p-values refer to two-sided tests.

Figure 2. Trust decisions and payoffs as a function of gossip honesty.



Notes: (a) graphs marginal means of trust in Agent as a function of reported trustworthiness; (b) graphs marginal means of payoffs as a function of gossip content and Agent type, where payoffs = 10 - tokens sent + tokens returned; Agent trustworthiness is that observed in the previous round and described by the gossiper; error bars denote 95% CIs.

As can be seen in Figure 2a, participants generally sent more tokens to Agents who were said to have been more trustworthy. Participants appeared to be most sensitive to messages in the range of moderately low trustworthiness to moderately high trustworthiness, with trust increasing significantly as Agents moved from 22-33% to 34-44%, *b*=1.70, χ^2 (1)=69.55, *p*<.001, and again as they moved from 34-44% to 45-55%, *b*=1.31, χ^2 (1)=86.22, *p*<.001. They were less responsive to messages claiming extremely low or extremely high trustworthiness, perhaps reflecting scepticism about extreme claims.

To test the second condition, we regressed participants' payoffs onto a dummy representing the Agent's previous trustworthiness (coded 1 if Agent returned at least 33% ECU, otherwise 0), a categorical variable representing the honesty of the gossip claiming to describe that trustworthiness (coded 1 if truth, 2 if positive lie, 3 if negative lie) and the two-way interaction between these variables. This showed that at least some lies decrease payoffs (see Figure 2b). In the case of previously untrustworthy Agents (who were generally disadvantageous to interact with), participants who acted on positive lies had significantly lower payoffs than those acting on the truth (*b*=-0.80, χ^2 (1)=17.44, *p*<.001), while those who acted on negative lies were not significantly disadvantaged (*b*=-0.32, χ^2 (1)=0.90, *p*=.344). In the case of previously trustworthy Agents (who were generally advantageous to interact with), participants who acted on positive lies had significantly lower payoffs than those acting on truth (*b*=-0.90, χ^2 (1)=9.12, *p*=.003), while those who acted on negative lies were again not significantly disadvantaged (*b*=-0.28, χ^2 (1)=2.51, *p*=.113). This pattern for previously trustworthy Agents differs somewhat from Study 1 and is due to the fact that the populations of Agents underlying each type of message were not identical (see Tables 2.4 and 2.5 in SI). We have no good reason for why this occurred, apart from random chance.

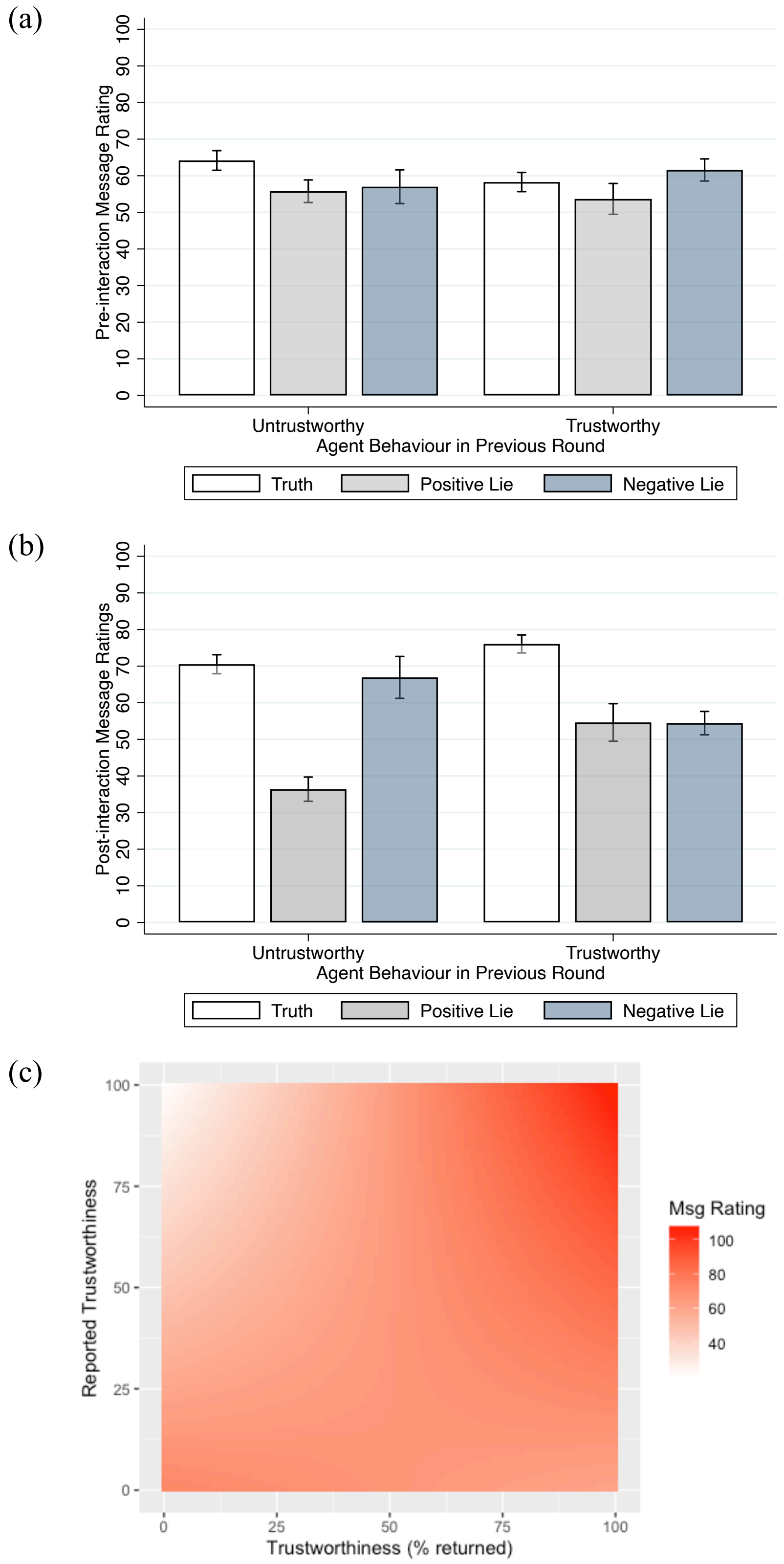
Detecting Dishonest Gossip

Our findings are again consistent with the possibility that dishonest gossip may erode the ability for gossip to bolster cooperation. We now test H2 and the claim that people will evaluate dishonest gossip as less accurate than honest gossip. To do this, we first regressed participants' ratings of message accuracy on their first reading of it (i.e., pre-interaction) onto the dummy representing Agent previous trustworthiness, the categorical variable representing the honesty with which this trustworthiness was described and their two-way interaction (codes as described above). The results are depicted in Figure 3a. In the case of previously untrustworthy Agents, participants rated honest messages as significantly more accurate than either positive ($\chi^2(1)=46.39$, *p*<.001) or negative lies ($\chi^2(1)=11.29$, *p*<.001). In the case of previously trustworthy Agents, participants again rated honest messages as more accurate than positive lies ($\chi^2(1)= 5.97$, *p*=.015) but as significantly less accurate than negative lies ($\chi^2(1)= 8.28$, *p*=.004). Additional analysis (see Supplementary materials) suggests that these findings are largely driven by participants' scepticism of behaviours that are said to deviate from the sample mean.

Next, we repeated this regression analysis for participants' ratings of message accuracy after their interaction with the Agent (see Figure 3b). In the case of previously untrustworthy Agents, participants rated honest messages as significantly more accurate than positive lies (*b*=34.13, $\chi^2(1)=417.70$, *p*<.001), but not negative ones (*b*=3.60, $\chi^2(1)=1.53$, *p*=.216). In the case of previously trustworthy Agents, participants rated honest messages as more accurate than either positive (*b*=21.44, $\chi^2(1)=69.54$, *p*<.001) or negative lies (*b*=21.62, $\chi^2(1)=192.00$, *p*<.001).

Thus, in line with H2, when participants had the chance to directly test the veracity of the gossip, they were reasonably adept at distinguishing between honest and dishonest gossip. Indeed, if we assume that participants would categorise gossip they rated above 50 percent as "honest" and that they rated below 50 percent as "dishonest" they would have been correct 62 percent of time (2,078/3,360) — significantly better than chance (*p*<.0001, binomial test).

Figure 3. Investor ratings of message accuracy as a function of gossip honesty.



Notes (a) graphs marginal means of accuracy ratings pre-interaction; (b) graphs marginal means of accuracy ratings post-interaction; (c) graphs fitted estimates of post-interaction accuracy as a function of reported (GossipPctReturned(t-1)) and actual trustworthiness (PctReturned(t-1)): Rating = b0 + b1 PctReturned(t-1) + b2 GossipPctReturned(t-1) + b3 PctReturned(t-1) x GossipPctReturned(t-1) + uj + vi + eijt; error bars denote 95% CIs.

Finally, in order to gain a better understanding of the way in which participants' ratings of gossip accuracy after their interaction were influenced by gossip about an Agent's *previous* trustworthiness as well as their *actual* trustworthiness in the game, we regressed post-interaction ratings onto actual Agent trustworthiness, the gossip about the Agent's trustworthiness in the previous round, and their two-way interaction. The fitted estimates from this regression are graphed in Figure 3c, where darker shades of red indicate higher ratings. From this, we can see that positive lies are detected more readily than negative lies (top left vs. bottom right). We also see a bias towards good news: honest messages about trustworthiness receive higher ratings than honest messages about untrustworthiness (top right vs. bottom left).

Treatment of Dishonest Gossipers

We have shown that participants are able to detect at least some dishonest gossip; we now ask whether, in line with H1, they treat these gossipers less favourably on the gossiper-trust game. To test this, we regressed the number of ECU sent to gossipers in Part 2 onto the dummy representing Agent previous trustworthiness (i.e., their behaviour in round 23), the categorical variable representing the honesty of the gossip describing this trustworthiness (received in round 24) and their two-way interaction (codes as described above). As participants only provided a single response, we did not include any random effects in this model (including Agent random effects did not change the results quantitatively). In this round, gossipers told 58 lies (36.25% of 160 messages), of which 31 concerned previously trustworthy Agents (3 positive and 28 negative) and 27 concerned previously untrustworthy Agents (22 positive and 5 negative). Results are summarised in Table 1.

This analysis provided little evidence for H1. In the case of previously trustworthy Agents, participants were more likely to trust honest gossipers than those who told negative lies, b=-2.39, F(1, 154)=9.43, p=.003; however, honest gossipers were not trusted more than the very small number of gossipers who told positive lies, *b*=-1.85, *F*(1, 154)=0.91, *p*=.341. In the case of previously untrustworthy Agents, participants were marginally *less* likely to trust gossipers who were honest than those who told positive lies, *b*=1.56, F(1, 154)=3.62, *p*=.059; however, honest gossipers were not trusted less than the very small number of gossipers who told negative lies, *b*=2.20, F(1, 154)=2.09, *p*=.150. Interestingly, participants were more likely to trust honest gossipers who described previously trustworthy Agents rather than previously untrustworthy ones, *b*=3.05, F(1, 154)=22.25, *p*<.001.

Table 1. Marginal effect estimates (SEs) of ECU sent to gossipers as a function of gossip content and Agent type.

	Truth	Positive Lie	Negative Lie
Trustworthy Agent	5.85 (0.44)	4.00 (1.88)	3.46 (0.62)
Untrustworthy Agent	2.80 (0.44)	4.36 (0.69)	5.00 (1.46)
N =160			

To better understand the basis for participants' treatment of gossipers, we regressed the number of ECU that participants sent to their gossipers in Part 2 onto the gossip about the Agent's previous trustworthiness that participants received in round 24 (calculated as the proportion of ECU returned), participants' post-interaction rating of the accuracy of that message, and their round 24 payoff (full regression results in Supplement). These variables accounted for 21.67% of the variance in gossiper trust, *F*(3,156)=14.39, *p*<.001. Importantly, this revealed a significant positive effect of gossip accuracy, *b*=0.02, *F*(1,156)=7.03, *p*=.009, such that participants trusted gossipers more if they *believed* the gossip — they could of course be mistaken in their belief (this coefficient is small, but because accuracy is measured on a 0 to 100 scale, it means that participants would send 2 tokens more to a participant if they were certain the gossip was accurate than if they said that there was no chance the gossip was accurate). Participants were also more likely to trust gossipers who said the Agent was more trustworthy, *b*=5.73, *F*(1,156)=24.67, *p*<.001, and (marginally) if they earned a higher payoff in the round, *b*=0.16, *F*(1,156)=3.66, *p*=.057. The latter two effects are consistent with the previously observed tendencies for participants to treat gossipers more favourably if they said an Agent had been more trustworthy or described an Agent who had actually been more trustworthy.

Discussion

People can detect deception. Specifically, in line with H2, we found that — after having interacted with their Agent — participants were more sceptical about dishonest than honest gossip. This scepticism was especially marked for positive lies, potentially because by encouraging trust they allowed people to test their expectations. In other words, it is hard to uncover lies that discourage the behaviour that would test them (i.e., negative lies). However, unlike Study 1, and contrary to H1, the ability to detect deception did not straightforwardly translate into gossiper (mis)trust. While participants were more trusting of gossipers they believed were honest, as well those who actually told the truth about previously trustworthy Agents, they appeared to be less trusting of gossipers who actually told the truth about previously untrustworthy Agents. This suggests that the social benefits of being (seen to be) honest may not outweigh the costs of sharing negative gossip. This conclusion must be caveated by the fact that the sample size for this analysis is rather small, relying as it did on a single round of spontaneous lies. Additionally, some types of lies were very infrequent in the data. It is therefore important in future work to subject H1 to better powered behavioural tests.

General Discussion

Social sanctions have been identified as a promising mechanism for incentivising honesty in gossipers. Specifically, if lying gossipers are treated less favourably than those who tell the truth, this undermines any personal incentives to lie. However, for this mechanism to work, it is first necessary for people to be able to detect deception. Across two studies, we provide evidence that people may be somewhat good at doing this. While the evidence from Study 1 was indirect, it did show that participants differed in their (hypothetical) treatment of honest and dishonest gossipers. Importantly, Study 2 showed that participants were on average more confident about honest gossip than dishonest gossip. We also found that people may be more likely to uncover positive lies than negative ones.

These findings suggest that the social sanctioning mechanism may be viable; the key question is whether people use it. The answer that our studies provide to this question is rather mixed. On the one hand, Study 1 showed that people were generally less willing to trust gossipers who lied than those who told the truth in a hypothetical environment. On the other hand, Study 2 showed that whether people were more willing to trust honest gossipers in a live interaction with material consequences may depend on the trustworthiness of the gossip target. That is, when Agents were trustworthy, honest gossipers were trusted more than dishonest gossipers, but when Agents were untrustworthy, honest gossipers were if anything trusted less. At the same time, participants' subjective perceptions of honesty did positively predict their gossiper trust. Together, this suggests that people's perceptions of gossipers' honesty may inform their decisions of how to treat them but are not alone in doing so. Both studies provide initial evidence that suggests that gossipers who share positive gossip, whether honestly or not, are trusted more than those who share negative gossip — despite the fact that participants appeared to be better at detecting positive lies.

This valence effect may in part reflect the fact that gossip about previously trustworthy Agents precedes more profitable interactions, and that people misattribute some part of their fortune to the gossiper. It may also reflect a tendency to make more favourable social judgments of gossipers who say nice things about others. Indeed, we found that the positivity of the gossip predicted trust even after accounting for confidence in its truthfulness and the profitability of the interaction. This finding is consistent with a body of work that has shown that gossipers who share more positive information about others are, generally, perceived to be more moral (and therefore trustworthy) than those who share negative information, because they are seen to be trying to help those they discuss [29, 30]. If this explanation is correct, then it suggests that people's social inferences about gossipers' motives for sharing positive and negative information may undermine social sanctions for some types of lies. The valence effect could also reflect a lack of ability of Investors to consider the counterfactual case, and appropriately appreciate that they could have done better out of an interaction had they attended to negative gossip and invested less.

What are the implications for the honesty of gossip? Our findings are broadly consistent with the possibility that there may be social incentives for honesty in everyday gossip but suggest that such a mechanism is imperfect. First, while participants are somewhat adept at detecting deception, both before and (especially) after testing their gossip-based expectations, they do make mistakes. This means that even if people *intend* to treat honest gossipers more favourably, it may not always be honest gossipers that are the recipients of this favourable treatment. Second, as noted above, it is possible that the social consequences of being seen as honest may not always outweigh those of being seen to do a target harm. If so, social incentives may work most effectively for promoting honest descriptions of trustworthy targets, but less well for promoting honest descriptions of untrustworthy targets. The absence of rewards for sharing negative gossip suggests that this gossip may provide a more honest signal of a target's reputation than positive gossip (which may say more about the gossiper than the target).

Of course, the above suppositions are limited by the fact that we did not actually test the impact of social sanctions on gossipers' tendencies to lie, and we focused on trust behaviours. In relation to the former, whether or not social sanctions increase the honesty of gossip is likely to depend not only on the existence of such sanctions but also on gossipers' beliefs that such sanctions are likely to occur (for evidence that target's beliefs can mitigate against gossip inaccuracy, see [10]). In relation to the latter limitation, trust decisions are informed by a range of considerations including the Agents' trustworthiness and the participants' distributional preferences; decisions to punish or reward may reflect a different set of considerations and could therefore be more (or less) effective at promoting honest signalling in gossip. At a more fundamental level, conclusions are also limited by a lack of power in Study 2, which may be responsible for the somewhat inconsistent findings across the two studies.

In short, we add to an increasing body of evidence that demonstrates that gossip can be an effective social mechanism to sustain cooperation in large and/or dispersed populations where reputation information is hard to acquire directly. We find that people are reasonably adept at detecting dishonest gossip, but that the efficacy of the mechanisms to discipline gossipers is hampered by an important valence effect: people do not like the bearers of bad news, even though bad news may help people to avoid bad interactions. Why that is the case, and how gossiper networks mitigate this negative effect is an open question.

References

1. Giardini, F. & Wittek, R. (2019). Gossip, reputation and sustainable cooperation. In F. Giardini & R. Wittek (Eds). *The Oxford Handbook of gossip and reputation*. New York, NY: Oxford University Press.

2. Alexander, R. D. (1987). *The biology of moral systems*. Routledge: London and New York.

3. Nowak, M. A. & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437: 1291.

4. Jiao, R., Przepiorka, W. & Buskens, V. (2021). Reputation Effects in Peer-to-Peer Online Markets: A Meta-Analysis. *Social Science Research*, forthcoming.

5. Przepiorka, W., Norbutas, L & Corten, R. (2017). Order without Law: Reputation Promotes Cooperation in a Cryptomarket for Illegal Drugs. *European Sociological Review,* 33 (6), 752-64.

6. Nowak, M.A. & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature,* 393, 573.

7. Barclay, P. (2012). Harnessing the power of reputation: Strengths and limits for promoting cooperative behaviors. *Evolutionary Psychology*, 10(5), 147470491201000509.

8. Smith, E.R. (2014). Evil acts and malicious gossip: a multiagent model of the effects of gossip in socially distributed person perception. *Personality and Social Psychology Review,* 18, 311-325.

9. Giardini, F. (2012). Deterrence and transmission as mechanisms ensuring reliability of gossip. *Cognitive Processing,* 13, 465-475.

10. Fonseca, M.A. & Peters, K. (2018). Will any gossip do? Gossip need not be perfectly accurate to promote trust. *Games and Economic Behavior,* 107, 253-281

11. Fehr, D., & Sutter, M. (2019). Gossip and the efficiency of interactions. *Games and Economic Behavior*, *113*, 448-460.

12. Hess, N.H. & Hagen, E.H. (2006). Psychological adaptations for assessing gossip veracity. *Human Nature,* 17, 337-354.

13. McAndrew, F. T., & Milenkovic, M. A. (2002). Of tabloids and family secrets: The evolutionary psychology of gossip. *Journal of Applied Social Psychology*, 32(5),

14. Peters, K. & Fonseca, M. A. (2020). Truth, Lies, and Gossip. *Psychological Science*, *31*(6), 702-714.

15. Zahavi, A., & Zahavi, A. (1999). *The handicap principle: A missing piece of Darwin's puzzle*. Oxford University Press.

16. Smith, E. R. (2014). Evil acts and malicious gossip: A multiagent model of the effects of gossip in socially distributed person perception. *Personality and Social Psychology Review*, 18(4), 311-325.

17. Lachmann, M., Számadó, S., & Bergstrom, C. T. (2001). Cost and conflict in animal signals and human language. *Proceedings of the National Academy of Sciences*, 98(23), 13189-13194.

18. Giardini, F., & Conte, R. (2012). Gossip for social control in natural and artificial societies. *Simulation*, 88(1), 18-32.

19. Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological Science*, *25*(3), 656-664.

20. Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Experimental Social Psychology*, *53*, 107-117.

21. Fehr, E. & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives,* 14, 159-181.

22. Gambetta, D. & Przepiorka, W. (2014). Natural and Strategic Generosity as Signals of Trustworthiness. PLoS ONE, 9(5), e97533.

23. Bond Jr, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin*, 134(4), 477.

24. Michaelian, K. (2010). In defence of gullibility: the epistemology of testimony and the psychology of deception detection. *Synthese*, *176*(3), 399-427.

25. Park, H. S., Levine, T., McCornack, S., Morrison, K., & Ferrara, M. (2002). How people really detect lies. *Communication Monographs*, *69*(2), 144-157.

26. Berg, J., Dickhaut, J. & McCabe, K. (1995). Trust, reciprocity and social history. *Games and Economic Behavior,* 10, 122-142.

27. Murphy, R. O., Ackermann, K. A., & Handgraaf, M. (2011). Measuring social value orientation. *Judgment and Decision Making*, *6*(8), 771-781.

28. Jones, D. N., & Paulhus, D. L. (2011). The role of impulsivity in the Dark Triad of personality. *Personality and Individual Differences*, *51*(5), 679-682.

29. Peters, K. & Kashima, Y. (2015). Bad habit or social good? How perceptions of gossiper morality are related to gossip content. *European Journal of Social Psychology,* 45, 784-798.

30. Caivano, O., Leduc, K., & Talwar, V. (2020). When is gossiping wrong? The influence of valence and relationships on children's moral evaluations of gossip. *British Journal of Developmental Psychology*, *38*(2), 219-238.

[1] These findings replicate those of [14], which find that misrepresentation lies are harmful but that exaggeration lies are not (and may even be beneficial to audiences).

[2] Ambiguous lies involved a participant who had avoided their Agent reporting that they had trusted their Agent who had then returned less than one-third of ECU received.

[3] Note, in rounds where participants avoided their Agent the Agent's actual trustworthiness was undefined; to include these rounds in our analysis, we gave them a trustworthiness score of zero (previous work supports this approach [14]).