Running Head: TRUTH AND LIES

Truth, Lies and Gossip

Kim Peters[12] and Miguel A. Fonseca[13]

[1] University of Exeter Business School,

Rennes Drive, Exeter, EX4 4PU, U.K.

[2] University of Queensland School of Psychology,

St Lucia 4072, Australia

[3] NIPE, University of Minho,

Campus de Gualtar, 4710-057 Braga, Portugal.

The authors contributed equally to this work and correspondence can be addressed to k.o.peters2@exeter.edu.au and/or m.a.fonseca@exeter.ac.uk

Word Count: Introduction and Discussion: 1313; Method and Results: 4511.

Abstract

It is widely assumed that people will share inaccurate gossip for their own selfish purposes. This assumption, if true, presents a challenge to the growing body of work that argues that gossip is a ready source of accurate reputational information and therefore is welfare improving. We test this inaccuracy assumption by examining the frequency and form of spontaneous lies shared between gossiping members of networks playing a series of one-shot trust games (N=320). We manipulate whether gossipers are or are not competing with each other. We show that lies make up a sizeable minority of messages, and are twice as frequent under gossiper competition. However, this has no discernible effect on trust levels. We attribute this to the finding that, one, gossip targets are insensitive to lies, and two, some lies are welfare enhancing. These findings suggest that lies need not prevent — and may help — gossip to serve reputational functions.

Word Count = 150

Key Words: Gossip, accuracy, lies, trust, competition, reciprocity

Truth, Lies and Gossip

> The man who comes with a tale about others has himself an axe to grind.
>
> Chinese Proverb

Within the body of cultural knowledge there are ample warnings, like that above, about the dangers of attending to gossip. Such warnings are present in the academic literature too. There, it has been argued that people will share inaccurate gossip for their own selfish purposes, such as undermining enemies, promoting allies or competing for mates (Hess & Hagen, 2006; McAndrew & Milenkovic, 2002; Mace, Thomas, Wu, He, Ji & Tao, 2018). However, this argument, if true, presents a challenge to the growing body of work that presumes that, as a ready source of *accurate* reputational information, gossip is able to bolster overall levels of cooperation (Dunbar, 1993).

Gossip is the class of communicated content that conveys information about the behaviours and characteristics of social actors (Smith, 2014; Peters & Kashima, 2015). As such, it has the potential to inform us about our social world, and the reputations of the people who inhabit it. If people act on the basis of the gossip they hear, cooperating more with those who are said to have behaved cooperatively in the past, then there are incentives for engaging in the costly cooperative acts that build positive reputation (Nowak & Sigmund, 1998; Wedekind & Milinski, 2000; Barclay, 2012). In other words, gossip may enable the indirect reciprocity that has been argued to boost cooperation in large social groups (e.g., neighbourhoods, organizations, online communities and

markets; Alexander, 1987; Nowak & Sigmund, 2005; for a discussion of indirect reciprocity theorising and gossip, see Giardini & Wittek, 2019).

However, if gossip is inaccurate, such that information about a person's cooperation in the past is a poor guide to their cooperation in the future, then this virtuous cycle could break down (Roberts, 2008; Smith, 2014; Giardini, 2012). In line with this possibility, a recent lab study (Fonseca & Peters, 2018; see also Fehr & Sutter, 2019) found that gossip was less effective at securing cooperation when there was (and was known to be) a high chance that messages would be misdelivered, and hence describe the wrong person's previous cooperation. Interestingly though, there was no evidence that gossip was less effective when this inaccuracy was spontaneously introduced by gossipers. Therefore, it is currently unclear whether inaccuracy — especially that which occurs naturally — is indeed one of the main threats to the capacity of gossip to fulfil reputational functions. To shed light on this, we need a better understanding of when and why gossipers lie.

In this paper, we summarise the results of a pre-registered behavioural study that was designed to build this understanding by examining the spontaneous lies shared between gossiping members of networks playing a series of one-shot trust games. We test whether competition between gossipers increases lying, a prediction supported by game theory (Crawford & Sobel, 1982), sparking a low-discrimination, low-trustworthiness and low-trust cascade. We test the following novel hypothesis:

H1: *As gossiper-audience competition increases, gossip will become less accurate*.

Further, replicating findings from a similar paradigm by Fonseca and Peters (2018), we also test the following hypothesis:

H2: *As gossip becomes less accurate, (a) there will be less gossip-based discrimination, (b) leading to less trustworthiness, (c) leading to less trust*.

We also examine the form and functions of gossipers' lies. This exploratory analysis reveals, for the first time, that lies are used to pursue negative and positive social welfare goals, and that some lies may serve reputational functions more effectively than truth.

Method

Participants

We recruited 320 participants (48% male, Age *M*=20.30, *SD*=3.93) from a pool of registered participants of the UQEEL lab at the University of Queensland and leaflets distributed on campus. We recruited the largest sample that was feasible in light of limits to participant availability and funding. The resulting sample included more than twice the number of participants per condition than previous work with this paradigm (Fonseca & Peters, 2018). Participants were paid an average of A$19.94 for their participation. This study received ethics approval from the University of Queensland. Pre-registration, materials, data, code and supplementary analyses are available at https://osf.io/k3jsk/?view_only=fc0ee0aad2254356b458c28c00cd32f9.

Procedure

The experiment was conducted in groups of 16 individuals who played 20 rounds of an anonymous trust game (Berg, Dickhaut & McCabe, 1995) on networked computers. Half of the participants in each group were allocated to the role of Investor and the remainder were allocated to the role of Agent. In each round, Investors decided how many of their 10-token endowment to send to their allocated Agent — a measure of trust. Agents received three times the number of tokens that were sent and decided how many

to return to the Investor — a measure of trustworthiness. Investors could, if they wanted to, avoid their Agent by sending 0 tokens. Participants knew that they would never play the same person twice in a row but that pairings were otherwise random.

At the end of each round, decisions and payoffs were presented on screen. To elicit gossip, Investors were asked to send a message to the Investor who would play with their Agent in the next round. In this message, they stated the number of tokens they had sent and the number their Agent had returned to them (the factual behaviour remained on screen for reference). From the second round onwards, Investors received the message that described how their new Agent had behaved in the previous round before deciding how many tokens to send. Agents knew that messages about their behaviour were exchanged, but never saw them (consistent with definitions of gossip as involving communications about *absent* third parties; Smith, 2014; Peters & Kashima, 2015).

To test whether competition could increase rates of lying, we allocated 10 groups of participants to the *competition* condition (the remaining 10 groups were in control). To achieve this, we assigned all participants a colour that they retained through the experiment (in each group, 4 Investors and 4 Agents were 'red' and 4 Investors and 4 Agents were 'blue'). In the competition condition, Investors competed along colour lines for a bonus. Specifically, the total payoffs accumulated by red and blue Investors over the course of the experiment were compared, and Investors belonging to the winning colour each received a bonus of A$5 (the losing colour received A$0). Colour had no payoff consequences for Investors in the control condition or for Agents in either condition (these participants received a flat bonus of A$2.50). Investors and Agents received different information about the bonuses in their group (neither was informed about the

existence of different conditions). In particular, Investors were always told about the basis for all bonuses (e.g., in control, that Investors and Agents receive $2.50), but Agents were only told about their own bonus allocation (i.e., $2.50) and informed that Investor bonuses may be calculated differently. Agents, therefore, had no reason for anticipating different Investor behaviour in competition and control.

After completing the experiment, participants were paid (payment equalled a show-up fee *plus* 3 randomly selected rounds *plus* a bonus). They were then asked to complete a short survey. Among other things, participants reported their social bonding with same and different colour Investors and Agents (4 sets of 3 items from Peters & Kashima, 2007; α=.86 to .92): I "had a social bond with [target]", "connected with [target]", "trusted [target]". Investors also responded to 3 open questions asking when and why they had described an Agent's behaviour "truthfully", "too positively" and "too negatively". Agents instead responded to 2 items about their reputation concern (r=.52, p<.001): "When deciding how many tokens to return, I thought about what the next Investor would think of me", "I returned more tokens than I wanted to in order to ensure that the next Investor would see me positively"; and 2 sets of 3 items about their expectations of discrimination from same and different colour Investors (collapsed into a 6-item scale: α=.92; note that here, and throughout the paper, discrimination refers to the extent to which an Investor bases their decision to trust an Agent on the information that they receive about this Agent's previous trustworthiness): "I think that [colour] Investors decided how many tokens to send me on the basis of the message they received about my behaviour", "I think that if [colour] Investor was told that I returned a small number of tokens, they would send me fewer", "I think that if [colour] Investor was told that I

returned a large number of tokens, they would send me more." Items were accompanied by 7-point scales (1=strongly disagree, 7=strongly agree).

Results

Descriptive Statistics

A slight majority of rounds involved a trusting Investor, who sent a positive number of tokens (74% rounds; sent *M*=6.97, *SD*=3.05), and a trustworthy Agent, who returned at least one-third of received tokens (71% of trust rounds; returned *M*=45.21%, *SD*=10.84). In the remaining rounds, Agents were either untrustworthy (harming the Investor by returning fewer than one-third of received tokens; returned *M*=11.36%, *SD*=9.85), or were expected to be untrustworthy and thus avoided: avoid rounds were usually preceded by a message stating that the Agent had been avoided (23% cases) or that the Agent had been untrustworthy (tokens returned *M*=19.00%, *SD*=28.81; 49% cases).

Any message that misstated tokens sent or returned was considered to be a lie. While Investors told the truth most of the time, as was observed in Fonseca and Peters (2018), a substantial minority of messages were lies (25.88%). These lies were substantial in size. In trust rounds, positive lies (*N*=171) claimed that an average of 7.93 tokens had been sent and 14.56 returned, when the actual values were 6.84 sent and 7.44 returned; negative lies (*N*=341) claimed an average of 7.87 tokens had been sent and 3.38 returned, when the actual values were 7.22 sent and 8.68 returned. In avoid rounds, when no tokens were sent, positive lies (*N*=150) claimed an average of 7.68 tokens sent and 10.48 returned, and negative lies (*N*=166) claimed an average of 8.07 tokens sent and 0

returned. For more detail relating to these analyses, and all those that follow, see Supplement.

Manipulation Check

To test of the effectiveness of our competition manipulation, we ran a 2 (condition: competition, control) x 2 (target colour: same, different) mixed ANOVA of Investors' reported social bonds. The main effects of target colour, $F(1,157) = 141.75$, $p<.001$, $\eta^2=.35$, and condition, $F(1,157) = 14.08$, $p<.001$, $\eta^2=.08$, were qualified by the two-way interaction, $F(1,157) = 69.74$, $p<.001$, $\eta^2=.20$. This revealed that the tendency for Investors to report a stronger social bond with ingroup (same-colour) than outgroup (different-colour) Investors was more pronounced in the competition condition (ingroup $M=5.29$, $SD=1.60$; outgroup $M=2.94$, $SD=1.30$) than in control (ingroup $M=3.60$, $SD=1.43$; outgroup $M=3.28$, $SD=1.45$). This suggests that the manipulation introduced a stronger intergroup dynamic in competition than control.

Hypothesis testing

We first report the tests of our hypotheses that participants would be especially likely to lie if they were competing with their audience and that this would lead to a breakdown in discrimination, trust and trustworthiness (pre-registered analyses produce identical findings and are reported in the Supplement). In order to understand how competition affected rates of lying, we ran a mixed effects regression of whether or not the message was a lie on a condition dummy, an audience group affiliation dummy and their interaction. In line with H1, lies were most prevalent when communicating with outgroups in competition. Specifically, in the competition condition, 43 percent of the messages that Investors sent to outgroup audiences were lies, while only 20 percent of

those they sent to ingroup audiences were, $\chi^2(1) = 192.19$, $p<.001$. In the control condition, in contrast, Investors were about as likely to lie to the ingroup as the outgroup (21 and 19 percent of messages, respectively; $\chi^2(1) = 1.03$, $p=.310$). It is interesting to note that even in the absence of incentives one out of every five messages was a lie — an almost identical figure to that observed by Fonseca and Peters (2018).

To see whether Investors were less likely to discriminate on messages when lies were more prevalent, we ran a mixed effects regression of Investors' trust on a condition dummy, a gossiper group affiliation dummy, the content of the message (expressed as proportion of tokens returned), as well as all two- and three-way interactions. In line with H2a, we found a significant three-way interaction between condition, gossiper group affiliation and message content, such that Investors' tendencies to act on gossip were more influenced by the gossiper's group affiliation in competition than control, $\chi^2(1)=5.18$, $p=.023$. Specifically, in the competition condition, a message that an Agent had returned half of received tokens (versus none) increased trust by an average of 3.39 tokens if it came from an ingroup gossiper, but only by 1.71 tokens if it came from an outgroup gossiper, $\chi^2(1)=22.94$, $p<.001$. In the control condition, there was no evidence that Investors responded differently to messages from ingroup or outgroup gossipers, $\chi^2(1)=.79$, $p=.375$, and the above message increased trust by 5.17 tokens on average.

This suggests that the association between an Agent's behaviour in one round and the number of tokens they were sent in the next was attenuated in the competition condition. To see if Agents took advantage of this opportunity to behave selfishly, we ran a mixed effects regression of the proportion of received tokens that Agents returned to Investors on a condition dummy. Unexpectedly, and contrary to H2b, Agents were about

as trustworthy in lower discrimination Competition networks (return M=34%) as they were in higher discrimination Control networks (return M=37%), $\chi^2(1)=1.50$, p=.220. This behaviour corresponded with our finding that Agents' self-reported concern for their reputation (control M=5.51, SD=1.44; competition M=5.31, SD=1.47) and expectations of discrimination (control M=4.89, SD=1.41; competition M=4.66, SD=1.52) were reasonably high and did not vary significantly with condition, all $t(156) \leq 0.98$, $p \geq .327$. In other words, the actual discrimination that Agents experienced did not appear to alter their beliefs about this discrimination. In light of our finding that competition Agents were no less trustworthy than control ones, it is unsurprising to find that, contrary to H2c, levels of trust did not appear to differ across condition either, $\chi^2(1)=0.21$, p=.647.

These findings are consistent with claims that lies are an important component of gossip, especially if there are material incentives for lying. However, while lies were 63% more prevalent in the competition condition (and levels of discrimination were commensurably lower), trustworthiness and trust were not significantly affected, which suggests that gossip targets are not especially adept at calibrating their levels of reputation concern to gossipers' tendencies to act on the gossip they receive. The non-trivial base rate of lies also suggests that a desire to mislead the audience is not the only reason that gossipers lie. Indeed, Fonseca and Peters (2018) noted that some gossipers provided unprompted descriptions of using lies for a range of purposes, including punishing or rewarding targets and boosting investment. To shed light on the range of purposes that lies may serve, we now systematically explore their form and function.

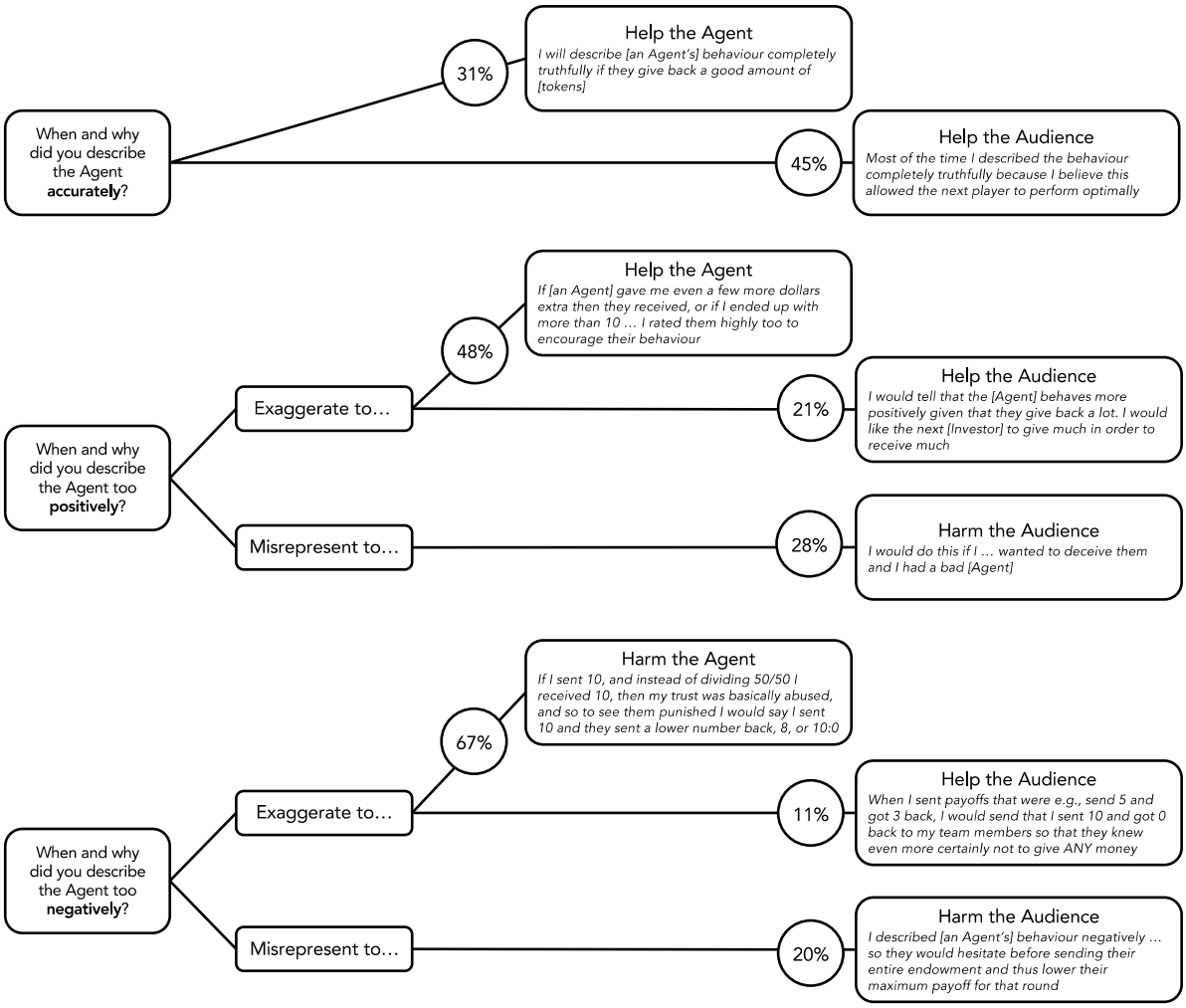
Form and function of gossipers' lies

We identified four main forms of lies. These reflected the interaction of two orthogonal dimensions: first, whether the Agent in question was trustworthy or untrustworthy (i.e., did they return at least 33% of received tokens, or did they return less than this), and second, whether the lie claimed that they were more or less trustworthy than they actually had been. *Positive misrepresentation lies* described untrustworthy or avoided Agents as more trustworthy than they actually were; *negative misrepresentation lies* described trustworthy Agents as less trustworthy than they actually were; *positive exaggeration lies* described trustworthy Agents as more trustworthy than they actually were; and *negative exaggeration lies* described untrustworthy or avoided Agents as less trustworthy than they actually were.

As a first step towards understanding why gossipers chose to share these different kinds of lies, we independently coded Investors' post-experimental explanations for their decisions to send accurate or inaccurate messages. This analysis revealed that between 80 and 98 percent of codable explanations related to social welfare motives — that is, a desire to send content to help or harm the Agent or audience (coder Kappas = .65 to .94). Social welfare motives representing more than 5 percent of explanations for a given type of message are summarised in Figure 1.

This analysis points to an important distinction between misrepresentation and exaggeration lies. The former type was solely justified by a desire to *harm* the audience (by encouraging behaviour likely to diminish their payoffs). In contrast, the latter type was either justified by a desire to *help* their audience (by encouraging behaviour likely to improve their payoffs) or to achieve *reciprocity* with the Agent (by encouraging

behaviour likely to improve the payoffs of trustworthy Agents and diminish those of untrustworthy Agents).



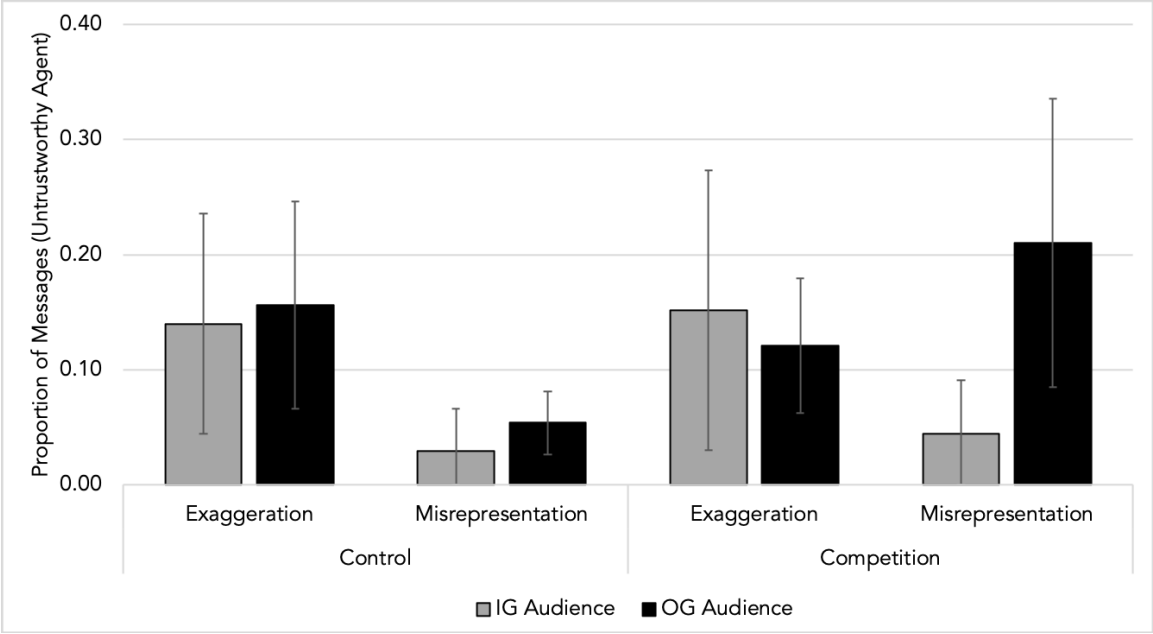
Notes: Percentage of participants using welfare themes when explaining when and why they used different content. Percentages calculated from Investors who reported sending a given type of message and provided a codable explanation: accurate $N = 122$, positive and negative inaccurate: Ns = 67, 79. Motives about the welfare of the Agent are on the left and motives about the welfare of the audience are on the right. Typical responses provided.

Figure 1. Social welfare themes in participants' explanations of their gossip content.

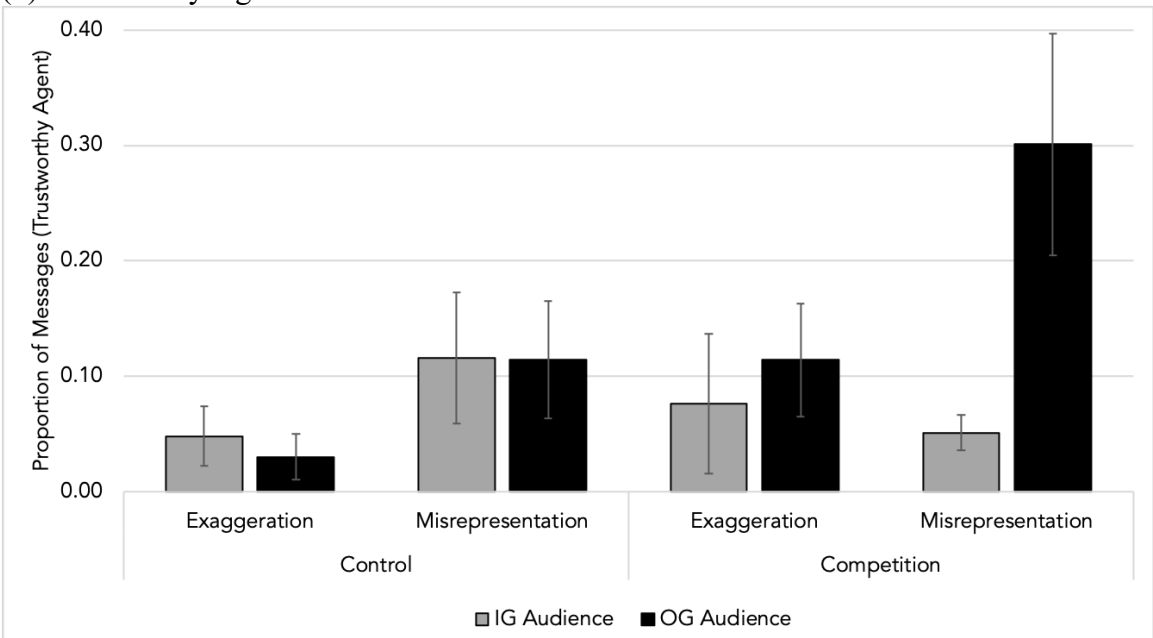
This suggests that the increased prevalence of lies in the competition condition should be primarily underpinned by an increase in misrepresentation (associated with the desire to harm competitor audiences). It also suggests that the lies that were sent to other (non-competitor) audiences should be primarily composed of exaggeration. To test this behavioural expectation, we used a multinomial logit regression of the type of lie on agent type (trustworthy vs. untrustworthy), a condition dummy, a gossiper group affiliation dummy, as well as all two-way and three-way interactions. We clustered standard errors at the session level to account for interdependencies (Wooldridge, 2003). Figure 2 displays the results of the estimation.

The analysis presented in text relates only to *trust* rounds, but the pattern in *avoid* rounds is generally consistent (see Supplement). In line with the possibility that misrepresentation lies are motivated by a desire to harm the audience, we found that gossipers were most likely to tell misrepresentation lies to outgroup competitors. Specifically, we found that Competition Investors made significantly more use of misrepresentation when messaging the outgroup than Control Investors did (untrustworthy Agents: $\chi^2(1) = 5.36$, $p=.021$; trustworthy Agents: $\chi^2(1) = 11.56$, $p<0.001$). Competition Investors also made significantly more use of misrepresentation lies when messaging the outgroup than the ingroup (untrustworthy Agents: $\chi^2(1) = 11.10$, $p<.001$; trustworthy Agents: $\chi^2(1) = 27.82$, $p<0.001$).

(a) Untrustworthy Agents



(b) Trustworthy Agents



Notes. Graphs summarise the estimated proportion of messages about (a) untrustworthy Agents and (b) trustworthy Agents that take the form of exaggeration or misrepresentation lies in trust rounds. Error bars represent 95% confidence intervals.

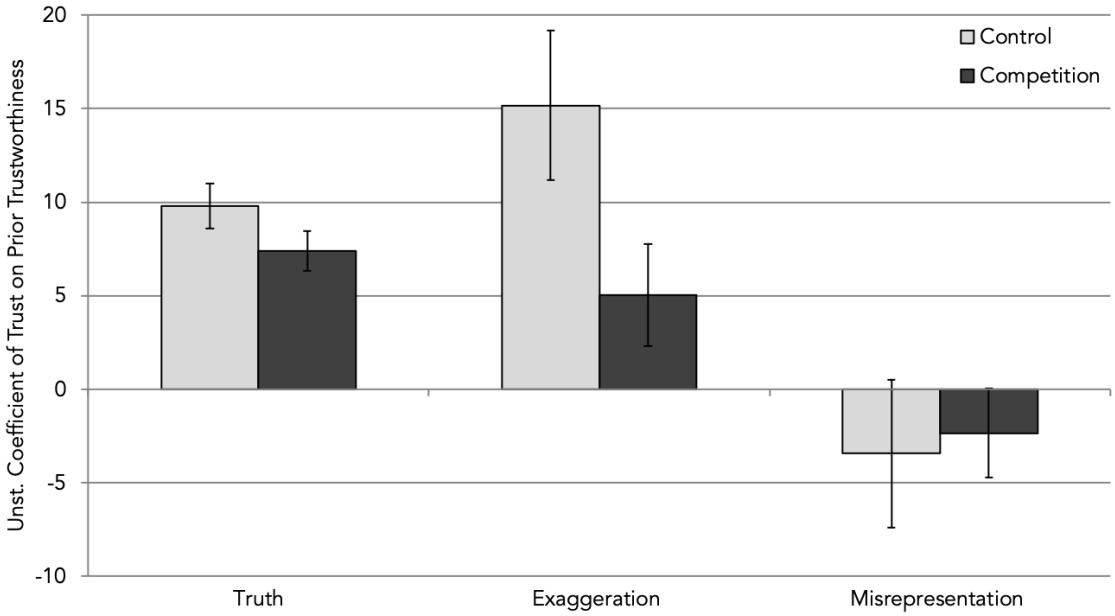
Figure 2. Prevalence of trust round lies as a function of lie type, condition and audience.

To examine the kinds of lies that were sent to non-competitor audiences, we first looked at ingroup audiences in the competition condition (with whom Investors reported having a strong social bond). Lies to this audience were never significantly more frequent than to outgroup competitors (tests for misrepresentation above; exaggerating trustworthiness: $\chi^2(1) = 2.47$, p=.116; exaggerating untrustworthiness: $\chi^2(1) = 0.32$, p=.574). However, we did find that when Competition Investors lied to the ingroup, they were more likely to exaggerate than misrepresent — although this difference was only significant when communicating about untrustworthy Agents, $\chi^2(1) = 5.61$, p=.018 (trustworthy Agents: $\chi^2(1) = 0.78$, p=.376). This is consistent with the possibility that Investors may have used exaggeration lies in an attempt to help the ingroup audience.

Next, we looked at the lies that Investors told in the control condition, which reveals a different pattern again. Here we find that Investors showed a preference for telling negative rather than positive lies (i.e., exaggerating untrustworthiness and misrepresenting trustworthiness) regardless of the identity of the audience (untrustworthy target: ingroup audience $\chi^2(1) = 4.80$, p=.029, outgroup audience $\chi^2(1) = 3.84$, p=.050; trustworthy target: ingroup audience $\chi^2(1) = 3.78$, p=.052, outgroup audience: $\chi^2(1) = 14.04$, p<.001). This negativity bias is robust across more stringent standards for trustworthiness (i.e., where an Agent had to return 40%, 45% or 50% tokens to be considered trustworthy), suggesting that it is not due to participants having higher standards for trustworthiness than we do (i.e. a 33% return rate).

Our analysis suggests that gossipers believe that misrepresentation impedes adaptive gossip-based discrimination, but that exaggeration facilitates it by helping audiences to reciprocate target behaviour. If true, this raises a novel possibility: that some

lies may actually support the indirect reciprocity that has been implicated in cooperation in large populations. To test this possibility, we analysed how the degree to which an Agent's trustworthiness in one round was reciprocated in the next was affected by whether the message about their behaviour was the truth or an exaggeration or misrepresentation lie. We ran a mixed effects regression of the tokens sent to an Agent on a condition dummy, two lie dummies (exaggeration or misrepresentation, truth as omitted category), the Agent's trustworthiness in the previous round (i.e., proportion tokens returned) and the two and three-way interactions between the condition, each lie dummy and trustworthiness. The reciprocation coefficients are graphed in Figure 3.



Notes. Error bars are 95% confidence intervals.

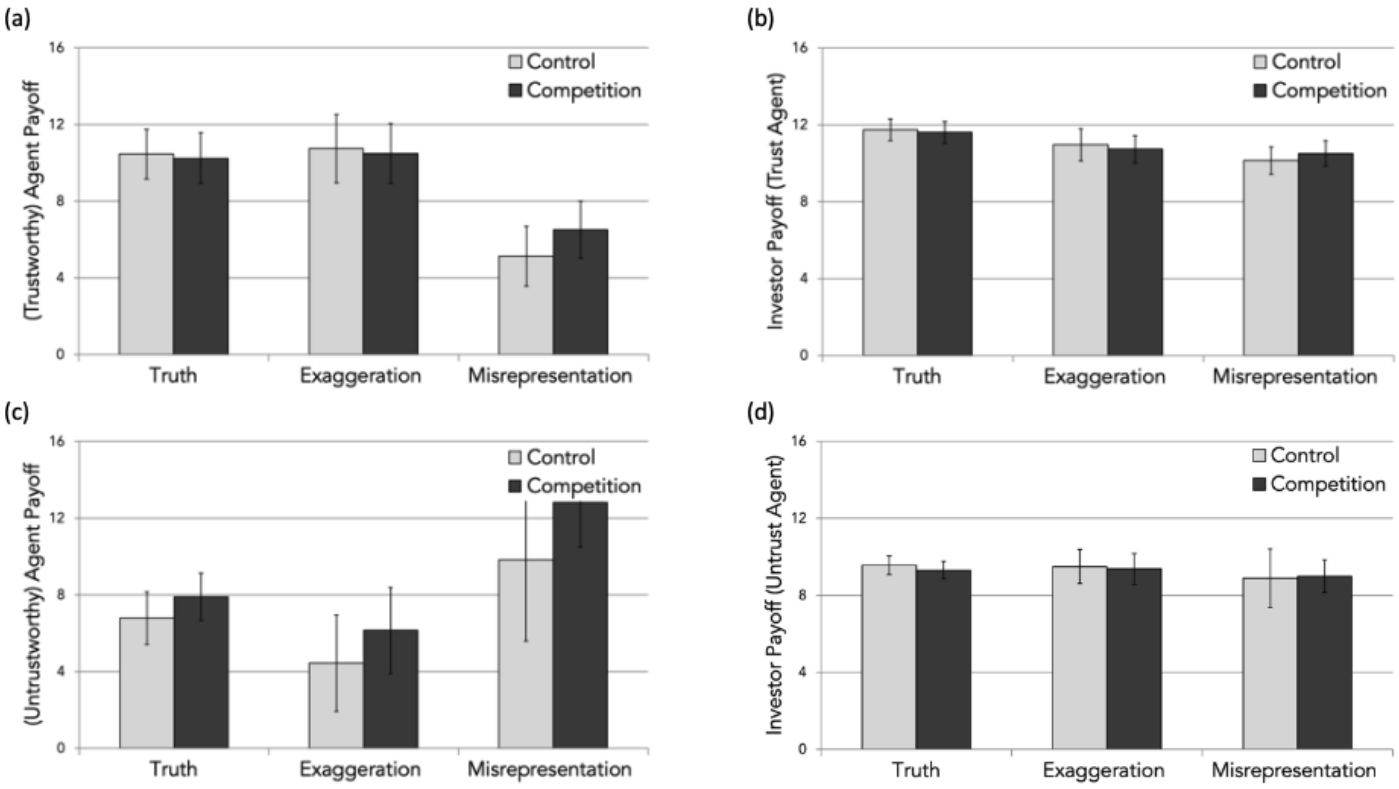
Figure 3. Reciprocation as a function of message type and condition.

Starting with the control condition, we find that the association between an Agent's trustworthiness and the extent to which they were subsequently trusted was

positive and significant when Investors told the *truth*, $\chi^2(1) = 253.50$, p<.001. Importantly, when Investors *exaggerated* the Agent's trustworthiness, this association was significantly stronger, $\chi^2(1) = 55.36$, p<.001. In contrast, when Investors *misrepresented* the Agent's trustworthiness, the association was weakly negative, $\chi^2(1) = 2.91$, p=.088. Turning to the competition condition, we find that the association between an Agent's trustworthiness and the extent to which they were subsequently trusted was positive and significant when Investors told the *truth*, $\chi^2(1) = 191.33$, p<.001, or *exaggerated*, $\chi^2(1) = 13.15$, p<.001; these do not significantly differ, $\chi^2(1) = 2.53$, p=.111. When Investors *misrepresented* the Agent's trustworthiness, the association was again weakly negative, $\chi^2(1) = 3.72$, p=.054.

In short, these results suggest that truthful gossip and exaggeration lies are both likely to ensure that Agents receive their just desserts, while misrepresentation lies are likely to prevent this. To test this possibility, we ran a mixed effects regression of Agent payoffs on a condition dummy, two lie dummies (see above), an Agent trustworthy type dummy and the two and three-way interactions between condition, each lie dummy and Agent type. The coefficients are graphed in Figure 4 (panels a and c).

This analysis reveals that payoffs to *trustworthy* Agents are significantly lower when their behaviour is misrepresented than when it is either truthfully described (control: $\chi^2(1) = 100.88$, p<.001; competition: $\chi^2(1) = 59.10$, p<.001) or exaggerated (control: $\chi^2(1) = 47.57$, p<.001; competition: $\chi^2(1) = 38.34$, p<.001). Payoffs from truth and exaggeration do not differ (control: $\chi^2(1) = 0.18$, p=.669; competition: $\chi^2(1) = 0.21$, p=.647).



Notes: Panel coefficients represent mean payoffs (a and c: Agent payoffs; b and d: Investor payoffs) as a function of Agent trustworthiness in the previous round (a and b: trustworthy Agents; c and d: untrustworthy Agents). Error bars are 95% confidence intervals.

Figure 4. Mean payoffs as a function of Agent previous round trustworthiness, message type and condition.

The reverse pattern is evident for payoffs to *untrustworthy* Agents in competition. Here, payoffs are significantly higher when their behaviour is misrepresented than when it is either truthfully described, $\chi^2(1) = 16.73$, $p<.001$, or exaggerated, $\chi^2(1) = 20.51$, $p<.001$; the latter messages do not differ, $\chi^2(1) = 2.30$, $p=.129$. In control, however, payoffs from misrepresentation are only significantly higher than those from exaggeration, $\chi^2(1) = 4.95$, $p=.026$, and not those from truth ($\chi^2(1) = 1.94$, $p=.163$).

Further, payoffs from truth are marginally higher than exaggeration, $\chi^2(1) = 3.19$, p=.074. Thus, exaggeration is as effective as truth at achieving positive reciprocity, and may be more effective at achieving negatively reciprocity.

As a final step, we ran this same analysis for Investor payoffs (Figure 4, panels b and d). When interacting with *trustworthy* Agents, Investors who acted on the truth received higher payoffs than those who acted on exaggeration (control: $\chi^2(1) = 5.59$, p=.018, competition: $\chi^2(1) = 10.53$, p=.001) or misrepresentation (control: $\chi^2(1) = 38.35$, p<.001, competition: $\chi^2(1) = 22.19$, p<.001). Investors who acted on exaggeration received higher payoffs than those acting on misrepresentation in control only, $\chi^2(1) = 4.16$, p=.041 (competition: $\chi^2(1) = 0.55$, p=.459). When interacting with *untrustworthy* Agents, Investor payoffs did not significantly differ on the basis of the content of the message in either control, all $\chi^2(1) \leq 0.74$, p≥.389, or competition, all $\chi^2(1) \leq 0.52$, p≥.471. Thus, truth (and, to a more limited extent, exaggeration) improves Investor payoffs relative to misrepresentation when Agents are trustworthy, but does not when Agents are untrustworthy.

These exploratory findings suggest that gossipers use lies to achieve a range of outcomes and challenge the widespread assumption that lies are necessarily malicious and harmful. Next, we describe the results of a study that aimed to independently verify gossipers' motives for telling the truth or exaggeration or misrepresentation lies.

Verifying the motives underlying gossipers' lies

We used an online experiment (*N*=81 UK-based adults, see Supplement for full details of method and results) that emulated the main study with a strategy method to independently verify the mapping of lies to social welfare motives. Participants

responded to an identical set of questions three times. In the first *neutral audience* iteration, participants were asked to imagine being an Investor in a population that played repeated trust games with Agents and exchanged messages about these interactions. Participants were then asked to imagine interacting with three Agents in turn: an untrustworthy Agent (sent 8, returned 3), a trustworthy Agent (sent 8, returned 10), and an Agent they chose to avoid (sent 0). In each case, they were asked to rate their social welfare motives towards (1) their Agent and (2) their audience (i.e., the next Investor to play their Agent; 7-point Likert scales: 1=strong desire to harm; 7=strong desire to help). They were also presented with three concrete messages that they could send to their audience (the truth, a positive lie and a negative lie) and asked to rate how much each message would allow them to jointly affect their Agent and audience as desired (8-point Likert scales; 0=not at all, 7=definitely). The content of the lies was based on the typical content sent about the three types of Agents in the main study, rounded to the nearest feasible integer.

This was followed by two *intergroup* iterations. Specifically, participants were informed that they had joined a new population that consisted of two teams of Investors that were competing to earn the most tokens. They were asked to respond to the above questions twice: once for non-competitor audiences (i.e., ingroup members), and once for competitor audiences (i.e., outgroup members). The order of audience was randomised.

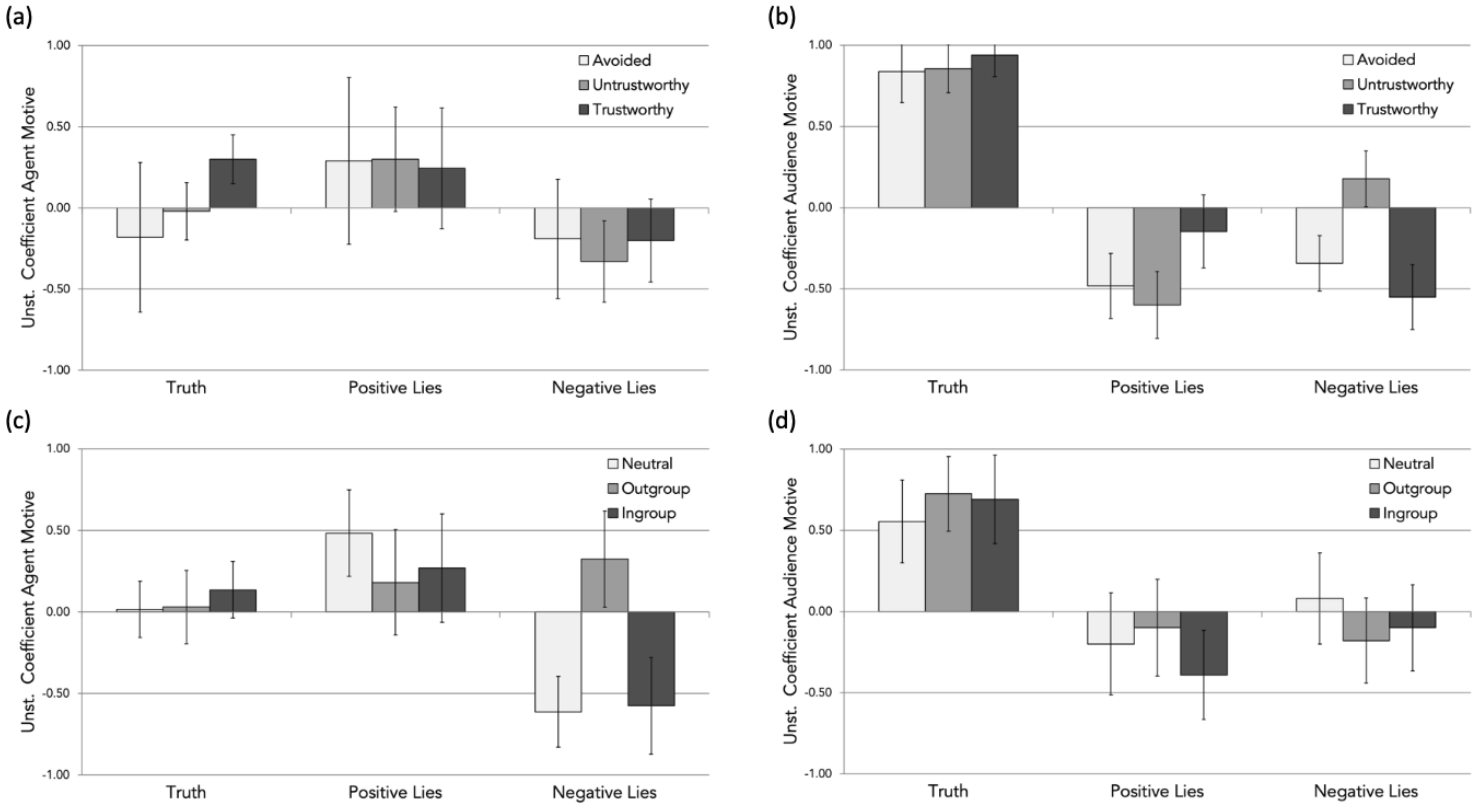
To check that the Agent and audience manipulations affected participants' social welfare motives in the expected ways, we first regressed participants' desire to help (versus harm) their Agent on two dummies representing Agent trustworthiness (trustworthy or untrustworthy, avoided as omitted category), clustering standard errors at

the level of the participant. This analysis revealed that relative to avoided Agents participants were less motivated to help untrustworthy Agents (b=-0.71, 95%C.I. -0.94 to -0.47, p<.001) and more motivated to help trustworthy ones (b=0.59, 95%C.I. 0.34 to 0.83, p<.001). We next regressed participants' desire to help (versus harm) their audience onto two dummies representing audience affiliation (ingroup or outgroup, neutral as omitted category), clustering standard errors at the participant level. This revealed that relative to neutral audiences participants were less motivated to help outgroup audiences (b=-2.16, 95%C.I. -2.53 to -1.80, p<.001) and more motivated to help ingroup ones (b=0.67, 95%C.I. 0.42 to 0.91, p<.001). The manipulations were therefore successful.

To understand participants' mapping of social welfare motives onto different messages, we ran regressions of participants' ratings that a given message would achieve their social motives for truthful messages and positive and negative lies in turn. To simplify interpretation, we ran the regressions once accounting for Agent type and a second time accounting for audience affiliation. In each case, we regressed message motive achievement onto participants' desire to help (versus harm) their Agent and help (versus harm) their audience, two dummies representing either Agent trustworthiness *or* audience affiliation, and the four two-way interactions between social welfare motives and the relevant dummies. Standard errors were clustered at the participant level.

The coefficients from these regressions are graphed in Figure 4. We first describe participants' perceptions that different messages can achieve their goals for the Agent's social welfare (panels a and c). The analysis reveals that the *truth* is seen to help trustworthy Agents, p<.001 (all other coefficients, p≥.133). *Positive lies* are seen to help Agents when the audience is neutral, p=.001 (all other coefficients p≥.074). *Negative lies*

are seen to harm Agents when they *exaggerate* their untrustworthiness, p=.012, or the audience is either neutral or an ingroup member, all p<.001. They are seen to help Agents when the audience is an outgroup member, p=.035 (all other coefficients $p \geq .126$).



Notes: Panel coefficients represent each message type's achievement of welfare motives (a and c: Agent motives; b and d: audience motives). Positive and negative coefficients represent message achievement of motives to help and harm, respectively. Panels (a) and (b) include Agent trustworthiness dummies; panels (c) and (d) include audience affiliation dummies. Error bars are 95% confidence intervals.

Figure 5. Unstandardised coefficients of message motive achievement as a function of motive target, message type, Agent trustworthiness and audience affiliation.

We turn now to participants' perceptions that different messages can achieve their goals for the audience's social welfare (panels b and d). The analysis reveals that the *truth* is seen to help the audience regardless of their affiliation or the Agent's trustworthiness, all $p<.001$. *Positive lies* are seen to harm the audience when they *misrepresent* an avoided or untrustworthy Agent, $p<.001$, or the audience is an ingroup member, $p=.006$ (other coefficients, $p\geq.206$). Finally, telling *negative lies* is seen to harm the audience when they *misrepresent* an avoided or trustworthy Agent, all $p<.001$, but to help the audience when they *exaggerate* an untrustworthy Agent, $p=.047$ (other coefficients, $p\geq.187$).

These findings are highly consistent with those of the main study. In particular, participants believed that truth would positively, and negative exaggeration negatively, reciprocate Agent behaviour. They also believed that truth and negative exaggeration would help, and misrepresentation harm, the audience. Participants appear to anticipate that outgroup members are less likely to act on their gossip than neutral and ingroup members, conditioning the ability of messages to achieve their social welfare motives.

Discussion

It is widely assumed that, given the opportunity and a selfish reason for doing so, gossipers will lie, and that this means that gossip is unlikely to support the indirect reciprocity that shores up cooperation. This study suggests that there may be many circumstances in which this assumption does not hold true.

Replicating findings from Fonseca and Peters (2018), we find that gossipers *do* lie (although not all gossipers, and not all the time), and that self-interested motives are not the only ones in play. Thus, Feinberg et al.'s (2012) finding about the importance of

prosocial motives for people's sharing of gossip extends to people's decisions to distort it. Specifically, when it comes to exaggeration lies, gossipers report using them to help the audience better discriminate between targets and thereby reciprocate the target's previous behaviour. In other words, it seems that exaggeration lies are the product of gossipers' attempts to actively engineer indirect reciprocity. This observation aligns with recent theorising on the evolution of communication that suggests that in noisy environments, where audiences may miss signals, exaggeration is an expected adaptation (Wiley, 2017). Misrepresentation lies, in contrast, are the prototypical harmful lie — believed to be useful for harming the audience, and therefore directed predominantly at competitors. The behavioural data generally supports gossipers' expectations about the impact of these different lies on the welfare of the target and audience.

The above analysis points to one important caveat to the aforementioned assumption. If lies take the form of exaggeration, then there is no reason to suppose that this should erode cooperation, and indeed it is possible that it may bolster it more than the truth. If, however, they take the form of misrepresentation, discriminating gossipers and their targets will experience paradoxical outcomes. Over time, gossipers should stop attending to gossip and targets should stop expecting rewards for cooperative behaviour. With this, the benefits of gossip should disappear. In terms of understanding the apparent robustness of cooperation in our study to the presence of lies, this presents one potential answer: the rate of misrepresentation may not have been high enough to make anomalous outcomes sufficiently frequent to produce behaviour change. Indeed, targets in this study did not even appear to have insight into the levels of discrimination in their network. Together with previous work that shows that targets are quick to exploit low

discrimination when it is explicitly pointed out to them (Fonseca & Peters, 2018; Fehr & Sutter, 2019), this suggests that the task of inferring discrimination on the basis of one's treatment is a difficult one.

These observations provide some basis for expecting that gossip may be more robust to lies in everyday life than is commonly supposed. First, it is unlikely that misrepresentation lies will dominate everyday gossip. Misrepresentation always comes with a cost: gossipers can either mislead the audience (and thereby do them harm) *or* reciprocate the target's behaviour; they cannot do both. In other words, the degree to which gossipers wish to harm their audience needs to be sufficiently high for them to sacrifice their powerful drive to achieve reciprocity (Fehr & Gächter, 2000). And second, as long as people believe that there is a sufficient chance that gossip is exchanged, is accurate and is attended to, they may adopt a risk averse strategy of being more cooperative than they would otherwise.

These claims are of course speculative, and are therefore worthy of future empirical attention. It is also important to acknowledge the limitations to our claims that are attached to our study's operationalisation. Specifically, our study was designed to test arguments that emerge from the literature on indirect reciprocity, which is primarily concerned with understanding cooperation in large networks that involve limited opportunities for repeated interaction. In everyday life, people belong to many networks that allow for repeated interaction with some (if not all) network members. In such networks, direct reciprocity may come into play, as well as a variety of other social motives. Specifically, in circumstances when gossipers are not anonymous, a gossiper's desire to enhance their status and build social bonds with their audience is, among other

things, likely to play an important role in their decisions to share gossip — and potentially to distort it (see Beersma & Van Kleef, 2012). How the multiplicity of motives that accompany gossip in many everyday circumstances will affect the dynamics that we describe here is an open question. It is also possible that patterns of lying that we observe for cooperative behaviour (which can be expected to vary over time, whether because of deliberate defection, error, or retaliation, e.g. Charness & Rabin, 2002) may differ from behaviours or characteristics that are less variable and therefore easier for audiences to verify. Further theorising in these domains is needed.

In sum, this paper suggests that gossip can be inaccurate, but that this is a far from fatal flaw. To understand how and why gossipers lie, and their social effects, we need to move beyond an assumption that these lies are necessarily malicious and harmful and consider the evident richness in their forms and functions.

Contributions

K. Peters and M.A. Fonseca contributed equally to all aspects of the study design, the collection and analysis of data and the manuscript write up. Both authors approved the final version of the manuscript for submission.

References

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

Barclay, P. (2012). Harnessing the power of reputation: strengths and limits for promoting cooperative behaviors. *Evolutionary Psychology,* 10, 868-883.

Beersma, B., & Van Kleef, G. A. (2012). Why people gossip: An empirical analysis of social motives, antecedents, and consequences. *Journal of Applied Social Psychology*, 42(11), 2640-2670.

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

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817-869.

Crawford, V.P. & Sobel, J. (1982). Strategic information transmission. *Econometrica,* 50,1431-51.

Dunbar, R.I.M. (1993). Coevolution of neocortical size, group size and language in humans. *Behavior and Brain Sciences,* 16, 681-735.

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

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

Feinberg, M., Willer, R., Stellar, J., & Keltner, D. (2012). The virtues of gossip: Reputational information sharing as prosocial behavior. *Journal of Personality and Social Psychology*, *102*(5), 1015.

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

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

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.

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

Kamenica, E. & Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review,* 101, 2590-2615.

Mace, R., Thomas, M.G., Wu, J., He, Q., Ji, T. & Tao, Y. (2018). Population structured by witchcraft beliefs. *Nature Human Behaviour,* 2, 39-44.

McAndrew, F.T. & Milenkovic, M.A. (2002). Of tabloids and family secrets: the evolutionary psychology of gossip. *Journal of Applied Social Psychology,* 32, 1064-1082.

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

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

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.

Roberts, G. (2008). Evolution of direct and indirect reciprocity. *Proceedings of the Royal Society B,* 275, 173-179.

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.

Wedekind, C. & Milinski, M. (2000). Cooperation through image scoring in humans *Science,* 288, 850-852.

Wiley, R. H. (2017). How noise determines the evolution of communication. *Animal Behaviour,* 124, 307-313.

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review*, 93(2), 133-138.