

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Posting and Reposting: The Reputational Costs of Spreading False Content Online

Permalink

<https://escholarship.org/uc/item/355145mx>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Marsili, Neri

Blond, Rocío Lana

Kneer, Markus

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Posting and Reposting: Investigating Reputation, Trust, and Deniability in Online Communication

Neri Marsili (neri.marsili@unito.it)¹, Rocío Lana Blond², & Markus Kneer³

¹Department of Philosophy and Education Science, University of Turin

²Faculty of Psychology, Complutense University of Madrid

³IDEA_Lab, University of Graz

Abstract

Reposting has been identified as a primary conduit for misinformation spread. We hypothesize that reposting is especially conducive to misinformation because it allows the sender to spread false content while incurring lower reputational costs compared to posting. Experiment 1 (N=293) tests this hypothesis by measuring how reposting falsehoods affects reputation. It compares the effects of *posting*, *reposting*, and *verbally reporting* the same false content. Sources that directly posted false content were judged significantly less reliable than sources that reposted or reported the same false content. Experiment 2 (N=256) tests robustness by replicating findings across platforms (Facebook and Twitter) and source types (private individuals rather than institutions). Original posters again suffered greater reputational damage than reposters. These results strongly support the idea that reposting comes with reduced accountability, suggesting that its reduced reputational costs might play an important role in facilitating misinformation spread on social media.

Keywords: misinformation; social media; reposting; reputation; plausible deniability

Introduction

The advent of digital communication has revolutionised the way we access information. As social media become a central hub for news consumption and content dissemination, there have been growing concerns about the perceived proliferation of misinformation in digital spaces (Allcott & Gentzkow, 2017; Lazer et al., 2018; O'Connor & Weatherall, 2018; Silverman, 2016). Although weighing positive and negative effects of these recent transitions has proven difficult (Acerbi et al., 2022; Haidt & Bail, ongoing; Scheufele et al., 2021) there is consensus that these changes pose distinctive threats to our information ecosystem.

Existing studies have focused primarily on documenting patterns of misinformation circulation, analysing the content characteristics that promote viral spread (such as negative valence), or the vulnerabilities that favour its consumption and intake (e.g. Acerbi, 2019; Pennycook & Rand, 2021). While this work is essential, there is a crucial driver of online misinformation spread that has been overlooked by researchers, legislators, and commentators alike: *the role played by the introduction of the reposting function, and its negative effects on the reputational incentives that motivate communicators to be truthful on social media.*

There is a trivial sense in which reposting enables the propagation of misinformation: on most platforms, reposting

is the most basic tool to redistribute and amplify a given message. Reposting facilitates spreading content regardless of its quality – it fuels both the spread of information and misinformation.

Nevertheless, there's some evidence suggesting that misinformation is more likely to make the cut: for example, we know that a majority of reposters (about 3 in 4) fail to scrutinize the veracity of the content that they repost (Sundar et al., 2025). But why would reposting facilitate the spread of misinformation in particular?

A potential explanation is that unlike direct posting, reposting comes with limited liability: in other words, REPOSTING IS CHEAP. If this is true, and reposting avoids some of *the costs* associated with spreading false messages, we should expect it to be conducive to misinformation. If reposting reduces the costs associated with sending a false message, its introduction on social platforms allows misinformation to circulate without the usual social attrition. This hypothesis can be made more precise with the aid of *costly signalling theory*, a framework that emerged in evolutionary biology to analyse the incentives that motivate senders to communicate truthfully.

The Costs of Communicating Falsehoods

As the name suggests, a key insight of costly signalling theory is that honest communication can be sustained by costs that deter dishonest signalling (Smith & Harper, 2003). Roughly speaking, signalling costs come in two flavours: material and reputational.

Material costs are tied to the production of a signal itself. A classic example is the peacock's vibrant tail. The tail signals fitness to potential mates: the more conspicuous the tail, the more attractive the peacock to peahens. Maintaining a huge and conspicuous tail, however, is metabolically expensive, and it increases risk of predation. Consequently, growing a huge tail is a signal that only a genuinely fit peacock can afford. Here, the costs explain why the signal is reliable, despite the incentive to deceive (Grafen, 1990; Zahavi & Zahavi, 1997). In human contexts, conspicuous consumption plays a similar role: owning a Lamborghini signals wealth precisely because it's a signal that only well-off individuals can afford (Griskevicius et al., 2007).

Reputational costs, by contrast, are tied to the *consequences* of sending a dishonest signal, typically consequences that are socially imposed (Webster et al., 2018). Paper wasps and house sparrows, for instance, appear to punish deceptive signallers by physical retaliation (Møller, 1987; Tibbetts & Dale, 2004; Tibbetts & Izzo, 2010). Other species, like ground squirrels and vervet monkeys, ignore or ostracise individuals who frequently emit unreliable signals (Cheney & Seyfarth, 1988; Hare & Atkins, 2001).

In human communication, too, reputational and social costs constitute key incentives to communicate truthfully and reliably, and play an important role in containing misinformation (Green, 2009, 2025; Reid et al., 2020; Graham, 2020; Kneer & Marsili, 2025; Marsili, 2026). We collectively invest a significant amount of social and cognitive energy in policing communication, holding speakers accountable for what they say, and keeping track of reliable sources (Sperber et al., 2010; Mercier, 2020). The threat of damaged credibility, social sanctions, and diminished influence within the community plays a key role in motivating speakers to be truthful and cooperative.

Strategic Deniability

A key point of departure between human and animal communication is the variety of strategies available to speakers who want to evade the reputational and social costs associated with sending a dishonest message. One in particular deserves attention here: *strategic deniability* (Pinker, 2007; Pinker et al., 2008; Lee & Pinker, 2010). A strategically deniable message is one that is sent to convey a certain meaning, only to later deny that that meaning was the one that the speaker intended to convey.

Paradigmatically, strategic deniability exploits indirect speech (or other forms of strategic ambiguity) to convey a message that remains “off the record”, so that the speaker can later disavow having meant it (Camp, 2018; Mazzarella, 2023). A speaker who chooses this strategy can reap the benefits of sending the intended message, while significantly diminishing the risk of incurring the relevant social and reputational costs (be it damaged credibility, legal repercussions, or other consequences) if their plan goes sideways.

Deniable speech comes in many forms, such as insinuation (Camp, 2018; Oswald, 2022), dogwhistles (Henderson & McCready, 2018; Saul, 2024), and literally true but misleading statements (Saul, 2012; Wiegmann et al., 2022, 2024). A strategy of particular interest involves *speech reports*. If I *claim* (e.g.) that a company is about to go bankrupt, I’m accountable for what I said. The same isn’t true if I merely *report* the same claim. If I say: “John told me that the company is going bankrupt”, I can later deny that I meant to communicate that it was true that they will go bankrupt – I

only reported John’s opinion, after all. Generalising, speech reports shield the speaker from some of the reputational costs associated with communicating false information (Bary, 2023; Faller, 2019; Pancheva & Rudin, 2019).

Crucially, the same seems to be true of reposts. During his first presidency retweeting a post containing false crime statistics about white people murdered by Afro-Americans (originally posted by a neo-Nazi account), Donald Trump faced public backlash. Confronted publicly on TV by Bill O’Reilly, he brushed off criticisms with confidence: “Bill, am I gonna check every statistic? All it was is a retweet. It wasn’t from me” (Colvin, 2016). Precisely like a speech report, a repost affords the speaker some degree of plausible deniability: in both cases, the sender can insist that they’re merely reporting someone else’s opinion (“It wasn’t from me”), and that they shouldn’t be held accountable for someone else’s mistakes (“a retweet is not an endorsement”, as the adagio goes).

Reposting: Cheap and Pervasive

We’re not the first to liken uncommented reposts to speech reports: the thesis that reposters merely “report” or “quote” the original post has already been explored in the literature (Arielli, 2018; Marsili, 2021). While there’s disagreement as to whether this classification best captures the main communicative point or function of reposting, there is consensus that reposting falls short of reasserting the target content, and that retweeters retain some degree of plausible deniability as to whether they meant to present the reposted message as content that they themselves endorse (Rini, 2017; Arielli, 2018; Pepp et al., 2019; Marsili, 2021; Frigerio & Tenchini, 2023; Michaelson et al., 2024).

Now, a crucial feature that distinguishes reposting from traditional forms of deniable speech like speech reports – and a feature that makes it particularly dangerous – is its extremely low cost of production. Explaining how a cabal of Satanist, cannibalistic paedophiles controls US politics (as the QAnon conspiracy theory has it) is a complex task, which requires time and cognitive investment. Reposting QAnon conspiracies or anti-vax propaganda, by contrast, takes a single click.

This practical affordability (i.e. minimal *material costs*) has led reposts to be strikingly **pervasive** on social media: in 2021, retweets constituted approximately half (49%) of all content posted on Twitter (now X, McClain et al., 2021), a proportion that has been steadily rising since the feature was first introduced on Twitter in 2009 (Liu et al., 2014). By contrast, reported speech constitutes a far smaller proportion of offline linguistic exchanges¹. The pervasiveness of reposting is bound to have an impact on the reputational infrastructure of online communication, because it creates a communicative environment where cheap signals (signals

¹ To get a sense of how much smaller this proportion is, we ran a frequency analysis on an existing Twitter corpus (Quintana et al., 2022) using a dictionary that detects evidential constructions, including speech reports. We found that only 0.11% of the items in

the corpus were evidential expressions: speech reports, in turn, are a subset of this proportion. This only gives us a rough estimate, but (with a ~400-fold difference) it’s enough to support the conjecture that reporting is way less frequent than reposting on social media.

associated with low reputational costs) are much more pervasive than in offline speech.

The deadly combination of **low costs of production, low reputational costs, and high pervasiveness** creates what we call a ***misinformation-prone reputational infrastructure***: a communicative environment where the usual reputational incentives against inaccurate signalling are significantly reduced – that is, an environment conducive to higher circulation of misinformation. Crucially, despite its potential to identify a key driver of online misinformation, this hypothesis has never been empirically explored.

Our hypothesis that REPOSTING IS CHEAP (in the sense that it comes at reduced reputational costs), in turn, was based on some further assumptions: (i) that deniable speech more generally comes at reduced reputational costs, and (ii) that reposting can be likened to other forms of deniable speech, in particular speech reports. While (i) already enjoys significant empirical support (Mazzarella et al., 2018; Bonalumi et al., 2019, 2020, 2023; Hall & Mazzarella, 2023) (ii) remains unverified.

In our first experiment, we tested and found support for our conjecture that REPOSTING IS CHEAP, in addition to its motivating premise (ii). Our interest, in other words, was in verifying the following two hypotheses (H1 was our main concern; H2 the broader underlying assumption):

H1: REPOSTING IS CHEAP: False *reposted* claims should damage one's reputation *less* than *direct claims*

H2: REPORTING IS CHEAP: False *verbally reported* claims should damage one's reputation *less* than *direct claims*

Experiment 1

Participants, Methods, and Materials

To test our hypotheses, we recruited 305 participants on Prolific, restricting participation to US and UK residents who reported English as their native language. In line with the preregistered criteria,² 12 participants who failed at least one of two comprehension checks were excluded, leaving a sample of 293 participants (female: 67%; age $M=42$ years, $SD=13$ years, range: 21–72 years).

In our experiment, we presented participants with a scenario and contrasted the perceived costs of posting and either verbally reporting (Blog version) or reposting (Facebook version) the same content. The Blog version of the scenario, in which the blog *Gaudía* posts the original claim, and the blog *Barcinews* merely reports it, reads:

Gaudía, a popular blog covering local news in Barcelona, has recently published a post about the closure of a historically important restaurant. The blog post contains the following sentence:

The beloved restaurant "La Masia" is closing its doors for good this Tuesday, after over 150 years of continuous activity.

Later that day Barcinews, another popular blog that covers local news in Barcelona, reports Gaudía's claim in one of their posts. Barcinews's blog post contains the following sentence:

The beloved restaurant "La Masia" is closing its doors for good this Tuesday, after over 150 years of continuous activity, according to a report by Gaudía.

However, it turns out that La Masia didn't even consider shutting its activities. But as a result of the rumour, La Masia ends up losing a lot of customers, and goes through a period of financial hardship.

Participants were then asked to assess the comparative *reliability* of the two sources, their *willingness to donate* to one of them, and – in an explorative vein – whether Barcinews should be understood as having *claimed* the relevant content (i.e. that the restaurant is closing). Below are the questions (excluding the labelling), with forced-choice response options in brackets.

- Q1: Based on what you have read so far, which source do you think is less reliable when it comes to information about restaurants in Barcelona? [Gaudía/Barcinews]
- Q2: You have been given a 10€ coupon that you can only donate to Barcinews or Gaudía. Who would you donate the money to? [Gaudía/Barcinews]
- Q3: Barcinews claimed that La Masia would close its doors for good on Tuesday. [Agree/Disagree]

The Facebook condition was near-identical, except that Gaudía and Barcinews are introduced as Facebook pages, and Barcinews *reposts* Gaudía's *post*. Accordingly, "Blog" was substituted by "Facebook page", and the characteristic visual depiction of Facebook posts was imitated.

In line with our hypotheses, we preregistered two predictions: that a significant majority of participants should deem the original poster less reliable (P1) and should prefer to donate to the second-hand source (the reporter in the Blog condition and the reposter in the Facebook condition) (P2).

Results

The proportions of responses for our main questions are visually represented in Figure 1. All three preregistered hypotheses found support in our results. Consistently with P1, the original poster, Gaudía, was perceived as less reliable by significantly more participants than expected by chance, both in the Facebook (repost) condition ($z = -7.51, p < .001$) and in the Blog (verbal report) condition, ($z = -5.87, p < .001$). The difference in perceived reliability across sources did not differ significantly across medium type (Facebook v. Blog), $\chi^2=1.32, p=.251$.

The same patterns emerged with respect to people's willingness to donate to one of the two sources. Consistently with P2, the original poster Gaudía elicited significantly fewer donations than expected by chance in the Facebook

² See <https://aspredicted.org/rqk4-7hgq.pdf>

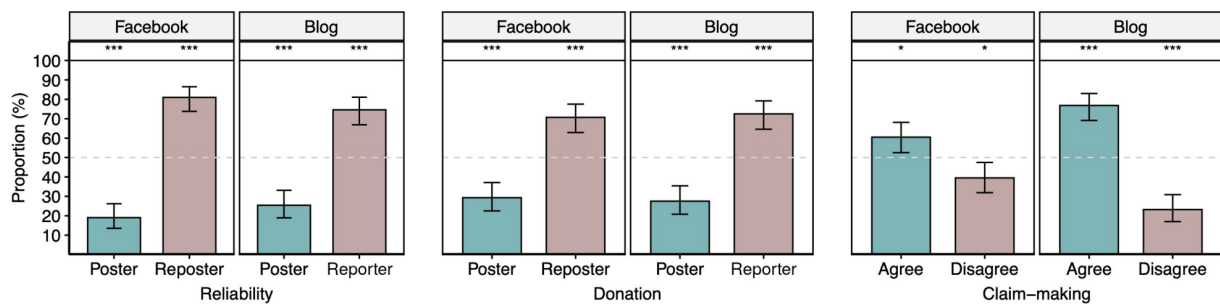


Figure 1 Proportions of responses split by medium type and dependent variable. Error bars denote 95% Agresti–Coull confidence intervals. Significance levels for the difference from chance. * $p < .05$; ** $p < .01$; *** $p < .001$.

condition ($z = -5.03$, $p < .001$) and the Blog condition ($z = 5.37$, $p < .001$). The difference in willingness to donate did not differ significantly across type of medium (Facebook v. Blog), $\chi^2 = .04$, $p = .837$.

Turning to our explorative question, i.e. whether people considered that Barcinews *claimed* the relevant content (that the restaurant was closing), a significant majority (69%) responded in the affirmative both in the Facebook scenario, where Barcinews reposted the content, ($z = 2.56$, $p = .011$) and the Blog scenario, where Barcinews verbally reported it ($z = 6.38$, $p < .001$), though participants were significantly more willing to consider the Blog report a claim than the Facebook repost, $\chi^2 = 8.07$, Cramér’s $V = .165$, $p = .005$.

Discussion

Our findings support our preregistered predictions. Reposting false content yields less reputational damage than directly posting it: indeed, REPOSTING IS CHEAP (as per H1), as evidenced by measured *perceived reliability* and *willingness to donate*. The same was found true about *reporting*, which supports our assumption that REPORTING IS CHEAP (H2). Finally, we found no difference between source type, lending support to our conjecture that the social costs of reposting and reporting are comparable.

It’s worth noting that even though speech reports can be marked in various forms, our experiment tested only one species: a sentence-ending parenthetical. Parentheticals in different linear positions (initial or medial) and different reporting constructions (e.g. *says Gaudia*, *Gaudia has claimed that*, etc.) might in principle deliver slightly different results. While testing alternative constructions is unlikely to challenge our finding that REPORTING IS CHEAP, it’s important to qualify our claim that reposting and reporting affect reputation in similar ways: we found that this holds for a particular kind of speech report, namely sentence-ending parentheticals.

Relatedly, we found a difference in how reports and reposts are classified: participants were less inclined to describe the Facebook repost as a claim than the reported text on the blog. Despite this difference, however, a significant majority still likened the Facebook repost to making a claim. This is somewhat puzzling: we saw that scholars agree that reposting *falls short* of claiming something. Could this surprising result simply be an artifact of the design? This possibility shouldn’t be ruled out. Perhaps participants took disagreeing with Q3 (“Barcinews claimed that La Masia would close its doors for

good on Tuesday”) to mean that the reposter *in no way communicated* that La Masia would close. If this is right, their preferred choice (“agree”) might simply indicate an unwillingness to suggest that the reposter is not responsible for disseminating the false claim – which is understandable. If this conjecture is correct, a more informative framing (for instance, one contrasting *stating* with *merely reporting*) should elicit different response patterns.

Another limitation of our vignettes is that they featured a specific kind of source: blogs and Facebook pages dedicated to local news. Now, it’s unclear whether news sources and professional pages are held to the same epistemic standards as individuals. If not, our findings might not generalise to the vast majority of social media users who share content in their personal capacities. In a similar spirit, we might also wonder whether our findings on reposting are specific to the reputational dynamics of Facebook, or whether they generalise to yet other platforms.

To address these concerns, we designed a second experiment, shifting focus from news sources to private individuals, and broadening scope from Facebook to X (formerly Twitter). We also revised the exploratory question to test whether laypeople preferentially classify reposts as claims also when better alternatives are available.

Experiment 2

Materials and Methods

We recruited 270 participants on Prolific, restricting participation to US and UK residents who reported English as their native language. In line with the preregistered criteria³, 14 participants who failed at least one of two comprehension checks or weren’t native English speakers were excluded, leaving a sample of 256 participants (female: 50%; age $M = 40.89$ years, $SD = 12.71$ years, range: 18–79 years).

The scenario again explored potential differences in *type of communicative act* (within-subjects: posting v. reposting) and *medium type* (between-subjects: Facebook v. X, formerly Twitter). There were thus two conditions, differing according to type of medium, to which participants were randomly assigned. The text of the vignette was adapted from the Facebook condition from experiment 1, introducing variations to match medium (for the Twitter condition) and source type: in this version of the scenario, a private individual (Giuseppe), introduced as a social media contact

³ See <https://aspredicted.org/4n2w-82b7.pdf>

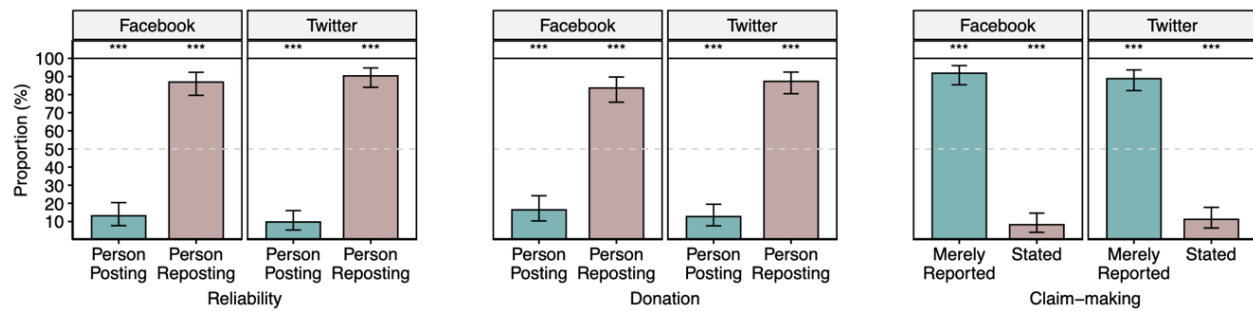


Figure 2: Proportions of responses split by medium type and dependent variable. Error bars denote 95% Agresti–Coull confidence intervals. Significance levels for the difference from chance. * $p < .05$; ** $p < .01$; *** $p < .001$.

of the participant, posts a message, and another individual (Francine) reposts it⁴. The same questions about responsibility and willingness to donate (Q1-Q2, adapted to individuals), were asked, followed by a new version of Q3:

Please complete the sentence by selecting the option that best describes what Francine did:

Francine [stated/merely reported] that “La Masia” was closing its doors for good.

We preregistered two predictions: that a significant majority of participants would deem the original poster less reliable (P1) and would instead prefer to donate to the reposter (P2).

Results

The proportions of responses for our key questions are visually represented in Figure 2. All our preregistered predictions were matched by our findings. In line with P1, the poster, Giuseppe, was perceived as reliable by significantly fewer participants than expected by chance in both the Facebook ($z = -8.15$, $p < .001$) and the Twitter condition ($z = 9.33$, $p < .001$). Perceived reliability of sources did not differ significantly across type of medium (Facebook v. Twitter), $\chi^2 = .44$, $p = .507$.

As regards people’s willingness to donate to one of the two sources (P2), the poster Giuseppe again elicited significantly fewer donations than expected by chance in the Facebook condition ($z = -7.42$, $p < .001$) and the Twitter condition ($z = 8.64$, $p < .001$). Again, willingness to donate did not differ significantly across type of medium (Facebook v. Twitter), $\chi^2 = .442$, $p = .506$.

Turning to the revised version of our explorative question, i.e. whether people considered the reposter as making a claim in the first place, we found that a significant majority thought that they merely reported a claim in the Facebook scenario ($z = 9.23$, $p < .001$) and the Twitter scenario ($z = 8.98$, $p < .001$). Participants were equally willing to consider the Facebook and the Twitter reposts as “mere reports” (rather than statements), $\chi^2 = .355$, $p = .551$.

Discussion

Experiment 2 addressed two potential concerns about the external validity of our findings. First, we tested whether the reduced reputational costs of reposting extend to private

individuals sharing content within their personal networks. Second, we examined whether the effect holds across different platforms, comparing Facebook with Twitter (now X). On both counts, the hypothesis that REPOSTING IS CHEAP proved robust: original posters suffered significantly greater reputational damage than reposters across all conditions, with no significant differences between platforms.

The revised exploratory question yielded notably different results from Experiment 1. Responses now aligned with the scholarly consensus that reposting falls short of claiming, with a (much larger) majority agreeing that the reposter “merely reported” the original claim. This reversal is not surprising in light of our conjecture that Experiment 1 failed to offer a sufficiently meaningful alternative to classifying the repost as a claim. Once participants are offered a way to acknowledge the reposter’s communicative contribution, their inclination to classify a repost as a statement dissipates. This framing appears to better capture participants’ intuitions, and lends some indirect support to theoretical accounts (like Marsili 2021) that characterise reposting as a form of quotation or reported speech.

General Discussion

Across two experiments, we found consistent support for our conjecture that REPOSTING IS CHEAP. Reposting false content incurs significantly lower reputational costs than directly posting the same content, as expressed by participants’ ascriptions of reliability and willingness to donate. Our findings were consistent across types of media, notably Facebook and Twitter/X, and proved robust across source type (public institutions and private individuals). Furthermore, we found evidence that people perceive reposting as a form of “mere reporting” (rather than reassertion), which can explain why we found that they both carry analogously reduced social costs.

Significance

Our findings have significant implications for understanding how and why misinformation propagates on social media. On top of carrying low costs of production and low reputational costs, reposts are strikingly pervasive online. We already noted that this is an particularly concerning combination: since reposts have become the most common form of

⁴ The (inferred) gender of the (re)posters was not counterbalanced. Across several pilots run with the gender reversed, however, the same pattern of results was obtained, indicating that

the reputational asymmetry does not depend on the inferred gender of either source.

communication on social media, these platforms have become environments where a dramatically high proportion of signals come with reduced reputational costs, compared to other environments (where reposting is either not present or not pervasive). This leads to a *misinformation-prone reputational infrastructure*: a communicative environment where the usual social incentives against inaccurate signalling are significantly weakened, so that misinformation can circulate with diminished social friction.

Note that this diagnosis doesn't postulate malevolent actors behind misinformation spread. In line with any analysis of human behaviour that emphasises the importance of social incentives in cooperation (Bicchieri, 2005, 2017), all that we assume is that individuals can be *more* prone to cooperate when incentives nudge them in the right direction. Of course, malevolent trolls and propagandists can reap the benefits of a communicative environment where sending misinformation becomes cheap. But removing social friction against communicating falsehoods also means, more mundanely, that negligent conduct is likely to become more common. If people have weaker motivations to verify the veracity of what they repost, they will likely do it less often, leading to accidental misinformation spread. This matches what empirical studies have found: misinformation sharing is often driven by bad habits rather than malevolence (Ceylan et al., 2023), such as not checking the links that one reposts (Sundar et al., 2025), leading to crowdsourced, fast misinformation spread (Vosoughi et al., 2018).

We lay no claim that this is the whole story about what drives misinformation on social media: even limiting our attention to social incentives, the picture is inevitably much more complex (Reimann, 2022; Marsili, 2026). Our warning, instead, is that any story that fails to appreciate the importance of the affordability and pervasiveness of reposts is bound to be dramatically incomplete. More worryingly, a story so incomplete runs the risk of failing to identify practicable interventions, like artificially reintroducing social friction. Far from being a theoretical possibility, interventions of this sort have already been explored (Jahn et al., 2025) – as when Twitter engineers in 2020 made quote tweets the default option for retweeting (Hatmaker, 2020), and introduced a pop-up window prompting users to read articles before sharing links they hadn't opened (Vincent, 2020), with some degree of success (Tameez, 2020).

Limitations and Future Avenues of Research

A limitation worth mentioning is that while our studies explicitly stipulated that the shared content was false, in real-world contexts receivers lack direct access to the facts of the matter. Instead, they evaluate reliability against their own beliefs: a source is deemed unreliable when their claims conflict with what the receiver takes to be true. This means that sending false messages might actually improve the sender's reputation in some contexts – for example, in polarised information environments where misinformation aligns with receivers' prior beliefs (Osmundsen et al., 2021).

Even in contexts where these complications hold, however, reputation will be scored following the patterns we observed – the only caveat is that people rely on their beliefs as a proxy for truth when they assess whether a message is truthful, and that communicators are aware of this when they calculate the expected reputational effects of the sender's communicative act. This in no way challenges our findings. It merely means that translating them into a full account of the incentives that motivate communicators to be truthful requires incorporating an additional consideration (namely, that senders are aware that false signals that match the beliefs of the receiver will be initially scored as truthful, and that in some contexts this will influence their cost-benefit assessment). That noted, the significance of this qualification should not be overstated either. People are less gullible than we might assume (Mercier, 2020), and they do hold a good proportion of true beliefs. Wherever truth and beliefs coincide, the complications generated by their discrepancy dissipate.

Building on these limitations, the future avenues of research are manifold. To increase external validity, different scenarios, contexts, experimental designs, social media platforms and response mechanisms should be employed to replicate our findings. Paradigms that move beyond hypothetical scenarios and vignette-based studies, to test real social costs in actual social media use, would be most interesting in this regard. Another avenue of research would be to test for a range of plausible moderators and mediators. For instance, the question arises whether reputational costs of reposting are mediated by perceived plausible deniability and/or the perceived degree of speaker commitment.

Conclusion

Misinformation can circulate rapidly on social media, and understanding the mechanisms that facilitate its spread is crucial for developing effective interventions. Our experiments provide the first empirical evidence that reposting false content incurs significantly lower reputational costs than directly posting the same content. This reduced accountability, combined with the minimal effort required to repost and the striking pervasiveness of this form of communication on platforms like X and Facebook, creates what we term a *misinformation-prone reputational infrastructure*: an environment where the social incentives that ordinarily discourage inaccurate communication are systematically weakened. Our findings suggest that platform design choices – such as the introduction of the repost function – have non-trivial epistemic consequences, and that interventions aimed at artificially introducing friction (such as prompts to read before sharing) deserve serious consideration (Kozyreva et al., 2024; Jahn et al., 2025). More broadly, we hope this work demonstrates the value of integrating insights from costly signalling theory into the study of online misinformation (Reimann, 2022; Marsili, 2026), opening new avenues for both theoretical development and practical intervention.

References

- Acerbi, A. (2019). Cognitive attraction and online misinformation. *Palgrave Communications*, 5(1), 15. <https://doi.org/10.1057/s41599-019-0224-y>
- Acerbi, A., Altay, S., & Mercier, H. (2022). Fighting misinformation or fighting for information? *Harvard Kennedy School Misinformation Review*. <https://doi.org/10.37016/mr-2020-87>
- Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2), 211–236.
- Arielli, E. (2018). Sharing as Speech Act. *Versus*, XLVII(2), 243–258. <https://doi.org/10.14649/91354>
- Bary, C. (2023). Speech reports and evidence. *Journal of Pragmatics*, 210, 172–189. <https://doi.org/10.1016/j.pragma.2022.10.013>
- Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511616037>
- Bicchieri, C. (2017). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.
- Bonalumi, F., Isella, M., & Michael, J. (2019). Cueing Implicit Commitment. *Review of Philosophy and Psychology*, 10(4), 669–688. <https://doi.org/10.1007/s13164-018-0425-0>
- Bonalumi, F., Mahr, J. B., Marie, P., & Pouscoulous, N. (2023). Beyond the Implicit/Explicit Dichotomy: The Pragmatics of Plausible Deniability. *Review of Philosophy and Psychology*. <https://doi.org/10.1007/s13164-023-00699-5>
- Bonalumi, F., Scott-Phillips, T., Tacha, J., & Heintz, C. (2020). Commitment and communication: Are we committed to what we mean, or what we say? *Language and Cognition*, 12(2), 360–384. <https://doi.org/10.1017/langcog.2020.2>
- Camp, E. (2018). Insinuation, Common Ground, and the Conversational Record. In D. Fogal, D. W. Harris, & M. Moss (Eds), *New work on speech acts* (pp. 40–66). Oxford University Press.
- Ceylan, G., Anderson, I. A., & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, 120(4), e2216614120. <https://doi.org/10.1073/pnas.2216614120>
- Cheney, D. L., & Seyfarth, R. M. (1988). Assessment of meaning and the detection of unreliable signals by vervet monkeys. *Animal Behaviour*, 36(2), 477–486. [https://doi.org/10.1016/S0003-3472\(88\)80018-6](https://doi.org/10.1016/S0003-3472(88)80018-6)
- Colvin, J. (2016, February 22). *Donald Trump's favorite excuse? It was just a 'retweet'*. AP News. <https://apnews.com/events-united-states-presidential-election-22ee825323ef41efa17bd547d431c6ad>
- Faller, M. T. (2019). The discourse commitments of illocutionary reportatives. *Semantics and Pragmatics*, 12, 8:1-53. <https://doi.org/10.3765/sp.12.8>
- Frigerio, A., & Tenchini, M. P. (2023). Sharing. *Ratio*, 36(2), 147–158. <https://doi.org/10.1111/rati.12375>
- Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, 144(4), 517–546. [https://doi.org/10.1016/S0022-5193\(05\)80088-8](https://doi.org/10.1016/S0022-5193(05)80088-8)
- Graham, P. J. (2020). Assertions, Handicaps, and Social Norms. *Episteme*, 17(3), 349–363. <https://doi.org/10.1017/epi.2019.53>
- Green, M. (2009). Speech Acts, the Handicap Principle and the Expression of Psychological States. *Mind & Language*, 24(2), 139–163. <https://doi.org/10.1111/j.1468-0017.2008.01357.x>
- Green, M. (2025). The cultural evolution of speech act norms. *Intercultural Pragmatics*, 22(1), 3–23. <https://doi.org/10.1515/ip-2025-0002>
- Griskevicius, V., Tybur, J. M., Sundie, J. M., Cialdini, R. B., Miller, G. F., & Kenrick, D. T. (2007). Blatant benevolence and conspicuous consumption: When romantic motives elicit strategic costly signals. *Journal of Personality and Social Psychology*, 93(1), 85–102. <https://doi.org/10.1037/0022-3514.93.1.85>
- Haidt, J., & Bail, C. (Ongoing). *Social Media and Political Dysfunction*. https://docs.google.com/document/d/1vVAtMCQnz8WVxtSNQev_e1cGmY9rnY96ecYuAj6C548/edit?usp=embed_fac ebook
- Hall, A., & Mazarella, D. (2023). Pragmatic inference, levels of meaning and speaker accountability. *Journal of Pragmatics*, 205, 92–110. <https://doi.org/10.1016/j.pragma.2022.12.007>
- Hare, J. F., & Atkins, B. A. (2001). The squirrel that cried wolf: Reliability detection by juvenile Richardson's ground squirrels (*Spermophilus richardsonii*). *Behavioral Ecology and Sociobiology*, 51(1), 108–112. <https://doi.org/10.1007/s002650100414>
- Hatmaker, T. (2020, October 9). Changing how retweets work, Twitter seeks to slow down election misinformation | TechCrunch. <https://techcrunch.com/2020/10/09/twitter-retweet-changes-quote-tweet-election-misinformation/>
- Henderson, R., & McCready, E. (2018). How Dogwhistles Work. In S. Arai, K. Kojima, K. Mineshima, D. Bekki, K. Satoh, & Y. Ohta (Eds), *New Frontiers in Artificial Intelligence* (Vol. 10838, pp. 231–240). Springer International Publishing. https://doi.org/10.1007/978-3-319-93794-6_16
- Jahn, L., Rendsvig, R. K., Flammini, A., Menczer, F., & Hendricks, V. F. (2025). A perspective on friction interventions to curb the spread of misinformation. *Npj Complexity*, 2(1), 31. <https://doi.org/10.1038/s44260-025-00051-1>
- Kneer, M., & Marsili, N. (2025). The truth about assertion and retraction: A review of the empirical literature. In A. Wiegmann (Ed.), *Lying, Fake News, and Bullshit* (pp. 81–116). Bloomsbury.
- Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J., Barzilai, S., Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Huang, H.,

- Larreguy, H., Maertens, R., ... Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, 8(6), 1044–1052. <https://doi.org/10.1038/s41562-024-01881-0>
- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news: Addressing fake news requires a multidisciplinary effort. *Science*, 359(6380), 1094–1096. <https://doi.org/10.1126/science.aao2998>
- Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, 117(3), 785–807. <https://doi.org/10.1037/a0019688>
- Liu, Y., Kliman-Silver, C., & Mislove, A. (2014). The Tweets They are a-Changin': Evolution of Twitter Users and Behavior. *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 8(1), 305–314. <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/download/8043/8131>
- Marsili, N. (2021). Retweeting: Its linguistic and epistemic value. *Synthese*, 198, 10457–10483. <https://doi.org/10.1007/s11229-020-02731-y>
- Marsili, N. (2026). How online misinformation works: A costly signalling perspective. In M. Popa-Wyatt (Ed.), *Misinformation and Other Epistemic Pathologies*. Cambridge University Press.
- Mazzarella, D. (2023). “I didn’t mean to suggest anything like that!”: Deniability and context reconstruction. *Mind & Language*, 38(1), 218–236. <https://doi.org/10.1111/mila.12377>
- Mazzarella, D., Reinecke, R., Noveck, I., & Mercier, H. (2018). Saying, presupposing and implicating: How pragmatics modulates commitment. *Journal of Pragmatics*, 133, 15–27. <https://doi.org/10.1016/j.pragma.2018.05.009>
- McClain, C., Widjaya, R., Rivero, G., & Smith, A. (2021). *The Behaviors and Attitudes of U.S. Adults on Twitter*. Pew Research Center. <https://www.pewresearch.org/internet/2021/11/15/the-behaviors-and-attitudes-of-u-s-adults-on-twitter/>
- Mercier, H. (2020). Not Born Yesterday: The Science of Who We Trust and What We Believe. In *Not Born Yesterday*. Princeton University Press. <https://doi.org/10.1515/9780691198842>
- Michaelson, E., Pepp, J., & Sterken, R. (2024). On Retweeting. In E. Lepore & L. Anderson (Eds), *The Oxford Handbook of Applied Philosophy of Language* (p. 0). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780192844118.013.49>
- Møller, A. P. (1987). Social Control of Deception among Status Signalling House Sparrows *Passer domesticus*. *Behavioral Ecology and Sociobiology*, 20(5), 307–311. <https://www.jstor.org/stable/4600025>
- O'Connor, C., & Weatherall, J. O. (2018). The Misinformation Age. In *The Misinformation Age*. Yale University Press. <https://doi.org/10.2307/j.ctv8j90hk>
- Osmundsen, M., Bor, A., Vahlstrup, P., Bechmann, A., & Petersen, M. (2021). Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. *American Political Science Review*, 115, 1–17. <https://doi.org/10.1017/S0003055421000290>
- Oswald, S. (2022). Insinuation is committing. *Journal of Pragmatics*, 198, 158–170. <https://doi.org/10.1016/j.pragma.2022.07.006>
- Pancheva, R., & Rudin, D. (2019). Speaker discourse roles and the discourse profile of reportative evidentials. *Proceedings of the Amsterdam Colloquium*, 327–336. <https://platform.openjournals.nl/PAC/article/view/22109>
- Pennycook, G., & Rand, D. G. (2021). The Psychology of Fake News. *Trends in Cognitive Sciences*, 25(5), 388–402. <https://doi.org/10.1016/j.tics.2021.02.007>
- Pepp, J., Michaelson, E., & Sterken, R. K. (2019). What’s New About Fake News? *Journal of Ethics and Social Philosophy*, XVII(2), 67–94.
- Pinker, S. (2007). The evolutionary social psychology of off-record indirect speech acts. *Intercultural Pragmatics*, 4(4), 437–461. <https://doi.org/10.1515/IP.2007.023>
- Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences of the United States of America*, 105(3), 833–838. <https://doi.org/10.1073/pnas.0707192105>
- Quintana, I. O., Reimann, R., Cheong, M., Alfano, M., & Klein, C. (2022). Polarization and trust in the evolution of vaccine discourse on Twitter during COVID-19. *PLOS ONE*, 17(12), e0277292. <https://doi.org/10.1371/journal.pone.0277292>
- Reid, A. S., Zhang, J., Andreson, G. L., & Keblusek, L. (2020). Costly Signaling in Human Communication. In *The Handbook of Communication Science and Biology*. Routledge.
- Reimann, R. W. P. (2022). Costly Displays in a Digital World: Signalling Trustworthiness on Social Media. *Social Epistemology*, 1–18. <https://doi.org/10.1080/02691728.2022.2150990>
- Rini, R. (2017). Fake News and Partisan Epistemology. *Kennedy Institute of Ethics Journal*, 27(2S), E-43–E-64. <https://doi.org/10.1353/ken.2017.0025>
- Saul, J. M. (2012). *Lying, Misleading, and What is Said: An Exploration in Philosophy of Language and Ethics* (Issue 1). Oxford University Press.
- Saul, J. M. (2024). *Dogwhistles and Figleaves: How Manipulative Language Spreads Racism and Falsehood*. Oxford University Press.
- Scheufele, D. A., Krause, N. M., & Freiling, I. (2021). Misinformed about the “infodemic?” Science’s ongoing struggle with misinformation. *Journal of Applied Research in Memory and Cognition*, 10(4), 522–526. <https://doi.org/10.1016/j.jarmac.2021.10.009>
- Silverman, C. (2016). This analysis shows how fake election news stories outperformed real news on Facebook. In *BuzzFeed News* (pp. 1–7). <https://doi.org/10.1038/260777a0>

Sperber, D., Clement, F., Heintz, C., Mascaro, O., Mercier, H., Origg, G., & Wilson, D. (2010). Epistemic vigilance. *Mind and Language*, 25(4), 359–393. <https://doi.org/10.1111/j.1468-0017.2010.01394.x>

Sundar, S. S., Snyder, E. C., Liao, M., Yin, J., Wang, J., & Chi, G. (2025). Sharing without clicking on news in social media. *Nature Human Behaviour*, 9(1), 156–168. <https://doi.org/10.1038/s41562-024-02067-4>

Tameez, H. (2020). Following successful experiments, Twitter will prompt all users to read the articles they're about to retweet. *Nieman Lab*. <https://www.niemanlab.org/2020/09/following-successful-experiments-twitter-will-prompt-all-users-to-read-the-articles-theyre-about-to-retweet/>

Tibbetts, E. A., & Dale, J. (2004). A socially enforced signal of quality in a paper wasp. *Nature*, 432(7014), 218–222. <https://doi.org/10.1038/nature02949>

Tibbetts, E. A., & Izzo, A. (2010). Social Punishment of Dishonest Signalers Caused by Mismatch between Signal and Behavior. *Current Biology*, 20(18), 1637–1640. <https://doi.org/10.1016/j.cub.2010.07.042>

Vincent, J. (2020, September 25). *Twitter is bringing its 'read before you retweet' prompt to all users*. The Verge. <https://www.theverge.com/2020/9/25/21455635/twitter-read-before-you-tweet-article-prompt-rolling-out-globally-soon>

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>

Webster, M. S., Ligon, R. A., & Leighton, G. M. (2018). Social costs are an underappreciated force for honest signalling in animal aggregations. *Animal Behaviour*, 143, 167–176. <https://doi.org/10.1016/j.anbehav.2017.12.006>

Wiegmann, A., Reins, L. M., Mizumoto, M., Erut, A., Li, Q., & Orr, S. (2024). *The role of implicit content for lie and truth judgments: A cross-cultural study*. OSF. <https://doi.org/10.31234/osf.io/gdc2z>

Wiegmann, A., Willemsen, P., & Meibauer, J. (2022). Lying, Deceptive Implicatures, and Commitment. *Ergo*, 8(0), Article 0. <https://doi.org/10.3998/ergo.2251>

Zahavi, A., & Zahavi, A. (1997). *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford University Press.