The invention of AI 'gaydar' could be the start of something much worse

James Vincent

Two weeks ago, a pair of researchers from Stanford University made a startling claim. Using hundreds of thousands of images taken from a dating website, they said they had trained a facial recognition system that could identify whether someone was straight or gay just by looking at them. The work was first covered by *The Economist*, and other publications soon followed suit, with headlines like "New AI can guess whether you're gay or straight from a photograph" and "AI Can Tell If You're Gay From a Photo, and It's Terrifying."

As you might have guessed, it's not as straightforward as that. (And to be clear, based on this work alone, AI *can't* tell whether someone is gay or straight from a photo.) But the research captures common fears about artificial intelligence: that it will open up new avenues for surveillance and control, and could be particularly harmful for marginalized people. One of the paper's authors, Dr Michal Kosinski, says his intent is to sound the alarm about the dangers of AI, and warns that facial recognition will soon be able to identify not only someone's sexual orientation, but their political views, criminality, and even their IQ.

> *"some warn we're replacing the calipers of physiognomy with neural networks"*

With statements like these, some worry we're reviving an old belief with a bad history: that you can intuit character from appearance. This

pseudoscience, physiognomy, was fuel for the scientific racism of the 19th and 20th centuries, and gave moral cover to some of humanity's worst impulses: to demonize, condemn, and exterminate fellow humans. Critics of Kosinski's work accuse him of replacing the calipers of the 19th century with the neural networks of the 21st, while the professor himself says he is horrified by his findings, and happy to be proved wrong. "It's a controversial and upsetting subject, and it's also upsetting to us," he tells *The Verge*.

But is it possible that pseudoscience is sneaking back into the world, disguised in new garb thanks to AI? Some people say machines are simply able to read more about us than we can ourselves, but what if we're training them to carry out our prejudices, and, in doing so, giving new life to old ideas we rightly dismissed? How are we going to know the difference?

Can AI really spot sexual orientation?

First, we need to look at the study at the heart of the recent debate, written by Kosinski and his co-author Yilun Wang. Its results have been poorly reported, with a lot of the hype coming from misrepresentations of the system's accuracy. The paper states: "Given a single facial image, [the software] could correctly distinguish between gay and heterosexual men in 81 percent of cases, and in 71 percent of cases for women." These rates increase when the system is given five pictures of an individual: up to 91 percent for men, and 83 percent for women.

On the face of it, this sounds like "AI can tell if a man is gay or straight 81 percent of the time by looking at his photo." (Thus the headlines.) But that's not what the figures mean. The AI wasn't 81 percent correct when being shown random photos: it was tested on a *pair* of photos, one of a gay person and one of a straight person, and then asked which individual was *more likely* to be gay. It guessed right 81 percent of the time for men and 71 percent of the time for women, but the structure of the test means it

started with a baseline of 50 percent — that's what it'd get guessing at random. And although it was significantly better than that, the results aren't the same as saying it can identify *anyone*'s sexual orientation 81 percent of the time.

> *"'"People are scared of a situation where [you're in a crowd] and a computer identifies whether you're gay."'"*

As Philip Cohen, a sociologist at the University of Maryland who wrote a blog post critiquing the paper, told *The Verge*: "People are scared of a situation where you have a private life and your sexual orientation isn't known, and you go to an airport or a sporting event and a computer scans the crowd and identifies whether you're gay or straight. But there's just not much evidence this technology can do that."

Kosinski and Wang make this clear themselves toward the end of the paper when they test their system against 1,000 photographs instead of two. They ask the AI to pick out who is most likely to be gay in a dataset in which 7 percent of the photo subjects are gay, roughly reflecting the proportion of straight and gay men in the US population. When asked to select the 100 individuals most likely to be gay, the system gets only 47 out of 70 possible hits. The remaining 53 have been incorrectly identified. And when asked to identify a top 10, nine are right.

If you were a bad actor trying to use this system to identify gay people, you couldn't know for sure you were getting correct answers. Although, if you used it against a large enough dataset, you might get mostly correct guesses. Is this dangerous? If the system is being used to target gay people, then yes, of course. But the rest of the study suggests the program has even further limitations.

What can computers really see that humans can't?

It's also not clear what factors the facial recognition system is using to make its judgements. Kosinski and Wang's hypothesis is that it's primarily identifying structural differences: feminine features in the faces of gay men and masculine features in the faces of gay women. But it's possible that the AI is being confused by other stimuli — like facial expressions in the photos.

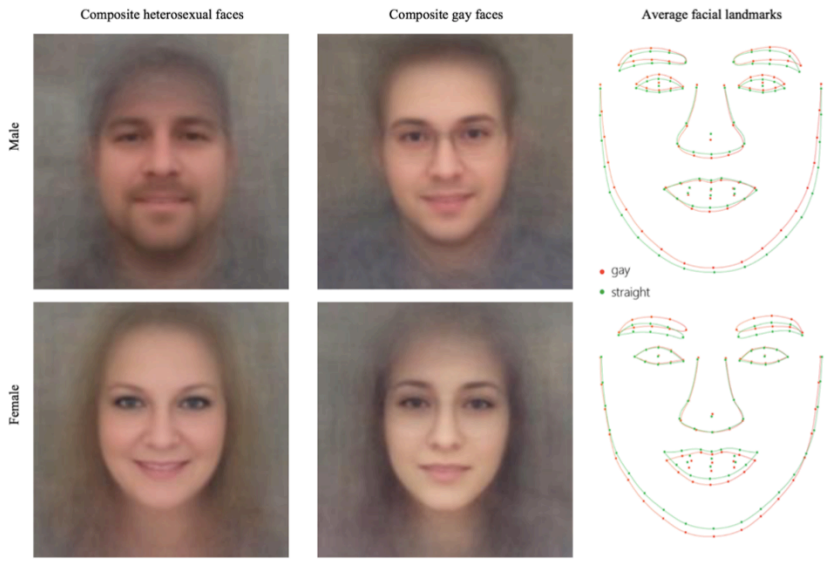
> *"The AI might be identifying stereotypes, not biological differences"*

This is particularly relevant because the images used in the study were taken from a dating website. As Greggor Mattson, a professor of sociology at Oberlin College, pointed out in a blog post, this means that the images themselves are biased, as they were selected specifically to attract someone of a certain sexual orientation. They almost certainly play up to our cultural expectations of how gay and straight people should look, and, to further narrow their applicability, all the subjects were white, with no inclusion of bisexual or self-identified trans individuals. If a straight male chooses the most stereotypically "manly" picture of himself for a dating site, it says more about what he thinks society wants from him than a link between the shape of his jaw and his sexual orientation.

To try and ensure their system was looking at facial structure only, Kosinski and Wang used software called VGG-Face, which encodes faces as strings of numbers and has been used for tasks like spotting celebrity lookalikes in paintings. This program, they write, allows them to "minimize the role [of] transient features" like lighting, pose, and facial expression.

But researcher Tom White, who works on AI facial system, says VGG-Face is actually very good at picking up on these elements. White pointed this out on Twitter, and explained to *The Verge* over email how he'd tested the

software and used it to successfully distinguish between faces with expressions like "neutral" and "happy," as well as poses and background color.



A figure from the paper showing the average faces of the participants, and the difference in facial structures that they identified between the two sets.
Image: Kosinski and Wang

Speaking to *The Verge*, Kosinski says he and Wang have been explicit that things like facial hair and makeup *could* be a factor in the AI's decision-making, but he maintains that facial structure is the most important. "If you look at the overall properties of VGG-Face, it tends to put very little weight on transient facial features," Kosinski says. "We also provide evidence that non-transient facial features seem to be predictive of sexual orientation."

The problem is, we can't know for sure. Kosinski and Wang haven't released the program they created or the pictures they used to train it. They do test their AI on other picture sources, to see if it's identifying some

factor common to all gay and straight, but these tests were limited and also drew from a biased dataset — Facebook profile pictures from men who liked pages such as "I love being Gay," and "Gay and Fabulous."

Do men in these groups serve as reasonable proxies for all gay men? Probably not, and Kosinski says it's possible his work is wrong. "Many more studies will need to be conducted to verify [this]," he says. But it's tricky to say how one could completely eliminate selection bias to perform a conclusive test. Kosinski tells *The Verge,* "You don't need to understand how the model works to test whether it's correct or not." However, it's the acceptance of the opacity of algorithms that makes this sort of research so fraught.

If AI can't show its working, can we trust it?

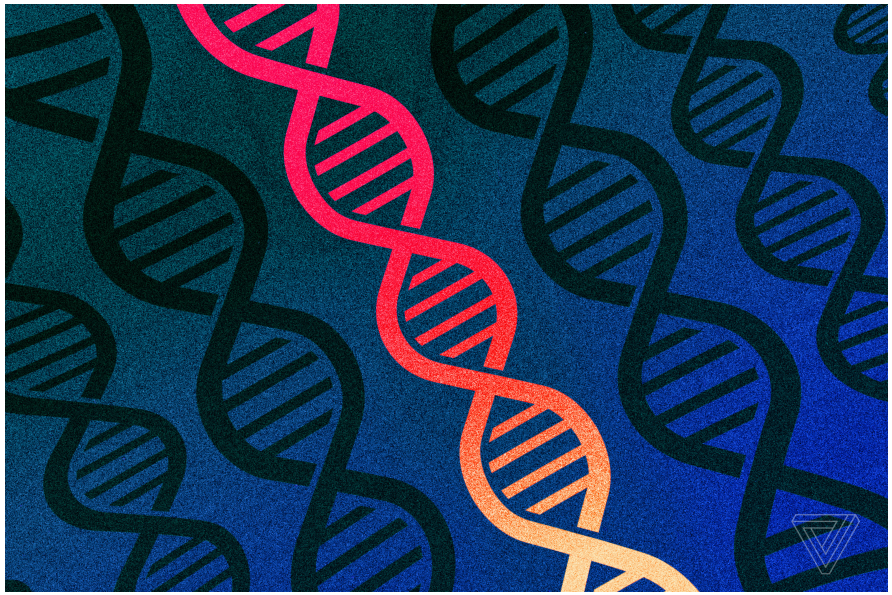
AI researchers can't fully explain why their machines do the things they do. It's a challenge that runs through the entire field, and is sometimes referred to as the "black box" problem. Because of the methods used to train AI, these programs can't show their work in the same way normal software does, although researchers are working to amend this.

In the meantime, it leads to all sorts of problems. A common one is that sexist and racist biases are captured from humans in the training data and reproduced by the AI. In the case of Kosinski and Wang's work, the "black box" allows them to make a particular scientific leap of faith. Because they're confident their system is primarily analyzing facial structures, they say their research shows that facial structures predict sexual orientation. ("Study 1a showed that facial features extracted by a [neural network] can be used to accurately identify the sexual orientation of both men and women.")

> *""Biology's a little bit more nuanced than we often give it credit for.""*

Experts say this is a misleading claim that isn't supported by the latest science. There may be a common cause for face shape and sexual orientation — the most probable cause is the balance of hormones in the womb — but that doesn't mean face shape reliably *predicts* sexual orientation, says Qazi Rahman, an academic at King's College London who studies the biology of sexual orientation. "Biology's a little bit more nuanced than we often give it credit for," he tells *The Verge*. "The issue here is the strength of the association."

The idea that sexual orientation comes primarily from biology is itself controversial. Rahman, who believes that sexual orientation is mostly biological, praises Kosinski and Wang's work. "It's not junk science," he says. "More like science someone doesn't like." But when it comes to *predicting* sexual orientation, he says there's a whole package of "atypical gender behavior" that needs to be considered. "The issue for me is more that [the study] misses the point, and that's behavior."



Is there a gay gene? Or is sexuality equally shaped by society and culture?

Reducing the question of sexual orientation to a single, measurable factor in the body has a long and often inglorious history. As Matton writes in his blog post, approaches have ranged from "19th century measurements of lesbians' clitorises and homosexual men's hips, to late 20th century claims to have discovered 'gay genes,' 'gay brains,' 'gay ring fingers,' 'lesbian ears,' and 'gay scalp hair.'" The impact of this work is mixed, but at its worst it's a tool of oppression: it gives people who want to dehumanize and persecute sexual minorities a "scientific" pretext.

Jenny Davis, a lecturer in sociology at the Australian National University, describes it as a form of biological essentialism. This is the belief that things like sexual orientation are rooted in the body. This approach, she says, is double-edged. On the one hand, it "does a useful political thing: detaching blame from same-sex desire. But on the other hand, it reinforces the devalued position of that kind of desire," setting up hetrosexuality as the norm and framing homosexuality as "less valuable … a sort of illness."

And it's when we consider Kosinski and Wang's research in this context that AI-powered facial recognition takes on an even darker aspect — namely, say some critics, as part of a trend to the return of physiognomy, powered by AI.

Your character, as plain as the nose on your face

For centuries, people have believed that the face held the key to the character. The notion has its roots in ancient Greece, but was particularly influential in the 19th century. Proponents of physiognomy suggested that by measuring things like the angle of someone's forehead or the shape of their nose, they could determine if a person was honest or a criminal. Last year in China, AI researchers claimed they could do the same thing using facial recognition.

Their research, published as "Automated Inference on Criminality Using Face Images," caused a minor uproar in the AI community. Scientists

pointed out flaws in the study, and concluded that that work was replicating human prejudices about what constitutes a "mean" or a "nice" face. In a widely shared rebuttal titled "Physiognomy's New Clothes," Google researcher Blaise Agüera y Arcas and two co-authors wrote that we should expect "more research in the coming years that has similar … false claims to scientific objectivity in order to 'launder' human prejudice and discrimination." (Google declined to make Agüera y Arcas available to comment on this report.)

An illustration of physiognomy from Giambattista della Porta's *De humana physiognomonia*

Kosinski and Wang's paper clearly acknowledges the dangers of physiognomy, noting that the practice "is now universally, and rightly, rejected as a mix of superstition and racism disguised as science." But, they continue, just because a subject is "taboo," doesn't mean it has no basis in truth. They say that because humans are able to read characteristics like personality in other people's faces with "low accuracy," machines should be able to do the same but more accurately.

Kosinski says his research isn't physiognomy because it's using rigorous scientific methods, and his paper cites a number of studies showing that we can deduce (with varying accuracy) traits about people by looking at them. "I was educated and made to believe that it's absolutely impossible that the face contains any information about your intimate traits, because physiognomy and phrenology were just pseudosciences," he says. "But the fact that they were claiming things without any basis in fact, that they were making stuff up, doesn't mean that this stuff is not real." He agrees that physiognomy is not science, but says there may be truth in its basic concepts that computers can reveal.

"AI's intelligence isn't artificial: it's human"

For Davis, this sort of attitude comes from a widespread and mistaken belief in the neutrality and objectivity of AI. "Artificial intelligence is not in fact artificial," she tells *The Verge*. "Machines learn like humans learn. We're taught through culture and absorb the norms of social structure, and so does artificial intelligence. So it will re-create, amplify, and continue on the trajectories we've taught it, which are always going to reflect existing cultural norms."

We've already created sexist and racist algorithms, and these sorts of cultural biases and physiognomy are really just two sides of the same coin: both rely on bad evidence to judge others. The work by the Chinese researchers is an extreme example, but it's certainly not the only one. There's at least one startup already active that claims it can spot terrorists and pedophiles using face recognition, and there are many others offering to analyze "emotional intelligence" and conduct AI-powered surveillance.

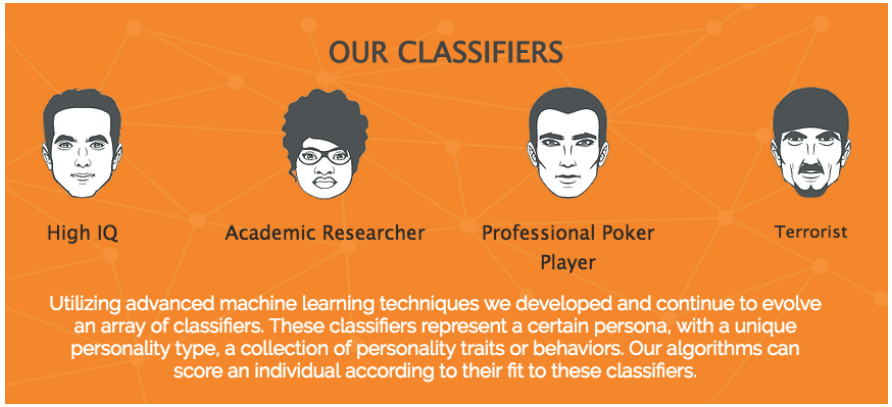
Facing up to what's coming

But to return to the questions implied by those alarming headlines about Kosinski and Wang's paper: is AI going to be used to persecute sexual

minorities?

This system? No. A different one? Maybe.

Kosinski and Wang's work is not invalid, but its results need serious qualifications and further testing. Without that, all we know about their system is that it can spot with some reliability the difference between self-identified gay and straight white people on one particular dating site. We don't know that it's spotted a biological difference common to all gay and straight people; we don't know if it would work with a wider set of photos; and the work doesn't show that sexual orientation can be deduced with nothing more than, say, a measurement of the jaw. It's not decoded human sexuality any more than AI chatbots have decoded the art of a good conversation. (Nor do its authors make such a claim.)



Startup Faception claims it can identify how likely people are to be terrorists just by looking at their face.
Image: Faception

The research was published to warn people, say Kosinski, but he admits it's an "unavoidable paradox" that to do so you have to explain how you did what you did. All the tools used in the paper are available for anyone to find and put together themselves. Writing at the deep learning education site Fast.ai, researcher Jeremy Howard concludes: "It is probably

reasonably [sic] to assume that many organizations have already completed similar projects, but without publishing them in the academic literature."

We've already mentioned startups working on this tech, and it's not hard to find government regimes that would use it. In countries like Iran and Saudi Arabia homosexuality is still punishable by death; in many other countries, being gay means being hounded, imprisoned, and tortured by the state. Recent reports have spoken of the opening of concentration camps for gay men in the Chechen Republic, so what if someone there decides to make their own AI gaydar, and scan profile pictures from Russian social media?

Here, it becomes clear that the accuracy of systems like Kosinski and Wang's isn't really the point. If people *believe* AI can be used to determine sexual preference, they will use it. With that in mind, it's more important than ever that we understand the limitations of artificial intelligence, to try and neutralize dangers before they start impacting people. Before we teach machines our prejudices, we need to first teach ourselves.