



Research Article

Breaking *Harmony Square*: A game that “inoculates” against political misinformation

We present Harmony Square, a short, free-to-play online game in which players learn how political misinformation is produced and spread. We find that the game confers psychological resistance against manipulation techniques commonly used in political misinformation: players from around the world find social media content making use of these techniques significantly less reliable after playing, are more confident in their ability to spot such content, and less likely to report sharing it with others in their network.

Authors: Jon Roozenbeek (1), Sander van der Linden (1)

Affiliations: (1) Department of Psychology, University of Cambridge, United Kingdom

How to cite: Roozenbeek, J., & van der Linden, S. (2020). Breaking *Harmony Square*: A game that “inoculates” against political misinformation. *Harvard Kennedy School (HKS) Misinformation Review*, 1(8).

Received: July 23rd, 2020. Accepted: September 15th, 2020. Published: November 6th, 2020.

Research questions

- Does playing *Harmony Square* make people better at spotting manipulation techniques commonly used in political misinformation?
- Does playing the game increase people’s confidence in their ability to spot such manipulation techniques in social media content?
- Does playing the game reduce people’s self-reported willingness to share manipulative social media content with people in their network?

Essay summary

- In collaboration with the Dutch media collective DROG, design agency Gusmanson, Park Advisors, the U.S. Department of State’s Global Engagement Center and the Department of Homeland Security, we created a 10-minute, free online browser game called *Harmony Square*.
- Drawing on “inoculation theory,” the game functions as a psychological “vaccine” by exposing people to weakened doses of the common techniques used in political misinformation especially during elections.
- The game incorporates active experiential learning through a perspective-taking exercise: players are tasked with spreading misinformation and fomenting internal divisions in the quiet, peaceful neighborhood of Harmony Square.

¹ A publication of the Shorenstein Center for Media, Politics and Public Policy, at Harvard University, John F. Kennedy School of Government.

- Over the course of 4 levels, players learn about 5 manipulation techniques commonly used in the spread of political media content: trolling, using emotional language, polarizing audiences, spreading conspiracy theories, and artificially amplifying the reach of their content through bots and fake likes.
- In a mixed randomized controlled trial (international sample, $N = 681$), we tested if playing *Harmony Square* improves (a) people’s ability to spot both “real” and “fictional” misinformation, (b) whether it increases their confidence in their own judgments, and (c) makes them less likely to report sharing such content within their network.
- Overall, we find that people who play the game find misinformation significantly less reliable after playing, are significantly more confident in their assessment, and are significantly less likely to report sharing misinformation, supporting *Harmony Square*’s effectiveness as a tool to inoculate people against online manipulation.

Implications

Harmony Square is an interactive social impact game about election misinformation². The goal of the game is to reveal the tactics and manipulation techniques that fake news producers use to mislead their audience, build a following, and exploit societal tensions to achieve a political goal. The game’s setting is Harmony Square, a peaceful place where residents have a healthy obsession with democracy. At the start of the game, players are hired as Chief Disinformation Officer. Their job is to ruin the square’s idyllic state by fomenting internal divisions and pitting its residents against each other, all while gathering as many “likes” as they can. In order to deliver sufficiently weakened doses of the informational “virus,” *Harmony Square* makes use of humor throughout the game. For example, players can share humorous messages in a fictional social network, and are shown entertaining headlines in a news ticker at the top of the screen (see the second panel in Figure 1). Aside from increasing the entertainment value of the game, the use of humor in inoculation interventions has the added benefit of potentially decreasing reactance, i.e., resistance to voluntarily engaging with the intervention (Compton, 2018; Vraga et al., 2019). Over the course of 4 different levels (*Trolling*, *Emotion*, *Amplification* and *Escalation*), the player’s misinformation campaign causes the square to gradually go from a peaceful state to full-blown mayhem. Figure 1 shows a number of screenshots of what the game’s landing page and game environment look like.

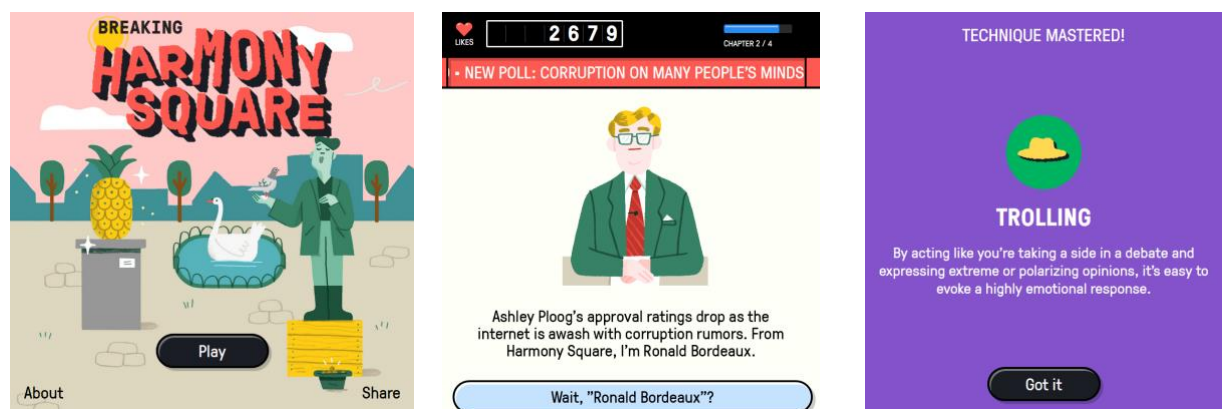


Figure 1. Screenshots of the *Harmony Square* landing page (left) and game environment.

² The game can be played in any browser at <https://www.harmonysquare.game>

The game was produced with and based on the US Department of Homeland Security's Cybersecurity and Infrastructure Agency's (CISA) approach to understanding foreign interference in elections. In an effort to "strengthen the national immune system for disinformation," CISA launched a campaign in 2019 that sought to expose common misinformation tactics through the lens of an ostensibly innocent and non-partisan issue: whether or not to put pineapple on pizza (Ward, 2019). In a similar vein to the "pineapple pizza" campaign, *Harmony Square* exposes the 5 steps of the election misinformation playbook: targeting divisive issues (in line with the "trolling" & "emotion" scenarios in *Harmony Square*, in which players learn how to turn an ostensibly neutral issue into a heated and polarizing debate); moving accounts into place; amplifying and distorting the conversation (the "amplification" scenario, in which players learn how bots and fake accounts can be used to amplify the reach of manipulative content); making the mainstream; and taking the conversation into the real world (the "escalation" scenario, in which players learn how to escalate online debates into real-world action) (CISA, 2019).

Harmony Square builds on the success of our other gamified anti-misinformation interventions, such as the *Bad News* game (Roozenbeek & van der Linden, 2019; Roozenbeek, van der Linden, et al., 2020).³ Similar to *Bad News*, playing *Harmony Square* builds cognitive resistance against common forms of manipulation that people may encounter online by preemptively warning and exposing people to weakened doses of these techniques in a controlled environment. Unlike *Bad News*, *Harmony Square* is an election game that focuses specifically on how misinformation can be used to achieve political polarization (Bessi et al., 2016; Shao et al., 2017), for example, by fueling outgroup hostility, a critical element of both organic misinformation and targeted disinformation campaigns, particularly during contentious political events such as the 2020 US Presidential elections (Groshek & Koc-Michalska, 2017; Hindman & Barash, 2018; Iyengar & Massey, 2018; Keller et al., 2020).

The idea of psychological "vaccines," coined in the 1960s by the psychologist William McGuire (McGuire, 1964; McGuire & Papageorgis, 1961b), is called inoculation theory. It follows a medical analogy: a vaccine is usually a weakened version of a particular pathogen which, after being introduced to the body, induces the production of antibodies, preventing an individual from becoming sick when exposed to the real illness. Inoculation theory states that the same can be achieved with (malicious) persuasion attempts: pre-emptively exposing someone to a weakened version of a particular misleading argument prompts a process that is akin to the production of "mental antibodies," which make it less likely that a person is persuaded by the "real" manipulation later on (Compton, 2013; van der Linden et al., 2017). During gameplay, players are exposed to weakened doses of manipulation techniques by stepping into the shoes of a fake news producer to trigger the production of psychological antibodies.

Instead of focusing on specific examples, also known as issue-based inoculation (which has been the standard in inoculation research), *Harmony Square* builds cognitive resistance against the *techniques* that underpin a whole range of political misinformation in an attempt to achieve broad-spectrum resistance against manipulation (Basol et al., 2020; Cook et al., 2017). The game functions as a perspective-taking exercise (i.e., by putting the player in the position of a fake news creator), an approach known as "active inoculation" (McGuire & Papageorgis, 1961a; Roozenbeek & van der Linden, 2018). Actively involving individuals in the inoculation process by playing a game—as opposed to subjecting them to a more passive reading exercise—has the potential advantage of increasing retention in memory and increasing the longevity of the inoculation effect (Compton & Pfau, 2005; Maertens et al., 2020; Roozenbeek & van der Linden, 2019).

By playing through the 4 levels in *Harmony Square*, players learn about 5 manipulation techniques, all of which are common features of political misinformation (van der Linden & Roozenbeek, 2020):

³ The *Bad News* game can be played at <https://www.getbadnews.com>

1. Trolling people, i.e., deliberately provoking people to react emotionally, thus evoking outrage see (McCosker, 2014; Roozenbeek & van der Linden, 2019).
2. Exploiting emotional language, i.e., trying to make people afraid or angry about a particular topic (Brady et al., 2017; Zollo et al., 2015).
3. Artificially amplifying the reach and popularity of certain messages, for example through social media bots or by buying fake followers (McKew, 2018; Shao et al., 2017).
4. Creating and spreading conspiracy theories, i.e., blaming a small, secretive and nefarious organization for events going on in the world (Lewandowsky et al., 2013; van der Linden, 2015).
5. Polarizing audiences by deliberately emphasizing and magnifying inter-group differences (Iyengar & Massey, 2018; Prior, 2013).

In this study, we find that people who played *Harmony Square* rated manipulative social media posts making use of the above techniques as less reliable after playing, were more confident in their ability to spot such content, and importantly, were less likely to report to share it in their social network. This finding held for both “real” manipulative content that has gone viral online in the past, as well as for “fictional” content that participants in our study had never seen before. These findings highlight the potential for social impact games as an effective approach to counter misleading, fake, or manipulative content proliferating online (Basol et al., 2020; Maertens et al., 2020; Roozenbeek, Schneider, et al., 2020).

In addition, we find that political ideology did not interact with the learning effect conferred by *Harmony Square*, meaning that the game was effective at teaching manipulation techniques for both liberals and conservatives. This finding that the intervention can be effective across partisan lines is particularly important in light of the polarization of not just the US media landscape, but of the term “fake news” itself (van der Linden et al., 2020).

Specifically, we tested if people became better at spotting troll posts (i.e., posts that “bait” people into responding emotionally), exploitative emotional language use, conspiratorial content, and content that deliberately seeks to polarize different groups. These are all important manipulation techniques used in political misinformation (Roozenbeek & van der Linden, 2019; van der Linden & Roozenbeek, 2020). In addition, these techniques are key components of many organized disinformation campaigns (Bertolin et al., 2017; Cook et al., 2017; Hindman & Barash, 2018). Building cognitive resistance against these techniques at scale is a powerful tool to reduce the risk of disinformation campaigns affecting the democratic process. Recent research with the *Bad News* game has looked into the longevity of the inoculation effect conferred by “fake news” games such as *Bad News* (Maertens et al., 2020). This research finds that significant inoculation effects remain detectable for at least one week, and much longer when participants are presented with very short reminders or “booster shots.”

In short, we show that *Harmony Square* is effective at reducing some of the harmful effects that manipulative content can have on individuals. Taking only around 10 minutes to complete, the game conveys critical information about key manipulation techniques in a fun and interactive manner. We note that *Harmony Square* focuses mostly on political misinformation, and the game’s setting is a fictional democratic society. Since (political) misinformation is a significant problem in non-democratic countries, one of the limitations of this game is its limited applicability in countries that lack free elections. In addition, although research has shown that gamified inoculation effects can persist for months (Maertens et al., 2020), the longevity of the effect of *Harmony Square* was not evaluated here.

Findings

The purpose of this study was to find out whether people who play *Harmony Square* 1) find manipulative social media content less reliable after playing; 2) are more confident in their ability to spot such

manipulative content; and 3) are less likely to indicate that they are willing to share manipulative social media content in their network, compared to a gamified control group (which played *Tetris* for around the same amount of time it takes to complete *Harmony Square*). To answer these questions, we first calculated the difference between the average score on all 3 of these questions for all 8 “real fake news” and all 8 “fictional fake news” social media posts that we used as measures (see the “methods” section) before and after the intervention, for each participant. We call this the “pre-post difference score.” We then checked if these difference scores for the treatment group were significantly different from the control group for each outcome variable, using an Analysis of Variance (ANOVA).

Finding 1: People who play Harmony Square find manipulative social media content significantly less reliable after playing compared to a control group.

For all the “real fake news” social media posts combined (see “methods” section), we find a significant main effect of the treatment condition (i.e. playing *Harmony Square*) on aggregate reliability judgments, meaning that playing the game significantly reduces the perceived reliability of “real fake news” compared to the control group ($F(1,679) = 43.21, p < .001, \eta^2 = .060, d = 0.51$, Figure 2).⁴ We find the same result for the “fictitious fake news” items ($F(1,679) = 48.79, p < .001, \eta^2 = .067, d = 0.54$).⁵ Importantly, the effect-sizes are nearly identical, which illustrates that using real or fictitious fake news does not matter much for assessment. The results are visualized in Figure 2.

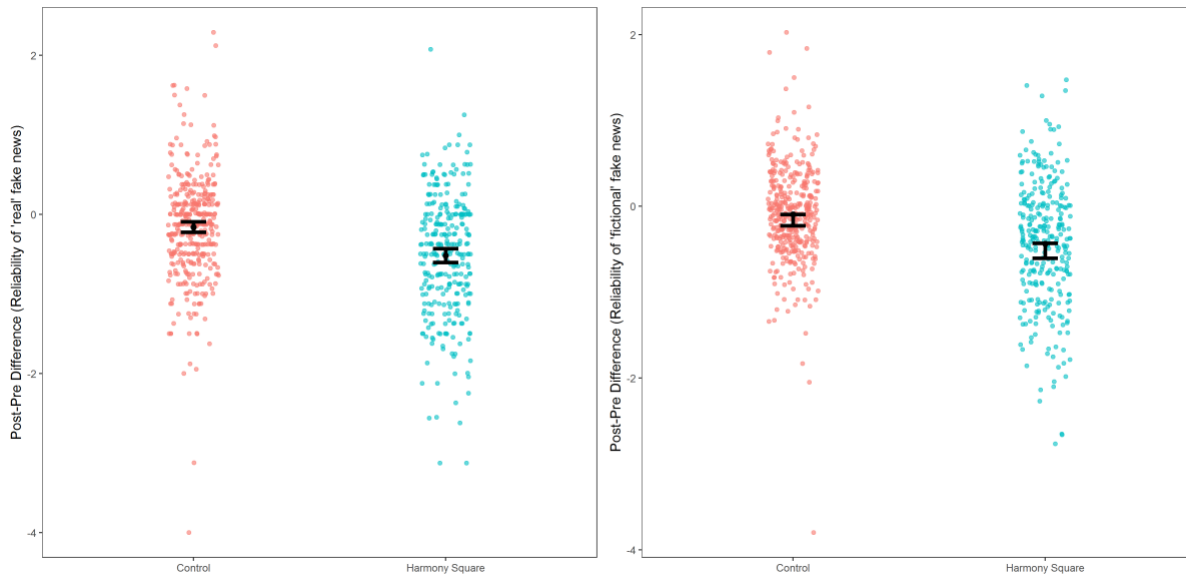


Figure 2. Pre-post difference scores for the reliability judgments for both “real” and “fictional” misinformation, for the treatment (panel A) and control groups (panel B). The dots represent the data points. The black dot in the middle indicates the mean value. Error bars represent 95% confidence intervals. In both plots, the distribution for the *Harmony Square* group is shifted downwards, indicating that people found fake news posts significantly less reliable after playing in the treatment group. Separate bars for pre and post scores (by condition) are provided in Supplementary Figure 1.

⁴ Specifically, compared to the control condition, the shift in post-pre difference scores for reliability judgments was significantly more negative for the treatment group ($M_{diff,control} = -0.16, SD_{diff,control} = 0.66$ vs $M_{diff,treatment} = -0.52, SD_{diff,treatment} = 0.77, d = 0.51$). We found no main effect ($F(2,675) = 0.486, p = 0.62$) nor an interaction effect ($F(2,675) = 0.663, p = 0.52$) for political ideology, meaning that ideology does not make a significant difference to the inoculation effect.

⁵ Similar to “real” fake news, the shift in post-pre difference scores for reliability judgments was significantly more negative for the treatment group ($M_{diff,control} = -0.095, SD_{diff,control} = 0.55$ vs $M_{diff,treatment} = -0.44, SD_{diff,treatment} = 0.72, d = 0.54$). Here, we also found no main effect ($F(2,675) = 1.154, p = 0.32$) nor an interaction effect ($F(2,675) = 0.810, p = 0.45$) for political ideology.

Finding 2: People who played Harmony Square are significantly more confident in their ability to spot manipulative content in social media posts, compared to a control group.

Next, we checked if playing *Harmony Square* increases people’s confidence in spotting manipulative content. We used the same method as above: first, we calculated the difference between the average scores before and after the intervention for both the treatment and control group, and then conducted an ANOVA to see if the differences between treatment and control were significant. Participants who played *Harmony Square* became significantly more confident in their ability to spot both “real” ($F(1,679) = 14.52, p < .001, \eta^2 = .021, d = 0.30$)⁶ and “fictional” ($F(1,679) = 14.55, p < .001, \eta^2 = .021, d = 0.30$)⁷ manipulative content. Figure 3 visualizes the results.

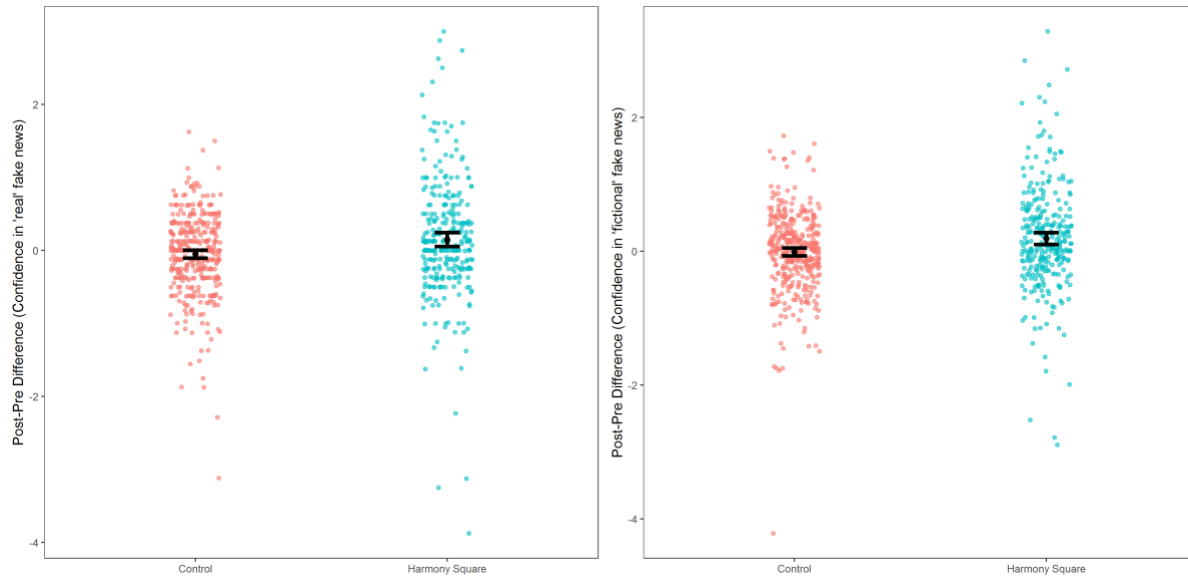


Figure 3. Pre-post difference scores for the confidence judgments for both “real” and “fictional” misinformation, for the treatment and control groups. The black dot in the middle indicates the mean value. Error bars represent 95% confidence intervals. In both plots, the distribution for the *Harmony Square* group is shifted upwards, indicating that people were significantly more confident in their judgments about misinformation after playing in the treatment group. Separate bars for pre and post scores (by condition) are provided in Supplementary Figure 1.

Finding 3: People who played Harmony Square are significantly less likely to report sharing manipulative social media content with others.

Finally, we checked whether playing *Harmony Square* reduces participants’ self-reported willingness to share manipulative content with people in their network. Using the same method as above, we found that people who played the game were significantly less likely to share both “real” ($F(1,679) = 12.85, p <$

⁶ Compared to the control condition, the shift in post-pre difference scores for confidence judgments was significantly more positive for the treatment group ($M_{diff,control} = -0.051, SD_{diff,control} = 0.55$ vs $M_{diff,treatment} = 0.15, SD_{diff,treatment} = 0.82, d = 0.30$).

⁷ Compared to the control condition, the shift in post-pre difference scores for confidence judgments was significantly more positive for the treatment group ($M_{diff,control} = -0.011, SD_{diff,control} = 0.59$ vs $M_{diff,treatment} = 0.19, SD_{diff,treatment} = 0.79, d = 0.30$).

.001, $\eta^2 = .019$, $d = 0.28$)⁸ and “fictional” ($F(1,679) = 12.62$, $p < .001$, $\eta^2 = .018$, $d = 0.27$)⁹ manipulative content that they encounter online. Figure 4 visualizes the results.

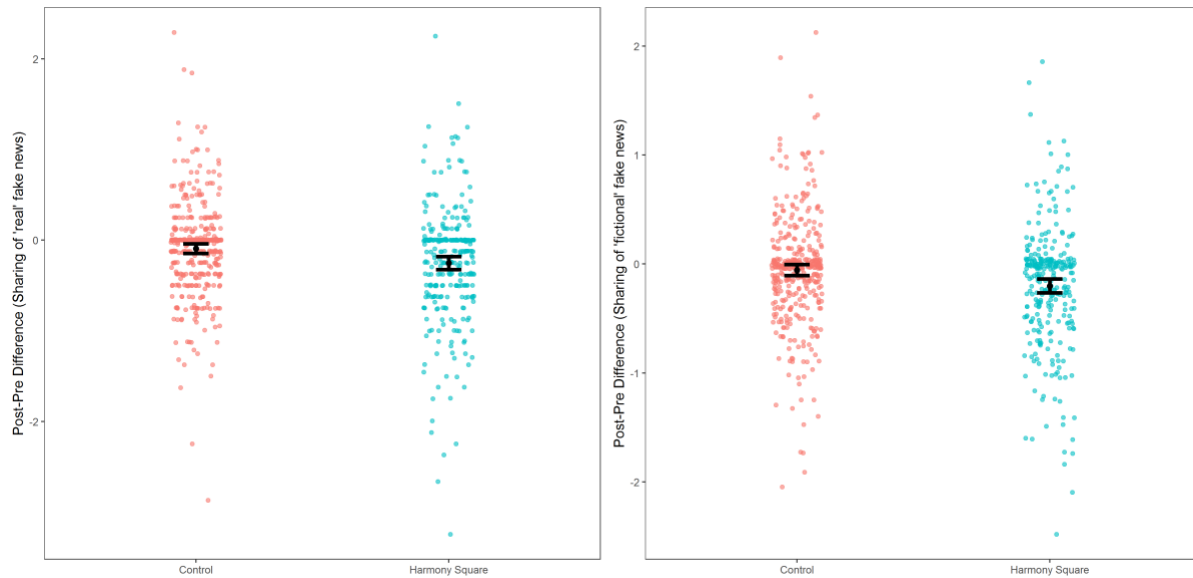


Figure 4. Pre-post difference scores for willingness to share for both “real” and “fictional” misinformation, for the treatment and control groups. The black dot in the middle indicates the mean value. Error bars represent 95% confidence intervals. In both plots, the distribution for the *Harmony Square* group is shifted downwards, indicating that people shared fake news posts significantly less after playing in the treatment group. Separate bars for pre and post scores (by condition) are provided in Supplementary Figure 1.

To restrict multiple testing, we only present results for the aggregated fake news indices here. However, when looking at the manipulation techniques featured in the game (trolling, emotion, conspiracy and polarization), we show that players improve on each technique as well. The full list of ANOVAs per manipulation technique, including effect size estimates, can be found in Supplementary Table S3. A bar plot which summarizes the results for the pre and post-test separately in a single figure can be found in Supplementary Figure S1. We conducted two robustness checks to verify the main analyses presented here: a linear regression using post-test as the dependent variable, pre-test as a covariate, and condition as a between-subject factor (see the Supplementary Methods and Analyses section and Supplementary Table S4); and a robust linear regression clustering scores at the participant and rating level (see the Supplementary Methods and Analyses section and Supplementary Table S5). Both approaches give the same results as what is presented above.

Methods

To test if *Harmony Square* improves people’s ability to spot manipulative online content, we conducted a 2 (treatment vs control) by 2 (pre vs post) mixed design randomized controlled trial.¹⁰ The treatment

⁸ Compared to the control condition, the shift in post-pre difference scores for willingness to share was significantly more negative for the treatment group ($M_{diff,control} = -0.09$, $SD_{diff,control} = 0.53$ vs $M_{diff,treatment} = -0.25$, $SD_{diff,treatment} = 0.64$, $d = 0.28$).

⁹ Compared to the control condition, the shift in post-pre difference scores for willingness to share was significantly more negative for the treatment group ($M_{diff,control} = -0.06$, $SD_{diff,control} = 0.51$ vs $M_{diff,treatment} = -0.20$, $SD_{diff,treatment} = 0.56$, $d = 0.27$).

¹⁰ The full dataset, “real” and “fictional” social media posts and R scripts used in this study are available on the OSF. Link: <https://osf.io/r89h3/>.

condition involved playing *Harmony Square* from beginning to end. The control condition played *Tetris* for about 10 minutes. We chose *Tetris* because it is in the public domain, most people know how it works without practicing, and it involves about the same amount of cognitive effort as playing *Harmony Square*. Following the methodology established in prior research on “fake news” games (Basol et al., 2020; Maertens et al., 2020; Roozenbeek, Maertens, et al., 2020; Roozenbeek & van der Linden, 2019), we measured reliability judgments of social media posts containing misinformation, both before and after the intervention on a 1-7 Likert scale.

Measures

We sought to answer three questions about the effectiveness of *Harmony Square* as an anti-misinformation tool:

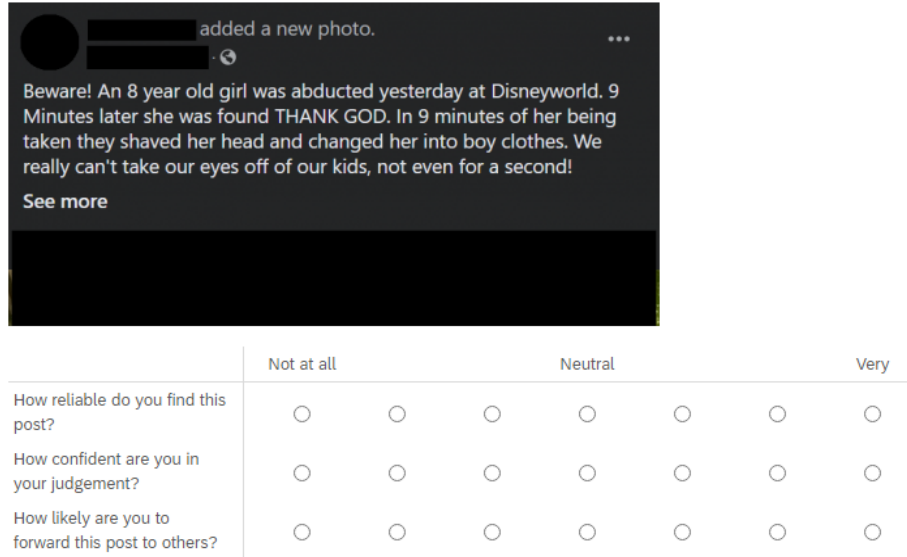
1. Does playing *Harmony Square* make people better at spotting manipulation techniques commonly used in political misinformation?
2. Does playing the game increase people’s confidence in their ability to spot such manipulation techniques in social media content?
3. Does playing the game reduce people’s self-reported willingness to share manipulative social media content with people in their network?

To address these questions, we showed the participants in our study 16 social media posts, each of which made use of one of 4 manipulation techniques learned while playing *Harmony Square*: trolling, using emotional language, conspiratorial reasoning, and group polarization. These posts were selected to be a mix of politically partisan and politically neutral content. Politically neutral items covered topics such as a kidnapping at an amusement park. Since the *Harmony Square* game is about political misinformation, we also included several items that were ideologically or politically charged. These items were balanced overall, with an equal number of right-leaning and left-leaning items. All items are available on our OSF page as well as in supplementary table S6.

In total, 8 of these posts were examples of “real” manipulative content found “in the wild” on social media and in fake news articles. The other 8 were social media posts that we created (“fictional fake news”), which were validated in previous research (Basol et al., 2020; Maertens et al., 2020; Roozenbeek, Maertens, et al., 2020). We did not hypothesize any significant differences between participants’ assessments of “real” and “fictional” misinformation, but chose to include both types for the following reasons: 1) including “real” items increases the ecological validity of the study, as participants are tested on information that they could have encountered “in the wild”; 2) including “fictional” items maximizes experimental control and thus allows us to better isolate each manipulation technique and ensure political neutrality, and 3) by including “fictional” items, we account for the possibility that participants may have seen the “real” manipulative content before, a memory confound which could bias their assessment (Roozenbeek & van der Linden, 2019).¹¹ Following Basol et al. (2020), we deliberately chose to only include manipulative content, as opposed to a mix of manipulative and non-manipulative content. The purpose of the *Harmony Square* game was not to learn how to distinguish high-quality and low-quality content, but rather to teach people how to spot common types of misinformation on social media. We therefore chose to focus on addressing the question of whether *Harmony Square* is effective at reducing susceptibility to political misinformation, rather than truth discernment (Pennycook et al., 2020) but we note that media literacy interventions can affect the rating of both credible and non-credible items (Guess et al., 2020). For a more detailed discussion on how gamified inoculation interventions affect people’s perception of “real” (high-quality) news we refer the reader to Roozenbeek, Maertens et al. (2020).

¹¹ The Supplementary Methods & Analyses appendix contains further information on the item selection procedure and a Principal Component Analysis for both the real and fake social media posts.

Figure 5 shows an example of what our items look like in the survey environment. The social media post in the figure is a real example of a rumor that went viral about a girl being kidnapped in a theme park, an example of the “emotional language” technique learned in the game (AFP Canada, 2019; Pennycook et al., 2019). The full list of items can be found on the OSF¹² and in the supplement (Table S6).



added a new photo.

Beware! An 8 year old girl was abducted yesterday at Disneyworld. 9 Minutes later she was found THANK GOD. In 9 minutes of her being taken they shaved her head and changed her into boy clothes. We really can't take our eyes off of our kids, not even for a second!

See more

	Not at all			Neutral			Very
How reliable do you find this post?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How confident are you in your judgement?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How likely are you to forward this post to others?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 5. Example of a social media post (item) used in the survey.

For each of the 16 items, we asked participants 3 questions, which they could answer on a 1-7 scale (1 being “not at all”, 4 being “neutral” and 7 being “very”):¹³

1. How reliable do you find this post?
2. How confident are you in your judgment?
3. How likely are you to forward this post to others?

We asked these questions for all 16 social media posts both before and after the intervention (*Harmony Square* for the treatment group, and *Tetris* for the control group). This allowed us to measure the difference between the “before score” and the “after score” for each group (the “pre-post difference score”). We thus arrive at our hypotheses: if *Harmony Square* is effective as an anti-misinformation tool, participants who played it should 1) find manipulative content significantly less reliable after playing, 2) be significantly more confident in their judgment, and 3) be significantly less likely to report sharing such content with people in their network, whereas the control group—who did not learn anything about manipulative content while playing *Tetris*—should show no significant differences for each of the three questions before and after playing. To control for multiple testing, we only evaluated the aggregate indices for each dependent variable but for completeness we report effects for all 4 manipulation techniques separately in the supplementary information (Supplementary Table S3). Descriptive statistics for each individual item can be found in Supplementary Table S2.

In total, 681 people were recruited in 2 separate data collections; a US-only sample ($n = 312$) and an international sample ($n = 369$). We pooled the results here (effect-sizes are slightly larger for the US-only

¹² See <https://osf.io/r89h3/>

¹³ A reliability analysis shows acceptable to good internal consistency for all 3 outcome measures, both for the “real” fake news item set and the “fictional” fake news item set ($M_{real, reliability} = 3.26$, $SD_{real, reliability} = 0.98$, $\alpha = 0.68$; $M_{fake, reliability} = 2.99$, $SD_{fake, reliability} = 1.06$, $\alpha = 0.78$; $M_{real, confidence} = 5.20$, $SD_{real, confidence} = 1.06$, $\alpha = 0.84$; $M_{fake, confidence} = 5.13$, $SD_{fake, confidence} = 1.08$, $\alpha = 0.86$; $M_{real, sharing} = 2.26$, $SD_{real, sharing} = 1.17$, $\alpha = 0.85$; $M_{fake, sharing} = 2.15$, $SD_{fake, sharing} = 1.19$, $\alpha = 0.88$).

sample). In total, 296 participants played *Harmony Square* (the treatment group), and 385 people played *Tetris* (the control group). A detailed overview of the sample selection process and study participants, as well as several robustness checks for the main analyses, can be found in the Supplementary Methods & Analyses appendix.

Bibliography

- AFP Canada. (2019). *Tale of thwarted child abduction returns with Canadian theme park twist*. Factcheck.afp. <https://factcheck.afp.com/tale-thwarted-child-abduction-returns-canadian-theme-park-twist>
- Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, 3(1), 2, 1–9. <http://doi.org/10.5334/joc.91>
- Bertolin, G., Agarwal, N., Bandeli, K., Biteniece, N., & Sedova, K. (2017). *Digital hydra: Security implications of false information online*. <https://www.stratcomcoe.org/digital-hydra-security-implications-false-information-online>
- Bessi, A., Zollo, F., Del Vicario, M., Puliga, M., Scala, A., Caldarelli, G., Uzzi, B., & Quattrociocchi, W. (2016). Users polarization on Facebook and YouTube. *PLoS ONE*, 11(8), 1–24. <https://doi.org/10.1371/journal.pone.0159641>
- Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28), 7313–7318. <https://doi.org/10.1073/pnas.1618923114>
- CISA. (2019). *The war on pineapple: Understanding foreign interference in 5 steps*. DHS. https://www.dhs.gov/sites/default/files/publications/19_0717_cisa_the-war-on-pineapple-understanding-foreign-interference-in-5-steps.pdf
- Compton, J. (2013). Inoculation theory. In J. P. Dillard & L. Shen (Eds.), *The SAGE Handbook of Persuasion: Developments in Theory and Practice* (2nd ed., pp. 220–236). <https://doi.org/10.4135/9781452218410>
- Compton, J. (2018). Inoculation against/with political humor. In J. C. Baumgartner & A. B. Becker (Eds.), *Political humor in a changing media landscape: A new generation of research* (pp. 95–113). London: Lexington Books.
- Compton, J., & Pfau, M. (2005). Inoculation theory of resistance to influence at maturity: Recent progress in theory development and application and suggestions for future research. *Annals of the International Communication Association*, 29(1), 97–145. https://doi.org/10.1207/s15567419cy2901_4
- Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS ONE*, 12(5), 1–21. <https://doi.org/10.1371/journal.pone.0175799>
- Groshek, J., & Koc-Michalska, K. (2017). Helping populism win? Social media use, filter bubbles, and support for populist presidential candidates in the 2016 US election campaign. *Information Communication and Society*, 20(9). <https://doi.org/10.1080/1369118X.2017.1329334>
- Hindman, M., & Barash, V. (2018). *Disinformation, “fake news” and influence campaigns on Twitter*. https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/238/original/KF-DisinformationReport-final2.pdf
- Iyengar, S., & Massey, D. S. (2018). Scientific communication in a post-truth society. *Proceedings of the National Academy of Sciences*, 116(16), 7656–7661. <https://doi.org/10.1073/PNAS.1805868115>

- Keller, F. B., Schoch, D., Stier, S., & Yang, J. (2020). Political astroturfing on Twitter: How to coordinate a disinformation campaign. *Political Communication*, 37(2), 256–280.
<https://doi.org/10.1080/10584609.2019.1661888>
- Lewandowsky, S., Oberauer, K., & Gignac, G. E. (2013). NASA faked the moon landing—Therefore, (climate) science is a hoax: An anatomy of the motivated rejection of science. *Psychological Science*, 24(5), 622–633. <https://doi.org/10.1177/0956797612457686>
- Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2020). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*. <https://doi.org/https://dx.doi.org/10.1037/xap0000315>
- McCosker, A. (2014). Trolling as provocation: YouTube’s agonistic publics. *Convergence*, 20(2), 201–217.
<https://doi.org/10.1177/1354856513501413>
- McGuire, W. J. (1964). Inducing resistance against persuasion: Some contemporary approaches. *Advances in Experimental Social Psychology*, 1, 191–229.
[https://doi.org/http://dx.doi.org/10.1016/S0065-2601\(08\)60052-0](https://doi.org/http://dx.doi.org/10.1016/S0065-2601(08)60052-0)
- McGuire, W. J., & Papageorgis, D. (1961a). Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. *Journal of Abnormal and Social Psychology*, 63, 326–332.
- McGuire, W. J., & Papageorgis, D. (1961b). The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *Journal of Abnormal and Social Psychology*, 62(2), 327–337.
- McKew, M. K. (2018, February 4). How Twitter bots and Trump fans made #ReleaseTheMemo go viral. *Politico*. <https://www.politico.com/magazine/story/2018/02/04/trump-twitter-russians-release-the-memo-216935>
- Pennycook, G., Martel, C., & Rand, D. G. (2019). *Knowing how fake news preys on your emotions can help you spot it*. CBC. <https://www.cbc.ca/news/canada/saskatchewan/analysis-fake-news-appeals-to-emotion-1.5274207>
- Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7), 770–780. <https://doi.org/10.1177/0956797620939054>
- Prior, M. (2013). Media and political polarization. *Annual Review of Political Science*, 16(1), 101–127.
<https://doi.org/10.1146/annurev-polisci-100711-135242>
- Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2020). Differentiating item and testing effects in inoculation research on online misinformation. *Educational and Psychological Measurement*, 1–23. <https://doi.org/10.1177/0013164420940378>
- Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L. J., Recchia, G., van der Bles, A.M., & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, 7(2011199). <https://doi.org/10.1098/rsos.201199>
- Roozenbeek, J., & van der Linden, S. (2018). The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research*, 22(5), 570–580.
<https://doi.org/10.1080/13669877.2018.1443491>
- Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Humanities and Social Sciences Communications*, 5(65), 1–10.
<https://doi.org/10.1057/s41599-019-0279-9>
- Roozenbeek, J., van der Linden, S., & Nygren, T. (2020). Prebunking interventions based on “inoculation” theory can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School (HKS) Misinformation Review*, 1(2). <https://doi.org/10.37016//mr-2020-008>
- Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2017). The spread of fake news by social bots. *ArXiv:1707.07592 (cs.SI)*. <http://arxiv.org/abs/1707.07592>

- van der Linden, S. (2015). The conspiracy-effect: Exposure to conspiracy theories (about global warming) decreases pro-social behavior and science acceptance. *Personality and Individual Differences*, 87, 171–173. <http://www.sciencedirect.com/science/article/pii/S0191886915005024>
- van der Linden, S., Panagopoulos, C., & Roozenbeek, J. (2020). You are fake news: The emergence of political bias in perceptions of fake news. *Media, Culture & Society*, 42(3), 460–470. <https://doi.org/10.1177/0163443720906992>
- van der Linden, S., & Roozenbeek, J. (2020). Psychological inoculation against fake news. In R. Greifenader, M. Jaffé, E. Newman, & N. Schwarz (Eds.), *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation*. <https://doi.org/10.4324/9780429295379-11>
- Vraga, E. K., Kim, S. C., & Cook, J. (2019). Testing logic-based and humor-based corrections for science, health, and political misinformation on social media. *Journal of Broadcasting & Electronic Media*, 63(3), 393–414. <https://doi.org/10.1080/08838151.2019.1653102>
- Ward, J. (2019, July 27). *U.S. cybersecurity agency uses pineapple pizza to demonstrate vulnerability to foreign influence*. NBC News. <https://www.nbcnews.com/news/us-news/u-s-cybersecurity-agency-uses-pineapple-pizza-demonstrate-vulnerability-foreign-n1035296>
- Zollo, F., Novak, P. K., Del Vicario, M., Bessi, A., Mozetič, I., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Emotional dynamics in the age of misinformation. *PLoS ONE*, 10(9), 1–22. <https://doi.org/10.1371/journal.pone.0138740>

Acknowledgements

We thank Cecilie Steenbuch Traberg for designing several items (social media posts) used in this study.

Funding

The Department of State's Global Engagement Center (GEC) has, in collaboration with DROG and the University of Cambridge, assisted in the development and tailoring of *Harmony Square* within the scope of addressing foreign adversarial propaganda and disinformation and its impact on foreign audiences and elections overseas. The GEC seeks to "facilitate the use of a wide range of technologies and techniques by sharing expertise among federal department and agencies, seeking expertise from external sources and implementing best practices" in order to "recognize, understand, expose, and counter foreign state and foreign non-state propaganda and disinformation" (NDAA section 1287).

In support of this mission, the GEC's Technology Engagement Team collaborated in tailoring *Harmony Square* to address the challenges of foreign adversarial propaganda and disinformation that could impact foreign elections. *Harmony Square* is an example of how the GEC's programmatic initiatives – the Tech Demo Series and the Testbed – assist in identifying and testing tools to facilitate unique technological approaches to countering propaganda and disinformation.

Harmony Square is freely available to play at: <https://www.harmonysquare.game>

Competing interests

J.R. and S.v.d.L. were responsible for the content of the *Harmony Square* game. J.R. was financially compensated for this work by DROG. Neither DROG nor GEC and DHS were involved in any way in this study's design, analysis, or write-up.

Ethics

This study was approved by the Department of Psychology Research Ethics Committee at the University of Cambridge (PRE2020.052).

We asked participants to indicate their gender in this study. In line with established survey research practice in psychology, participants could indicate whether they self-identified as male, female, or other (in which case they could fill in a text box to specify how they self-identify).

Copyright

This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are properly credited.

Data Availability

The data, items and scripts used in this study is available in the OSF: <https://osf.io/r89h3/>

Appendix: Supplementary methods & analyses

Supplementary information on sample selection and study participants

For this study, we recruited participants via the online platform Prolific Academic (Palan & Schitter, 2018; Peer et al., 2017). Based on previous research (Basol et al., 2020), we first conducted an a priori power analysis using G* power, with $\alpha = 0.05$, $f = 0.26$ ($d = 0.52$), power of 0.95, and 2 experimental conditions. The minimal sample size required for detecting the main effect was 258. In total, 681 people were recruited in 2 separate data collections; a US-only sample ($n = 312$) and an international sample ($n = 369$). We pooled the results here (effect-sizes are slightly larger for the US-only sample). In total, 296 participants played *Harmony Square* (the treatment group), and 385 people played *Tetris* (the control group). This discrepancy is explained by the fact that we only included participants in the treatment group that played through the game in its entirety; following quality-control practices from previous research (Basol et al., 2020; Maertens et al., 2020). Specifically, participants in the treatment group were required to fill in a code before proceeding to the next stage of the study, which only appeared after finishing the game. Some participants ($n = 78$) entered the wrong code or no code at all, and were thus excluded from the dataset¹⁴. This is important because otherwise we cannot ensure that participants played through the whole game. No other exclusions were applied.

In total, 46.3% of our participants were from the United States, 11.0% from Portugal, 10.9% from Poland, 6.8% from the United Kingdom, 5.6% from Italy, 4.1% from Mexico, and another 15.4% from elsewhere. 43.2% of participants identified as female, 55.7% as male, and 1.2% as other (e.g. non-binary or agender). Participants were mostly younger, with 41.4% being between 18 and 24 years of age. The average education level was high, with 62.4% of participants indicating that they have at least a Bachelor’s degree. The sample also skewed somewhat left in terms of political ideology, with the average score on the 1-7 political ideology scale (1 being “very left-wing” and 7 being “very right-wing”) $M = 3.13$, $SD = 1.44$. On average, participants were paid £2.42 (or US \$3.12). The average completion time was around 20 minutes. Supplementary Table S1 gives a detailed overview of the sample that was recruited for this study; it also shows that the sample of individuals that did not enter the correct completion code after playing *Harmony Square* and were thus excluded ($n = 78$) does not differ meaningfully from the rest of the sample, aside from their political ideology (which skews slightly more to the right for excluded participants).

Supplementary analyses & robustness checks

We conducted two separate robustness checks to validate the main analyses. First, we ran a linear regression analysis for each of the 3 outcome variables above, with the post-test as the dependent (outcome) variable, the condition (control or treatment) as a dummy variable, and the pre-test as the independent variable, for the reliability judgments, confidence judgments, as well as participants’ willingness to share manipulative content. This analysis gives the same result as the ANOVA analysis that we ran for the difference scores above. The linear regression models for each outcome variable can be found in Supplementary Table S4. Second, following Pennycook et al. (2020), we also conducted a multi-level analysis with robust standard errors at the rating level, clustered on study participants and all 16 items (pre- and post-intervention). We find a significant interaction between pre-post differences and the treatment (inoculation) condition for the reliability, confidence and sharing measures, further validating the results reported above. These results are reported in Supplementary Table S5.

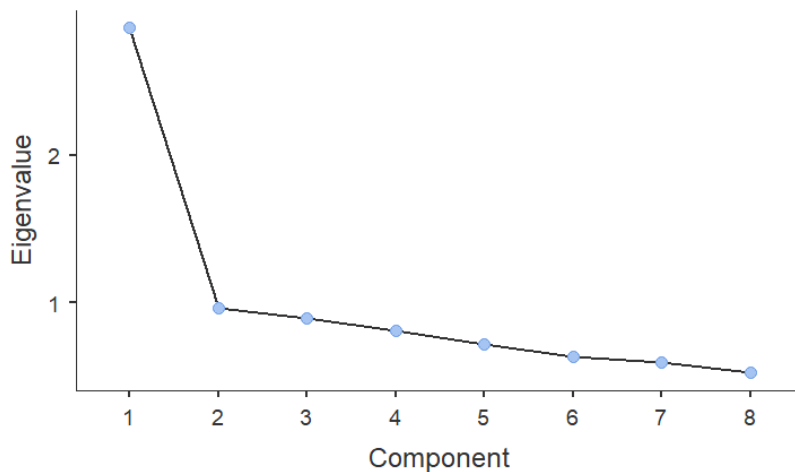
Items (social media posts) selection procedure & Principal Component Analysis

To maintain balance, we selected 4 posts per manipulation technique (2 fictional and 2 “real”), for a total of 2 sets of 8 items (16 items in total). We conducted an exploratory principal component analysis (PCA)

¹⁴ The dataset for the excluded participants ($n = 78$) is available on the OSF: <https://osf.io/r89h3/>.

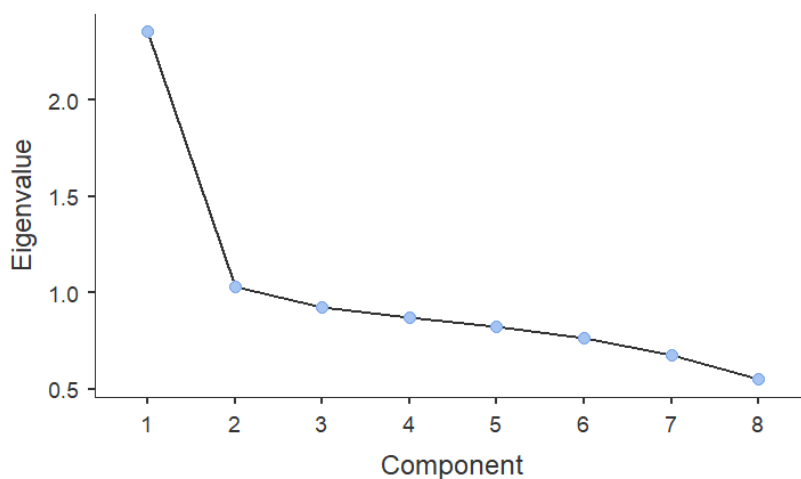
on both the “real” and the “fictional” item sets. Both sets loaded on a single dimension, with an eigenvalue of 2.35 for the “real” item set (accounting for 29.4% of the variance), and 2.86 for the “fictional” item set (accounting for 35.7% of the variance). Thus, for ease of interpretation and to limit multiple testing, both item sets were collapsed and treated as two measures, which we report throughout the paper. See Supplementary Figures S2 and S3 for the scree plots. To check for technique-level results, we also report the results for each individual manipulation technique taught in the game (both for the “real” and the fictional misinformation items) in Supplementary Table S3. In addition, descriptive statistics for each of the 16 items can be found in Supplementary Table S2.

Scree Plot



Supplementary Figure S2. Scree plot for reliability judgments following PCA for the “fictional” misinformation items.

Scree Plot



Supplementary Figure S3. Scree plot for reliability judgments following PCA for the “real” misinformation items.

Appendix: Table S1: Sample composition

Variable	N	%	N	%
	Included participants (N = 681)		Excluded (N = 78)	
Age				
18-24	282	41.4 %	20	25.6%
25-34	228	33.5 %	35	44.9%
35-44	96	14.1 %	11	14.1%
45-54	55	8.1 %	8	10.3%
55 or older	20	2.9 %	4	5.1%
Gender				
Female	294	43.2 %	38	48.7%
Male	379	55.7 %	40	51.3%
Other	8	1.2 %	0	0.00%
Education				
No formal education above age 16	5	0.7 %	0	0.00%
Professional or technical qualifications above age 16	25	3.7 %	3	3.8%
School education up to age 18	226	33.2 %	26	33.3%
Degree (Bachelor's) or equivalent	296	43.5 %	35	44.9%
Degree (Master's) or other postgraduate qualification	112	16.4 %	13	16.7%
Doctorate	17	2.5 %	1	1.3%
Country				
Italy	38	5.6 %	6	7.7%
Mexico	28	4.1 %	12	15.4%
Poland	74	10.9 %	12	15.4%
Portugal	75	11.0 %	3	3.8%
United Kingdom	46	6.8 %	5	6.4%
United States	315	46.3 %	29	37.2%
Other	105	15.4 %	11	14.1%
Other variables				
	N	SD	N	SD
How often do you check the news (1-5)	3.68	0.91	3.78	0.95
How often do you use social media? (1-5)	4.08	0.97	4.10	1.11
How interested are you in politics? (1-5)	3.34	1.19	3.28	1.12
Political ideology (1-7)	3.13	1.44	3.60	1.41

Appendix: Table S2: Descriptive statistics for social media posts (items), per outcome variable, averaged and by item

Note: The word “Real” in Supplementary Tables S2, S3, S4 and S6 refers to “real fake social media posts”, i.e., fake stories that have previously appeared online. It does not refer to “real news” (in the sense of truthful information). The word “fake” refers to fictional fake news social media posts (see also Supplementary Table S6).

Outcome variable	Condition	<i>N</i>	<i>M_{pre}</i>	<i>M_{post}</i>	<i>M_{diff}</i>	<i>SD_{pre}</i>	<i>SD_{post}</i>
Reliability judgments							
<i>Averaged per category</i>							
Real-Reliability	Control	385	3.279	3.12	-0.159	1.009	1.116
	Harmony Square	296	3.228	2.71	-0.518	0.951	1.145
Fake-Reliability	Control	385	3.034	2.939	-0.095	1.052	1.135
	Harmony Square	296	2.934	2.499	-0.435	1.065	1.189
<i>Per individual item</i>							
Real-Trolling-1-Reliability	Control	385	3.358	3.031	-0.327	2.052	1.976
	Harmony Square	296	3.22	2.439	-0.781	1.918	1.825
Real-Trolling-2-Reliability	Control	385	3.356	3.177	-0.179	1.918	1.841
	Harmony Square	296	3.341	2.905	-0.436	1.892	1.932
Real-Emotion-1-Reliability	Control	385	3.403	3.265	-0.138	1.663	1.668
	Harmony Square	296	3.297	2.764	-0.533	1.681	1.619
Real-Emotion-2-Reliability	Control	385	2.855	2.855	0	1.824	1.795
	Harmony Square	296	2.791	2.453	-0.338	1.763	1.681
Real-Conspir-1-Reliability	Control	385	2.291	2.369	0.078	1.595	1.623
	Harmony Square	296	2.203	1.922	-0.281	1.56	1.425
Real-Conspir-2-Reliability	Control	385	3.236	2.987	-0.249	1.715	1.709
	Harmony Square	296	3.182	2.78	-0.402	1.785	1.808
Real-Polariz-1-Reliability	Control	385	4.223	3.94	-0.283	1.902	1.902
	Harmony Square	296	4.149	3.301	-0.848	1.96	1.922
Real-Polariz-2-Reliability	Control	385	3.506	3.221	-0.285	1.676	1.703
	Harmony Square	296	3.642	2.959	-0.683	1.661	1.661
Fake-Trolling-1-Reliability	Control	385	2.771	2.748	-0.023	1.98	1.88
	Harmony Square	296	2.736	2.446	-0.29	1.853	1.786
Fake-Trolling-2-Reliability	Control	385	3.278	3.213	-0.065	1.567	1.604
	Harmony Square	296	3.358	2.774	-0.584	1.675	1.665
Fake-Emotion-1-Reliability	Control	385	2.836	2.818	-0.018	1.596	1.621
	Harmony Square	296	2.726	2.351	-0.375	1.627	1.475
Fake-Emotion-2-Reliability	Control	385	3.587	3.538	-0.049	1.69	1.754
	Harmony Square	296	3.497	2.726	-0.771	1.739	1.712
Fake-Conspir-1-Reliability	Control	385	2.987	2.831	-0.156	1.712	1.728
	Harmony Square	296	2.959	2.524	-0.435	1.646	1.577

Fake-Conspir-2-Reliability	Control	385	3.187	2.964	-0.223	1.759	1.726
	Harmony Square	296	2.959	2.503	-0.456	1.705	1.588
Fake-Polariz-1-Reliability	Control	385	2.688	2.686	-0.002	1.575	1.55
	Harmony Square	296	2.537	2.27	-0.267	1.615	1.519
Fake-Polariz-2-Reliability	Control	385	2.94	2.717	-0.223	1.688	1.523
	Harmony Square	296	2.699	2.399	-0.3	1.62	1.583

Confidence judgments

Averaged per category

Real-Pre-Confidence	Control	385	5.147	5.096	-0.051	1.069	1.179
	Harmony Square	296	5.265	5.415	0.15	1.027	1.189
Fake-Pre-Confidence	Control	385	5.068	5.057	-0.011	1.062	1.193
	Harmony Square	296	5.201	5.392	0.191	1.088	1.205

Per individual item

Real-Trolling-1-Pre-Confidence	Control	385	5.374	5.379	0.005	1.625	1.635
	Harmony Square	296	5.392	5.608	0.216	1.637	1.582
Real-Trolling-2-Pre-Confidence	Control	385	5.187	5.119	-0.068	1.558	1.535
	Harmony Square	296	5.378	5.547	0.169	1.456	1.463
Real-Emotion-1-Pre-Confidence	Control	385	4.958	4.878	-0.08	1.462	1.48
	Harmony Square	296	5.014	5.176	0.162	1.482	1.526
Real-Emotion-2-Pre-Confidence	Control	385	5.203	5.171	-0.032	1.519	1.565
	Harmony Square	296	5.179	5.412	0.233	1.63	1.564
Real-Conspir-1-Pre-Confidence	Control	385	5.239	5.195	-0.044	1.625	1.649
	Harmony Square	296	5.446	5.676	0.23	1.544	1.561
Real-Conspir-2-Pre-Confidence	Control	385	5.049	4.995	-0.054	1.452	1.615
	Harmony Square	296	5.324	5.453	0.129	1.432	1.488
Real-Polariz-1-Pre-Confidence	Control	385	5.265	5.132	-0.133	1.501	1.485
	Harmony Square	296	5.439	5.439	0	1.467	1.513
Real-Polariz-2-Pre-Confidence	Control	385	4.899	4.958	0.059	1.483	1.505
	Harmony Square	296	4.949	5.149	0.2	1.407	1.502
Fake-Trolling-1-Pre-Confidence	Control	385	5.626	5.403	-0.223	1.606	1.598
	Harmony Square	296	5.706	5.709	0.003	1.491	1.492
Fake-Trolling-2-Pre-Confidence	Control	385	4.769	4.8	0.031	1.523	1.489
	Harmony Square	296	4.777	5.152	0.375	1.535	1.53
Fake-Emotion-1-Pre-Confidence	Control	385	4.961	5.039	0.078	1.578	1.492
	Harmony Square	296	5.091	5.416	0.325	1.572	1.542
Fake-Emotion-2-Pre-Confidence	Control	385	4.888	4.891	0.003	1.414	1.459
	Harmony Square	296	5.084	5.206	0.122	1.434	1.55
Fake-Conspir-1-Pre-Confidence	Control	385	5.094	5.117	0.023	1.507	1.549
	Harmony Square	296	5.253	5.409	0.156	1.509	1.468
Fake-Conspir-2-Pre-Confidence	Control	385	5.073	5.062	-0.011	1.484	1.575
	Harmony Square	296	5.111	5.432	0.321	1.54	1.483

Fake-Polariz-1-Pre-Confidence	Control	385	5.112	5.096	-0.016	1.528	1.51
	Harmony Square	296	5.368	5.446	0.078	1.488	1.533
Fake-Polariz-2-Pre-Confidence	Control	385	5.018	5.047	0.029	1.528	1.492
	Harmony Square	296	5.22	5.365	0.145	1.519	1.523

Willingness to share

Averaged per category

Real-Pre-Sharing	Control	385	2.271	2.179	-0.092	1.187	1.261
	Harmony Square	296	2.255	2.003	-0.252	1.151	1.243
Fake-Pre-Sharing	Control	385	2.175	2.119	-0.056	1.208	1.257
	Harmony Square	296	2.106	1.904	-0.202	1.162	1.201

Per individual item

Real-Trolling-1-Pre-Sharing	Control	385	2.117	2.013	-0.104	1.668	1.606
	Harmony Square	296	2.152	1.902	-0.25	1.764	1.607
Real-Trolling-2-Pre-Sharing	Control	385	2.613	2.439	-0.174	1.921	1.808
	Harmony Square	296	2.804	2.324	-0.48	1.947	1.812
Real-Emotion-1-Pre-Sharing	Control	385	2.208	2.164	-0.044	1.663	1.629
	Harmony Square	296	2.071	1.899	-0.172	1.493	1.504
Real-Emotion-2-Pre-Sharing	Control	385	2.291	2.19	-0.101	1.791	1.658
	Harmony Square	296	2.236	2.007	-0.229	1.767	1.661
Real-Conspir-1-Pre-Sharing	Control	385	1.945	1.886	-0.059	1.5	1.492
	Harmony Square	296	1.858	1.723	-0.135	1.478	1.394
Real-Conspir-2-Pre-Sharing	Control	385	2.436	2.273	-0.163	1.698	1.655
	Harmony Square	296	2.358	2.041	-0.317	1.693	1.636
Real-Polariz-1-Pre-Sharing	Control	385	2.426	2.369	-0.057	1.84	1.801
	Harmony Square	296	2.318	2.088	-0.23	1.665	1.677
Real-Polariz-2-Pre-Sharing	Control	385	2.135	2.109	-0.026	1.549	1.532
	Harmony Square	296	2.243	2.024	-0.219	1.603	1.478
Fake-Trolling-1-Pre-Sharing	Control	385	2	2.01	0.01	1.599	1.584
	Harmony Square	296	2.051	1.922	-0.129	1.686	1.517
Fake-Trolling-2-Pre-Sharing	Control	385	2.135	2.096	-0.039	1.585	1.534
	Harmony Square	296	2.037	1.878	-0.159	1.528	1.387
Fake-Emotion-1-Pre-Sharing	Control	385	2.143	2.187	0.044	1.672	1.629
	Harmony Square	296	2.068	1.868	-0.2	1.616	1.542
Fake-Emotion-2-Pre-Sharing	Control	385	2.457	2.356	-0.101	1.729	1.72
	Harmony Square	296	2.378	2.007	-0.371	1.659	1.58
Fake-Conspir-1-Pre-Sharing	Control	385	2.369	2.171	-0.198	1.724	1.656
	Harmony Square	296	2.216	1.922	-0.294	1.651	1.451
Fake-Conspir-2-Pre-Sharing	Control	385	2.275	2.19	-0.085	1.719	1.65
	Harmony Square	296	2.182	1.963	-0.219	1.637	1.521
Fake-Polariz-1-Pre-Sharing	Control	385	2.013	1.987	-0.026	1.52	1.471
	Harmony Square	296	1.932	1.838	-0.094	1.455	1.466
Fake-Polariz-2-Pre-Sharing	Control	385	2.01	1.956	-0.054	1.472	1.476
	Harmony Square	296	1.983	1.834	-0.149	1.458	1.344

Appendix: Table S3: ANOVAs for difference scores, averaged by type and by each of the 4 manipulation techniques.

Note: We supply these results for completeness but p -values are unadjusted.

One-Way ANOVA (Fisher's)						
Variable	F	$df1$	$df2$	p	η^2	d
Reliability judgments						
<i>Averaged per category</i>						
Real-Reliability-Diff	43.205	1	679	< .001	0.06	0.51
Fake-Reliability-Diff	48.788	1	679	< .001	0.067	0.54
<i>Averaged per technique</i>						
Real-Trolling-Reliability-Diff	18.82	1	679	< .001	0.027	0.33
Real-Emotion-Reliability-Diff	24.22	1	679	< .001	0.034	0.38
Real-Conspiracy-Reliability-Diff	11.36	1	679	< .001	0.016	0.26
Real-Polarization-Reliability-Diff	25.97	1	679	< .001	0.037	0.39
Fake-Trolling-Reliability-Diff	24.91	1	679	< .001	0.035	0.38
Fake-Emotion-Reliability-Diff	48.04	1	679	< .001	0.066	0.53
Fake-Conspiracy-Reliability-Diff	11.01	1	679	< .001	0.016	0.26
Fake-Polarization-Reliability-Diff	4.64	1	679	0.032	0.007	0.17
Confidence judgments						
<i>Averaged per category</i>						
Real-Confidence-Diff	14.518	1	679	< .001	0.021	0.29
Fake-Confidence-Diff	14.547	1	679	< .001	0.021	0.29
<i>Averaged per technique</i>						
Real-Trolling-Confidence-Diff	8.99	1	585	0.003	0.014	0.24
Real-Emotion-Confidence-Diff	9.83	1	534	0.002	0.015	0.25
Real-Conspiracy-Confidence-Diff	7.39	1	522	0.007	0.012	0.22
Real-Polarization-Confidence-Diff	2.36	1	565	0.125	0.004	0.13
Fake-Trolling-Confidence-Diff	12.44	1	608	< .001	0.018	0.27
Fake-Emotion-Confidence-Diff	4.71	1	587	0.03	0.007	0.17
Fake-Conspiracy-Confidence-Diff	8.34	1	603	0.004	0.012	0.22
Fake-Polarization-Confidence-Diff	1.56	1	585	0.212	0.002	0.09
Willingness to share						
<i>Averaged per category</i>						
Real-Sharing-Diff	12.85	1	679	< .001	0.019	0.28
Fake-Sharing-Diff	12.619	1	679	< .001	0.018	0.27
<i>Averaged per technique</i>						
Real-Trolling-Sharing-Diff	8.83	1	679	0.003	0.013	0.23
Real-Emotion-Sharing-Diff	3.03	1	679	0.082	0.004	0.13
Real-Conspiracy-Sharing-Diff	3.4	1	679	0.066	0.005	0.14
Real-Polarization-Sharing-Diff	6.28	1	679	0.012	0.009	0.19
Fake-Trolling-Sharing-Diff	3.66	1	679	0.056	0.005	0.14
Fake-Emotion-Sharing-Diff	14.19	1	679	< .001	0.02	0.29
Fake-Conspiracy-Sharing-Diff	2.88	1	679	0.09	0.004	0.13
Fake-Polarization-Sharing-Diff	1.51	1	679	0.22	0.002	0.09

Appendix: Table S4: Linear regression model robustness check

Note: Post-test is used as the dependent variable, condition (control-treatment) as the dummy, and pre-test as covariate. Regression models are displayed by outcome variable.

Outcome measure				
Real fake news - reliability				
Model fit	<i>R</i>	<i>R</i> ²		
	0.792	0.627		
Model Coefficients - Real-Post-Reliability				
Predictor	<i>Estimate</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept ^a	0.17	0.0964	1.76	0.079
Harmony Square – Control	-0.364	0.0542	-6.72	< .001
Real-Pre-Reliability	0.90	0.0273	32.92	< .001
Fictional fake news - reliability				
Model fit	<i>R</i>	<i>R</i> ²		
	0.848	0.72		
Model Coefficients - Fake-Post-Reliability				
Predictor	<i>Estimate</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept ^a	0.138	0.0758	1.81	0.07
Harmony Square – Control	-0.348	0.0483	-7.19	< .001
Fake-Pre-Reliability	0.923	0.0227	40.73	< .001
Real fake news - confidence				
Model fit	<i>R</i>	<i>R</i> ²		
	0.824	0.679		
Model Coefficients - Real-Post-Confidence				
Predictor	<i>Estimate</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept ^a	0.34	0.1317	2.58	0.01
Harmony Square – Control	0.21	0.0524	4	< .001
Real-Pre-Confidence	0.924	0.0247	37.41	< .001
Fictional fake news - confidence				
Model fit	<i>R</i>	<i>R</i> ²		
	0.829	0.687		
Model Coefficients - Fake-Post-Confidence				
Predictor	<i>Estimate</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept ^a	0.39	0.1275	3.06	0.002
Harmony Square – Control	0.212	0.0525	4.04	< .001
Fake-Pre-Confidence	0.921	0.0242	38.02	< .001

Real fake news - sharing

Model fit	<i>R</i>	<i>R</i> ²		
	0.889	0.79		
Model Coefficients - Real-Post-Sharing				
Predictor	<i>Estimate</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept ^a	0.0202	0.0519	0.39	0.697
Harmony Square – Control	-0.1611	0.0445	-3.618	< .001
Real-Pre-Sharing	0.9505	0.0189	50.379	< .001

Fictional fake news - sharing

Model fit	<i>R</i>	<i>R</i> ²		
	0.905	0.82		
Model Coefficients - Fake-Post-Sharing				
Predictor	<i>Estimate</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept ^a	0.077	0.0456	1.69	0.092
Harmony Square – Control	-0.1499	0.0407	-3.69	< .001
Fake-Pre-Sharing	0.9388	0.017	55.27	< .001

^a Represents reference level

Appendix: Table S5: Linear regression model at the rating level following Pennycook et al. (2020)

Note: The code used for the analysis (in STATA) can be found on the OSF. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$. Coefficients are standardized.

Difference (reliability/confidence/sharing)	Reliability	Confidence	Sharing
Difference (pre=0, post=1)	-0.134	-0.0269	-0.737
Condition (Control=1, Treatment=2)	-0.0754*	0.126***	-0.0428
Difference * Condition	-0.352***	0.206***	-0.154***
Constant	3.156***	5.107***	2.223***
Observations	21,792	21,792	21,792
Subject clusters	681	681	681
Item clusters	32	32	32
R ²	0.0138	0.0069	0.0037
F (3, 21788)	101.63	50.78	27.26

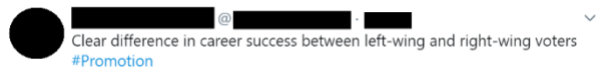
Appendix: Table S6: Items (social media posts)

Note: “Real” stands for real-life examples of misinformation, “fake” stands for fictional misinformation. Throughout the supplement, items from this table are referred to as “Real-Trolling-1”, “Fake-Trolling-1”, et cetera.

Item name	Real	Fake
Trolling-1		
Trolling-2		
Emotion-1		
Emotion-2		
Conspiracy-1		
Conspiracy-2		



Polarization-1



Polarization-2

Appendix: Figure S1: Supplementary bar plots

Note: Bar plots show the pre- and post-scores for the reliability, confidence, and sharing measures, for both “real fake news” (labelled “real” in the figure) and “fictional fake news” (labelled “fake”) by group. Error bars represent 95% confidence intervals. The figure shows that while perceived reliability for both “real” and “fictional” fake news goes down after playing for the treatment group, this is not the case for the control group. The pattern is the same for the confidence and sharing measures.

