Harvard Kennedy School Misinformation Review[1]
September 2022, Volume 3, Issue 5

Reprints and permissions: misinforeview@hks.harvard.edu
DOI: https://doi.org/10.37016/mr-2020-104
Website: misinforeview.hks.harvard.edu

Research Note

Research note: This salesperson does not exist: How tactics from political influence operations on social media are deployed for commercial lead generation

Researchers of foreign and domestic influence operations document tactics that frequently recur in covert propaganda campaigns on social media, including backstopping fake personas with plausible biographies or histories, using GAN-generated images as profile photos, and outsourcing account management to paid organizations. These tactics, however, can be applied outside of the political realm. In this paper, we describe how these three tactics are leveraged to serve a commercial purpose: lead generation for sales. We conducted the first study of fake accounts with GAN-generated images on LinkedIn that engage in lead generation (n = 1,003) and offer recommendations for grappling with fake persona creation, generally, GAN-generated imagery, specifically, and outsourcing.

Authors: Josh A. Goldstein (1), Renée DiResta (1)
Affiliations: (1) Stanford Internet Observatory, Stanford University, USA
How to cite: Goldstein, J. A., & DiResta, R. (2022). Research note: This salesperson does not exist: How tactics from political influence operations on social media are deployed for commercial lead generation. *Harvard Kennedy School (HKS) Misinformation Review*, *3*(5).
Received: April 28th, 2022. Accepted: August 12th, 2022. Published: September 15th, 2022.

Research questions

- How are tactics observed in political influence operations—such as creating fake personas with plausible histories, using profile photos generated by Generative Adversarial Networks (GANs) to backstop fake accounts, and outsourcing disinformation campaigns to third-party firms—used for commercial purposes?
- Once a researcher has found a LinkedIn profile with a GAN-generated image, what methods can they use to find other accounts with GAN-generated images related to the account of interest?
- What steps can LinkedIn take to combat the deceptive use of fake accounts, AI-generated images, and outsourcing in the future?

Research note summary

- We conducted what we believe to be the first study of the use of fake accounts on LinkedIn with GAN-generated profile photos. Our research surfaced more than 1,000 such inauthentic accounts on LinkedIn, which violate LinkedIn's terms of service.

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

- We observed tactics frequently deployed in political disinformation campaigns on Facebook and Twitter being used on LinkedIn for business purposes: in addition to the AI-generated profile pictures, these accounts had fake histories that included university attendance and employment history, joined LinkedIn Groups, and liked posts created by real users. Some of the accounts were created by third-party firms that offer lead generation services and are analogous to marketing or PR firms creating fake accounts for propaganda campaigns.
- Our research pilots several methods for researchers investigating the use of manipulative networks on LinkedIn, including using LinkedIn's "People Also Viewed" feature, investigating companies' "People" lists, and finding third-party firms that claim on their websites to offer AI-enabled lead generation services.
- Our investigation was possible because GAN-generated images currently have characteristics that are often identifiable with the naked eye; as AI-generated images improve, we suspect the problem will grow and fake accounts with AI-generated images will become increasingly difficult to identify.

Implications

Research in the disinformation field has documented recurring tactics in political influence operations on social media (e.g., Bradshaw et al., 2020; Goldstein & Grossman, 2021; Martin et al., 2020). However, these deceptive online practices are also used in other areas, including for business purposes (Stanford Internet Observatory, 2020). In this study, we outline how tactics used for political disinformation campaigns—fake persona creation, AI-generated imagery, and outsourcing—have similar or analogous benefits in the economic sphere. Specifically, we highlight possible benefits of these three tactics for a particular business application: lead generation for sales.[2] We then offer a case study of 1,003 fake accounts with AI-generated profile pictures on LinkedIn, in what we believe to be the first academic study of GAN-generated images on the platform.[3]

This research makes three contributions to the disinformation literature. First, relatively little research documents deceptive tactics on LinkedIn. The platform is difficult to study because LinkedIn does not offer a convenient tool for researchers and there are restrictions on viewing accounts' connections.[4] Moreover, LinkedIn itself does not provide routine detailed takedown reports that could give specific insight into malicious campaigns. Although our case study reflects only a tiny fraction of actioned accounts, we provide evidence of several types of deception on the platform.[5] Second, the emerging disinformation field is primarily focused on political messaging and campaigns. However, fraudsters and economically motivated spammers have long piloted deceptive online practices. Our theory and case study bridge the political and economic domains, encouraging researchers to learn from other fields. Third, we document the deceptive use of AI-generated images at a critical time, since hallmark signs of automation can still be detected with the naked eye. Implementing mitigations will be important as more sophisticated capabilities become widely available.

[2] Lead generation is the process of identifying and forming relationships with individuals who may have interest in a product, in hopes of translating that interest into a sale.

[3] GANs, or generative adversarial networks, are an approach to generative modeling that have been used to create highly realistic images. See, for example, Karras et al. (2019).

[4] For example, LinkedIn does not offer researchers a public insight tool such as CrowdTangle (offered by Meta), nor API access (such as that offered by Twitter).

[5] LinkedIn does provide aggregate numbers of actioned accounts in semi-annual transparency reports. See: https://about.linkedin.com/transparency/community-report.

How deceptive tactics from political influence operations may apply to lead generation:

Fake persona creation

After Russia's social media influence operations targeting American politics from 2015-2018, a body of research outlined how Internet Research Agency propagandists created fake personas to infiltrate target communities (DiResta et al., 2019; Howard et al., 2018; Linvill & Warren, 2020a; Linvill & Warren, 2020b). Research has shown that people are more receptive to information from in-group members and from sources they deem credible and trustworthy (Pornpitakpan, 2004). The process of enhancing a persona with a plausible history or credentials (a practice known as "backstopping") can include creating content that is not directly tied to the mission at hand to increase the persuasive effect of a campaign (Golovchenko et al., 2020).

A similar logic for persona creation may apply to sales: if people are more likely to respond favorably to salespeople they deem as credible, trustworthy, or in-group members, then businesses could benefit from creating fake personas with plausible and relatable histories for outreach efforts. This can enhance outreach efficiency and effectiveness if social media platforms like LinkedIn limit the number of messages an account can send, charge to send unsolicited or out-of-user-network messages, or if companies do not have salespeople who are members of the communities they wish to solicit.

AI-generated imagery

In December 2019, researchers from the network firm Graphika and the Atlantic Council's Digital Forensic Research Lab published an investigation into a network of accounts on Facebook and Instagram titled "#OperationFFS: Fake Face Swarm." The authors noted the network was the first time they had seen AI-generated pictures "deployed at scale to generate a mass collection of fake profile pictures deployed in a social media campaign" (Nimmo et al., 2019, p. 15). Since then, researchers in the disinformation field have observed a rise in the number of takedowns of accounts by Facebook and Twitter that use AI-generated profile pictures (Goldstein & Grossman, 2021). Such operations include efforts to counter West Papua's independence movement, to spread vox populi commentary about American politics, and to pose as American journalists (Graphika, 2020; Stanford Internet Observatory, 2020; Strick, 2020).

GANs offer several possible advantages when creating fake accounts. When a profile uses a stock or stolen photo, it can often be identified as such via reverse image search; GAN-generated images are unique and often undiscoverable through this process. Moreover, GAN-generated images allow propagandists to develop personas that blend into the desired target community. Websites such as generated[dot]photos—a website that offers "unique, worry-free model photos" that are AI-generated—allow users to select age, sex, ethnicity, and emotion.[6]

Compounding this, recent research suggests that AI-generated images are found to be believable and trustworthy; humans struggle to distinguish between real faces and synthetic ones (Nightingale & Farid, 2021; Shen et al., 2021). In experimental settings, users perceived social media accounts with AI-generated images as trustworthy and reported that they would likely accept a LinkedIn connection from such accounts (Mink et al., 2022). AI-generated photos may help fake accounts used for lead generation give a trustworthy impression and foster connections with real users.

[6] These websites offer access to AI-generated images for those with no specialized background in machine learning. Researchers suspected that GANs would become more widespread in influence operations because of the "ease with which threat actors can now use publicly available services to generate fake profile pictures…" (Nimmo et al., 2020, p. 2).

Outsourcing

Political actors frequently outsource their disinformation campaigns to third-party marketing or PR firms (DiResta et al., 2022; Goldstein & Grossman, 2021). Comparing Russian operations run in-house by Russian military intelligence with operations outsourced to the Internet Research Agency, DiResta et al. (2022) offer four reasons why a state may choose to outsource to digital mercenaries: "they give the state access to top talent trained in the latest social media campaign techniques, they can save the state money, they may make campaigns more difficult to discover, and they afford plausible deniability if they are discovered" (p. 4). These rationales likewise apply to sales: companies may outsource lead generation to third parties with specialized expertise who can identify prospective customers more cost-effectively.[7] If the third party leverages fake accounts to further decrease costs, the hiring company has plausible deniability.

Table 1. *Summary of hypothesized benefits of three tactics observed in the political disinformation space that may apply to lead generation.*[8]

Tactic	**Possible benefits for propagandist or sales team**
Fake accounts with plausible histories	● Increase persuasiveness of message ● Manufacture consensus ● Blend into target community (shared background/interests)
GAN-generated profile pictures	● Decrease detection relative to stolen pictures ● Blend into target community (shared demographics)
Outsourcing	● Lend expertise ● Economically efficient ● Decrease detection ● Offer plausible deniability

In this paper, we describe 1,003 accounts that use GANs that listed at least 63 different current employers at the time they were removed.[9] The accounts we surfaced violate LinkedIn terms of service because they backstopped with fake histories—falsely claiming to have attended colleges and worked for other companies—and because they claim to represent people that, to the best of our knowledge, do not exist (LinkedIn User Agreement). Our research shows that the three tactics described above in the political context are already used for business on LinkedIn.

[7] Using third-party outsourced sales and marketing teams is a legitimate business practice. Our focus here is on firms that leverage deception specifically, and the possible benefits for propagandists or sales teams of outsourcing deceptive work as opposed to running it in-house.

[8] Our case study shows that these three tactics are used in economic sales for lead generation, but does not isolate and test whether each tactic is beneficial compared to a relevant baseline. Future research could consider, for example, whether survey respondents would be more likely to accept connection requests from accounts with plausible histories (compared to no histories) or whether firms benefit from plausible deniability when those they contract are caught running fake accounts.

[9] We stopped our research after documenting 1,003 accounts because we felt we had sufficient evidence to provide proof of fake accounts with GAN-generated images on LinkedIn. Our investigation was meant to be a case study and provide proof of issue, rather than an exhaustive study. Even after we reached this mark, we continued to see additional accounts with GAN-generated images but faced limitations in how often we could update our analysis.

Our investigation establishes proof-of-issue on LinkedIn and raises two concerns. First, fake account detection is currently aided by artifacts present in AI-generated photos such as irregular ears, pupils, teeth, glasses, and jewelry (Gershgorn, 2018). As technology improves, it will become more difficult, if not impossible, to detect AI-generated images with the naked eye. Users may waste time connecting with fake salespeople or recruiters, or be defrauded by deceptive operations with fake accounts. When people begin to question the humanness of accounts on LinkedIn, they may be less inclined to network through the platform and grow less trusting of the broader online information environment. Second, our investigation demonstrates that those behind backstopped fake accounts can incorporate unwitting companies and universities into a manipulative operation.

Possible recommendations for LinkedIn to mitigate these concerns include:

1. Create a GAN-specific policy to minimize rule-skirting. LinkedIn could offer a clear policy of whether (or under what conditions) the use of AI-generated profile pictures is permissible on its website. (This might be informed by a platform-directed or research-directed study of the conditions under which LinkedIn users find AI-generated imagery acceptable.) While current LinkedIn platform rules appear to prohibit their use since they constitute misrepresenting one's identity,[10] LinkedIn could address this directly and decrease the likelihood of third-party companies claiming they are unaware of the prohibition. This recommendation is unlikely to deter a determined user, but could complement a broader effort to educate LinkedIn users about platform abuses. For example, LinkedIn could pair updates to its current platform policies with an overhaul of its transparency reports that would help make the public more aware of recent deceptive campaigns.

2. Integrate GAN-detection models. LinkedIn could integrate AI models that attempt to detect whether an image or profile picture is computer-generated to flag accounts upon creation or profile picture posting.[11] To be effective in surfacing AI-generated imagery, these models will require updating. As GAN technology improves, detection models will have to improve as well to remain effective.

In addition to using detection models to surface accounts with fake profile photos, this effort would offer LinkedIn trust and safety employees an opportunity to study networks of accounts with AI-generated images for other technical indicators of inauthenticity, possibly updating other detection processes based on any findings. To avoid punishing false positives, flagged accounts could be subject to manual review by the trust and safety team.

3. Create notification and verification options for companies to better manage affiliated profiles. LinkedIn users can claim to have worked for a company without that company receiving a notification or confirming that the listed employment history is factual, opening the door to fraud.[12] We suspect this problem is much more widespread than accounts that use GAN-generated images and can undermine both brand reputation and user trust in the platform. Additional research could explore these questions directly.

We recommend LinkedIn explore two *optional* features to mitigate this deceptive behavior: a notification system and a verification feature. A notification system could entail LinkedIn notifying a designated HR account whenever a new LinkedIn account lists the company as their employer. This would help companies more quickly notice fake activity, but should be optional given that it could add a burden on the user in charge of the company's LinkedIn presence. An optional verification system could entail

[10] A full-on ban may be ostensibly appealing, but not necessarily the right course of action. For instance, AI artists may have legitimate reasons to include AI-generated content in a profile picture. LinkedIn developing an explicit policy would clarify how to treat such edge cases, as well as the conditions under which GANs would be permissible more generally.

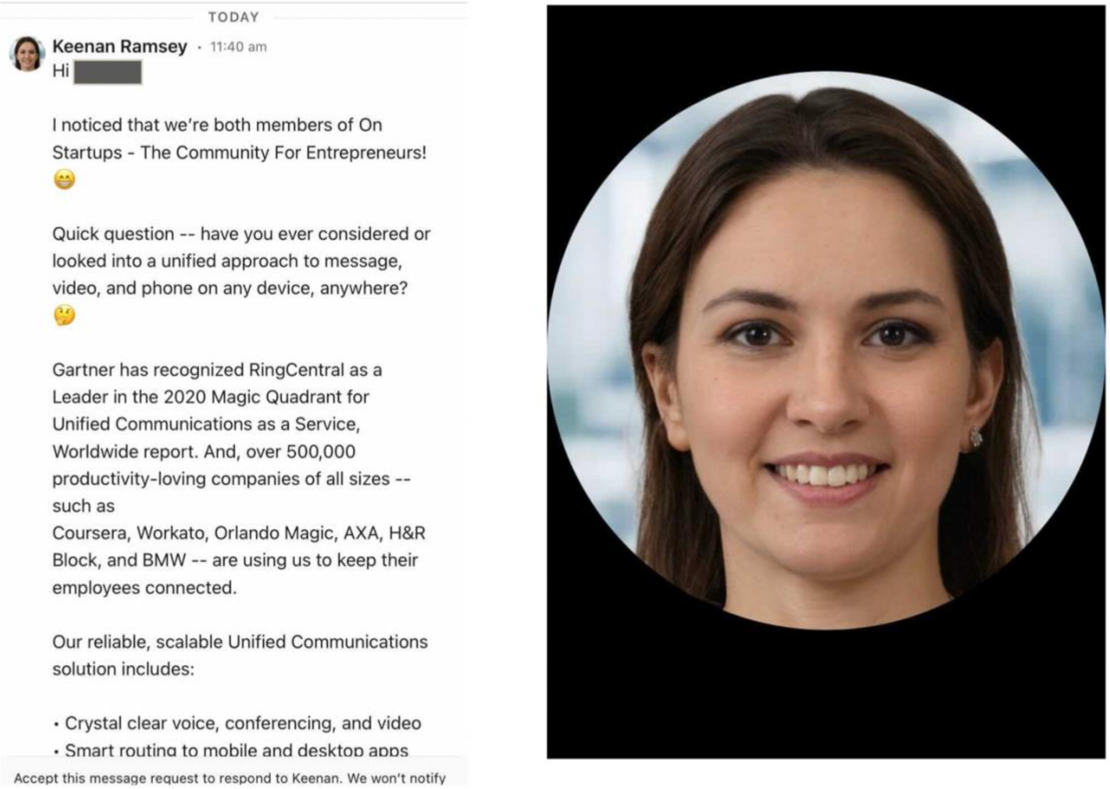
[11] For a recent review of the literature on GAN-generated faces detection, see Wang et al. (2022).

[12] In our investigation, this allowed fake accounts to create plausible histories by falsely backfilling past employers and education.

LinkedIn offering a symbol for work histories that have been verified by the company page.[13] This feature may increase trust among ordinary LinkedIn users who receive InMail messages or connection requests if they can have higher confidence that a user actually worked for a company listed in their employment history.

Evidence

On December 13, 2021, one of the authors received a LinkedIn message from an account with the name Keenan Ramsey that claimed to work for RingCentral (a cloud-based communications solution company). The account claimed to be a Sales Development Representative, a sales role that generates leads for more senior account executives. The LinkedIn message was a typical pitch: establishing a common connection by referencing a shared group membership and highlighting the benefits of RingCentral (Figure 1). However, its profile picture bore the hallmarks of a GAN-generated image: an earring on one ear but not the other, hair that fades into the background, and the blurred background typical of StyleGAN2 images. In fact, we found an identical face on a website, generated[dot]photos, that offers AI-generated images (Figure 2).



TODAY

Keenan Ramsey · 11:40 am

Hi [redacted]

I noticed that we're both members of On Startups - The Community For Entrepreneurs! 😁

Quick question -- have you ever considered or looked into a unified approach to message, video, and phone on any device, anywhere? 🤔

Gartner has recognized RingCentral as a Leader in the 2020 Magic Quadrant for Unified Communications as a Service, Worldwide report. And, over 500,000 productivity-loving companies of all sizes -- such as Coursera, Workato, Orlando Magic, AXA, H&R Block, and BMW -- are using us to keep their employees connected.

Our reliable, scalable Unified Communications solution includes:

- Crystal clear voice, conferencing, and video
- Smart routing to mobile and desktop apps

Accept this message request to respond to Keenan. We won't notify

Figure 1. Initial pitch by a LinkedIn account Keenan Ramsey, with signs that the image was generated by AI. *The image has an earring on one ear but not another, eyes that are perfectly centered, and strands of hair that blend into the background.*

[13] LinkedIn currently offers a feature where companies can register a domain, and employees must verify their work email address from that domain to gain access to content on the "My Company" page and to post jobs on behalf of the company. However, LinkedIn does not offer outward-facing symbols of verification that would be seen by other users.

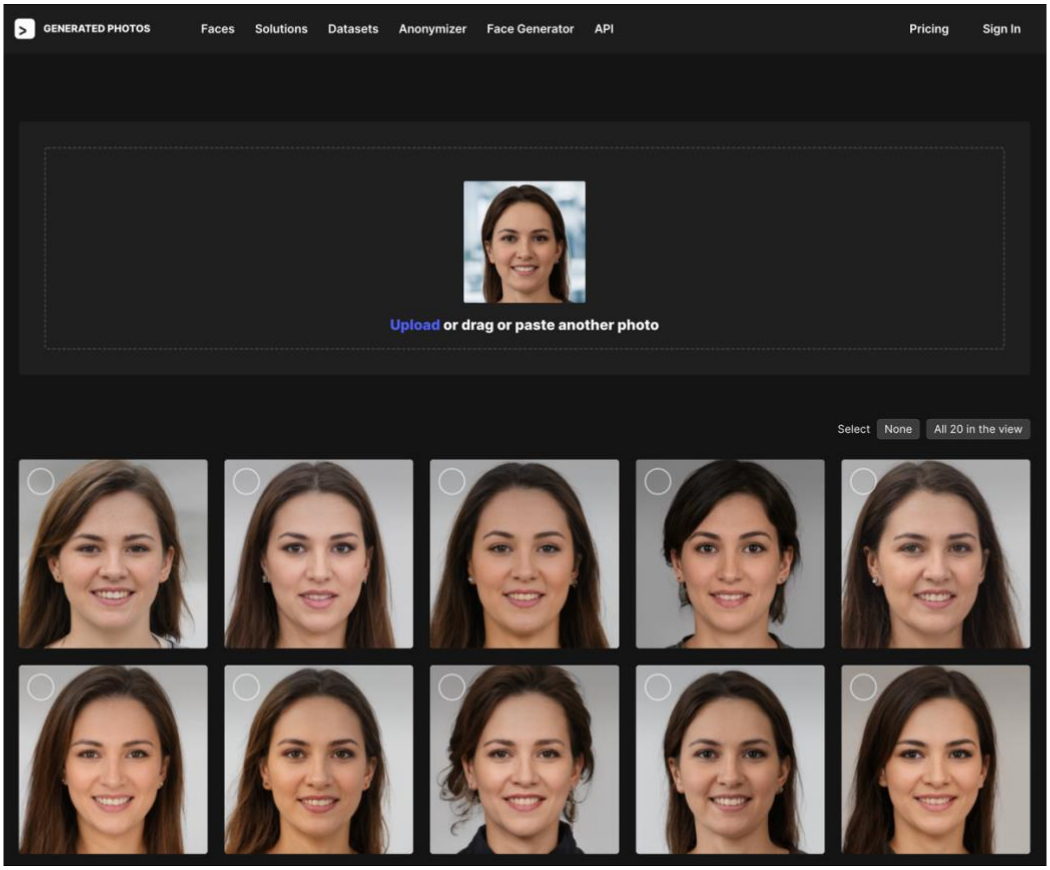


Figure 2. A screenshot from generated[dot]photos. *When we searched for Ramsey's photo, an identical face with a different background appears on generated[dot]photos (second row, second to the right), indicating the source of the AI-generated image.*

On RingCentral's "People" page on LinkedIn, we observed dozens of other accounts that claimed to be employees in sales roles (e.g., Sales Development Representative, Growth Specialist) and appeared to have GAN-generated profile pictures. The current state of GAN technology is such that unmodified images generated from models such as StyleGAN2 have eyes that consistently align. In Figure 3, we show the profile pictures of 15 accounts that purported to work for RingCentral, with consistent eye placement. Upon investigation, we noticed that the accounts that listed content in their "About" section all used the same stock description of RingCentral and listed universities attended and purported employment history. However, the accounts did not offer any other unique or personally identifiable information.

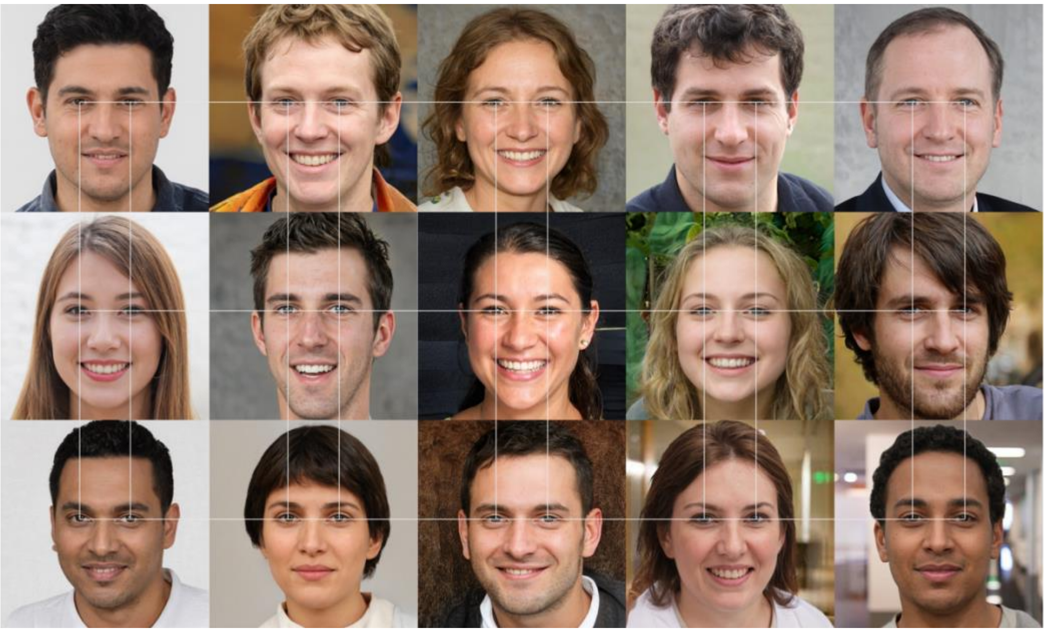


Figure 3. 15 images from accounts that purported to work for RingCentral. *These images have overlapping eye placement—a hallmark of GAN creation.*[14]

Days later, a real employee from RingCentral followed up on their "colleague" Keenan Ramsey's message, establishing that the original outreach was tied—knowingly or unknowingly—to RingCentral. We omit the name of this employee for privacy concerns and note that, according to their LinkedIn account, their employment at RingCentral halted within several months of the message. A journalist, Shannon Bond, searched for evidence of Ramsey's existence. She found that the employers Ramsey listed (e.g., RingCentral, Language I/O) had no record of her, and that her purported alma mater (New York University) said they have no record of anyone named Keenan Ramsey receiving an undergraduate degree (Bond, 2022).[15]

When viewing the RingCentral employee profiles with suspected GAN-generated images, we noticed in the "People Also Viewed" feature, where LinkedIn recommends other accounts, additional images that seemed GAN-generated. We used the "People Also Viewed" function and a company crawl (described in the Methods section below) to surface additional accounts with GAN-generated images, ultimately finding more than 1,000 profiles with GAN-generated images that purported to work for more than 63 companies at the time the accounts were removed.

Some of these images have clear signs of inauthenticity. In Figure 4, the image on the left has glasses that connect to one ear but not the other; the image on the right has a rendering of a second person that is disproportionate.

[14] RingCentral told Shannon Bond, a journalist at NPR, that they did not have a record of anyone by the names of these accounts working for the company. According to Sensity analysis (described below), each of these images has greater than a .99999 probability of being GAN-generated. These 15 accounts, along with dozens of others that claimed to work for RingCentral, were removed from the platform after Bond shared our findings with RingCentral and we shared our findings with LinkedIn.

[15] For more information on journalists who work on understanding misinformation online, see McClure Haughey et al. (2020).

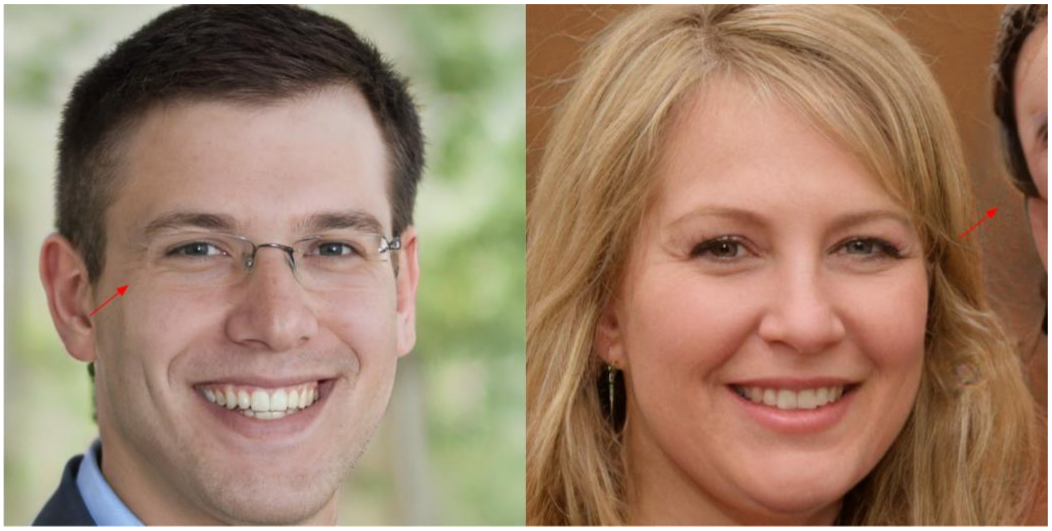
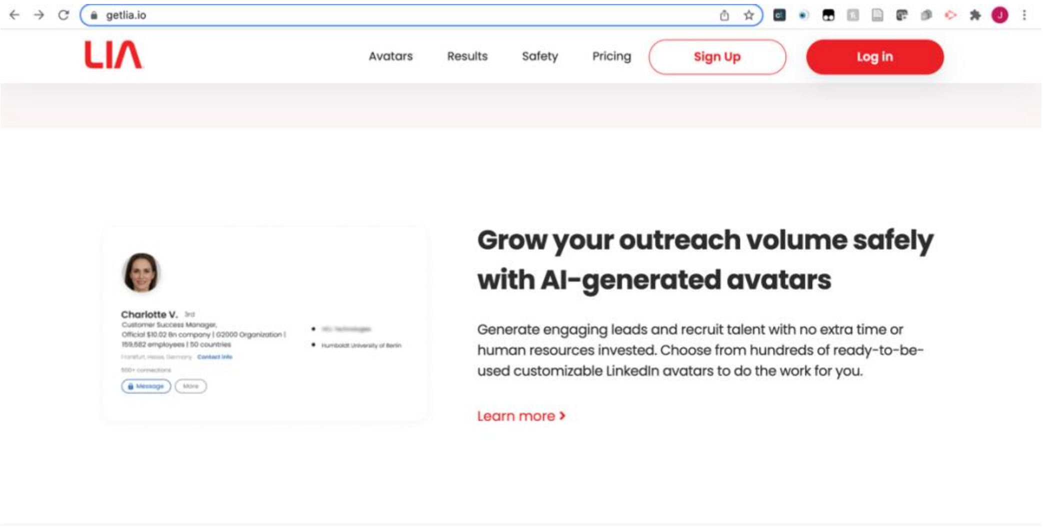


Figure 4. Examples of GAN-generated images from LinkedIn accounts with clear signs of inauthenticity. *The image on the left has glasses that blend into a face, and the image on the right has partial rendering of a second face.*

Through the "People Also Viewed" feature, we also found a company that advertised its services generating "engaging leads" with "AI-generated avatars" (Figure 5). Analogous to propaganda account outsourcing described above, this suggests that companies are outsourcing lead generation to third-party firms, which, in turn, are running fake accounts on behalf of the original companies. In Figure 5, we show a screenshot from the website getlia[dot]io, a company that offers LinkedIn automation tools as a service. We surfaced additional third-party firms from a simple Google search as well.



Charlotte V. 3rd
Customer Success Manager,
Official $10.02 Bn company | G2000 Organization |
159,582 employees | 50 countries
Frankfurt, Hesse, Germany · Contact info
500+ connections

- [redacted]
- Humboldt University of Berlin

Grow your outreach volume safely with AI-generated avatars

Generate engaging leads and recruit talent with no extra time or human resources invested. Choose from hundreds of ready-to-be-used customizable LinkedIn avatars to do the work for you.

Figure 5. LIA website screenshot. *Screenshot from the website of a third-party firm that offers clients LinkedIn outreach services with "AI-generated avatars."*

The accounts with GAN-generated images were designed to be indistinguishable from real people. Behaviorally, they joined groups (presumably for the purpose of gaining visibility into a larger percentage of LinkedIn users in order to generate new leads), then interacted with real target users via messages. The accounts also engaged in other activities such as liking posts; the accounts liked more than 60,000 distinct posts, bestowing more than 83,000 “Likes” in total.[16] The posts that the accounts liked were a mix of public content ranging from personal or professional commentary to job-related posts.[17]

To generate plausible identities, the accounts claimed to have attended university and worked at other companies in the past. There were over 250 different universities claimed as alma maters for the personas; the majority of these affiliations were appropriated once, but 17 schools appeared more than a dozen times, and New York University was listed in 53 of the 1,003 profiles. (New York City was the most popular declared location). The profiles claimed to work or to have worked in the past, for 594 distinct companies. A number of profiles appeared to be repurposed by their operators for multiple clients, either listing multiple jobs at companies that appeared to be clients or removing the prior client before listing a new client in its place. Other accounts listed what appeared to be fabricated ties to prominent companies (such as AT&T, Johnson & Johnson, and Coca-Cola). We suspect that, as with political influence operations, fake persona creation was an attempt to give the impression of a real person behind the account, in this environment, to decrease the likelihood of detection and increase the persuasiveness of messages to real LinkedIn users.

Methods

We surfaced the LinkedIn profiles described above using two different processes: 1) following recommendations from LinkedIn’s “People Also Viewed” feature, and 2) surfacing new leads via Google search. We outline these two methods below.

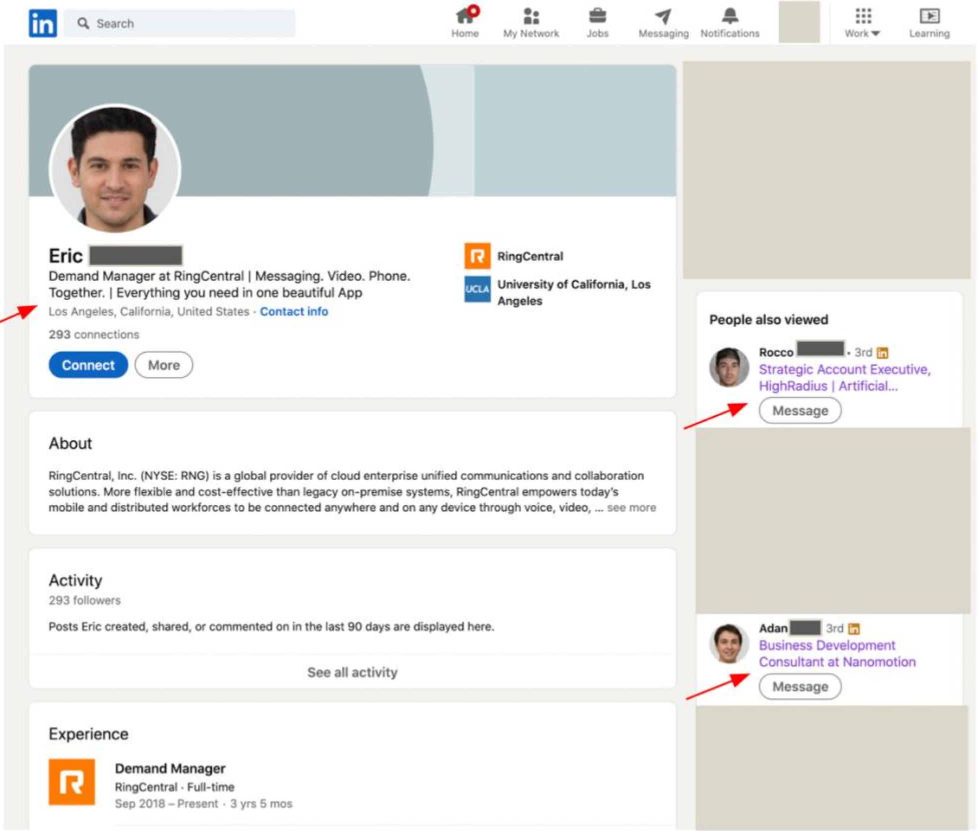
First, starting with the initial LinkedIn account that purported to work for RingCentral, we surfaced additional accounts with GAN-generated images by following LinkedIn’s “People Also Viewed” recommendations. In Figure 6, we provide an example. When viewing an account that claims to work for RingCentral with a GAN-generated image, under LinkedIn’s “People Also Viewed” we saw accounts that purported to work for HighRadius and Corrisoft with suspected GAN-generated images. We then examined those accounts and looked at other employees on the companies’ respective “People” pages on LinkedIn. Using this pivot approach, we surfaced dozens of different companies listed on LinkedIn where accounts with GAN-generated images claimed to work.[18] In Figure 7 we provide a stylized depiction of this search process.[19]

[16] To calculate these summary statistics, we used the open-source intelligence software Maltego.

[17] The post liked by the greatest number of the fake accounts (78 of the 1,003) was a highly popular post with over 500,000 likes. Most of the 60,000 posts were liked by only one of the accounts.

[18] For each suspected GAN profile, we recorded information in two ways: First, we used the Google Chrome “SingleFile” plug-in to capture the entire page as an HTML. This allowed us to archive information such as the account’s name, “About” description, educational history, and past work experience. Second, we downloaded the profile picture separate from the SingleFile to retain a higher resolution image. This higher resolution image allowed us to enlarge the picture for manual inspection, and to partner with a company with a GAN detection model (described further below) to assess whether the image was, in fact, AI-generated.

[19] This is similar to a snowball sampling or a chain-referral strategy. Instead of asking study participants to recommend others, we used their company listing or LinkedIn’s “People Also Viewed” feature to surface additional accounts.



Eric [redacted]
Demand Manager at RingCentral | Messaging. Video. Phone. Together. | Everything you need in one beautiful App
Los Angeles, California, United States · Contact info
293 connections

Connect   More

RingCentral
University of California, Los Angeles

People also viewed

Rocco [redacted] · 3rd
Strategic Account Executive, HighRadius | Artificial...
Message

Adan [redacted] 3rd
Business Development Consultant at Nanomotion
Message

About

RingCentral, Inc. (NYSE: RNG) is a global provider of cloud enterprise unified communications and collaboration solutions. More flexible and cost-effective than legacy on-premise systems, RingCentral empowers today's mobile and distributed workforces to be connected anywhere and on any device through voice, video, ... see more

Activity
293 followers

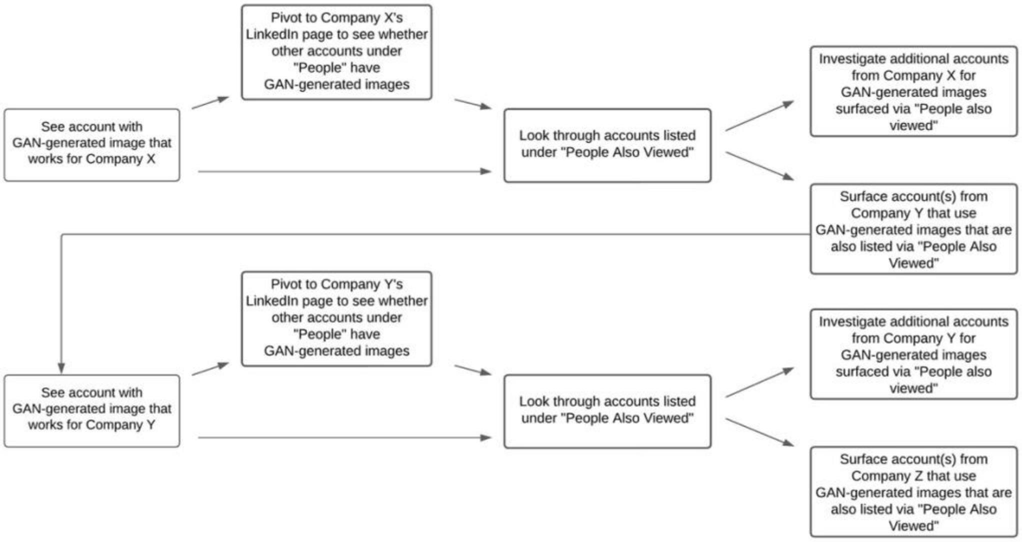
Posts Eric created, shared, or commented on in the last 90 days are displayed here.

See all activity

Experience

Demand Manager
RingCentral · Full-time
Sep 2018 – Present · 3 yrs 5 mos

***Figure 6. Screenshot highlighting "People Also Viewed" search strategy.** When viewing Eric's account that purports to work for RingCentral, accounts Rocco and Adan that purport to work for different companies were listed under "People Also Viewed." We used this search strategy for surfacing additional accounts and additional companies with GAN-generated images.*



See account with GAN-generated image that works for Company X
Pivot to Company X's LinkedIn page to see whether other accounts under "People" have GAN-generated images
Look through accounts listed under "People Also Viewed"
Investigate additional accounts from Company X for GAN-generated images surfaced via "People also viewed"
Surface account(s) from Company Y that use GAN-generated images that are also listed via "People Also Viewed"
See account with GAN-generated image that works for Company Y
Pivot to Company Y's LinkedIn page to see whether other accounts under "People" have GAN-generated images
Look through accounts listed under "People Also Viewed"
Investigate additional accounts from Company Y for GAN-generated images surfaced via "People also viewed"
Surface account(s) from Company Z that use GAN-generated images that are also listed via "People Also Viewed"

Figure 7. A stylized depiction of our first primary search strategy for surfacing GAN-generated images.

Second, we surfaced additional companies with GAN-generated images on LinkedIn by searching on Google for companies that advertised AI-generated LinkedIn services. In several cases, accounts with GAN-generated images were listed as employees of these companies on LinkedIn. When viewing these accounts, additional persona accounts (purportedly from other companies) again appeared on "People Also Viewed."

After surfacing the profiles, we contacted Sensity AI, a private firm with a model that detects GAN-generated images.[20] We ran 975 profile pictures from the LinkedIn accounts through Sensity (omitting images from accounts that had been removed between our initial investigation and our archiving process).[21] For 968 of those profile pictures, the model had over 90% confidence that the image was GAN-generated, and for 900 of the profile pictures, the confidence rating was above 99.9%. The overwhelming majority of those images were attributed to StyleGAN2.[22] Of the seven images that the Sensity model was not highly confident were GAN-generated, several had watermarks that may have distorted the Sensity analysis. While we could not independently audit Sensity's model, Sensity's analysis provided an independent data point to support the notion that the images were, in fact, generated by AI. While these models can detect StyleGAN/StyleGAN2 images, we acknowledge that they may not provide a solution to other AI-synthesized images and that as these detection technologies become more accessible, they may become easier to circumvent (the "detection dilemma") (Leibowicz et al., 2021).

Limitations

Our research has limitations related to exhaustiveness, replicability, and data access. Our research into GAN-generated images on LinkedIn was not exhaustive but meant to document a set of uses for the disinformation field. Because we detected suspected GAN-generated images by eye, there are likely GAN-generated images we did not include in our list of accounts, resulting in an undercount. In many cases, we saw suspected GAN-generated images but were restricted by LinkedIn from clicking on the account, perhaps because we were too many degrees of connections away. A second product of searching by eye is that the research may not be replicable. This is particularly the case now that LinkedIn has actioned almost all of the accounts.[23]

Bibliography

Bradshaw, S., Campbell-Smith U., Henle, A., Perini, A., Shalev, S., Bailey, H., & Howard, P. N. (2020). *Country case studies industrialized disinformation: 2020 global inventory of organized social media manipulation.* Oxford Internet Institute. http://demtech.oii.ox.ac.uk/wp-content/uploads/sites/127/2021/03/Case-Studies_FINAL.pdf

DiResta, R., Grossman, S., & Siegel, A. (2022). In-house vs. outsourced trolls: How digital mercenaries shape state influence strategies. *Political Communication, 39*(2), 222–253*.* https://doi.org/10.1080/10584609.2021.1994065

[20] We chose to use Sensity's model after experimenting with their previous publicly-available detection tool.

[21] When we first discovered the accounts with suspected GAN-generated images, we created a spreadsheet of account URLs, including information on the name of the account and the company the account claimed to work for. We did not immediately archive the accounts. When we archived the accounts several weeks later, 28 of them had been removed (either by LinkedIn or by the operators). So, we could not test the full 1,003 accounts using Sensity's model.

[22] StyleGAN and StyleGAN2 are used by thispersondoesnotexist[dot]com and generated[dot]photos (Metz 2019), two websites that produce GAN-generated images easily accessible to the public.

[23] Although we provided in-text examples of accounts with GAN-generated images in this paper, we do not provide a full set of archived pages in an online appendix. Although the risk of a false positive is reduced by using Sensity's model to independently analyze the nature of the images, because we could not independently audit the model, we do not risk exposing a false positive. For researchers interested in additional examples, please contact us.

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., Albright, J., & Johnson, B. (2019). *The tactics & tropes of the Internet Research Agency.* New Knowledge. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1003&context=senatedocs

Gershgorn, D. (2018, December 28). *AI can generate fake faces now. Here's how to spot them*. Quartz. https://qz.com/1510746/how-to-identify-fake-faces-generated-by-ai/

Goldstein, J., & Grossman, S. (2021, January 4). *How disinformation evolved in 2020.* Brookings TechStream. https://www.brookings.edu/techstream/how-disinformation-evolved-in-2020/

Golovchenko, Y., Buntain, C., Eady, G., Brown, M. A., & Tucker, J. A. (2020). Cross-platform state propaganda: Russian trolls on Twitter and YouTube during the 2016 U.S. presidential election. *The International Journal of Press/Politics*, *25*(3), 357–389. https://doi.org/10.1177/1940161220912682

Graphika. (2020, October). *Step into my Parler: Suspected Russian operation targeted far-right American users on platforms including Gab and Parler, resembled recent IRA-linked operation that targeted progressives.* https://public-assets.graphika.com/reports/graphika_report_step_into_my_parler.pdf

Generated.photos website. https://generated.photos/

Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (2018). *The IRA, social media and political polarization in the United States, 2012–2018.* Oxford Internet Institute. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1004&context=senatedocs

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4401–4410). IEEE. https://openaccess.thecvf.com/content_CVPR_2019/papers/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.pdf

LinkedIn User Agreement. Retrieved March 20, 2022, from https://www.linkedin.com/legal/user-agreement

LinkedIn Community Report. Retrieved June 25, 2022, from https://about.linkedin.com/transparency/community-report

Linvill, D. L. & Warren, P. L. (2020a). Troll factories: Manufacturing specialized disinformation on Twitter. *Political Communication, 37*(4), 447–467. https://doi.org/10.1080/10584609.2020.1718257

Linvill, D. L. & Warren, P. L. (2020b). Engaging with others: How the IRA coordinated information operation made friends. *Harvard Kennedy School (HKS) Misinformation Review, 1*(2). https://doi.org/10.37016/mr-2020-011

Leibowicz, C. R., McGregor, S., & Ovadya, A. (2021). The deepfake detection dilemma: A multistakeholder exploration of adversarial dynamics in synthetic media. In *Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society* (pp. 736–744). Association for Computing Machinery. https://doi.org/10.1145/3461702.3462584

Martin, D. A., Shapiro, J. N., & Ilhardt, J. (2020). *Trends in online influence efforts.* Empirical Studies of Conflict Project. https://esoc.princeton.edu/publications/trends-online-influence-efforts

McClure Haughey, M., Muralikumar, M. D., Wood, C. A., & Starbird, K. (2020). On the misinformation beat: Understanding the work of investigative journalists reporting on problematic information online. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), 1–22. https://doi.org/10.1145/3415204

Metz, R. (2019, February 28). *These people do not exist. Why websites are churning out fake images of people (and cats).* CNN. https://www.cnn.com/2019/02/28/tech/ai-fake-faces/index.html

Mink, J., Luo, L., Barbosa, N. M., Figueira, O., Wang, Y., & Wang, G. (2022). DeepPhish: Understanding user trust towards artificially generated profiles in online social networks. In *Proceedings of the 31st USENIX security conference* (pp.1669–1686). USENIX Association. https://www.usenix.org/system/files/sec22-mink.pdf

Nightingale, S. J., & Farid, H. (2021). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences, 119*(8). https://doi.org/10.1073/pnas.2120481119

Nimmo, B., Eib, C. S., Tamora, L., Johnson, K., Smith, I., Buziashvili, E., Kann, A., Karan, K., Ponce de Leon Rosas, E., & Rizzuto, M. (2019, December). *#OperationFFS: Fake face swarm.* Graphika and the Atlantic Council's Digital Forensics Research Lab. https://public-assets.graphika.com/reports/graphika_report_operation_ffs_fake_face_storm.pdf

Nimmo, B., François, C., Eib, C. S., & Ronzaud, L. (2020, August). *Spamouflage goes to America: Pro-Chinese inauthentic network debuts English-language videos*. Graphika. https://public-assets.graphika.com/reports/graphika_report_spamouflage_goes_to_america.pdf

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology, 34*(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x

RingCentral website. https://www.ringcentral.com/why-us-enterprise.html

Shen, B., RichardWebster, B., O'Toole, A., Bowyer, K., & Scheirer, W. J. (2021). *A study of the human perception of synthetic faces*. arXiv. https://arxiv.org/abs/2111.04230

Stanford Internet Observatory. (2020, October 8). *Reply-Guys go hunting: An investigation in a U.S. astroturfing operation on Facebook, Twitter, and Instagram*. https://cyber.fsi.stanford.edu/io/news/oct-2020-fb-rally-forge

Strick B. (2020, November 11). *West Papua: New online influence operation attempts to sway independence debate.* Bellingcat. https://www.bellingcat.com/news/2020/11/11/west-papua-new-online-influence-operation-attempts-to-sway-independence-debate/

Wang, X., Guo, H., Hu, S., Chang, M. C., & Lyu, S. (2022). *GAN-generated faces detection: A survey and new perspectives.* arXiv. https://arxiv.org/pdf/2202.07145.pdf

Acknowledgements

We thank Shannon Bond for investigating offline on our online findings. We thank SensityAI for running the images we found through their GAN-detection model. For feedback on earlier versions of this manuscript, we thank Elena Cryst, Shelby Grossman, Claire Leibowicz, Micah Musser, Riana Pfefferkorn, and three anonymous reviewers. All errors remain our own.

Funding

No funding has been received to conduct this research.

Competing interests

The authors declare no competing interests.

Ethics

We relied exclusively on publicly available data and did not seek IRB approval.

Copyright

Data availability

We are not making screenshots of the now-removed accounts publicly available, for risk of false-positive exposure. Researchers working on related topics can contact us for additional examples.