www.theguardian.com /technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies

‘There is no standard’: investigation finds AI algorithms objectify women’s bodies

Gianluca Mauro ⁝ 14-18 minutes ⁝ 2/8/2023

Images posted on social media are analyzed by artificial intelligence (AI) algorithms that decide what to amplify and what to suppress. Many of these algorithms, a Guardian investigation has found, have a gender bias, and may have been censoring and suppressing the reach of countless photos featuring women’s bodies.

These AI tools, developed by large technology companies, including Google and Microsoft, are meant to protect users by identifying violent or pornographic visuals so that social media companies can block it before anyone sees it. The companies claim that their AI tools can also detect “raciness” or how sexually suggestive an image is. With this classification, platforms – including Instagram and LinkedIn – may suppress contentious imagery.

Two Guardian journalists used the AI tools to analyze hundreds of photos of men and women in underwear, working out, using medical tests with partial nudity and found evidence that the AI tags photos of women in everyday situations as sexually suggestive. They also rate pictures of women as more “racy” or sexually suggestive than comparable pictures of men. As a result, the social media companies that leverage these or similar algorithms have suppressed the reach of countless images featuring women’s bodies, and hurt female-led businesses – further amplifying societal disparities.

Even medical pictures are affected by the issue. The AI algorithms were tested on images released by the US National Cancer Institute demonstrating how to do a clinical breast examination. Google’s AI gave this photo the highest score for raciness, Microsoft’s AI was 82% confident that the image was “explicitly sexual in nature”, and Amazon classified it as representing “explicit nudity”.

📷Microsoft’s AI was 82% confident that this image demonstrating how to do a breast exam was ‘explicitly sexual in nature’, and Amazon categorized it as ‘explicit nudity’. Photograph: National Cancer Institute/Unsplash

Pregnant bellies are also problematic for these AI tools. Google’s algorithm scored the photo as “very likely to contain racy content”. Microsoft’s algorithm was 90% confident that the image was “sexually suggestive in nature”.

📷Images of pregnant bellies are categorized as ‘very likely to contain racy content’. Photograph: Dragos Gontariu/Unsplash

“This is just wild,” said Leon Derczynski, a professor of computer science at the IT University of Copenhagen, who specializes in online harm. “Objectification of women seems deeply embedded in the system.”

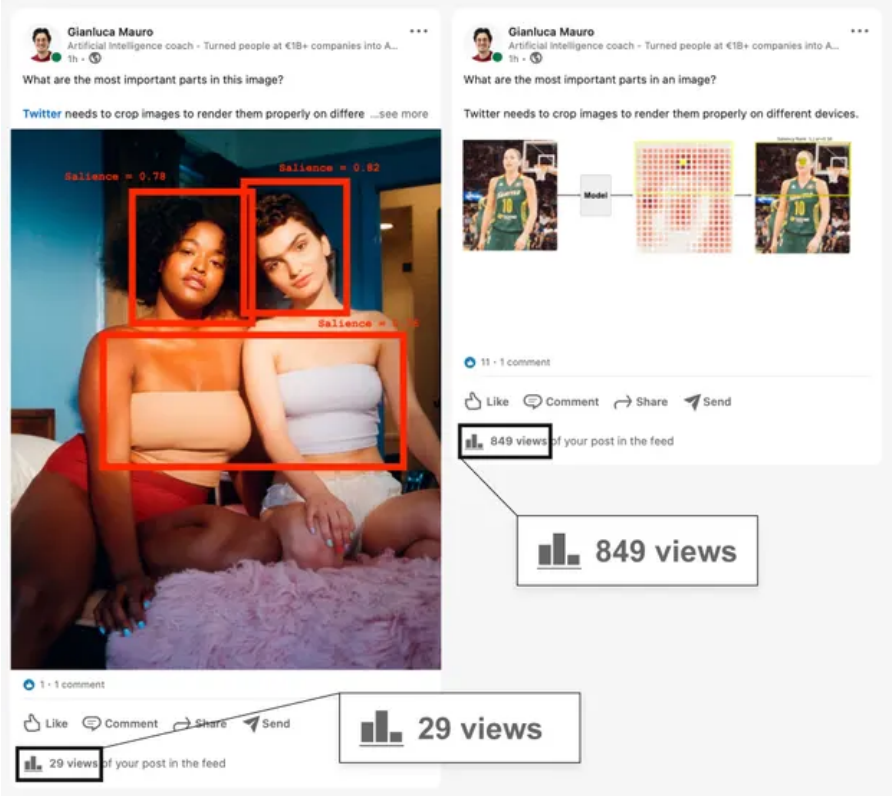
One social media company said it did not design its systems to create or reinforce biases and classifiers are not perfect.

“This is a complex and evolving space, and we continue to make meaningful improvements to SafeSearch classifiers to ensure they stay accurate and helpful for everyone,” a Google spokesperson said.

Getting shadowbanned

In May 2021, Gianluca Mauro, an AI entrepreneur, adviser and co-author of this article, published a LinkedIn post and was surprised it had just been seen 29 times in an hour, instead of the roughly 1,000 views he usually gets. Maybe the picture of two women wearing tube tops was the problem?

He re-uploaded the same exact text with another picture. The new post got 849 views in an hour.

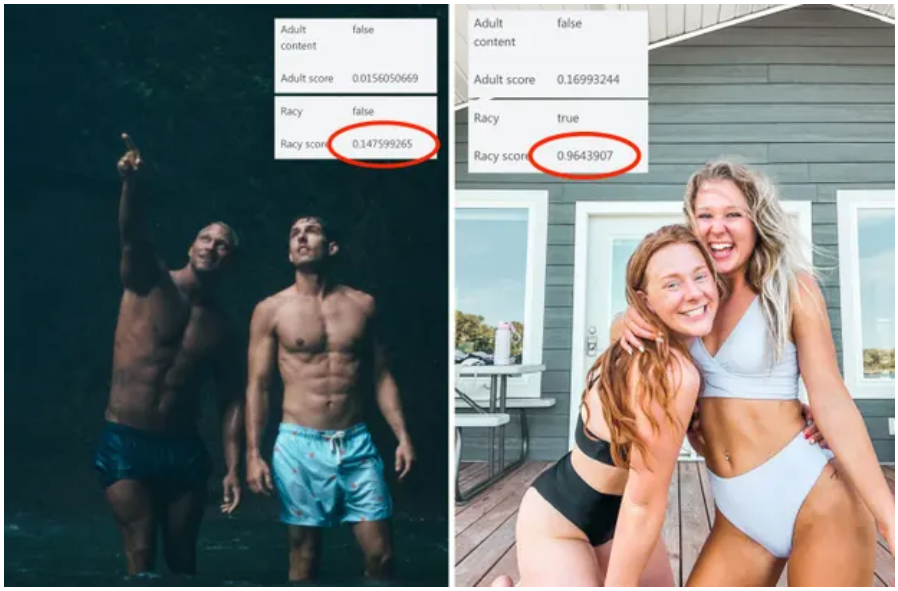


Mauro's LinkedIn post showing two women in tube tops received only 29 views in one hour compared to 849 views when a different image was used. Composite: Gianluca Mauro/The Guardian

It seemed like his post had been suppressed or "shadowbanned". Shadowbanning refers to the decision of a social media platform to limit the reach of a post or account. While a regular ban involves actively blocking a post or account and notifying the user, shadowbanning is less transparent - often the reach will be suppressed without the user's knowledge.

The Guardian found that Microsoft, Amazon and Google offer content moderation algorithms to any business for a small fee. Microsoft, the parent company and owner of LinkedIn, said its tool "can detect adult material in images so that developers can restrict the display of these images in their software".

Another experiment on LinkedIn was conducted to try to confirm the discovery.



The photo of the women got eight views in one hour, while the picture with the men received 655 views, suggesting the women's photo was either suppressed or shadowbanned. Composite: Gianluca Mauro/The Guardian

In two photos depicting both women and men in underwear, Microsoft's tool classified the picture showing two women as racy and gave it a 96% score. The picture with the men was classified as non-racy with a score of 14%.

The photo of the women got eight views within one hour, and the picture with the two men received 655 views, suggesting the photo of the women in underwear was either suppressed or shadowbanned.

Shadowbanning has been documented for years, but the Guardian journalists may have found a missing link to understand the phenomenon: biased AI algorithms. Social media platforms seem to leverage these algorithms to rate images and limit the reach of content that they consider too racy. The problem seems to be that these AI algorithms have built-in gender bias, rating women more racy than images containing men.

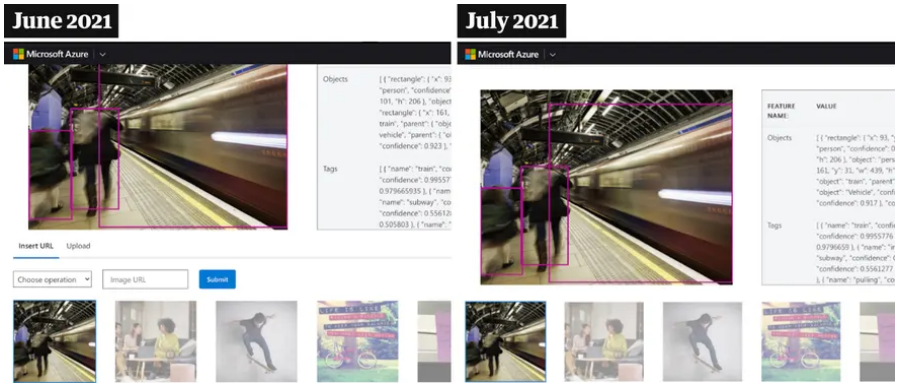
"Our teams utilize a combination of automated techniques, human expert reviews and member reporting to help identify and remove content that violates our professional community policies," said a LinkedIn spokesperson, Fred Han, in a statement. "In addition, our feed uses algorithms responsibly in order to surface content that helps our members be more productive and successful in their professional journey."

Amazon said content moderation was based on a variety of factors including geography, religious beliefs and cultural experience. However, "Amazon Rekognition is able to recognize a wide variety of content, but it does not determine the appropriateness of that content," an Amazon spokesperson said. "The service simply returns labels for items it detects for further evaluation by human moderators."

Digging deeper

Natasha Crampton, Microsoft's chief responsible AI officer, and her team began investigating when journalists notified her about the labeling of the photos.

“The initial results do not suggest that those false positives occur at a disproportionately higher rate for women as compared with men,” Crampton said. When additional photos were run through the tool, the demo website had been changed. Before the problem was discovered, it was possible to test the algorithms by simply dragging and dropping a picture. Now an account needed to be created and code had to be written.



Screenshots of Microsoft’s platform in June 2021 (left), and in July 2021 (right). In the first version, there is a button to upload any photo and test the technology, which has disappeared in the later version. Composite: Gianluca Mauro/The Guardian

But what are these AI classifiers actually analyzing in the photos? More experiments were needed, so Mauro agreed to be the test subject.

When photographed in long pants and with a bare chest, Microsoft’s algorithm had a confidence score lower than 22% for raciness. When Mauro put on a bra, the raciness score jumped to 97%. The algorithm gave a 99% score when the bra was held next to me.

“You are looking at decontextualized information where a bra is being seen as inherently racy rather than a thing that many women wear every day as a basic item of clothing,” said Kate Crawford, professor at the University of Southern California and the author of Atlas of AI.

Abeba Birhane, a senior fellow at the Mozilla Foundation and an expert in large visual datasets, said raciness is a social concept that differs from one culture to the other.

“These concepts are not like identifying a table where you have the physical thing and you can have a relatively agreeable definition or rating for a certain thing,” she said. “You cannot have one single uncontested definition of raciness.”

Why do these systems seem so biased?

Modern AI is built using machine learning, a set of algorithms that allow computers to learn from data. When developers use machine learning, they don’t write explicit rules telling computers how to perform a task. Instead, they provide computers with training data. People are hired to label images so that computers can analyze their scores and find whatever pattern helps it replicate human decisions.

Margaret Mitchell, chief ethics scientist at the AI firm Hugging Face and former co-head of Google’s Ethical AI research group, believes that the photos used to train these algorithms were probably labeled by straight men,

who may associate men working out with fitness, but may consider an image of a woman working out as racy. It's also possible that these ratings seem gender biased in the US and in Europe because the labelers may have been from a place with a more conservative culture.

Ideally, tech companies should have conducted thorough analyses on who is labeling their data, to make sure that the final dataset embeds a diversity of views, she said. The companies should also check that their algorithms perform similarly on photos of men v women and other groups, but that is not always done.

"There's no standard of quality here," Mitchell said.

This gender bias the Guardian uncovered is part of more than a decade of controversy around content moderation on social media. Images showing people breastfeeding their children and different standards for photos of male nipples, which are allowed on Instagram, and female nipples, which have to be covered, have long garnered outcries about social media platforms' content moderation practices.

Now Meta's oversight board – an external body including professors, researchers and journalists, who are paid by the company – has asked the tech giant to clarify its adult nudity and sexual activity community standard guidelines on social media platforms "so that all people are treated in a manner consistent with international human rights standards, without discrimination on the basis of sex or gender".

Meta declined to comment for this story.

'Women should be expressing themselves'

Bec Wood, a 38-year-old photographer based in Perth, Australia, said she was terrified of Instagram's algorithmic police force.

After Wood had a daughter nine years ago, she started studying childbirth education and photographing women trying to push back against societal pressures many women feel that they should look like supermodels.

"I was not having that for my daughter," she said. "Women should be expressing themselves and celebrating themselves and being seen in all these different shapes and sizes. I just think that's so important for humanity to move forward."

Wood's photos are intimate glimpses into women's connections with their offspring, photographing breastfeeding, pregnancy and other important moments in an artful manner. Her business is 100% dependent on Instagram: "That's where people find you," Wood said. "If I don't share my work, I don't get work."


Google and Microsoft rated Wood's photos as likely to contain explicit sexual content. Amazon categorized the image of the pregnant belly on the right as 'explicit nudity'.

Since Wood started her business in 2018, for some of her photos she got messages from Instagram that the company was either taking down some of her pictures or that they were going to allow them on her profile but not on the explore tab, a section of the app where people can discover content from accounts they don't follow. She hoped that Instagram was going to fix the issue over time, but the opposite happened, she said. "I honestly can't believe that it's gotten worse. It has devastated my business." Wood described 2022 as her worst year business-wise.

She is terrified that if she uploads the "wrong" image, she will be locked out of her account with over 13,000 followers, which would bankrupt her business: "I'm literally so scared to post because I'm like, 'Is this the post that's going to lose everything?'" she said.

To avoid this, Wood started going against what made her start her work in the first place: "I will censor as artistically as possible any nipples. I find this so offensive to art, but also to women," she said. "I almost feel like I'm part of perpetuating that ridiculous cycle that I don't want to have any part of."

Running some of Wood's photos through the AI algorithms of Microsoft, Google and Amazon, including those featuring a pregnant belly got rated as racy, nudity or even explicitly sexual.

Wood is not alone. Carolina Are, an expert on social media platforms and content moderation and currently an Innovation fellow at the Centre for Digital Citizens at Northumbria University said she has used Instagram to promote her business and was a victim of shadowbanning.

Are, a pole dance instructor, said some of her photos were taken down, and in 2019, she discovered that her pictures did not show up in the explore page or under the hashtag #FemaleFitness, where Instagram users can search content from users they do not follow. "It was literally just women working out in a very tame way. But then if you looked at hashtag #MaleFitness, it was all oily dudes and they were fine. They weren't shadowbanned," she said.

📷Carolina Are, a pole dance instructor, found that some of her photos were not showing up on social media. Photograph: Rachel Marsh/Courtesy of @ray.marsh

For Are, these individual problems point to larger systemic ones: many people, including chronically ill and disabled folks, rely on making money through social media and shadowbanning harms their business.

Mitchell, the chief ethics scientist at Hugging Face, these kinds of algorithms are often recreating societal biases: “It means that people who tend to be marginalized are even further marginalized – like literally pushed down in a very direct meaning of the term marginalization.”

It’s a representational harm and certain populations are not adequately represented, she added. “In this case, it would be an idea that women must cover themselves up more than men and so that ends up creating this sort of social pressure for women as this becomes the norm of what you see, ” Mitchell said.

The harm is worsened by a lack of transparency. While in some cases Wood has been notified that her pictures were banned or limited in reach, she believes Instagram took other actions against her account without her knowing it. “I’ve had people say ‘I can’t tag you,’ or ‘I was searching for you to show my friend the other day and you’re not showing up,’” she said. “I feel invisible.”

Because she might be, said computer scientist Derczynski: “The people posting these images will never find out about it, which is just so deeply problematic.” he said. “They get a disadvantage forced upon them and they have no agency in this happening and they’re not informed that it’s happening either.”