Harvard Kennedy School Misinformation Review[1]
April 2025, Volume 6, Issue 2

DOI: https://doi.org/10.37016/mr-2020-175
Website: misinforeview.hks.harvard.edu

Research Note

Feedback and education improve human detection of image manipulation on social media

This study investigates the impact of educational interventions and feedback on users' ability to detect manipulated images on social media, addressing a gap in research that has primarily focused on algorithmic approaches. Through a pre-registered randomized and controlled experiment, we found that feedback and educational content significantly improved participants' ability to detect manipulated images on social media. However, the educational content did not result in a significantly greater improvement compared to feedback alone. These findings underscore feedback as a powerful tool for enhancing digital literacy, with practical implications for combating misinformation.

Authors: Adnan Hoq (1), Matthew J. Facciani (1), Tim Weninger (1)
Affiliations: (1) Department of Computer Science and Engineering, University of Notre Dame, USA
How to cite: Hoq, A., Facciani, M. J., & Weninger, T. (2025). Feedback and education improve human detection of image manipulation on social media. *Harvard Kennedy School (HKS) Misinformation Review*, 6(2).
Received: November 15th, 2024. Accepted: March 13th, 2025. Published: April 2nd, 2025.

Research questions

- Does feedback and education improve users' accuracy in detecting manipulated images compared to no intervention?
- Does accuracy in detecting different types of image manipulation (copy-move, erase-fill, touch-up, etc.) vary across treatment and control groups?
- Does spending more time on image classification predict higher accuracy in detecting manipulation?

Essay summary

- This study employed a pre-registered randomized and controlled experiment to evaluate the impact of feedback and educational interventions on improving participants' ability to detect manipulated images on social media.
- Task feedback and feedback combined with educational content both significantly improved participants' ability to detect manipulated images. However, there was no significant difference between the two groups, which was somewhat unexpected.
- These findings suggest that feedback is an effective tool for enhancing digital literacy and combating misinformation. Given that the educational intervention did not additionally improve

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

accuracy, future interventions should focus on optimizing feedback systems and consider alternative educational strategies to improve image manipulation detection.

Implications

Image manipulation and media literacy interventions

In recent years, there has been a surge in manipulated content on the Internet (Shen et al., 2019; Wang et al., 2022). This sudden increase has made automatic image manipulation detection (e.g., DeepFake detection) a prominent area of research in computing and social sciences (Fridrich, 2009; Novozamsky et al., 2020; Thakur & Rohilla, 2020; Tyagi & Yadav, 2023). Much of the work in computing focuses on developing algorithms that utilize artificial intelligence (AI) systems to detect altered imagery (Bayar & Stamm, 2018; Cao et al., 2012; Mahdian & Saic, 2007; Wang et al., 2022; Zanardelli et al., 2023) with various technological approaches that work to varying degrees on different types of image and video alteration (Armas Vega et al., 2021; Chen et al., 2021; Cozzolino & Verdoliva, 2016; Huang & Ciou, 2019; Liu et al., 2022; Rossler et al., 2019; Yang et al., 2020; Zhang et al., 2020; Zhang et al., 2016; Zhou et al., 2017).

Image manipulation plays a significant role in spreading misinformation and disinformation, as altered visuals can mislead viewers by distorting reality, fueling biased narratives, or amplifying false messages (Ghai et al., 2024; Newman & Schwarz, 2023; Weikmann & Lecheler, 2023). This form of visual misinformation has profound societal impacts, influencing public opinion and trust in media (Langguth et al., 2021; Matatov et al., 2022; Yang et al., 2023); especially because social media users are typically unaware of the scale of image manipulations and are not good at detecting when manipulations are present (Nightingale et al., 2017; Schetinger et al., 2017).

Although advancements in algorithmic detection of image manipulation are essential, they must be paired with a deeper understanding of the user dynamics involved. There is a growing body of research demonstrating the effectiveness of online games that teach media and information literacy to users (Basol et al*.,* 2020; Facciani et al*.,* 2024; Roozenbeek & van der Linden, 2020). To effectively empower users to interact responsibly on social media, interface design decisions should work in tandem with media literacy efforts. This combination enables individuals to navigate online content with discernment, critically evaluate information, and make informed decisions for themselves and their followers.

Epstein and colleagues (2021) evaluated the impact of several different interventions on participant's willingness to share true versus false headlines. The results showed that simple interventions, such as asking participants to reflect on the accuracy of headlines or providing basic digital literacy tips, significantly improved discernment between real and false news. In March of 2025, Reddit instituted a similar feedback mechanism by issuing warnings to users who repeatedly upvoted violent content, employing behavioral feedback to discourage interaction with policy-violating material.[2] The findings from Epstein and colleagues (2021) are consistent with other work showing that simple accuracy nudges improve discernment (Pennycook et al*.,* 2020) along with media literacy training (Guess et al*.,* 2020). As research evolves in this area, researchers should strive to understand if certain interventions are more effective for certain contexts, audiences, or topics.

This study examines how user behavior, interface design, and media literacy interventions intersect to improve detection of image manipulation. We evaluate the impact of feedback and educational strategies on participants' ability to identify manipulated images, analyzing their interactions with a

[2] https://www.reddit.com/r/RedditSafety/comments/1j4cd53/warning_users_that_upvote_violent_content/

curated dataset of real (non-manipulated) and altered (manipulated) visuals to identify common errors and decision-making challenges. Figure 1 highlights key interventions implemented. This work contributes to the integration of technological tools with human-centered approaches to support critical media literacy skills.

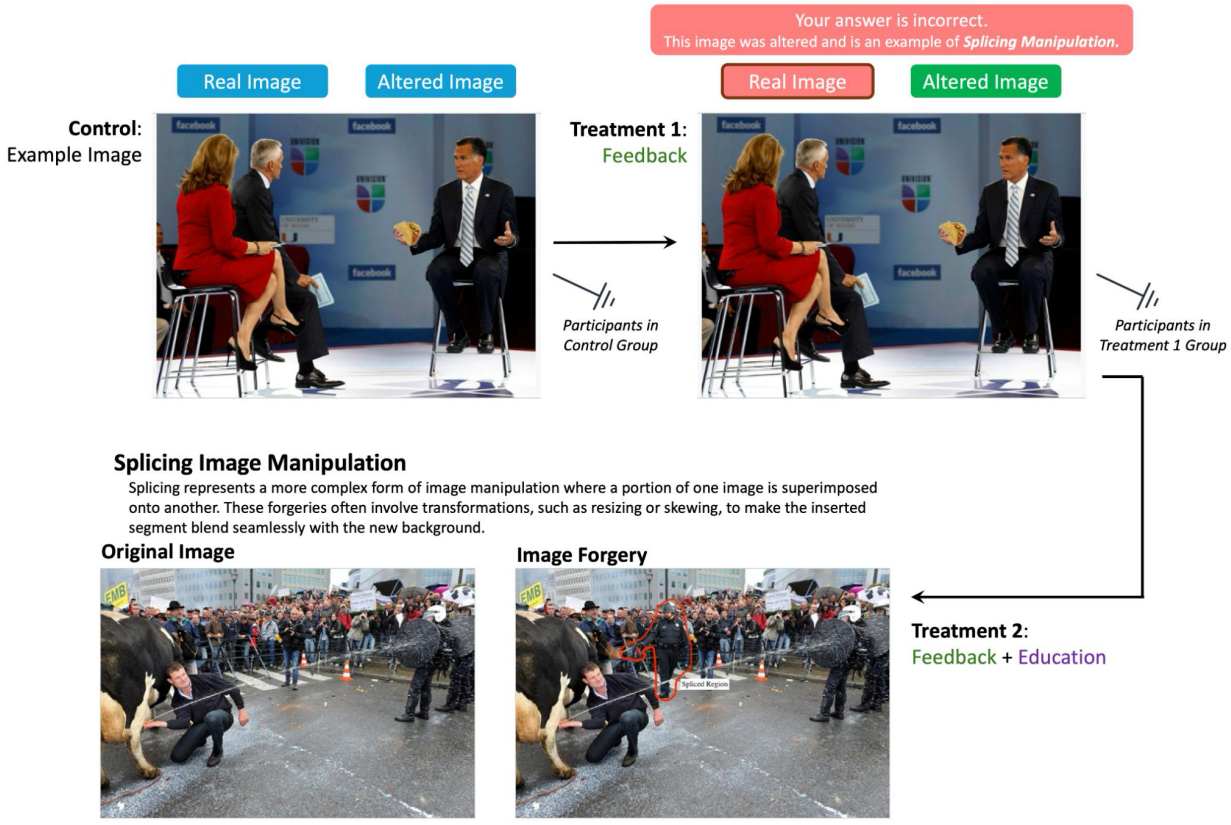

Figure 1. Illustration of the experimental design. *The design includes (1) the control group, which asks for participants to determine if an image is real (non-manipulated) or altered (manipulated), with no feedback; (2) the first treatment group, which notifies participants if their answer was correct or not; and (3) the second treatment group, which, in addition to the feedback treatment, provides additional educational lessons on the type of alteration that was performed.*

Implications from the present study

This study suggests that rapid, structured feedback significantly enhances users' ability to detect manipulated content. These findings may inform user interface design, media literacy education, and public policy, emphasizing three actionable strategies:

1. Enhanced user feedback systems. Real-time feedback systems, such as user-friendly notifications and algorithmic prompts, can effectively flag suspicious content. Features like tooltips on Instagram or links on Twitter/X could provide context for users. Transparency and nonpartisanship are essential to foster trust in these systems. Tailored interventions should consider political motivations behind manipulations and users' varying levels of media literacy. Periodic reminders could further encourage critical engagement.

2. Educational integration. Incorporating hands-on workshops and structured feedback into school curricula can strengthen digital literacy. Gamified and personalized learning experiences can increase user engagement while fostering collaborative and critical evaluation skills.

3. Policy and platform recommendations. Our findings suggest that policymakers could consider supporting digital literacy campaigns and encouraging platforms to adopt enhanced feedback tools. Social media companies could benefit from prioritizing user empowerment by offering accessible educational features and tools that encourage discerning content evaluation, which may contribute to a more reliable information-sharing environment.

By combining these approaches, stakeholders can address misinformation more effectively and create a media landscape that supports informed and critical users.

Findings

Finding 1: Providing feedback and media literacy improves user detection of manipulated images.

We used ordinary least squares (OLS) regression analysis, a statistical method that examines how different factors influence an outcome, to evaluate the impact of feedback and media literacy interventions. Participants were shown images and asked to classify them as real (*non-manipulated*) or altered (*manipulated*). The outcome (i.e., dependent variable [DV]) was the total number of correct classifications, while the factors (i.e., independent variables [IVs]) represented each treatment condition compared to a control group. As shown in Figure 2, the OLS results indicated that participants in both the *feedback* and the *feedback+education* groups achieved significantly higher accuracy than those in the control group ($p < .001$). Additionally, the OLS regression results revealed that familiarity with manipulated content had a small yet statistically significant positive effect on accuracy, while political orientation did not have a significant effect.

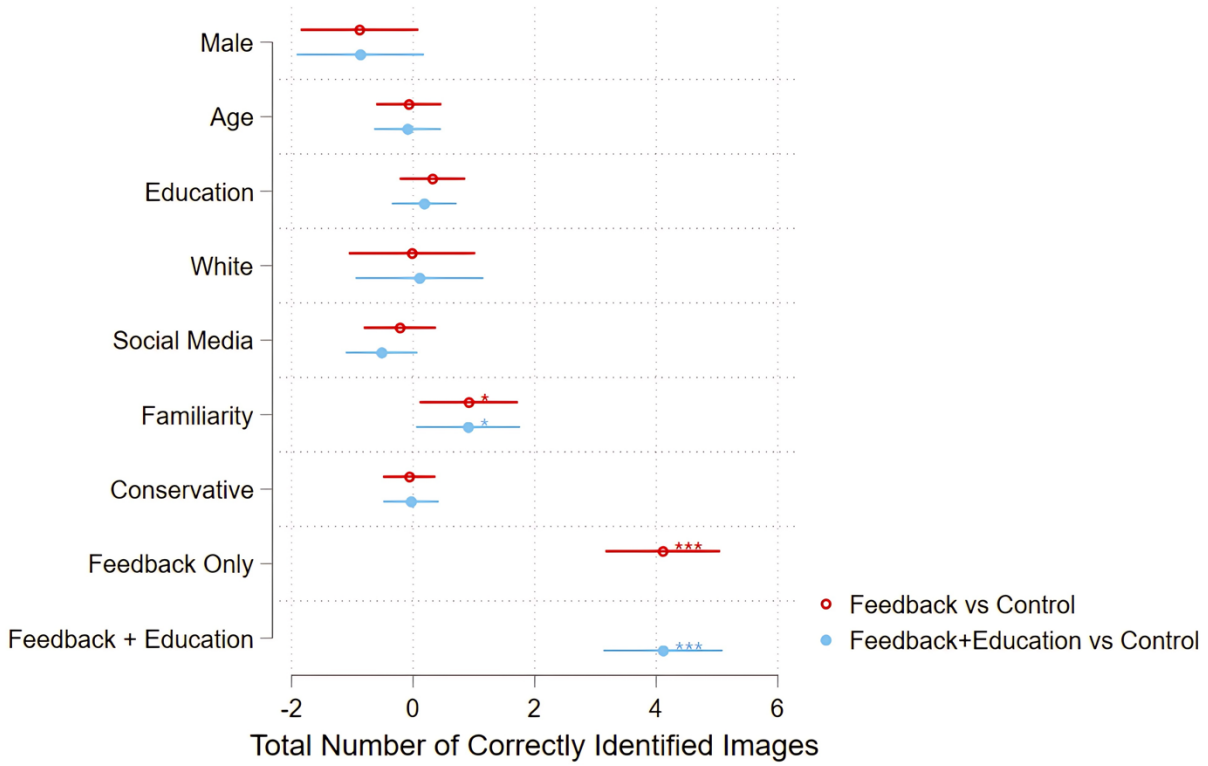

Figure 2. Participants in the feedback (treatment 1) and feedback+education (treatment 2) conditions had significantly higher correct responses compared to the control group. *Error bars show the variability within each group, illustrating the consistency of the observed improvements.*

A one-way analysis of variance (ANOVA) was conducted to compare the mean number of correct classifications among the three groups, yielding significant differences, $F(2, 267) = 67.29$, $p < .001$. ANOVA is a statistical test used to compare the means of multiple groups to determine if any significant differences exist among them. The control group achieved a mean of correct classifications of 18.02 out of 32. The *feedback* and *feedback+education* treatment groups achieved means of 22.09 and 22.25, respectively. Thus, our treatment groups increased image classification by about 4 images out of 32 possible images. Importantly, both treatment groups appeared to improve performance equally well and there was no significant difference between their image classification scores ($p > .05$). We also conducted an exploratory analysis that found this treatment effect significantly improved image accuracy at the $p < .001$ level when only comparing the scores of the first four images, suggesting a rapid learning effect from our treatments. The results of ANOVA, regression, and power analysis can be found in the Appendix.

Both treatment groups significantly improved participants' accuracy in identifying manipulated images compared to the control group, with an average increase of about four correctly classified images out of 32. However, there was no significant difference between the *feedback* and *feedback+education* groups, suggesting that both interventions were equally effective. This result was somewhat unexpected, as we anticipated education would have a more pronounced effect. One possible explanation is that the educational intervention was too brief and simple, which did not lead to any additional effect. Feedback, which is itself a type of educational intervention (Epstein et al., 2021), appeared to be sufficient to improve accuracy. These findings underscore the value of feedback and suggest that more robust or long-term educational interventions may be needed to enhance participants' detection abilities. Future research could explore the role of different types of educational interventions to better understand their impact.

Finding 2: The type of image manipulation matters.

We conducted one-way ANOVAs to evaluate how well participants classified images based on manipulation types (*erase-fill*, *splicing*, *copy-move*, and *Photoshop touch-up*) and intervention groups (*control*, *feedback*, and *feedback+education*). The analysis revealed a significant effect of manipulation type within the control group, $F(3, 1485) = 19.38$, $p < .001$, meaning that both the type of image manipulation and the intervention influenced participants' accuracy.

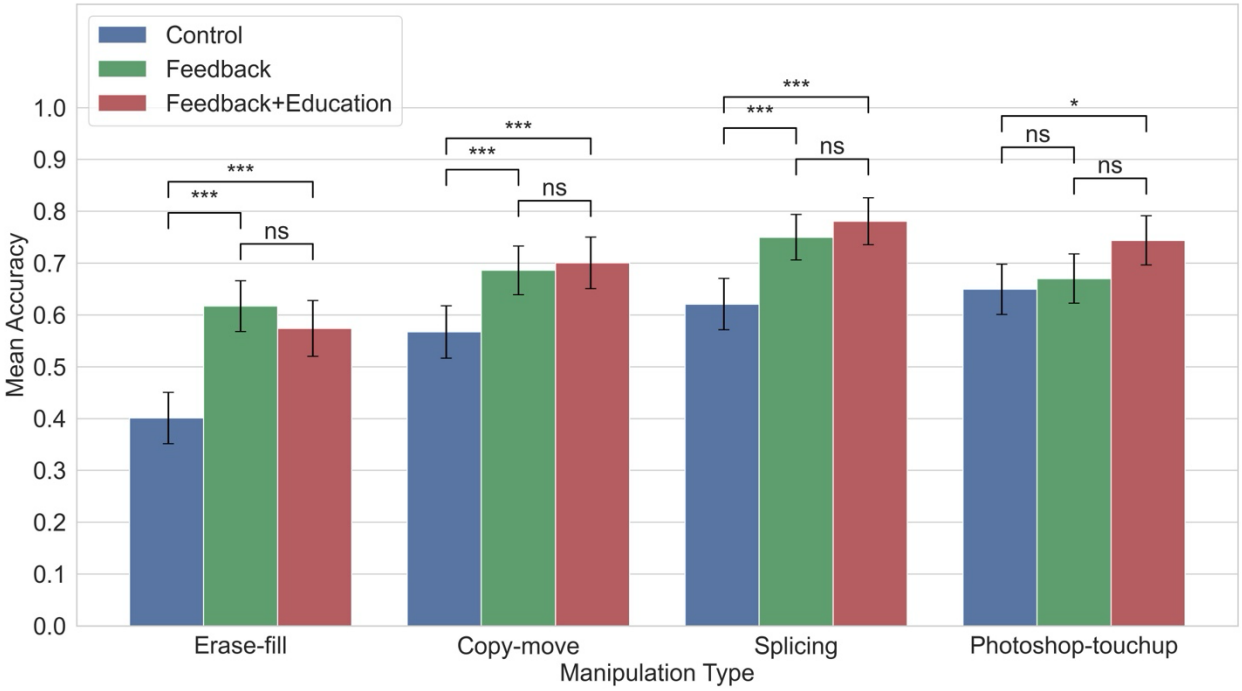

Figure 3. This figure compares mean accuracy across manipulation types and intervention groups. *The x-axis represents the manipulation types, and the y-axis shows the average accuracy scores. Significant improvements are indicated by asterisks. Note: ***p < .001, *p < .01, and ns = nonsignificant. Both feedback and feedback+education were significantly higher than the control group for every manipulation type except for the Photoshop touchup. Erase-fill was the most challenging for the participants. In the Photoshop-touchup condition, the feedback+education had a p-value of < .01 (*) compared to control and the feedback compared to control was nonsignificant.*

To better understand these results, we performed post hoc Bonferroni tests, which are follow-up analyses used to pinpoint specific differences between groups while accounting for multiple comparisons. As shown in Figure 3, these tests revealed that erase-fill manipulations were significantly harder to detect compared to splicing, copy-move, and Photoshop touch-up manipulations ($p < .001$). These findings highlight that some manipulation techniques, like erase-fill, pose a greater challenge for users, even when interventions are in place.

Feedback significantly improved user accuracy in detecting image manipulations across all types, with no additional benefit from the educational intervention. Erase-fill manipulations remained the most challenging for users, even with feedback, while splicing, copy-move, and Photoshop touch-up were relatively easier to classify. Bonferroni tests confirmed that feedback enhanced accuracy for more difficult manipulations, highlighting its role as the primary driver of improvement. These results suggest that tailoring feedback systems to address particularly challenging manipulation types, like erase-fill, could maximize their effectiveness.

Finding 3: Longer response times are associated with lower accuracy.

Our results show that each additional second spent on a decision slightly reduced the likelihood of a correct response (b = -0.034, 95% CI [-0.040, -0.027]; Table A12). Correlation analysis confirmed this negative relationship (r_{pb}= -0.12, $p < 0.001$). One explanation is that overthinking led to poorer judgments (Junghaenel et al., 2022). Alternatively, participants may have simply spent more time on difficult images, which were harder to classify regardless of time spent. In other words, rather than increased time causing lower accuracy, both prolonged response times and reduced accuracy may stem from the underlying difficulty of detecting certain manipulations.

Cognitive overload may also play a role—complex manipulations like erase-fill edits likely required more mental effort, leading to decision fatigue and lower accuracy. Additionally, distractions such as smartphone use may have further impacted performance (Ward et al., 2017). Future research could explore how task difficulty, cognitive load, and environmental factors interact to influence accuracy in detecting image manipulation.

Our study is not without limitations. It focused on images from a specific online community (/r/photoshopbattles), and the intervention's effectiveness for politically motivated or other specialized manipulations remains uncertain. Notably, there was no significant difference in accuracy between participants receiving feedback alone and those receiving feedback plus an educational intervention, suggesting that reflecting on the accuracy of their classifications during feedback may enhance critical thinking (Epstein et al., 2021). Additionally, as prior research has shown, media literacy interventions often have short-lived effects (Maertens et al., 2021). While we did not assess long-term retention, this remains an important consideration for future studies. Tailoring interventions to evolving manipulation techniques and measuring their lasting impact could further enhance detection accuracy over time.

Methods

Procedure

This pre-registered study employed a between-subjects design with one factor (educational intervention) across three groups. As shown in Figure 4, participants were randomly assigned to one of three groups: a control group (*no feedback*), Treatment Group 1 (*feedback*), or Treatment Group 2 (*feedback plus an educational intervention*). The educational intervention covered four image manipulation techniques: copy-move, splicing, erase-fill, and Photoshop touch-up. To ensure a balanced and randomized sample, each participant was shown 16 real (non-manipulated) and 16 altered (manipulated) images, with four from each manipulation type. This setup reflects the variety and unpredictability of social media content, making the findings more applicable to real-world scenarios. Instructions clarified the terms *real image* as non-manipulated and *altered image* as manipulated before participants began the task. Accuracy in detecting manipulated images was the primary outcome measure, evaluated based on responses to specific manipulation techniques.

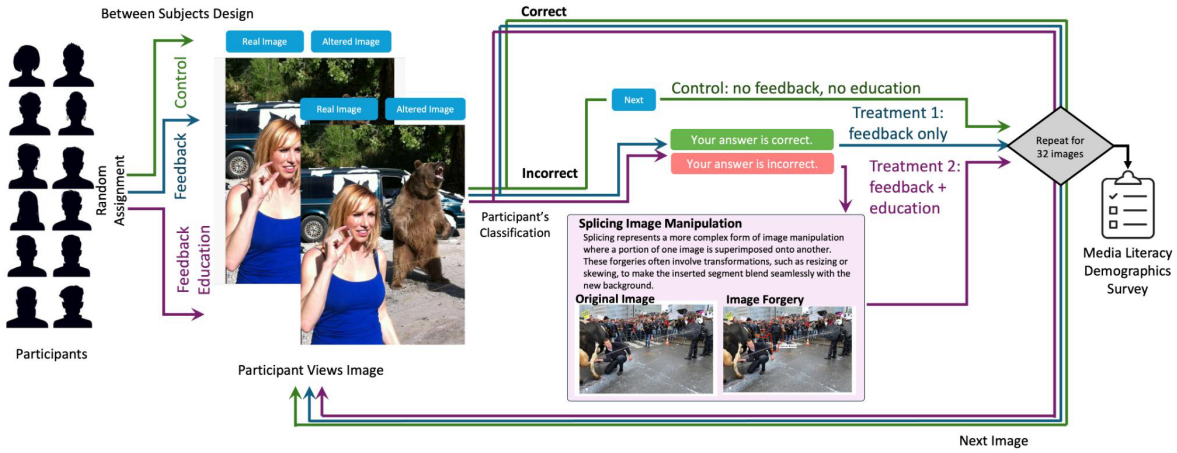

Figure 4. Illustration of the experimental procedure. *The between-subjects design assigned participants to either the control group, treatment 1 group, or treatment 2 group at the outset. Participants were provided with a real or altered image (randomly). If they correctly classified the image, then the next image was shown. If they were incorrect, then the treatment effects were applied as applicable.*

Participants

A total of 300 participants were recruited via Prolific, a platform for crowdsourcing tasks for academic research. Of those, 268 participants completed the survey. Participants were compensated $1.75 for completing the task. Upon completion, they provided demographic information, including race, age, educational background, and experience with digital image manipulation, allowing us to understand how different groups engaged with the task. The sample was 67% White and 33% non-White. Gender demographics were 42% women and 58% men or other identities. Ages were 53% under 35 and 47% over 35. Education levels were 64% with college or graduate degrees and 36% without. Social media usage was 78% daily and 22% less frequent. Familiarity with image manipulation was 93% somewhat familiar or higher and 7% not familiar. Politically, 56% were liberal or somewhat liberal and 44% moderate, conservative, or apolitical. A detailed breakdown is in the Appendix.

Image sourcing

This study used a dataset of 202 images from the subreddit /r/photoshopbattles, an online community focused on image manipulation. This dataset was chosen over deepfake datasets because it better represents the types of manipulations commonly encountered on social media, ranging from subtle edits to complex composites. The dataset included 79 authentic images (which provided enough manipulated images to ensure even distribution across types, enabling balanced random sampling) and 123 manipulated ones, categorized into four manipulation types:

1. Erase-fill (28 images): Removing parts of an image and seamlessly filling the gaps using tools like content-aware fill (Thakur & Rohilla, 2020).
2. Copy-move (29 images): Duplicating and repositioning regions within the same image, often with scaling or rotation (Khudhair et al., 2021).
3. Splicing (36 images): Combining portions of multiple images into a seamless final product, making detection challenging (Thakur & Rohilla, 2020).
4. Photoshop touchup (30 images): Subtle adjustments such as color correction, smoothing, or selective edits (Swerzenski, 2021; Wang et al., 2019).

Bibliography

Armas Vega, E. A., González Fernández, E., Sandoval Orozco, A. L., & García Villalba, L. J. (2021). Copy-move forgery detection technique based on discrete cosine transform blocks features. *Neural Computing and Applications*, *33*, 4713–4727. https://doi.org/10.1007/s00521-020-05433-1

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). https://doi.org/10.5334%2Fjoc.91

Bayar, B., & Stamm, M. C. (2018). Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security*, *13*(11), 2691–2706. https://doi.org/10.1109/TIFS.2018.2825953

Cao, Y., Gao, T., Fan, L., & Yang, Q. (2012). A robust detection algorithm for copy-move forgery in digital images. *Forensic Science International*, *214*(1–3), 33–43. https://doi.org/10.1016/j.forsciint.2011.07.015

Chen, X., Dong, C., Ji, J., Cao, J., & Li, X. (2021). Image manipulation detection by multi-view multi-scale supervision. In *2021 IEEE/CVF international conference on computer vision* (ICCV) (pp. 14185–14193). IEEE. https://doi.org/10.1109/ICCV48922.2021.01392

Cozzolino, D., & Verdoliva, L. (2016). Single-image splicing localization through autoencoder-based anomaly detection. In *2016 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–6). IEEE. https://doi.org/10.1109/WIFS.2016.7823921

Epstein Z., Berinsky A. J., Cole, R., Gully, A., Pennycook, G., & Rand, D. G. (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School (HKS) Misinformation Review, 2*(3). https://doi.org/10.37016/mr-2020-71

Facciani, M. J., Apriliawati, D., & Weninger, T. (2024). Playing Gali Fakta inoculates Indonesian participants against false information. *Harvard Kennedy School (HKS) Misinformation Review, 5*(4). https://doi.org/10.37016/mr-2020-152

Fridrich, J. (2009). Digital image forensics. *IEEE Signal Processing Magazine*, 26(2), 26–37. https://doi.org/10.1109/MSP.2008.931078

Ghai, A., Kumar, P., & Gupta, S. (2024). A deep-learning-based image forgery detection framework for controlling the spread of misinformation. *Information Technology & People*, *37*(2), 966–997. https://doi.org/10.1108/ITP-10-2020-0699

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Huang, H. Y., & Ciou, A. J. (2019). Copy-move forgery detection for image forensics using the superpixel segmentation and the Helmert transformation. *EURASIP Journal on Image and Video Processing*, 2019, 68. https://doi.org/10.1186/s13640-019-0469-9

Junghaenel, D. U., Schneider, S., Orriens, B., Jin, H., Lee, P. -J., Kapteyn, A., Meijer, E., Zelinski, E., Hernandez, R., & Stone, A. A. (2022). Inferring cognitive abilities from response times to web-administered survey items in a population-representative sample. *Journal of Intelligence*, *11*(1), 3. https://doi.org/10.3390/jintelligence11010003

Khudhair, Z. N., Mohamed, F., & Kadhim, K. A. (2021, April). A review on copy-move image forgery detection techniques. *Journal of Physics: Conference Series, 1892*(1), 012010. https://doi.org/10.1088/1742-6596/1892/1/012010

Langguth, J., Pogorelov, K., Brenner, S., Filkuková, P., & Schroeder, D. T. (2021). Don't trust your eyes: Image manipulation in the age of DeepFakes. *Frontiers in Communication*, *6*, 632317. https://doi.org/10.3389/fcomm.2021.632317

Liu, X., Liu, Y., Chen, J., & Liu, X. (2022). PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(11), 7505–7517. https://doi.org/10.1109/TCSVT.2022.3189545

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1–16. https://doi.org/10.1037/xap0000315

Mahdian, B., & Saic, S. (2007). Detection of copy–move forgery using a method based on blur moment invariants. *Forensic Science International*, 171(2–3), 180–189. https://doi.org/10.1016/j.forsciint.2006.11.002

Matatov, H., Naaman, M., & Amir, O. (2022). Stop the [Image] steal: The role and dynamics of visual content in the 2020 US election misinformation campaign. *Proceedings of the ACM on human-computer interaction*, *6*(CSCW2), 1–24. https://doi.org/10.1145/3555599

Newman, E. J., & Schwarz, N. (2023). Misinformed by images: How images influence perceptions of truth and what can be done about it. *Current Opinion in Psychology*, *56*, 101778. https://doi.org/10.1016/j.copsyc.2023.101778

Nightingale, S. J., Wade, K. A., & Watson, D. G. (2017). Can people identify original and manipulated photos of real-world scenes? *Cognitive research: Principles and Implications*, *2*, 30. https://doi.org/10.1186/s41235-017-0067-2

Novozamsky, A., Mahdian, B., & Saic, S. (2020). IMD2020: A large-scale annotated dataset tailored for detecting manipulated images. In *2020 IEEE winter applications of computer vision workshops* (pp. 71–80). IEEE. https://doi.org/10.1109/WACVW50321.2020.9096940

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,* 3*1*(7), 770–780. https://doi.org/10.1177/0956797620939054

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). https://doi.org/10.37016/mr-2020-47

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *2019 IEEE/CVF international conference on computer vision* (pp. 1–11). https://doi.org/10.1109/ICCV.2019.00009

Schetinger, V., Oliveira, M. M., da Silva, R., & Carvalho, T. J. (2017). Humans are easily fooled by digital images*. Computers & Graphics*, *68*, 142–151. https://doi.org/10.1016/j.cag.2017.08.010

Shen, C., Kasra, M., Pan, W., Bassett, G. A., Malloch, Y., & O'Brien, J. F. (2019). Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New Media & Society*, *21*(2), 438–463. https://doi.org/10.1177/1461444818799526

Swerzenski, J. D. (2021). Fact, fiction or Photoshop: Building awareness of visual manipulation through image editing software. *Journal of Visual Literacy*, *40*(2), 104–124. https://doi.org/10.1080/1051144X.2021.1902041

Thakur, R., & Rohilla, R. (2020). Recent advances in digital image manipulation detection techniques: A brief review. *Forensic Science International*, *312*, 110311. https://doi.org/10.1016/j.forsciint.2020.110311

Tyagi, S., & Yadav, D. (2023). A detailed analysis of image and video forgery detection techniques. *The Visual Computer*, *39*(3), 813–833. https://doi.org/10.1007/s00371-021-02347-4

Wang, J., Wu, Z., Chen, J., Han, X., Shrivastava, A., Lim, S. N., & Jiang, Y. G. (2022). ObjectFormer for image manipulation detection and localization. In *2022 IEEE/CVF conference on computer vision and pattern recognition* (pp. 2364–2373). IEEE. https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00240

Wang, S. Y., Wang, O., Owens, A., Zhang, R., & Efros, A. A. (2019). Detecting photoshopped faces by scripting Photoshop. In *2019 IEEE/CVF international conference on computer vision* (pp. 10072–10081). IEEE. https://doi.org/10.1109/ICCV.2019.01017

Ward, A. F., Duke, K., Gneezy, A., & Bos, M. W. (2017). Brain drain: The mere presence of one's own smartphone reduces available cognitive capacity. *Journal of the Association for Consumer Research*, *2*(2), 140–154. https://doi.org/10.1086/691462

Weikmann, T., & Lecheler, S. (2023). Visual disinformation in a digital age: A literature synthesis and research agenda. *New Media & Society*, *25*(12), 3696–3713. https://doi.org/10.1177/14614448221141648

Yang, Y., Davis, T., & Hindman, M. (2023). Visual misinformation on Facebook. *Journal of Communication*, *73*(4), 316–328. https://doi.org/10.1093/joc/jqac051

Yang, C., Li, H., Lin, F., Jiang, B., & Zhao, H. (2020, July). Constrained R-CNN: A general image manipulation detection model. In *2020 IEEE International conference on multimedia and expo (ICME)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICME46284.2020.9102825

Zanardelli, M., Guerrini, F., Leonardi, R., & Adami, N. (2023). Image forgery detection: A survey of recent deep-learning approaches. *Multimedia Tools and Applications*, *82*(12), 17521–17566. https://doi.org/10.1007/s11042-022-13797-w

Zhang, Y., Goh, J., Win, L. L., & Thing, V. (2016). Image region forgery detection: A deep learning approach. In A. Mathur, & A. Roychoudhury (Eds.), *Proceedings of the Singapore cyber-security conference (SG–CRC) 2016* (pp. 1–11). IOS Press. https://doi.org/10.3233/978-1-61499-617-0-1

Zhang, D., Chen, X., Li, F., Sangaiah, A. K., & Ding, X. (2020). Seam-Carved Image Tampering Detection Based on the Cooccurrence of Adjacent LBPs. *Security and Communication Networks*, 2020(1), 8830310. https://doi.org/10.1155/2020/8830310

Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2017). Two-stream neural networks for tampered face detection. *In 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 1831–1839). IEEE. https://doi.org/10.1109/CVPRW.2017.229

Acknowledgments
We thank the reviewers for their feedback.

Funding
This work was funded cooperative agreement from USAID, award 7200AA18CA00059.

Competing interests
The authors declare no competing interests.

Ethics
This study was conducted in accordance with ethical guidelines and principles governing research involving human participants. We obtained ethics review approval from Notre Dame Institutional Review Board (IRB) to ensure that our research adhered to ethical standards (Protocol #24-03-8475).

Prior to participation, all individuals were required to read and sign an informed consent form, which provided detailed information about the study's purpose, procedures, potential risks, and benefits. Participants were assured that their involvement was voluntary and that they could withdraw from the study at any time without penalty.

To protect participant privacy and confidentiality, no personally identifying information (PII) was collected during the study. Data were anonymized and stored securely to prevent unauthorized access.

Data availability
Pre-registration available at https://osf.io/pvq4d/?view_only=a111c2cfa40f4e05a615bf47040a3c64. All materials needed to replicate this study are available via the Harvard Dataverse: https://doi.org/10.7910/DVN/CKMBCG

Appendix: Statistical results

Table A1. *Mean of correct responses by each condition group.*

Group	*M* (*SD*)	*n*
Control	18.02 (3.34)	93
Feedback	22.09 (2.55)	94
Feedback+Education	22.26 (2.30)	81

Table A2*. ANOVA results for correct responses by condition.*

Source	*SS*	*df*	*MS*	*F*	*Prof > F*
Between Groups	1047.01	2	523.50	67.29	0.001
Within Groups	2061.65	265	7.78		
Total	3108.66	267	11.64		

Power analysis

An *a priori* power analysis was conducted via GPower software (Faul et al., 2009) for our OLS regression comparing mean differences in correct responses as the dependent variable with our treatment vs control variable as the independent variable. We had eight predictor variables in total (seven control variables plus our independent variable), and our power analysis was set at an alpha level of .05. Our power analysis found that to achieve .95 power with a medium (.15) effect size (see Cohen, 1988), we would need at least 160 participants. Our sample size exceeded 160 in our main analyses, which provided a satisfactory number of participants for our study.

Table A3. *OLS regression tables for treatment 1 vs. control and treatment 2 vs. control.*

	Feedback vs. Control	**Feedback+Education vs. Control**
Male	-0.883*	-0.869
	(0.490)	(0.530)
Age	-0.0715	-0.0927
	(0.271)	(0.278)
Education	0.317	0.182
	(0.273)	(0.270)
White	-0.0195	0.105
	(0.527)	(0.533)
Social media	-0.219	-0.520*
	(0.300)	(0.299)
Familiarity	0.915**	0.904**
	(0.409)	(0.432)
Conservative	-0.0640	-0.0342
	(0.218)	(0.232)
Feedback	4.106***	
	(0.478)	
Feedback + Education		4.112***
		(0.496)
Constant	16.95***	18.63***
	(2.138)	(2.060)
Observations	164	146
R^2	0.347	0.384

*Note: Standard errors in parentheses; *** p < .01, ** p < .05, * p < .10*

Table A4. *Exploratory analysis of first four images.*

	Accuracy of first four images
Male	0.203
	(0.122)
Age	0.026
	(0.066)
Education	0.033
	(0.064)
White	0.215
	(0.128)
Social media	-0.134
	(0.075)
Familiarity	0.119
	(0.103)
Conservative	-0.023
	(0.053)
Treatment conditions combined vs. Control	0.530***
	(0.122)
Constant	1.887***
	(0.519)
Observations	228
R^2	0.123

*Note: Standard errors in parentheses; *** p < .01, ** p < .05, * p < .1*

All Mean Differences Between Image Types Across Each Condition Group

Table A5. *Control descriptive statistics by group.*

Group	*M*	*SD*	*n*
Erasing	0.401	0.491	374
Copy-paste	0.567	0.496	372
Splicing	0.621	0.486	372
Photoshop touchup	0.650	0.478	371

Table A6. *Analysis of variance.*

Source	*SS*	*df*	*MS*	*F*	prob > *F*
Between groups	13.826	3.0	4.609	19.380	0.0
Within groups	353.163	1485.0	0.238		
Total	366.990	1488.0	0.247		

Table A7. *Control post hoc Bonferroni test results (control).*

				95% CI		
Group 1	Group 2	*MD*	*p*-adj	Lower bound	Upper bound	Reject
copy-paste	**erasing**	-0.166	.000	-0.258	-0.074	True
copy-paste	**Photoshop touchup**	0.082	.098	-0.010	0.174	False
copy-paste	**splicing**	0.054	.436	-0.038	0.146	False
erasing	**Photoshop touchup**	0.249	.000	0.157	0.340	True
erasing	**splicing**	0.220	.000	0.128	0.312	True
Photoshop touchup	**splicing**	-0.029	.854	-0.121	0.063	False

Table A8. *Treatment 1 (feedback) descriptive statistics by group.*

Group	*M*	*SD*	*n*
Erasing	0.617	0.487	376
Copy-paste	0.686	0.465	376
Splicing	0.750	0.434	376
Photoshop touchup	0.670	0.471	376

Table A9. *Post hoc Bonferroni test results (treatment 1).*

				95% CI		
Group 1	**Group 2**	***MD***	***p*-adj**	**lower bound**	**upper bound**	**Reject**
copy-paste	**erasing**	-0.069	.173	-0.156	0.018	False
copy-paste	**Photoshop touchup**	-0.016	.965	-0.103	0.071	False
copy-paste	**splicing**	0.064	.235	-0.023	0.151	False
erasing	**Photoshop touchup**	0.053	.396	-0.034	0.140	False
erasing	**splicing**	0.133	.001	0.046	0.220	True
Photoshop touchup	**splicing**	0.080	.086	-0.007	0.167	False

Table A10. *Treatment 2 (feedback + education) descriptive statistics by group.*

Group	***M***	***SD***	***n***
Erasing	0.574	0.495	324
Copy-paste	0.701	0.459	324
Splicing	0.781	0.414	324
Photoshop touchup	0.744	0.437	324

Table A11. *Post hoc Bonferroni test results (treatment 2).*

				95% CI		
Group 1	**Group 2**	***MD***	***p*-adj**	**lower bound**	**upper bound**	**Reject**
copy-paste	**erasing**	-0.127	.002	-0.218	-0.035	True
copy-paste	**Photoshop touchup**	0.043	.617	-0.048	0.135	False
copy-paste	**splicing**	0.080	.109	-0.011	0.172	False
erasing	**Photoshop touchup**	0.170	.000	0.078	0.261	True
erasing	**splicing**	0.207	.000	0.115	0.298	True
Photoshop touchup	**splicing**	0.037	.037	-0.054	0.129	False

Table A12. *Correlation between time spent and image classification accuracy for all images.*

					95% CI	
		SE	***z***	***p* > |*z*|**	**[0.025**	**0.975]**
Const	.945	0.039	24.357	.000	0.869	1.021
Time Spent	-.034	0.003	-10.589	.000	-0.040	-0.027

Note: r_{pb}*: -0.12, p-value < .001*

Table A13. *Demographic data.*

Demographic Category	*n*
Gender	
Women	112
Men	95
Preferring not to say	33
Other genders	28
Age	
18-24	36
25-34	103
35-50	80
50+	44
Education Level	
High school	32
Some college	62
College degree	118
Graduate degree	53
Social Media Use	
Daily	208
Weekly	42
Monthly	3
Less than monthly	7
Never	6
Familiarity with Image Manipulation	
Not familiar	18
Somewhat familiar	165
Very familiar	79
Experts	4
Political Affiliation	
Liberal	76
Somewhat liberal	74
Moderate	59
Somewhat conservative	27
Conservative	15
Apolitical	16
Racial Demographics	
White	178
Black	32
Asian	22
Latino/a/x	19
Other	15