DOI:10.1145/3445972

Technologies for manipulating and faking online media may outpace people's ability to tell the difference.

BY MATTHEW GROH, ZIV EPSTEIN, NICK OBRADOVICH, MANUEL CEBRIAN, AND IYAD RAHWAN

Human Detection of Machine-Manipulated Media

THE RECENT EMERGENCE of artificial intelligence (AI)-powered media manipulations has widespread societal implications for journalism and democracy,[7] national security,[1] and art.[8,14] AI models have the potential to scale misinformation to unprecedented levels by creating various forms of synthetic media.[21] For example, AI systems can synthesize realistic video portraits of an individual with full control of facial expressions, including eye and lip movement;[11,18,34–36] clone a speaker's voice with a few training samples and generate new natural-sounding audio of something the speaker never said;[2] synthesize visually indicated sound effects;[28] generate high-quality, relevant text based on an initial prompt;[31] produce photorealistic images of a variety of objects from text inputs;[5,17,27] and generate photorealistic videos of people expressing emotions from only a single image.[3,40] The technologies for producing machine-generated, fake media online may outpace the ability to manually detect and respond to such media.

We developed a neural network architecture that combines instance segmentation with image inpainting to automatically remove people and other objects from images.[13,39] Figure 1 presents four examples of participant-submitted images and their transformations. The AI, which we call a "target object removal architecture," detects an object, removes it, and replaces its pixels with pixels that approximate what the background should look like without the object. This architecture operationalizes one of the oldest forms of media manipulation, known in Latin as *damnatio memoriae*, which means erasing someone from official accounts.

The earliest known instances of *damnatio memoriae* were discovered in ancient Egyptian artifacts, and similar patterns of removal have appeared since.[10,37] Historically, visual and audio manipulations required both skilled experts and a significant investment of time and resources. Our architecture can produce photo-

» key insights

- **The speed at which misinformation can be produced is faster than it has ever been. By combining instance segmentation with image inpainting, we present an AI model that can automatically and plausibly disappear objects such as people, cars, and dogs from images.**
- **Exposure to manipulated content can prepare people to detect future manipulations. After seeing examples of manipulated images produced by the target object removal architecture, people learn to more accurately discern between manipulated and original images. Participant performance improves more after being exposed to subtle manipulations than blatant ones.**

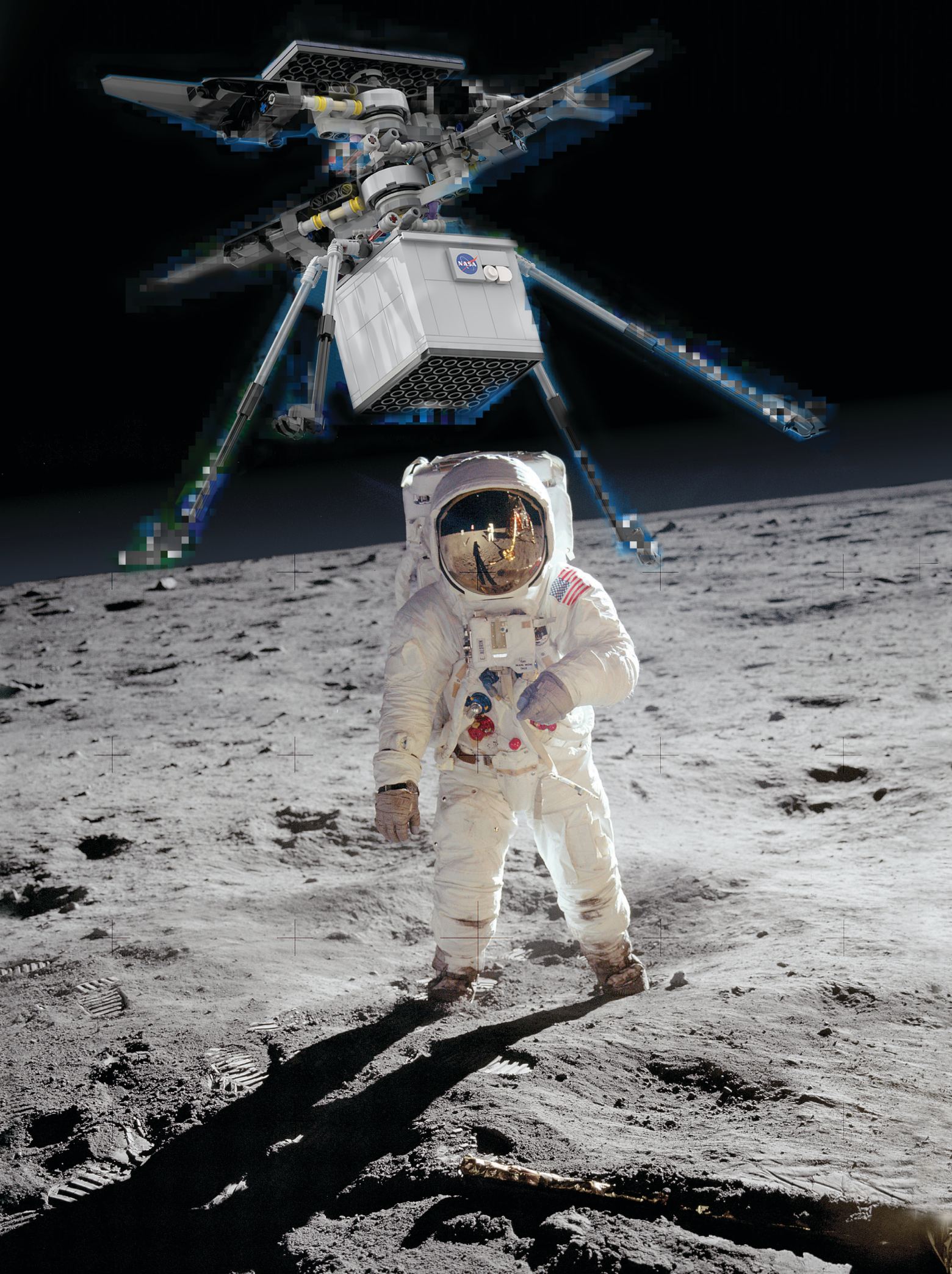


Figure 1. Examples of original images on the top row and manipulated images on the bottom row.

realistic manipulations nearly instantaneously, which magnifies the potential scale of misinformation. This new capacity for scalable manipulation raises the question of how prepared people are to detect manipulated media.

To publicly expose the realism of AI-media manipulations, we hosted a website called Deep Angel, where anyone in the world could examine our neural-network architecture and its resulting manipulations. Between August 2018 and May 2019, 110,000 people visited the website. We integrated a randomized experiment based on a two-alternative, forced-choice design within the Deep Angel website to examine how repeated exposure to machine-manipulated images affects an individual's ability to accurately identify manipulated imagery.

Two-Alternative, Forced-Choice Randomized Experiment

On the Deep Angel website's "Detect Fakes" page, participants are presented with two images consistent with standard two-alternative, forced-choice methodology and are asked a single question: "Which image has something removed by Deep Angel?" The pair of images contains one image manipulated by AI and one unaltered image. After the participant selects an image, the website reveals the manipulation and asks the participant to try again. The MIT Committee on the Use of Humans as Experimental Subjects (COUHES) approved IRB 1807431100 for this study on July 26, 2018.

The manipulated images are drawn from a population of 440 images submitted by participants to be shared publicly. The population of unaltered images contains 5,008 images from the MS-COCO dataset.[23] Images are randomly selected with replacements from each population of images. By randomizing the order of images that participants see, this experiment can causally evaluate the effect of image order on participants' ability to recognize fake media. We test the causal effects with the following linear probability models:

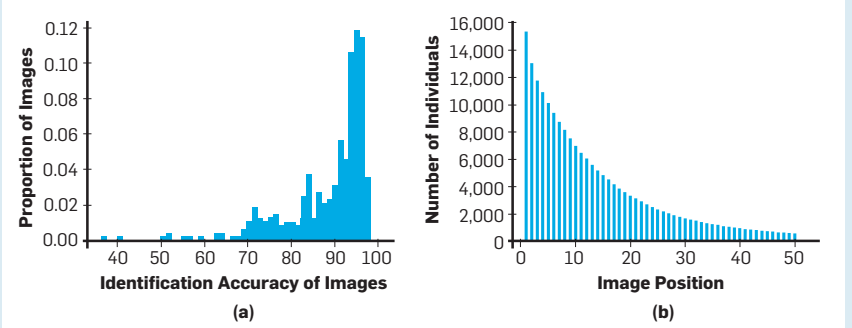
$$y_{i,j} = \boldsymbol{X}\alpha + \beta\log(T_{i_{n,}j}) + \mu_i + \nu_j + \epsilon_{i,j,n} \quad (1)$$

and

$$y_{i,j} = \boldsymbol{X}\alpha + \beta_1 T_{i_{1,}j} + \beta_2 T_{i_{2,}j} + \ldots + \beta_{10} T_{i_{10,}j} + \mu_i + \nu_j + \epsilon_{i,j,n} \quad (2)$$

where $y_{i,j}$ is the binary accuracy (correct or incorrect guess) of participant j on manipulated image i. X is a matrix of covariates indexed by i and j, T_{i_n} represents the order n in which manipulated image i appears to participant j, μ_i represents the manipulated image-fixed effects, ν_j represents the participant-fixed effects, and $\epsilon_{i,j}$ represents the error term. The first model fits a logarithmic transformation of T_{i_n} to $y_{i,j}$. The second model estimates treatment effects separately for each image position. Both models use Huber-White (robust) standard errors, and errors are clustered at the image level.

Figure 2. (a) Histogram of mean identification accuracies by participants per image (b) Bar chart plotting number of individuals over image position.



Results

Participation and average accuracy. From August 2018 to May 2019, 242,216 guesses were submitted from 16,542 unique IP addresses with a mean identification accuracy of 86%. The website did not require participant sign-in, so we study participant behavior under the assumption that each IP address represents a unique individual. The majority of participants participated in the two-alternative, forced-choice experiment multiple times, and 7,576 participants submitted at least 10 guesses.

Each image appears as the first image an average of 35 times and the

tenth image an average of 15 times. The majority of manipulated images were identified correctly more than 90% of the time. In the sample of participants who saw at least 10 images, the mean percentage correct classification is 78% on the first image seen and 88% on the tenth image seen. Figure 2a shows the distribution of identification accuracy across images, while Figure 2b shows the distribution of how many images each participant saw. The interquartile range of the number of guesses per participant is from three to 18 with a median of eight.

Figure 3a plots participant accuracy on the *y*-axis and image order on the *x*-axis, revealing a logarithmic relationship between accuracy and exposure to manipulated images. In this plot showing scores for all participants, accuracy increases rapidly over the first 10 images and plateaus around 88%.

Learning rate. With 242,216 observations, we run an ordinary least-squares regression with participant- and image-fixed effects on the likelihood of correctly guessing the manipulated image. The results of these regressions are presented in Tables 1 and 2 in the online appendix (https://dl.acm.org/doi/10.1145/3445972). Each column in Tables 1 and 2 adds an incremental filter to offer a series of robustness checks. The first column shows all observations. The second column drops all participants who submitted fewer than 10 guesses and removes all control images where nothing was removed. The third column drops all observations where a participant has already seen an image. The fourth column drops all images qualitatively judged as below very high quality.

Across all four robustness checks with and without fixed-effects, our models show a positive and statistically significant relationship between T_n and $\hat{y}_{i,j}$. In the linear-log model, a one-unit increase in $\log(T_{in})$ is associated with a 3% increase in $\hat{y}_{i,j}$. This effect is significant at the p<.01 level. In the model that estimates Equation 2, we find a 1% average marginal treatment effect size of image position on $\hat{y}_{i,j}$. This effect is also significant at the p<.01 level. In other words, participants improve their ability to guess by 1% for each of the first 10 guesses. Figure 3b shows these results graphically.

Within the context of object removal manipulations, exposure to media manipulation and feedback on what has been manipulated improves a participant's ability to recognize faked media. After getting feedback on 10 pairs of images for an average of 1 min., 14 sec., a participant's ability to detect manipulations improves by 10%. With clear evidence that human detection of machine-manipulated media can improve, the next question is: what is the mechanism that drives participant learning rates? How do feedback, image characteristics, and participant qualities affect learning rates?

Potential Explanatory Mechanisms

We can explore what drives the learning rate by examining heterogeneous effects of image characteristics and participant qualities. Figure 4 presents 10 plots of heterogeneous learning rates based on image-fixed effects regressions with errors clustered at the participant level.

We evaluate the quality of a manipulation across five measures: (a) a subjective quality rating, (b) 1st and 4th quartile image entropy, (c) 1st and 4th quartile proportion of area of the manipulated image, (d) 1st and 4th quartile mean identification accuracy per image, and (e) number of objects disappeared. The subjective quality rating is based on ratings provided by an outside party and is a binary rating based on whether obvious artifacts were created by the image manipulation.

Image entropy is measured based on delentropy, an extension of Shannon entropy for images.[20] To help understand delentropy, Figure 4 presents three pairs of images subjectively rated as high quality. Their corresponding entropy scores are included, along with the proportion of the image transformed, mean accuracy of participants' first guesses, and mean accuracy of subsequent participant guesses to exemplify what study participants learned.

For images subjectively marked as high quality, participants correctly discern 75% for the first image and 83% for the tenth image seen. In contrast, participant accuracy on the low-quality images is higher, at 82% and 94% for the first and tenth image seen, respectively. Table 3 (see online appendix) shows that the difference in means across the subjective quality measure is statistically significant at the 99% confidence level (p <.01), but we do not find a statistically significant difference in learning rates.

As seen in Figure 4a, there is evidence that participants learn to identify low-quality images faster than high-quality images if only looking at the first five images seen. When examining the first 10 images seen, we do not find a statistically significant difference in the interaction between subjective quality and the logarithm of the image position. These results indicate that the

Figure 3. Participants' overall and marginal accuracy by image order.

Error bars show a 95% confidence interval for each image position:
(a) overall accuracy for all participants with no fixed effects
(b) marginal accuracy (relative to the first image position) for all participants who saw at least 10 images controlling for participant- and image-fixed effects and clustering errors at the image level.
In (b), the 11th position includes all image positions beyond the 10th.

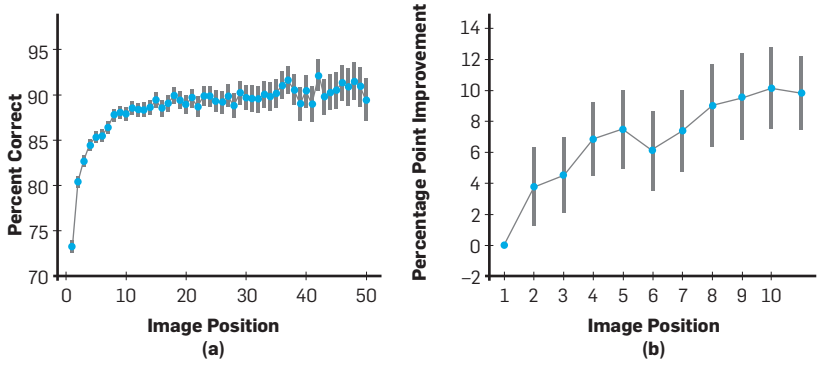
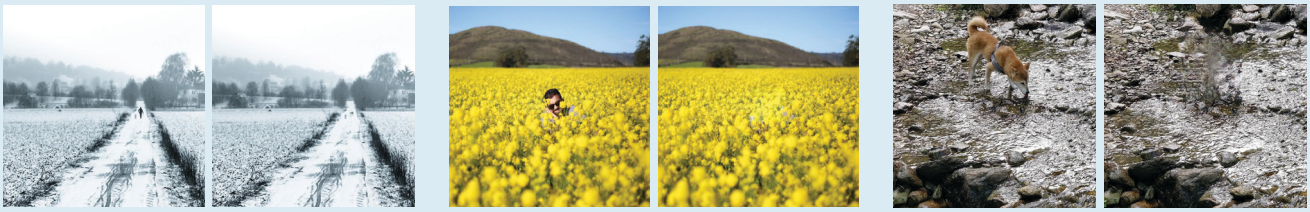
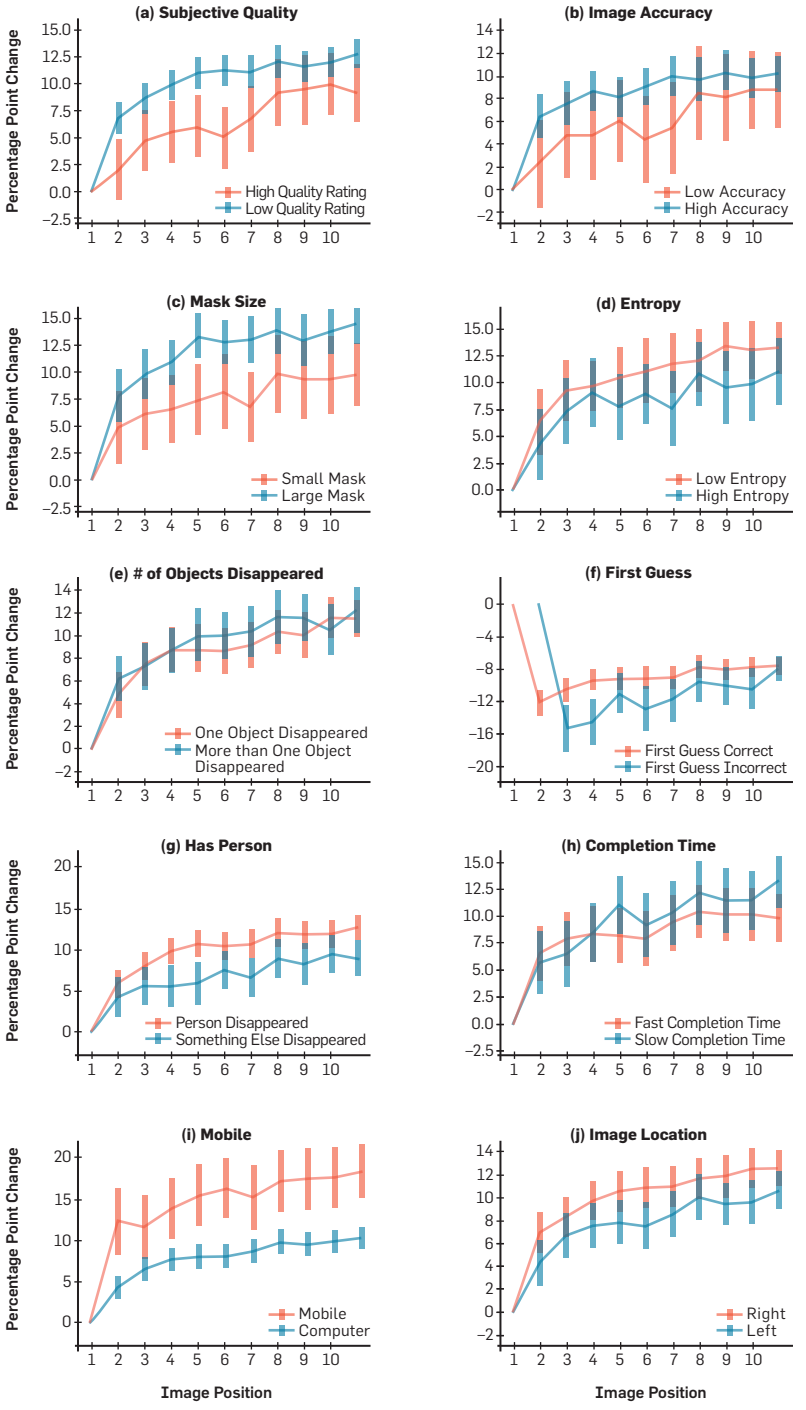


Figure 4. Ten plots of heterogeneous learning rates based on image-fixed effects regressions.



(i) From left to right, images (cropped into squares for display purposes) are increasing in entropy (3.5, 6.0, and 8.5), varying in percent of the image transformed (0.7%, 2.4%, and 1.9%), similar in accuracy on the first guess (70%, 71%, 70%), and varying in accuracy beyond the first guess (88%, 82%, 92%).

(ii) Ten plots display heterogeneous effects of image and participant characteristics on learning while controlling for participant- and image-fixed effects (a) whether the subjective image quality was judged as high by a third party, (b) whether the original image was in the 1st to 25th percentile of accuracy or 75th to 99th, (c) whether the original image was in the 1st to 25th percentile of image mask proportion or 75th to 99th, (d) whether the original image was in the 1st to 25th percentile of entropy or 75th to 99th, (e) whether there were one or multiple objects disappeared (f) whether the participant's first answer was correct (the omitted position for each learning curve represents perfect accuracy), (g) whether the image contained a person, (h) whether the original image was in the 1st to 25th percentile of time to evaluate 10 images or 75th to 99th, (i) whether the participant viewed the images on a mobile device or computer, and (j) whether the image was placed on the left or right side of the screen. The error bars represent the 95% confidence interval for each image position and errors are clustered at the image level.



main effect is not simply driven by participants becoming proficient at guessing low-quality images in our data.

The other proxies for image quality provide insight into how subtleties play a role in discerning image manipulations. Participants learn to identify low-entropy images faster than high-entropy images, and they recognize images with a large masked area faster than images with a small masked area. Table 3 shows that this difference in learning rates is statistically significant at the 95% ($p < .05$) and 90% ($p < .10$) levels, respectively.

Smaller masked areas and lower entropy is associated with less stark and more subtle changes between original and manipulated images. This relationship may indicate that participants learn more from subtle changes than more obvious manipulations. It may even mean people are learning to detect which kinds of images are hard to discern and, therefore, potentially likely to contain a manipulation when no obvious manipulation is apparent. It is important to note that neither the split between the 1st and 4th quartiles of mean accuracy per image nor the split between one object and many disappeared objects has a statistically significant effect on the learning rates. This means we find no association between overall manipulation discernment difficulty and learning rates.

A participant's initial performance is indicative of his or her future performance. In Figure 4, we compare subsequent learning rates of participants who correctly identified a manipulation on their first attempt to participants who failed on their first attempt and succeeded on their second. In this comparison, the omitted position for each learning curve represents perfect accuracy, which makes the marginal effects of subsequent image positions negative relative to these omitted image positions. On the first three of four image positions in this comparison, which correspond to the third through sixth image positions, we find that initially successful participants learn faster than participants who were initially unsuccessful. This heterogeneous effect does not persist in the seventh position or beyond. Overall, this heterogeneous effect is statistically significant at the 99% level ($p < .01$), suggesting that people who are better at discerning manipulations are also faster at learning to discern manipulations.

This new capacity for scalable manipulation raises the question of how prepared people are to detect manipulated media.

We find participants learn to discern manipulations involving disappeared people faster than images with any other object removed. This difference is statistically significant at the 95% confidence interval ($p < .05$) in the log-linear regression as shown in Table 3. Figure 4 also shows this difference as statistically significant in two of the 10 image positions, suggesting that participants may be learning to detect the kinds of images that are conducive to plausible object removals.

There is a clear difference in the learning rate of participants based on whether they participated with mobile phones or computers. Participants on mobile phones learn at a consistently faster rate than participants on computers, and this difference is statistically significant as shown in Table 3 and displayed across nine of 10 image positions in Figure 4. It is possible that the seamlessness of the zoom feature on a phone relative to a computer enables mobile participants to inspect each image more closely. We do not find evidence that image placement on the website correlates with overall accuracy.

No strong evidence suggests that the speed with which a participant rated 11 images is related to the learning rate, but we do find evidence of an interaction between answering speed upon wrong guesses of high-quality images. In Table 4 (see online appendix), we present a regression of current and lagged features on participant accuracy. It is important to note that we find high-quality images reduce participant accuracy by 4%, which is significant at the 99% confidence interval ($p < 0.1$), but we do not find a relationship between whether the previous image was high quality and participant accuracy on the current image. However, the interaction of seconds, guessing the previous answer incorrectly, and the previous image being high quality, is associated with a 0.3% increase in participant accuracy for every marginal second ($p < .05$). This correlational evidence suggests that when participants slow down after guessing incorrectly on high-quality, harder-to-guess images, they perform better.

Discussion

While AI models can improve clinical diagnoses[9,19,30] and enable autonomous driving,[6] they also have the potential to scale censorship,[32] amplify polarization,[4] and spread fake news and manipulated media.[38] We present results from a large-scale, randomized experiment showing that the combination of exposure to manipulated media and feedback on which media has been manipulated improves an individual's ability to detect media manipulations.

Direct interaction with cutting-edge technologies for content creation might enable more discerning media consumption across society. In practice, the news media has exposed high-profile, AI-manipulated media, including fake videos of the Speaker of the House of Representatives Nancy Pelosi and Facebook CEO Mark Zuckerberg, which serves as feedback to everyone on what manipulations look like.[24,25] Our results build on recent research showing that people can detect low-quality news,[29] human intuition can be a reliable source of information about adversarial perturbations to images,[42] and familiarizing people with how fake news is produced may confer them with cognitive immunity when they are later exposed to misinformation.[33] Our results also offer suggestive evidence for what drives learning to detect fake content. In this experiment, presenting participants with low-entropy images with minor manipulations on mobile devices increased learning rates at statistically significant levels. Participants appear to learn best from the most subtle manipulations.

Our results focus on a bespoke, custom-designed, neural-network architecture in a controlled, two-alternative, forced-choice experimental setting. The external validity of our findings should be further explored in different domains, using different generative models, and in settings where people are not instructed explicitly to look out for fakes, but rather encounter them in a more naturalistic social media feed, and in the context of reduced attention span. Likewise, future research in human perception of manipulated media should explore to what degree an individual's ability to adaptively detect manipulated media comes from learning by doing, direct feedback, and awareness that anything is manipulated at all.

With clear evidence that human detection of machine-manipulated media can improve, what is the mechanism that drives participants' learning rates?

Our results suggest a need to re-examine the precautionary principle that is commonly applied to content-generation technologies. In 2018, Google published BigGAN, which can generate realistic-appearing objects in images, but while the company hosted the generator for anyone to explore, it explicitly withheld the discriminator for its model.[5] Similarly, OpenAI restricted access to its GPT-2 model, which can generate plausible long-form stories given an initial text prompt, by only providing a pared-down model of GPT-2 trained with fewer parameters.[31] If exposure to manipulated content can prepare people to detect future manipulations, then censoring dissemination of AI research on content generation may prove harmful to society by leaving it unprepared for a future of ubiquitous AI-mediated content.

Methods

We developed a *Target Object Removal* architecture, combining instance segmentation with image inpainting to remove objects in images and replace those objects with a plausible background. Technically, we combine a convolutional neural network (CNN) trained to detect objects with a generative adversarial network (GAN) trained to inpaint missing pixels in an image.[12,13,16,22] Specifically, we generate object masks with a CNN based on a RoIAlign bilinear interpolation on nearby points in the feature map.[13] We crop the object masks from the image and apply a generative inpainting architecture to fill in the object masks.[15,39] The generative inpainting architecture is based on dilated CNNs with an adversarial loss function, allowing the generative inpainting architecture to learn semantic information from large-scale datasets and generate missing content that makes contextual sense in the masked portion of the image.[39]

Target Object Removal Pipeline

Our end-to-end, targeted object removal pipeline consists of three interfacing neural networks:

▸ **Object Mask Generator (G):** This network creates a segmentation mask $X' = G(X, y)$ given an input image X and a target class y. In our experiments, we initialize **G** from a semantic

segmentation network trained on the 2014 MS-COCO dataset following the Mask-RCNN algorithm.[13] The network generates masks for all object classes present in an image, and we select only the correct masks based on input y. This network was trained on 60 object classes.

▸ **Generative Inpainter (I):** This network creates an inpainted version $Z = I(X', X)$ of the input image X and the object mask X'. **I** is initialized following the DeepFill algorithm trained on the MIT Places 2 dataset.[39,41]

▸ **Local Discriminator (D):** The final discriminator network takes in the inpainted image and determines its validity. Following the training of a GAN discriminator, **D** is trained simultaneously on **I**, where X are images from the MIT Places 2 dataset and X' are the same images with randomly assigned holes following.[39,41]

Live Deployment

The Deep Angel website enabled us to make the Target Object Removal architecture publicly available.[a] We hosted the architecture API with a single Nvidia Geforce GTX Titan X; anyone could upload an image to the site and select an object to be removed from the image.

Participants uploaded 18,152 unique images from mobile phones and computers; they also directed the crawling of 12,580 unique images from Instagram. We can surface the most plausible object removal manipulations by examining the images with the lowest guessing accuracy. The Target Object Removal architecture can produce plausible content, but the plausibility is largely image dependent and constrained to specific domains, where objects are a small portion of the image, and the background is natural and uncluttered by other objects.

Data availability: The data and replication code are available at: https://github.com/mattgroh/human-detection-machine-manipulated-media-data-code.

Acknowledgments. We thank Abhimanyu Dubey, Mohit Tiwari, and David McKenzie for their helpful comments and feedback. C

a We retained the Cyberlaw Clinic from Harvard Law School and Berkman Klein Center for Internet & Society for advice on copyright protection of manipulated images.

References

1. Allen, G. and Chan, T. Artificial intelligence and national security. *Belfer Center for Science and International Affairs*, Cambridge, MA (2017).
2. Arik, S.O., Chen, J., Peng, K., Ping, W., and Zhou, Y. Neural voice cloning with a few samples. *arXiv preprint arXiv:1802.06006* (2018).
3. Averbuch-Elor, H., Cohen-Or, D., Kopf, J., and Cohen, M.F. Bringing portraits to life. *ACM Transactions on Graphics (TOG) 36*, 6 (2017), 196.
4. Bakshy, E., Messing, S., and Adamic, L.A. Exposure to ideologically diverse news and opinion on Facebook. *Science 348*, 6239 (2015), 1130–1132.
5. Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* (2018).
6. Chen, C., Seff, A., Kornhauser, A., and Xiao, J. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proc. of the IEEE Intern. Conf. on Computer Vision* (2015), 2722–2730.
7. Chesney, R. and Citron, D.K. Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review 107*, 6 (Dec. 2019).
8. Epstein, Z., Boulais, O., Gordon, S., and Groh, M. Interpolating GANs to scaffold autotelic creativity. *arXiv preprint arXiv:2007.11119* (2020).
9. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature 542*, 7639 (2017), 115.
10. Freedberg, D. The power of images: Studies in the history and theory of response. University of Chicago Press (1989).
11. Garrido, P., Valgaerts, L., Sarmadi, H., Steiner, I., Varanasi, K., Perez, P., and Theobalt, C. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. *Computer Graphics Forum 34*. Wiley Online Library (2015), 193–204.
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. (2014), 2672–2680.
13. He, K., Gkioxari, G., Dollár, P., and Girshick, R.B. Mask R-CNN. *CoRR abs/1703.06870* (2017).
14. Hertzmann, A. Can computers create art? In *Arts 7, Multidisciplinary Digital Publishing Institute* (2018), 18.
15. Iizuka, S., Simo-Serra, E., and Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. on Graphics (Proc. of SIGGRAPH 2017) 36*, 4 (2017).
16. Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).
17. Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948* (2018).
18. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollhöfer, M., and Theobalt, C. Deep video portraits. *arXiv preprint arXiv:1805.11714* (2018).
19. Kooi, T., Litjens, G., Van Ginneken, B., Gubern-Mérida, A., Sánchez, C.I., Mann, R., den Heeten, A., and Karssemeijer, N. Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis 35* (2017), 303–312.
20. Larkin, K.G. Reflections on Shannon information: In search of a natural information-entropy for images. *arXiv preprint arXiv:1609.01117* (2016).
21. Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., et al. The science of fake news. *Science 359*, 6380 (2018), 1094–1096.
22. LeCun, Y., Bengio, Y., and Hinton, G.E. Deep learning. *Nature 521*, 7553 (2015), 436–444.
23. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C.L. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, Springer. (2014), 740–755.
24. Mervosh, S. Distorted videos of Nancy Pelosi spread on Facebook and Twitter, helped by Trump. (May 2019). https://www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html.
25. Metz, C. A fake Zuckerberg video challenges Facebook's rules. (June 2019). https: //www.nytimes.com/2019/06/11/technology/fake-zuckerberg-video-facebook.html.
26. Molodetskikh, I., Erofeev, M., and Vatolin, D. Perceptually motivated method for image inpainting comparison. *arXiv preprint arXiv:1907.06296* (2019).
27. Nguyen, A., Yosinski, J., Bengio, Y., Dosovitskiy, A., and Clune, J. Plug & play generative networks: Conditional iterative generation of images in latent space. *CoRR abs/1612.00005* (2016).
28. Owens, A., Isola, P., McDermott, J.H., Torralba, A., Adelson, E.H., and Freeman, W.T. Visually indicated sounds. *CoRR abs/1512.08512* (2015).
29. Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A.A., Eckles, D., and Rand, D.G. Understanding and reducing the spread of misinformation online. (2019).
30. Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., and Webster, D.R. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering 2*, 3 (2018), 158.
31. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. (2019).
32. Roberts, M.E. *Censored: Distraction and Diversion Inside China's Great Firewall*. Princeton University Press (2018).
33. Roozenbeek, J. and van der Linden, S. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications 5* (2019).
34. Saito, S., Wei, L., Hu, L., Nagano, K., and Li, H. Photorealistic facial texture inference using deep neural networks. *CoRR abs/1612.00523* (2016).
35. Suwajanakorn, S., Seitz, S.M., and Kemelmacher-Shlizerman, I. Synthesizing Obama: Learning lip sync from audio. *ACM Trans. on Graphics (TOG) 36*, 4 (2017), 95.
36. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. Face2face: Real-time face capture and reenactment of RGB videos. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. (2016), 2387–2395.
37. Varner, E.R. Mutilation and transformation: Damnatio memoriae and Roman imperial portraiture. *Monumenta Graeca et Romana 10*, Brill (2004).
38. Vosoughi, S., Roy, D., and Aral, S. The spread of true and false news online. *Science 359*, 6380 (2018), 1146–1151.
39. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T.S. Generative image inpainting with contextual attention. *arXiv preprint arXiv:1801.07892* (2018).
40. Zakharov, E., Shysheya, A., Burkov, E., and Lempitsky, V. Few-shot adversarial learning of realistic neural talking head models (2019).
41. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).
42. Zhou, Z., and Firestone, C. Humans can decipher adversarial images. *Nature Communications 10*, 1 (2019), 1334.

Matthew Groh is a Ph.D. candidate in the Massachusetts Institute of Technology (MIT) Media Lab, Cambridge, MA, USA.

Ziv Epstein is a Ph.D. candidate in the Massachusetts Institute of Technology (MIT) Media Lab, Cambridge, MA, USA.

Nick Obradovich is a senior research scientist and principal investigator in the Center for Humans & Machines at Max Planck Institute for Human Development, Berlin, Germany.

Manuel Cebrian is the Max Planck Research Group Leader of the Digital Mobilization Research Group in the Center for Humans & Machines at Max Planck Institute for Human Development, Berlin, Germany.

Iyad Rahwan is director in the Center for Humans & Machines at Max Planck Institute for Human Development, Berlin, Germany.

Author contributions. M.G. implemented the methods, M.G., Z.E., N.O. analyzed data and wrote the article. All authors conceived the original idea, designed the research, and provided critical feedback on the analysis and manuscript.

Watch the authors discuss this work in the exclusive *Communications* video. https://cacm.acm.org/videos/machine-manipulated-media