Center for Countering Digital Hate

FACEBOOK'S AI FAILURE

How Meta's platform is failing to deal with AI-generated images of fake Americans in the US elections

The Center for Countering Digital Hate works to stop the spread of online hate and disinformation through innovative research, public campaigns and policy advocacy.

Our mission is to protect human rights and civil liberties online.

Social media platforms have changed the way we communicate, build and maintain relationships, set social standards, and negotiate and assert our society's values. In the process, they have become safe spaces for the spread of hate, conspiracy theories and disinformation.

Social media companies erode basic human rights and civil liberties by enabling the spread of online hate and disinformation.

At CCDH, we have developed a deep understanding of the online harm landscape, showing how easily hate actors and disinformation spreaders exploit the digital platforms and search engines that promote and profit from their content.

We are fighting for better online spaces that promote truth, democracy, and are safe for all.

If you appreciate this report, you can donate to CCDH at counterhate.com/donate. In the United States, Center for Countering Digital Hate Inc is a 501(c)(3) charity. In the United Kingdom, Center for Countering Digital Hate Ltd is a nonprofit company limited by guarantee.

Table of Contents

Political endorsements by fake Americans accrue 2 million interactions on Facebook

AI-generated images of fake Americans appearing to make political endorsements, some of which portray fake military veterans, are gaining millions of likes, comments and shares on Facebook in the run up to the US presidential elections.

Comments on the posts show that many Facebook users appear to find the images credible. They include photorealistic images of fake military veterans, police officers and protestors appearing to endorse political candidates or policies.

Despite Facebook's promise to label AI-generated images, none of the images identified by researchers were labelled. Transparency information shows that Pages posting the images are in some cases being run from Morrocco or Pakistan.

Researchers identified 169 posts published since July 1 with a combined total of 2.4 million interactions. These posts were shared 476,014 times. All posts contained an image of a fake person offering political opinions relevant to the US election and were assessed by researchers to be likely made using generative AI.

Examples of suspected AI-generated posts treated credibly by Facebook users include:

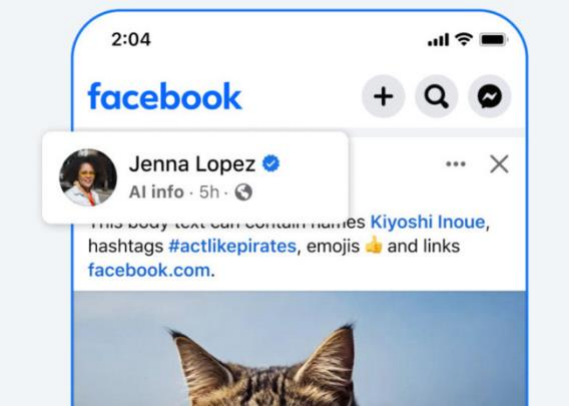
- A fake veteran behind text saying, "They'll hate me for this but learning English should be a requirement for citizenship" with 168,000 likes, comments and shares.[i]
- Fake police officers carrying a sign saying, "Let's give police officers a raise before paying off student loans" with 28,700 likes, comments and shares.[ii]
- A fake veteran behind text saying, "Pride month but no veteran's month? Our priorities need fixing!" with 120,300 likes, comments and shares.[iii]
- A fake man and child holding a sign saying: "Kamala Harris is so extreme, she'd even tax your lemonade stand!" with 3,009 likes, comments and shares.[iv]
- A fake veteran holding an incorrectly designed and folded American flag with the caption "Veterans deserve better than being second to student loans" with 64,875 likes, comments, and shares.[v]
- A fake woman holding a sign saying, "I am a voice for the unborn, advocating for their right to life" with 2,891 likes, comments and shares.[vi]

Often these images had subtle markers of AI generation such as distorted hands, distorted writing or vague backgrounds. The posts were identified from across ten separate Facebook Pages, each of which were selected for the study for having posted at least three suspected AI-generated images on political themes with at least 100 likes since July 1.

Facebook failed to apply AI labels to a single misleading image

Facebook claims to utilize technology that detects images made using AI so that users can be informed in cases where a photorealistic image is artificial.[vii] The company says it detects the use of AI using "industry standard AI image indicators" and marks images where generative AI is detected with an 'AI info' label.[viii]

The image below shows Facebook's example of how these labels look: a small 'AI info' label should appear above an image, next to the image upload date.[ix]



When the policy was first announced in February, Facebook specifically noted that it was being rolled out during a year in which "a number of important elections are taking place around the world".[x] The policy also states that in cases where an image has been assessed to hold a "particularly high risk of materially deceiving the public on a matter of importance", a more prominent label "may" be added.

Despite these policies, we found that none of the images in our sample featured an AI label – either large or small – leaving users with no context that the images are artificial.

There is no clear way of reporting content as AI using Facebook's reporting tools

Researchers were unable to identify a clear route for users to report deceptive images that bypass Facebook's AI detection systems.

Facebook's reporting tools enable users to report posts that violate Facebook's policies such as hate speech, but there is no option for users to report AI-generated content or 'manipulated media', despite referencing this content in its policies.[xi]

It appears that users would only be able to report posts featuring AI-generated images in cases that violate other Facebook policies, such as those covering misinformation.

Posts from Pages with admins outside the US accrued 1.5 million interactions

Over half of the Pages analyzed in this study, responsible for over a million likes, shares and comments on AI-generated images about the election, have administrators from outside the US including Morrocco, Pakistan, Indonesia, Belgium and the UK.

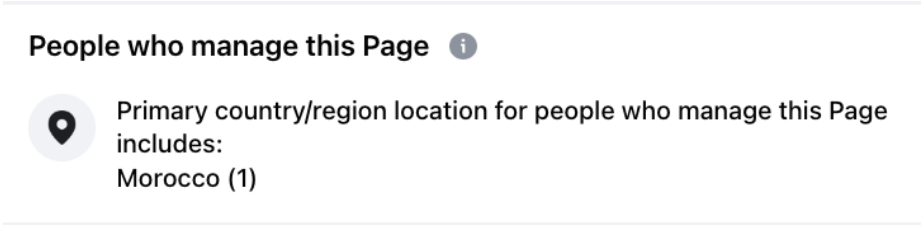
For each of the ten Pages analyzed in this study, researchers identified the locations of the users who manage the Page via the "Page transparency" section, which provides information on the primary country location for people who manage each Page.

This analysis showed that six of the ten Pages – which accumulated 1.5 million likes, shares and comments on AI-generated election posts in total – were administered by at least one user based abroad. It shows that users outside the US are generating images of fake Americans to promote highly politicized views during the US elections.

Page	Location of Managers	Total Interactions
Page 1	Morocco, USA	828,430
Page 2	USA	571,878
Page 3	Morocco	293,743
Page 4	USA, Indonesia	221,670
Page 5	Morocco	178,817
Page 6	USA	148,902
Page 7	USA	146,933
Page 8	USA	18,372
Page 9	Pakistan	15,579
Page 10	USA, Belgium, UK	7,915

The most common country listed was Morocco, with three Pages in the study having Morocco listed as the location for at least one user managing the Page. Pages with at least one manager listed as based in Morrocco were responsible for 1.2 million interactions on images in the study.

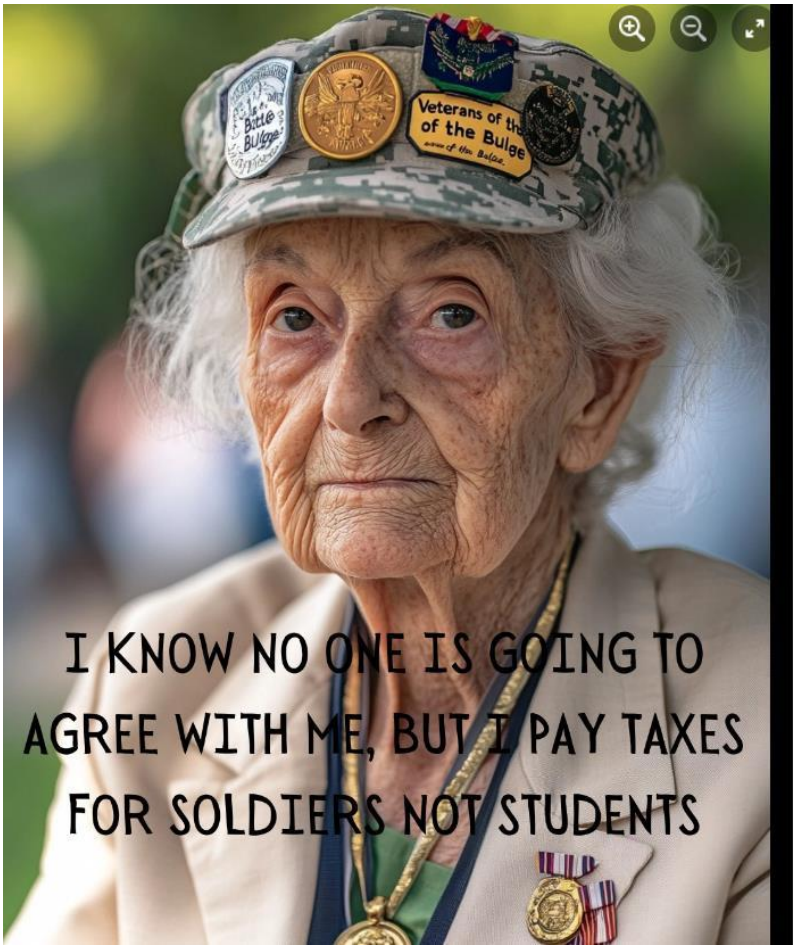
The below screenshot shows the Transparency section of one of the Pages in the study which has posted numerous AI images of fake veterans alongside political slogans. The only location listed for the Page's managers is Morocco.



Examples: Fake military veterans endorsing political candidates or policies

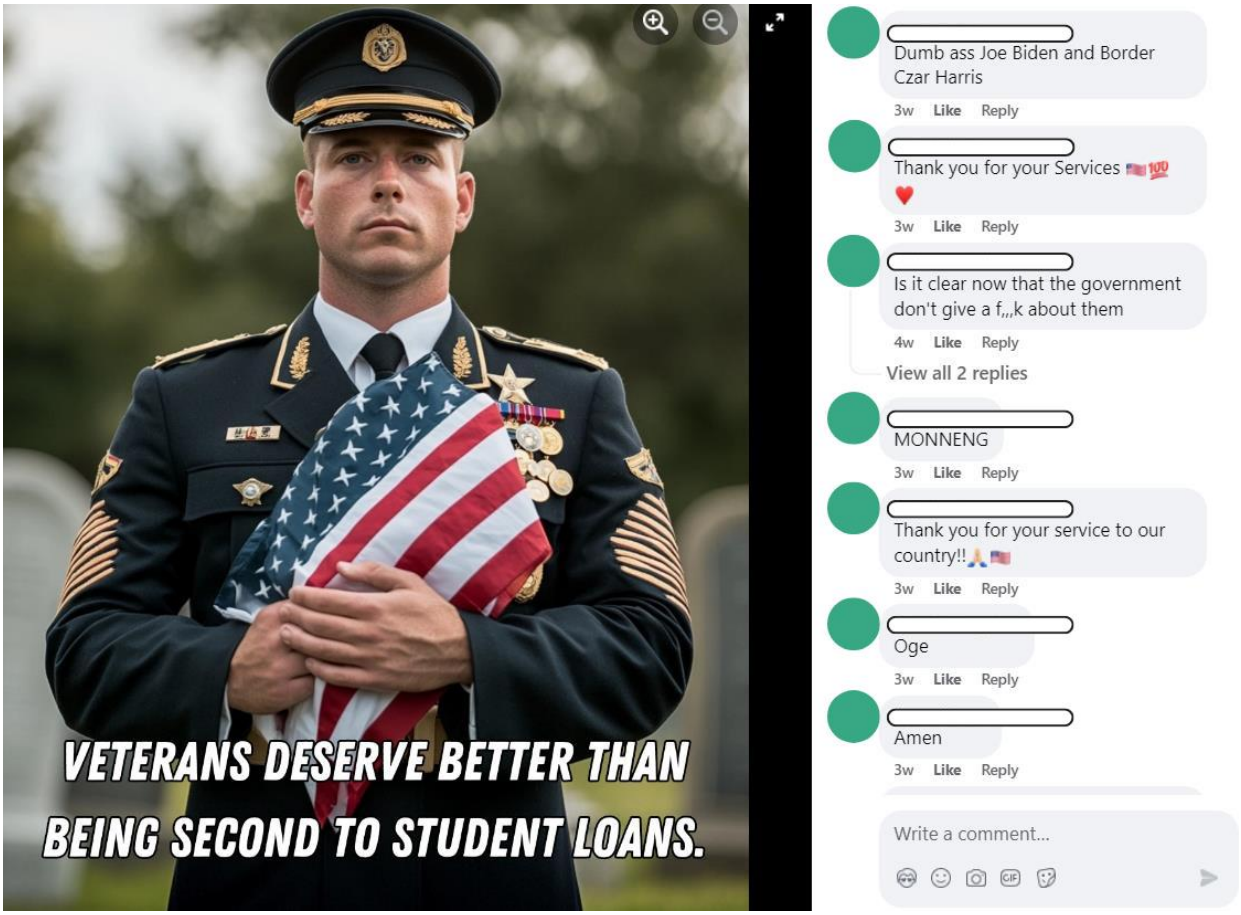
This image with over 128k reactions depicts a veteran with a caption endorsing the view that learning English should be a requirement for US citizenship.[xii] Multiple people in the comments are seen commenting "I agree". One sign of AI is that many of the ribbons on her uniform do not correspond to actual military honors and are not aligned properly.

This image depicts a veteran along with a caption implying they oppose student loan forgiveness.[xiii] One of the user comments "Thank you so very much for your service and sacrifice for our country!!!". On closer examination, the patches on her hat contain distortions and nonsensical text that is typical of AI-generated images.



This image depicts a veteran with the caption “Pride month but no veteran month? Our priorities need fixing!”.[xiv] One commenter said “I agree with this statement!” and another said “For sure thank you all for your service and sacrifices”. Signs of AI distortions can be seen in the design of the uniform and medals.

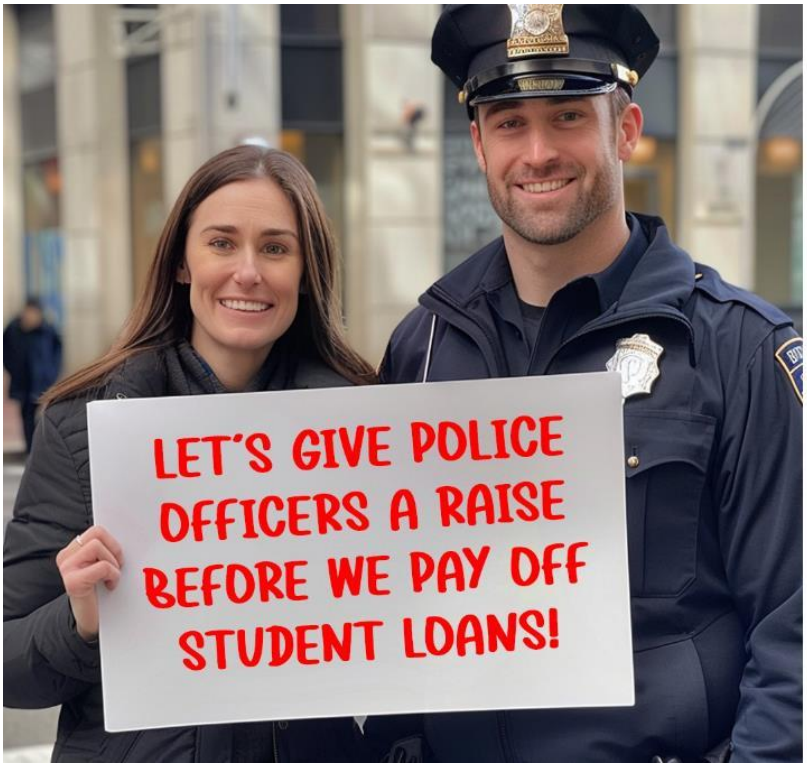
This image depicts a veteran holding an American flag in a cemetery with the caption "Veterans deserve better than being second to student loans".[xv] Two different users can be seen thanking the veteran for their service. Distortions in the uniform and flag, including how the flag is incorrectly folded, point to AI.



This image depicts a veteran holding a small flag and wearing an older military uniform with the caption “Maybe it’s just me, but I strongly believe veterans deserve far better benefits than those offered to undocumented immigrants”.[xvi] Several users commented “Amen” while others thanked the man for his service. The hands and incorrectly shaped stars on the flag point to AI.

Examples: Fake people endorsing political candidates or policies

This image depicts a police officer and a woman protesting about student loan forgiveness.[xvii] The post has several comments saying “I agree” and one saying “God bless and thank you all for your service”. The illegible text on the hat of the police officer points to AI.



This image depicts a friendly-looking man and child protest Kamala Harris for being “extreme”.[xviii] One person commented "Love it!" while another commented “You got that right.” Signs of AI in this photo include the eyes of the man and the ear of the young girl.

This image depicts a woman with a sign at an apparent protest on the topic of abortion.[xix] One commenter says “God bless you” and another says “An honorable cause. Standing up for a helpless baby, created by the good Lord”. The incorrect hand anatomy of the protestor is a classic sign of AI.

Appendix: Methodology

This appendix outlines in more detail how we identified Facebook Pages that post AI content relevant to the US elections and identified the locations of their administrators.

How we identified Facebook Pages posting election-relevant AI content

Researchers used a Facebook account set up with a false name, an adult date of birth and a US location. They then used this account to scroll the platform's News Feed, collecting examples of suspected AI-generated images and the Pages that posted them.[xx]

Pages were selected for further study if they had posted suspected AI-generated images related to the US elections on at least two occasions. These images were typically of fake people endorsing political messages.

How we collected suspected AI-generated images about the US elections

For each Facebook Page selected for further study, researchers gathered all posts made July 1 and October 22, 2024 with more than 100 likes that were suspected to be AI-generated. Posts were included if they contained photorealistic images of fake people endorsing political messages.

Pages were excluded from further analysis if they had fewer than three posts matching these criteria, leaving us with ten Facebook Pages.

How we identified and categorized AI-generated images in our final analysis

Posts were sorted into two categories in our final analysis: confirmed AI-generated images, and suspected AI-generated images.

Confirmed AI-generated images were assessed as being "likely to contain AI-generated or deepfake content" by Hive, a tool for detecting AI-generated content that has performed relatively well in comparisons.[xxi] These confirmed images were also independently assessed by two researchers as containing features typical of AI images, such as distorted hands and text, or vague backgrounds. 64% of images in our analysis are in this category.

Suspected AI-generated images were not assessed as being AI-generated images by Hive but were posted from Pages that we know to have posted images meeting our definition of "confirmed AI-generated images", and were independently assessed by two researchers as containing features typical of AI images, such as distorted hands and text, or vague backgrounds. 36% of images in our analysis are in this category.

Images falling into either of these two categories have been included in our final set of election-relevant AI-generated images. We then calculated the total number of likes,

comments and shares on posts in our set to arrive at our finding that 169 posts featuring these images had amassed 2.4 million interactions between July 1 and 22 October, 2024.

We designed this approach to maximize our capture of relevant images from Pages known to repeatedly post AI-generated images. At present, researchers simply do not have the means to make wholly reliable assessment of whether images are AI-generated. Notably, while AI-detection tools can be useful in giving an indication that images were made using an AI, they are not perfect and can be less effective at spotting the signs of AI in cases where images have been blurred, screenshotted or compressed.[xxii]

Our categorization of some images as "suspected AI-generated images" accepts the possibility that some of our assessments could be false positives, although our researchers agreed with assessments by Hive for 64% of the posts in our analysis.

How we determined the locations of Page administrators

For each of the 10 Pages analyzed in this study, researchers identified the locations of the users who manage the Page via the "Page transparency" section, which provides information on the primary country location for people who manage each Page.

End notes

[i] American pride, Facebook, 6 October 2024, https://www.facebook.com/photo?fbid=524908300297119&set=pb.100083338601008.-2207520000
[ii] Liberty Powered News, Facebook, 30 September 2024, https://www.facebook.com/photo/?fbid=12218990916822349 4&set=pb.61556704837966.-2207520000
[iii] CEPRU Unsaac, Facebook, 25 September 2024, https://www.facebook.com/photo/?fbid=551888240732065&set=pb.100077326822729.-2207520000
[iv] Liberty Powered News, Facebook, 25 July 2024, https://www.facebook.com/photo/?fbid=122162904848223494&set=pb.61556704837966.-2207520000
[v] Honoring Our Heroes, Facebook, 21 September 2024, https://www.facebook.com/photo/?fbid=122112628670501866&set=pb.61565055994917.-2207520000
[vi] Liberty Powered News, Facebook, 19 September 2024, https://www.facebook.com/photo/?fbid=122185263062223494&set=pb.61556704837966.-2207520000
[vii] "Our Approach to Labeling AI-Generated Content and Manipulated Media", Meta, 5 April 2024, https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/
"Labeling AI-Generated Images on Facebook, Instagram and Threads", Meta, 6 February 2024, https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/
[viii] "Our Approach to Labeling AI-Generated Content and Manipulated Media", Meta, 5 April 2024, https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/
[ix] "Our Approach to Labeling AI-Generated Content and Manipulated Media", Meta, 5 April 2024, https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/
[x] "Labeling AI-Generated Images on Facebook, Instagram and Threads", Meta, 6 February 2024, https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/
[xi] "Misinformation", Meta, accessed 25 October 2024, https://transparency.meta.com/es-la/policies/community-standards/misinformation
[xii] American pride, Facebook, 6 October 2024, https://www.facebook.com/photo?fbid=524908300297119&set=pb.100083338601008.-2207520000
[xiii] Honoring Our Heroes, Facebook, 4 October 2024, https://www.facebook.com/photo/?fbid=122116587710501866&set=pb.61565055994917.-2207520000
[xiv] CEPRU Unsaac, Facebook, 25 September 2024, https://www.facebook.com/photo/?fbid=551888240732065&set=pb.100077326822729.-2207520000
[xv] Honoring Our Heroes, Facebook, 21 September 2024, https://www.facebook.com/photo/?fbid=122112628670501866&set=pb.61565055994917.-2207520000
[xvi] Honoring Our Heroes, 5 September 2024, https://www.facebook.com/photo/?fbid=122105299466501866&set=pb.61565055994917.-2207520000
[xvii] Liberty Powered News, Facebook, 30 September 2024, https://www.facebook.com/photo/?fbid=12218990916822349 4&set=pb.61556704837966.-2207520000
[xviii] Liberty Powered News, Facebook, 25 July 2024, https://www.facebook.com/photo/?fbid=122162904848223494&set=pb.61556704837966.-2207520000
[xix] Liberty Powered News, Facebook, 19 September 2024, https://www.facebook.com/photo/?fbid=122185263062223494&set=pb.61556704837966.-2207520000
[xx] Researchers were asked to draw upon reputable guides for identifying AI-generated images, such as the following guide from McGill University.
"How to Spot AI Fakes (For Now)", McGill University, 14 March 2024, https://www.mcgill.ca/oss/article/critical-thinking-technology/how-spot-ai-fakes-now

[xxi] "Machine learning models to detect AI-generated content", Hive, accessed 25 October 2024, https://hivemoderation.com/ai-generated-content-detection
"Is it easy to fool detection tools?", The New York Times, 28 June 2023, https://www.nytimes.com/interactive/2023/06/28/technology/ai-detection-midjourney-stable-diffusion-dalle.html
[xxii] "Spotting the deepfakes in this year of elections: how AI detection tools work and where they fail", Reuters Institute, 15 April 2024, https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail