Last updated on Mar 18, 2022

Author Note

Deep neural networks are more accurate than humans at detecting sexual orientation from facial images

Journal of Personality and Social Psychology
by Michal Kosinski and Yilun Wang
Correspondence to michalk@stanford.edu

Your FREE copy of the article is here. Anonymized data used in the study is here.

The study has been successfully replicated by two research teams using independently collected data:

- Leuner (2019), arXiv.org
- Wang (2022), JPSP.

Press articles whose authors have actually bothered to read the paper:

1. The Economist
2. Vice.com
3. The Legal Implications of Sexual Orientation-Detecting Facial Recognition Technology (thesis by P. Wilkinson; awarded with Dukeminier Award)
4. Psychology Today

This document has three main sections:

1. Summary of the findings
2. You must be wrong—this is pseudoscience! (common points of criticism of this paper)
3. Our response to an irresponsible press release by GLAAD and HRC or a much better response by *LGBTQ Nation*

Summary of the findings

We did not build a privacy-invading tool. We studied existing facial recognition technologies, already widely used by companies and governments,[1] to see whether they can detect sexual orientation more accurately than humans.

We were terrified to find that they did. This presents severe risks to the privacy of LGBTQ people.

Our work is limited in many ways: We only looked at white people who self-reported to be gay or straight. We discuss those limitations at length in our paper and below. Those limitations do not, however, invalidate the findings or the core message of the study: that widely used technologies present a risk to the privacy of LGBTQ individuals.

[1] https://patents.google.com/patent/WO2014068567A1/en, https://patents.google.com/patent/US20160019411A1/en. Also see this patent by Alexander Todorov https://patents.google.com/patent/US20210089759A1

Our work is not the first one to show that sexual orientation can be detected from the human face. It is well established that humans can, with some accuracy, detect sexual orientation from a still facial image. It has also been previously shown that computers outcompete humans at many visual tasks, including detecting sexual orientation.

We invite you to consider the evidence before dismissing it.

Note: This study is not about sexual orientation or its origins, despite some people trying to interpret it in this way.

This study was peer-reviewed and published in the *Journal of Personality and Social Psychology*, the leading academic journal in psychology. In addition, before it was sent for a formal peer review, the manuscript was reviewed by over a dozen experts in sexuality, psychology, and artificial intelligence. Stanford's Internal Review Board has approved the research.

What are the main findings?

Across seven studies, we show that a computer algorithm can accurately detect sexual orientation from people's faces. When presented with a pair of participants, one gay and one straight, the algorithm could correctly distinguish between them 91% of the time for men and 83% for women.

This is comparable with the accuracy of mammograms (85%) or modern diagnostic tools for Parkinson's disease (90%). (Also see this section.)

We trained the algorithm on a sample of over 35,000 facial images of self-identified gay and straight individuals obtained from a publicly available database. The accuracy was verified on a subset of images the algorithm had not seen before. We ensured that the predictions were not affected by differences in age and ethnicity.

We also tested the algorithm on an independent sample of Facebook profile pictures and achieved similar results.

In contrast, human judges were not much more accurate than random guesses. This is yet another example of artificial intelligence (AI) outperforming humans.

What this study is not about

As the title indicates, our study aims to show that "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images."

This study is *not* about sexual orientation or its origins. In the process of studying the features employed by the classifier to distinguish between gay and straight faces, we noted that the former tend to be gender atypical (or, in other words, that faces of gay men tended to be slightly more feminine than those of straight men, while faces of lesbians tended to be slightly more masculine than those of straight women). This is consistent with one of the most widely accepted theories explaining the origins of sexual orientation (prenatal hormone theory).[2]

[2] See this comprehensive review if you are interested in the origins of sexual orientation.

Our findings' consistency with a widely accepted theory of the origins of sexual orientation does not prove that theory. Instead, it provides additional support for the validity of our findings.

This study was neither designed nor intended to explore prenatal hormone theory or the origins of sexual orientation.

So, basically, you built a gaydar?!

No, we did not. We showed that a widely used facial recognition technology inadvertently exposes people's sexual orientation.

Facial recognition software examines a face and turns it into a numerical representation or a bunch of numbers. Those numbers are typically compared across different facial images to see which one contains the same face.

We noticed that these numbers differ in predictable ways between gay and straight faces, allowing for the detection of sexual orientation and invasion of people's privacy. In other words, the information about your sexual orientation is already built into the results produced by facial recognition software. It is not explicit—the software does not say, "Michal is gay," but, as we show in our paper, this intimate information can be relatively easily extracted.

This might be news to you and the leaders of the LBTQ community, but technology companies and governmental institutions are all well aware of the fact that sensitive traits can be easily extracted from the numbers produced by facial recognition software.[3]

You must be wrong—this is pseudoscience!

See Section "This must be wrong!" below.

What are the implications of these findings for privacy?

The fact that algorithms can predict sexual orientation from human faces has serious privacy implications. The ability to control when and to whom to reveal one's sexual orientation is crucial not only for one's well-being but also for one's safety.

In some cases, losing the privacy of one's sexual orientation can be life-threatening. The members of the LGBTQ community still suffer physical and psychological abuse at the hands of governments, neighbors, and even their own families. The laws in many countries criminalize same-gender sexual behavior, and in some places, it is punishable by death.

The growing digitalization of our lives and rapid progress in AI continue to erode our privacy. As this and our previous studies illustrate, willingly shared digital footprints can be used to reveal intimate traits. Michal's 2013 paper warned that algorithms can accurately reveal people's intimate traits from their Facebook "Likes," which were, at that time, publicly visible by default. His 2015 paper showed that algorithms can predict one's behavior more accurately than a friend or spouse.

[3] https://patents.google.com/patent/WO2014068567A1/en

Those papers raised enough alarm to effect policy change. For example, within a few weeks of Michal's 2013 paper's publication, Facebook switched off Likes's public visibility. Michal's work was also discussed by lawmakers in the U.S. and the EU in the context of the new privacy legislation.

Our work is aimed at informing privacy policies and technology.

Unfortunately, even the best privacy-protecting laws and technologies are unlikely to guarantee privacy at all times and for everybody. One can regulate Facebook, Google, and U.S. intelligence agencies, but one cannot expect that hackers, startups with little to lose, or foreign intelligence services will also follow the rules. The digital environment is challenging to police; data can be easily moved across borders, stolen, or recorded without users' consent. Also, most people want some of their social media posts, blogs, or profiles to be public. Few would be willing to cover their faces while interacting with others. This and other studies show that this is enough to invade their privacy.

Consequently, the safety of gay people and other minorities hinges not on their right to privacy (which can be maliciously invaded) but on the protection of their human rights and tolerance of societies and governments. For the post-privacy world to be safer and hospitable, it must be inhabited by well-educated people who are radically intolerant of intolerance.

Even if we are wrong and privacy could be preserved, a tolerant world—where losing your privacy does not put you at risk—would be a much better place.

Were the authors concerned about publishing these results?

We were disturbed by these results and spent much time considering whether they should be made public. We did not want to enable the very risks that we were warning against.

Governments and corporations are already using tools aimed at revealing intimate traits from faces.[4] Facial images of billions of people are stockpiled in digital and traditional archives, including dating platforms, photo-sharing websites, and government databases. Profile pictures on Facebook, LinkedIn, and Google Plus are public by default. CCTV cameras and smartphones can be used to take photos of others' faces without their permission.

We felt that there is an urgent need to make policymakers and LGBTQ communities aware of the risks they face. Tech companies and government agencies are well aware of the potential of computer vision algorithm tools. We believe that people deserve to know about these risks and have the opportunity to take preventive measures.

We ensured that our work did not offer any advantage to those who may want to invade others' privacy. We used widely available off-the-shelf tools, publicly available data, and standard methods well-known to computer vision practitioners. We did not create a privacy-invading tool but rather showed that basic and widely used methods pose serious privacy threats.

What facial features were employed by the algorithm to detect sexual orientation?

The average faces most likely to belong to gay men (see Figure 1) were more feminine, while the faces most likely to belong to lesbians were more masculine. Typically, men have larger jaws, shorter noses,

[4] See, for instance, these *Wall Street Journal* and *Business Insider* articles, or this patent.

and smaller foreheads. Gay men, however, tended to have narrower jaws, longer noses, larger foreheads, and less facial hair. Conversely, lesbians tended to have more masculine faces (larger jaws and smaller foreheads) than heterosexual women.

The gender atypicality of gay faces extended beyond morphology. Lesbians tended to use less eye makeup, had darker hair, and wore less revealing clothes (note the higher neckline)—indicating less feminine grooming and style. Furthermore, although women tended to smile more in general, lesbians smiled less than their heterosexual counterparts.

Additionally, consistent with the association between baseball caps and masculinity in American culture, heterosexual men and lesbians tended to wear baseball caps (see the shadow on their foreheads in Figure 1; a manual inspection of individual images also confirmed this).

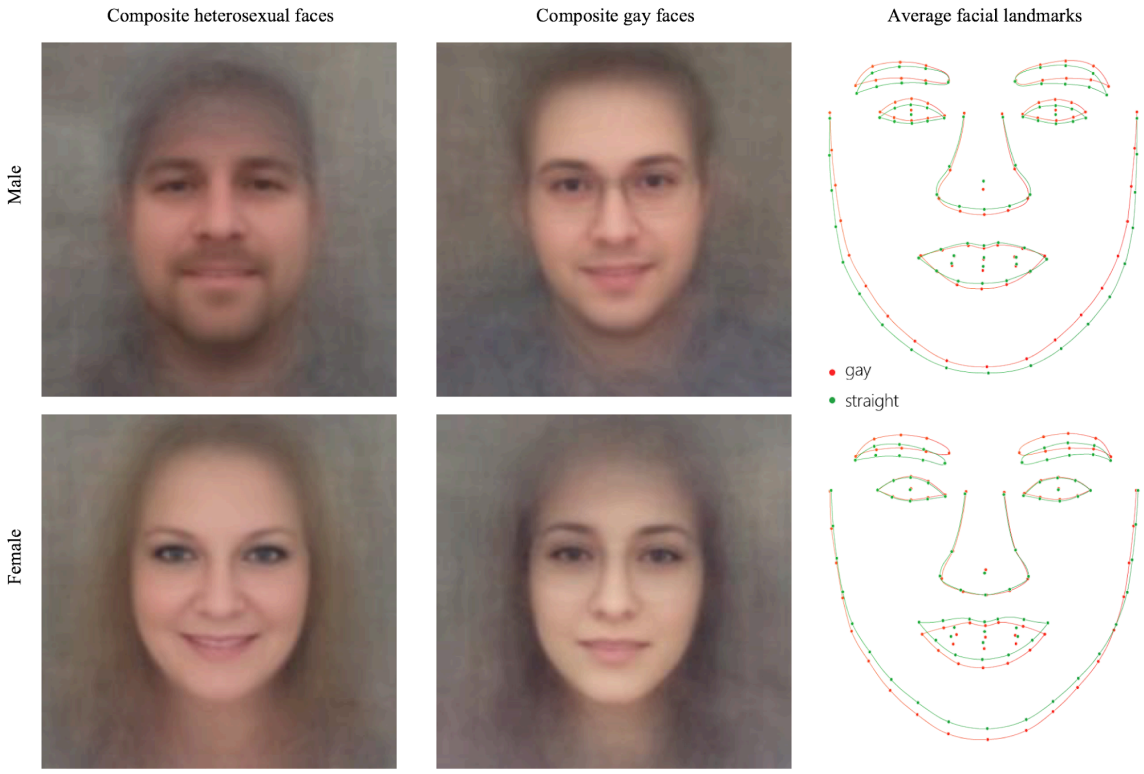


Figure 1. Composite faces and average face outlines produced by averaging faces/outlines classified as most likely to be gay or straight.

Gender atypicality of the faces of gay men and women is consistent with a large number of previous studies. Previous findings (see the section on adult gender nonconformity in this review) showed gender atypicality in occupations, hobbies, patterns of movement (i.e., gestures and walking), speech (i.e., articulation), physical presentation (i.e., clothing choices and hairstyles), and facial appearance. Perhaps

the most widely accepted theory used to account for gender atypicality is the prenatal hormone theory (PHT) of sexual orientation.[5]

The fact that the facial features predictive of sexual orientation are consistent with the well-established theory yields support for the validity of the classifier.

How does one interpret the accuracy of the algorithm?

Our work is intended as a warning that predictions of this kind can be made with worrying accuracy rather than an attempt to estimate the maximum accuracy of such predictions. We used basic tools and images in low resolution. Those deploying such methods in practice use more sensitive DNN models and devices.

How accurate was the classifier in our study? Very accurate: It is comparable with the accuracy of mammograms (85%) or modern diagnostic tools for Parkinson's disease (90%).

Interpreting classification accuracy is not trivial and is often counterintuitive. Let us illustrate it with a few examples.

Imagine a group of 1,000 men, including 70 gay men, whose faces were assessed by the classifier with an accuracy of AUC=.91 (comparable to the one our study achieved for males with five images per person).

The classifier does not tell you which person is gay but labels each person with a probability of being gay. The non-trivial decision that you need to make now is to decide where to set the cut-off point—the probability above which you classify someone as gay.

If you wanted to select a small sample of gay men and make as few mistakes as possible - label as gay only a few cases with top probabilities. You will get high precision (e.g., the fraction of gay people among those classified as gay will be high) but low recall (e.g., you will “miss” many gay men). If you prefer to cast a wider net, you will “catch” more gay men but also erroneously label more straight men as gay (so-called “false positives”). In other words, aiming for high precision reduces recall and vice versa.

Back to the group of 1,000 men, including 70 gay men: If one selected 100 random males from this sample, only seven are expected to be gay; a random draw offers a precision of 7% (seven out of 100 selected men were gay).

Let’s turn to the classifier. Among the 100 individuals with the highest probability of being gay according to the classifier, 47 were gay (precision: 47/100 = 47%). In other words, the classifier improved precision nearly seven-fold over a random selection. There were also 53 “false positives”—straight men classified as gay. Note, however, that as there are only 70 gay men in the examined population, there would be 30 “false positives” even if the classifier was perfect.

---

[5] According to the PHT, same-gender sexual orientation stems from the underexposure of male fetuses and overexposure of female fetuses to androgens that are responsible for sexual differentiation. As the same androgens are responsible for the sexual dimorphism of the face and the brain, the PHT predicts that gay people will tend to have gender-atypical facial morphology and gender-atypical preferences (including gender-atypical sexual preferences).

The number of false positives could be decreased, and the precision increased by narrowing the targeted subsample. Among 30 males with the highest probability of being gay, 23 were gay, an eleven-fold improvement in precision over a random draw. (Only 2.1 men would be expected to be gay in a random subset of 30 males.) Finally, among the top 10 individuals with the highest probability of being gay, nine were indeed gay: a 13-fold improvement in precision over a random draw.

What are the potential mechanisms linking intimate traits with facial features?

There are three types of such mechanisms. First, the character can influence one's facial appearance. For example, women who scored high on extraversion early in life tend to become more attractive with age.

Second, facial appearance can alter one's character. Good-looking people, for example, receive more positive social feedback and thus tend to become even more extroverted.

Third, many factors affect both facial appearance and one's traits. Those include prenatal and postnatal hormonal levels, developmental history, environmental factors, and genes. Testosterone, for instance, significantly affects both behavior (e.g., dominance) and facial appearance (e.g., facial width and facial hair).

How did this research come about?

Our project came up by accident. We were browsing through profile images submitted by volunteers (along with their personality scores) to see if we could identify correlates of personality in the background. We were surprised to notice that we seem to be able to infer personality from the face itself. We could not believe it and started investigating this issue more closely. Soon, we realized that a simple facial recognition algorithm could do a much better job than us (and other human judges) and that it could accurately infer intimate traits ranging from political views to sexual orientation. We found it very disturbing.

***

You must be wrong—this is pseudoscience!

We get a lot of feedback along these lines. And quite frankly, we would be delighted if our results were wrong. Humanity would have one less problem, and we could get back to writing self-help bestsellers about how power posing makes you bolder, smiling makes you happier, and seeing pictures of eyes makes you more honest.[6]

This study, as virtually any other study, has many limitations. We discuss some of them below. Also, see this great article from the *LGBTQ Nation*.

[6] Unfortunately, these findings do not seem to replicate.

This must be wrong; classifying everyone as straight would give higher accuracy!”

That’s correct; the most accurate model had an accuracy of 91%. Yet, classifying everyone as “straight” would offer an accuracy of about 93% (about 7% of people are gay, so one would be wrong only in 7% of cases).

This issue nicely illustrates how counterintuitive the interpretation of classification accuracy could be. Any classifier with an accuracy lower or equal to the prevalence of the target category has an accuracy lower than what could be achieved by classifying all cases as belonging to the target category. This issue is typically discussed early in stat 101 courses. In a classic example, a 99% accurate test for a rare disease (1% prevalence in the population) performs no better than classifying everyone as healthy. The students marvel at the fact that simply classifying everyone as “healthy” yields the same accuracy (99%) as the nominal accuracy of the classifier (also 99%) or that about half of those classified as “sick” are, in fact, “healthy” (precision of 50%).

As this example shows, classification accuracy is a very poor measure of the benefits of a classifier applied to populations with unequal base rates. Identifying a subset of cases with a 50% prevalence of an otherwise 1-in-100 disease offers a stunning 50-fold improvement over a dumb classifier in both precision and recall (while remaining no better in terms of raw accuracy). Students quickly discover that the point of the classification is not beating the dumb classifier but maximizing precision and/or recall (and solving the trade-off between them) and that there are many other coefficients, such as F1 or AUC, that are much better at expressing the performance of such classifiers.

“This must be wrong; the study is based on only white people!”

Despite our attempts to obtain a more diverse sample, we were limited to studying white participants from the U.S.

This does not invalidate our results showing that you can distinguish between white gay and straight individuals. Naturally, while it is possible that the same does not apply to other ethnicities, our findings suggest that this, unfortunately, is likely. Similar biological, developmental, and cultural factors—which are responsible for differences between gay and straight white people—are likely to affect people of other ethnicities as well.

“This must be wrong; bisexual people were excluded from the analysis.”

That’s true; we did not check if one can predict whether someone is bisexual from their face.

This does not invalidate the results in any way. We still show that you can distinguish between gay and straight individuals. It is possible that some of the users categorized as heterosexual or gay were, in fact, bisexual. Correcting such errors, however, would likely boost the accuracy of the classifiers examined here.

Importantly, excluding bisexual or non-binary people does not mean that we are denying their existence.

"This must be wrong; you used a dating website sample of openly gay/straight people!"

That is a serious limitation, which we discuss at length in our paper. It is reasonable to expect that the images obtained from a dating website could be especially revealing of sexual orientation; this, however, did not seem to be the case.

First, we tested our classifier on an external sample of Facebook photos. It achieved comparable accuracy as on the dating website sample, suggesting that the images from the dating website were not more revealing than Facebook profile pictures.

Second, we asked humans to judge the sexual orientation of these faces. Human accuracy was no better than in past studies, where humans judged sexual orientation from carefully standardized images taken in the lab. This shows that the images used here were not especially revealing of sexual orientation—at least, not to humans.

Finally, the deep neural network used here was specifically trained to focus on fixed facial features that cannot be easily altered, such as the shape of facial elements. This helped reduce the risk of the classifier discovering some superficial and non-face-related differences between the facial images of gay and straight people used in this study.

"This must be wrong; it is widely known that there are no links between faces and character traits!"

Unfortunately, this belief is not supported by evidence. Quite the contrary, many studies have shown that people can determine others' political views, personality, sexual orientation, honesty, and many other traits from their faces.[7] Also, humans' low accuracy when judging such traits does not necessarily mean that those traits are not prominently displayed on the face. Instead, people may have a limited ability to detect or interpret the cues—a limitation that does not necessarily apply to algorithms.

"This must be wrong; I read on *Wired* that physiognomists believed that criminals were part ape!"

Well, it seems that physiognomists were at least partially correct, as we are all 100% ape.

Without a doubt, physiognomy was based on unscientific studies, superstition, anecdotal evidence, and racist pseudo-theories. However, the fact that its claims were unsupported does not automatically mean that they are all wrong. Some of the physiognomists' claims may have been correct, perhaps by mere accident.

Physiognomists were clearly wrong when claiming they could *accurately* judge characters based on facial appearance. Modern scientific studies have shown that we are *not* very accurate at this task. The same studies, however, consistently show that we are better than chance, revealing that faces contain at least some information about one's character.

Thus, physiognomists' main claim—that the character is, to some extent, displayed on one's face—is supported by modern science.

---

[7] See the many references in our paper, and read this.

“This must be wrong; your classifier is certainly picking on something unrelated to the face when making predictions!”

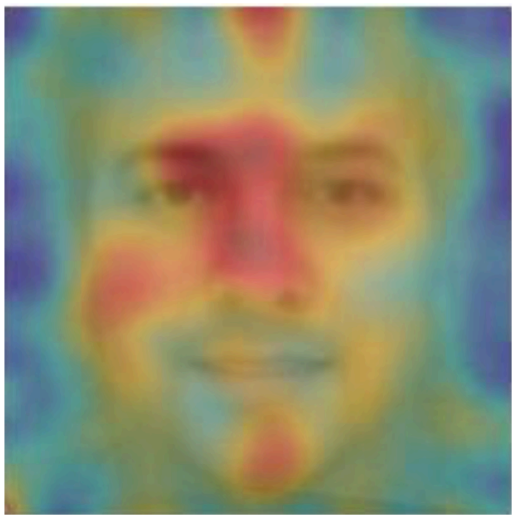
We thought about this a lot, and we hope that future studies will help to prove or disprove the predictability of sexual orientation from human faces. However, we have put much effort into controlling this issue.

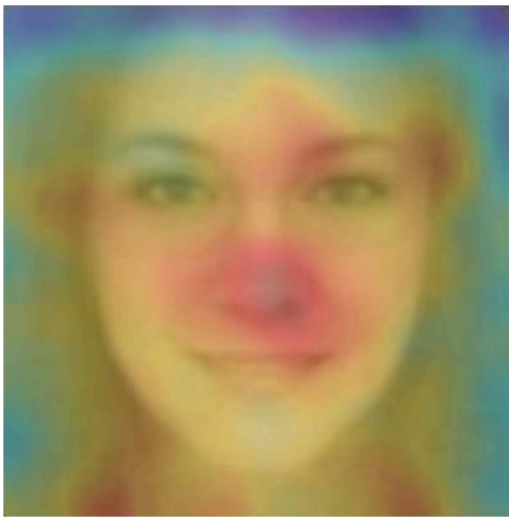
First, our models were specifically trained to focus on fixed facial features that cannot be easily altered, such as the shape of facial elements. The deep neural network used here was trained for a completely different task: recognizing the same person across images. This helped us to reduce the risk of the classifier discovering some superficial and not-face-related differences between the facial images of gay and straight people used in this study.

Second, we validated the findings on an external sample.

Third, we investigated which elements of the facial image were predictive of sexual orientation to ascertain that it was, in fact, facial features (and not other factors). As you can read in the paper, even if all of the visual information is removed, the classifier can still be quite accurate based on merely the outline of the face.

Fourth, we revealed only the facial area to the classifier and removed the background of the images. We also checked that the classifier focused on facial features, not the background, while making the prediction. Heatmaps below (taken from Figure 3 in the paper) clearly show that the classifier focused on facial areas (red) and ignored the background (blue).





Finally, and perhaps most importantly, the differences between gay and straight faces picked up by the classifier are consistent with and predicted by the prenatal hormone theory—the most widely accepted theory explaining the origins of sexual orientation.

“This must be wrong; your findings show that gay people tend to be gender atypical—and I know many gender-typical gay men and women!”

We also know many very masculine gay men and very feminine gay women. We also know many very old men, which does not invalidate the statement that women tend to live longer. The fact that the faces of gay men are more feminine on average (as they tended to be in our study) does not imply that *all* gay men are more feminine than *all* heterosexual men or that there are no gay men with very masculine faces (and vice versa for lesbians).

The differences in femininity/masculinity observed in this study were subtle and spread across many facial features: enough to be apparent to a sensitive algorithm but imperceptible to humans.

Also, please read “The causes of sexual orientation: An interim summary” in this review article.

“This must be wrong; many of your participants must have lied about their sexual orientation!”

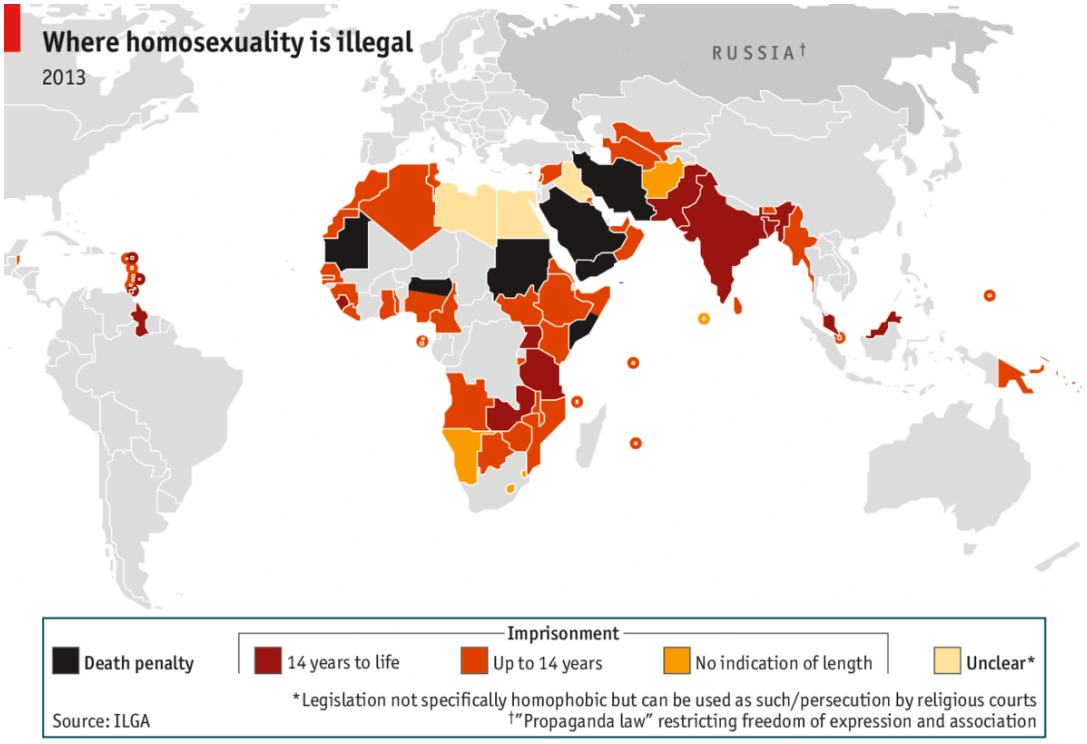
It is certainly possible that some of the participants who told us that they were straight were, in fact, gay (or vice versa). We believe, however, that people voluntarily posting and seeking partners on dating websites have little incentive to lie about their sexual orientation.

Also, if some of our participants were mislabeled, correcting such errors would most likely further *increase* the classification accuracy.

“This must be wrong; the only reason that it works is because gay men have better style or take better pictures!”

We could be easily convinced that gay men (*our* gay friends, for sure!) have better hairstyles and facial hairstyles and take better pictures. As we discuss in our paper, gay and straight faces do differ in terms of grooming. However, they also seem to differ in terms of morphology. Facial contour alone provided an accuracy of over 70% among men and above 60% among women.

Even if the differences between gay and straight faces are exclusively due to differences in grooming, lifestyle, or fashion (i.e., nurture), this does not necessarily reduce the privacy threats faced by gay men and women. Many grooming or fashion choices are made unconsciously; removing other revealing features might require changing someone’s lifestyle.



Economist.com/graphicdetail