“The study has been approved by the IRB”: Gayface AI, research hype and the pervasive data ethics gap

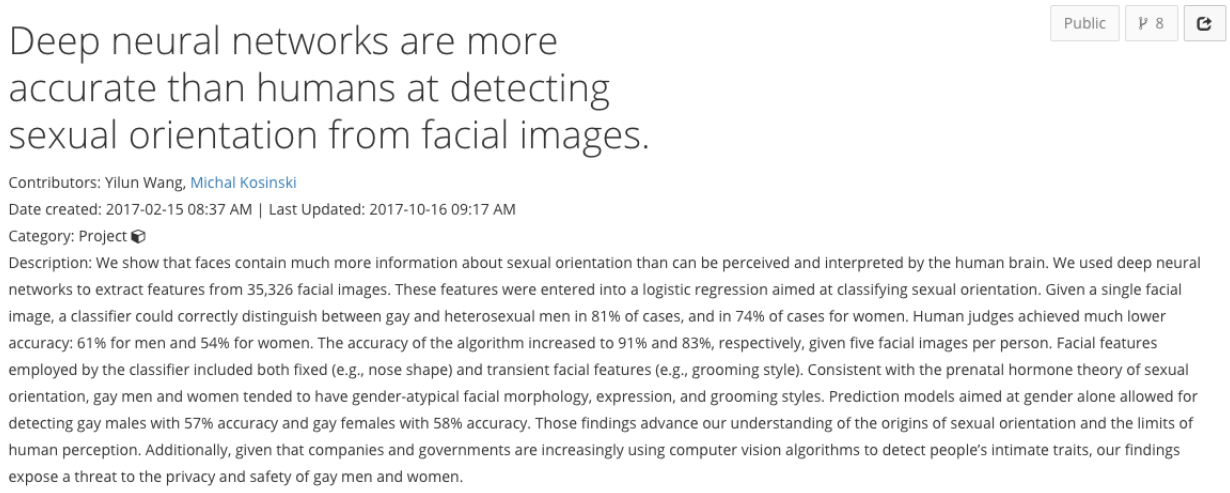
medium.com/@pervade_team/the-study-has-been-approved-by-the-irb-gayface-ai-research-hype-and-the-pervasive-data-ethics-3b36c5a53eec

November 30, 2017

Just as it has changed the methods of science and engineering, the tools of large scale data analytics have caused major shifts in how we judge the ethical consequences of scientific research. And our current methods are not keeping up. Historically, research ethics has been animated by a core set of questions, such as how do you decide if a scientific experiment is justified given the potential risks and benefits to the people being studied, or to society at large? How do you track who has to bear those risks and who gets the benefits?

Now we are faced with the problem of what happens to the methods we have developed for answering those questions when the number of people affected by a study jumps by multiple orders of magnitude over the historical norm.

When we talk about research ethics, we’re ultimately addressing two distinct questions: 1) what are the correct/ideal norms for judging the experimental methods, and 2) what determination is made by the institutions (such as IRBs) tasked with regulating researchers and protecting research subjects from harms. It turns out that too often when we discuss data science the answers to those two questions diverge, resulting in a situation where we cannot effectively track and mitigate the ethical and social consequences of the research. Thus when data scientists cite IRB review as a certificate of ethical methods, they are misreading the purpose and scope of IRBs.



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

Public 8

Contributors: Yilun Wang, Michal Kosinski

Date created: 2017-02-15 08:37 AM | Last Updated: 2017-10-16 09:17 AM

Category: Project

Description: We show that faces contain much more information about sexual orientation than can be perceived and interpreted by the human brain. We used deep neural networks to extract features from 35,326 facial images. These features were entered into a logistic regression aimed at classifying sexual orientation. Given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 81% of cases, and in 74% of cases for women. Human judges achieved much lower accuracy: 61% for men and 54% for women. The accuracy of the algorithm increased to 91% and 83%, respectively, given five facial images per person. Facial features employed by the classifier included both fixed (e.g., nose shape) and transient facial features (e.g., grooming style). Consistent with the prenatal hormone theory of sexual orientation, gay men and women tended to have gender-atypical facial morphology, expression, and grooming styles. Prediction models aimed at gender alone allowed for detecting gay males with 57% accuracy and gay females with 58% accuracy. Those findings advance our understanding of the origins of sexual orientation and the limits of human perception. Additionally, given that companies and governments are increasingly using computer vision algorithms to detect people's intimate traits, our findings expose a threat to the privacy and safety of gay men and women.

The recent controversy over a pre-print paper (in press at the journal *Personality and Social Psychology*)by Michal Kosinski and Yilun Wang of Stanford illustrates a number of the new potential risks of social and behavioral data science. What makes this study so interesting from a research ethics perspective is that the scientists and their critics alike agree it would be highly troublesome if their results were applied in the wild; where they disagree is whether the study is justifiable under those conditions. This study falls precisely in the gap opened by data analytics between what we would hope are ideal ethical conditions for scientific experiments and the decisions made by institutions tasked with protecting research subjects.

In their paper, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images," Kosinski and Wang describe a machine learning study in which they trained a deep neural network to sort human faces according to sexual preference at a greater accuracy than their control group of human sorters.* Put colloquially, they built an early prototype of an artificial intelligence "gaydar" using off-the-shelf machine-learning components and "publicly-available" data. Many accounts have referred to it as "gayface AI," using the slang term for an exaggerated stereotypically gay male facial expression.

> What makes this study so interesting from a research ethics perspective is that the scientists and their critics alike agree it would be highly troublesome if their results were applied in the wild; where they disagree is whether the study is justifiable under those conditions.

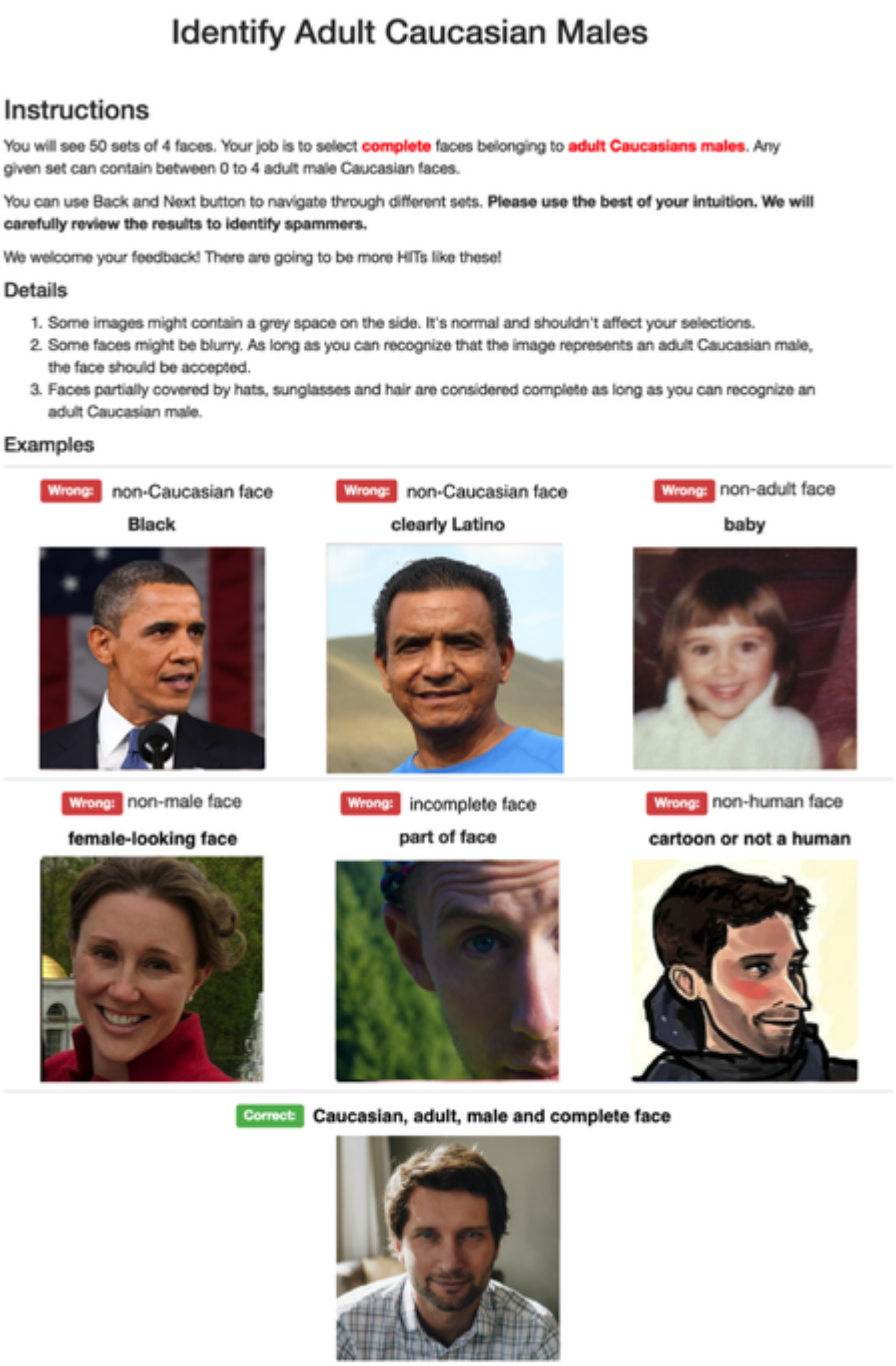
The press ate this up as a it deftly hits on a number of cultural hot spots: social media privacy, artificial intelligence, and sexual difference. Here is a sampling of coverage from mid-September 2017 when the paper was first noticed by the press:

Skeptics of Big Tech and machine learning were alongside LGBT advocates in expressing strong reservations about this study. There are a number of solid methodological critiques of Kosinski and Wang's study, most notably from Greggor Mattson from Oberlin College and Carl Bergstrom and Jevin West of the Calling Bullshit blog. Additionally, LGBT groups like GLAAD and HRC were quick to point to the consequences of the methodological blind spots in the paper.

Here I want to explore the particularly toxic brew around research ethics and research hype which has largely not been covered elsewhere, because it illustrates how research ethics regulations struggle to address the methods and consequences of pervasive data research. In particular, data science has the potential to weaponize general knowledge about a population (e.g., algorithms can predict sexual orientation from photographs with X degree of certainty) as a source of leverage in the lives of specific individuals outside of the study (e.g., the algorithm says *this person* is homosexual). General knowledge about a population is the definitional hallmark of researchin our ethics regulations, but harm done to people outside of the study is entirely invisible to those same regulations.

The study

Kosinski and Wang describe a study in which they received 35,326 “publicly available” pictures of 14,776 individuals enrolled on a popular American dating site on which users self-identify as seeking heterosexual or homosexual romantic partnerships. Half of the photos are from people seeking heterosexual relationships, half from people seeking homosexual relationships. Faces were sorted into gender categories, with roughly half men and women. The study was limited to Caucasian faces, due to what the authors claimed was a lack of other racial/ethnic groups in the available training set. (Reducing the diversity of faces also increases the likelihood the machine will identify a strong pattern.) Faces were controlled for transient factors such as lighting, head tilt and pitch, and whether the photo was large and full enough. The machine’s sorting was confirmed by Mechanical Turk workers who were instructed to sort photos by ethnicity and gender using criteria that would be familiar to contemporary American audiences. Kosinski used his and his girlfriend’s pictures as the prototypical white male and female faces, Barack Obama’s face as Black even though he is biracial, and a stock photo of someone who is “clearly” Latino.



Instructions given to Mechanical Turk workers for identifying Caucasian males, pg. 46 of Wang and Kosinki 2017. Kosinski is the prototypical white male.

The researchers divided the sample into 20 subsets, reserving one for a test set and using the others for training an open-source deep neural network optimized for facial recognition, called VGGFace. Setting self-identified sexual orientation as the dependent variable and 500 facial features as the independent variables, they trained an algorithm to recognize patterns of facial features of self-identified heterosexuals and homosexuals. Again using Mechanical Turk workers, the same test set was offered to human testers to gauge sexual orientation based on facial photos alone.

When the algorithm developed from the learning set was used on the test set, its predictions about sexual orientation were accurate in 81% of cases for men, and in 71% of cases for women. When the algorithm was given 5 different photos of the same face

the accuracy of the algorithm increased to 91% and 83%, respectively. The MTurk judges were significantly less accurate: 61% for men and 54% for women. Remembering that 50% accuracy is what you would expect with just random guesses when presented with binary options, it turns out that *untrained* humans are not all that great at judging sexual orientation with facial cues alone, and that *trained* machines are better but not perfect. Furthermore, the machine did much worse when it did not choose between binaries with one person known to be gay and ther other known to be straight, but rather had to make a decision one face at a time.

That is in a nutshell what the study demonstrated: untrained humans are significantly less accurate than a trained computer vision algorithm at discerning patterns in facial structures correlated with self-identified sexual preferences among an artificially constrained group of people.

The claims

Ultimately the core findings of the study are interesting but modest. The reason this study was explosive, especially to the press, are the explicit and implicit assumptions about the intrinsic nature of sexual orientation.

There is a long history of the search for a "scientific gaydar," as my PERVADE colleague Matt Bietz put it. A scientific gaydar could only work if there were intrinsic traits that provide a signal — genes, morphology, biochemistry, etc. — strongly correlated with sexual orientation, which is otherwise only observable as a behavior. Some scientists have long been certain that sexual orientation is so significant that there must be a signal to be found (Mattson's blog post cited above has a brief history of this phenomenon and this review article cited by Kosinski and Wang is a thorough look at research on the causes of homosexuality). Find the signal, and it can likely be traced to a cause. Find two signals that are correlated and then you have a fairly robust causal claim. Notably, the search for a scientific gaydar often takes the form of looking for *what makes homosexuals different* rather than *what makes all people have a sexual orientation at all*.

Along the way, the authors methodologically reduce the complexity and diversity of the biosocial phenomena they claim to be studying: gender, sexual preference and facial morphology. Such methodological points may seem to stray from the concerns of human subjects research ethics, but it has a direct consequence for how the study was constructed and its potential downstream effects. Controlling such variables is certainly a legitimate approach to empirical research, but it should properly cause scientists to dial down the implications of their outcomes. *It should take a lot of conceptual and empirical work to reduce a complex phenomenon like sexual preference and gender to a binary dependent variable and then build it back up again to make even a hedged generalizable claim about the nature of the complex phenomenon*.

Instead, Kosinski and Wang swing for the fences by claiming their study supports one of the plausible theories about biological causes sexual orientation: prenatal hormone theory (PHT). PHT holds that sexual orientation is determined at least in part by the level of androgens that a fetus is exposed to *in utero*. The typical way of framing it is that low

androgens cause homosexuality (defined as gender atypicality) in male fetuses and high androgen causes homosexuality in female fetuses. PHT is an unproven but entirely mainstream theory about biological roots of sexual preference. Because androgens are also known to affect facial structures in fetal development (and when adults take testosterone), it is plausible that facial structures may be a machine-readable signal of sexual preference due to a common cause.

Thus, the core premise, though not stated explicitly, of Kosinski and Wang's paper is that if sophisticated facial recognition software can correlate subtle facial structures with sexual orientation with a reasonable degree of accuracy, then there is support for a common intrinsic cause for both facial structure and sexual orientation, *i.e.*, PHT (see chart 1). Despite their claims to the contrary in the author's notes, the paper only holds together if they are at least implying that the deep neural network found the traces of an inherent basis of human sexual preference.

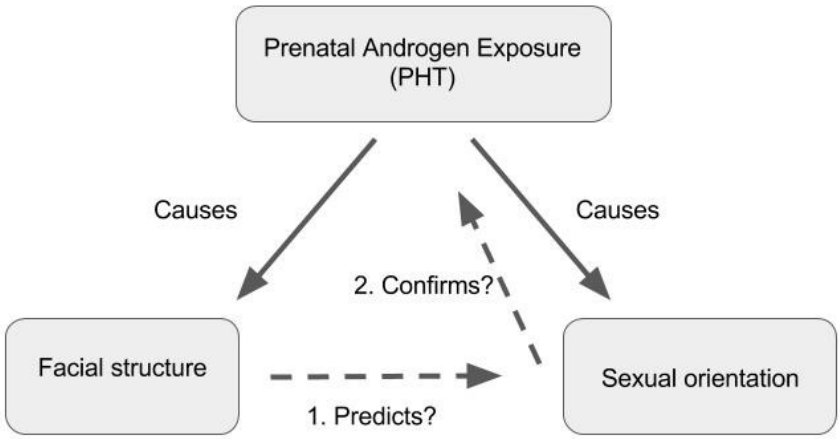


Chart 1: My explanation of the causal scheme presupposed by Kosinski and Wang.

But do they actually do the work to architect this reduced phenomenon (AI can sometimes correlate facial structures with binary categories of sexual orientation) into support for a global explanation of sexual orientation (prenatal androgen exposure determines sexual orientation)? In short, no.

The easiest explanation as to why not: there is no embryological, biochemical or social psychology research done in this paper.

The actual research done in this paper claims to correlate the machine-detected facial structures of Caucasian binary-gendered men and women with self-identified binary sexual orientation (1. In chart 1). Yet a surprising amount of the paper and subsequent press focusses on claims about PHT (2. in chart 1). In fairness, the authors are careful about using the terminology "is consistent with PHT" throughout the paper, which is an

appropriately hedged claim. On the other hand, for a research project that didn't measure a single micrometer of human blood (but did study facial hair grooming habits and baseball caps) their paper uses a surprising amount of ink discussing androgen levels. Logically, the only way that gayface AI points to intrinsic traits rather than transient factors is reference to biology, but the paper is not presenting biological research. That a study utilizing deep neural networks to analyze social media data makes *any* claims — even if strictly hedged — about intrinsic causes of sexual behavior in humans ought to be surprising.

"It passed the IRB"

In the paper, interviews and social media the authors raise the specter of discrimination in a post-privacy world as an ethical justification for their study. In the general discussion section of the paper, they write:

> Such pictures are often easily accessible; Facebook, LinkedIn, and Google Plus profile pictures, for instance, are public by default and can be accessed by anyone on the Internet. Our findings suggest that such publicly available data and conventional machine learning tools could be employed to build accurate sexual orientation classifiers. As much of the signal seems to be provided by fixed morphological features, such methods could be deployed to detect sexual orientation without a person's consent or knowledge. …
>
> Some people may wonder if such findings should be made public lest they inspire the very application that we are warning against. We share this concern. However, as the governments and companies seem to be already deploying face-based classifiers aimed at detecting intimate traits (Chin & Lin, 2017; Lubin, 2016), there is an urgent need for making policymakers, the general public, and gay communities aware of the risks that they might be facing already. Delaying or abandoning the publication of these findings could deprive individuals of the chance to take preventive measures and policymakers the ability to introduce legislation to protect people.

In other words, Kosinski and Wang were motivated to conduct this study not to create machine learning tools *for* discrimination, but to show that off-the-shelf machine learning tools can be used to facilitate discrimination because Internet data discloses innate, private characteristics that we cannot hide. On the one hand, it is banally predictable that the consequences of machine-learning-enabled surveillance will fall disproportionately on demographic minorities. On the other hand, queer folks hardly need data scientists scrutinizing their jawlines and hairstyles to warm them about this. They have always known this.

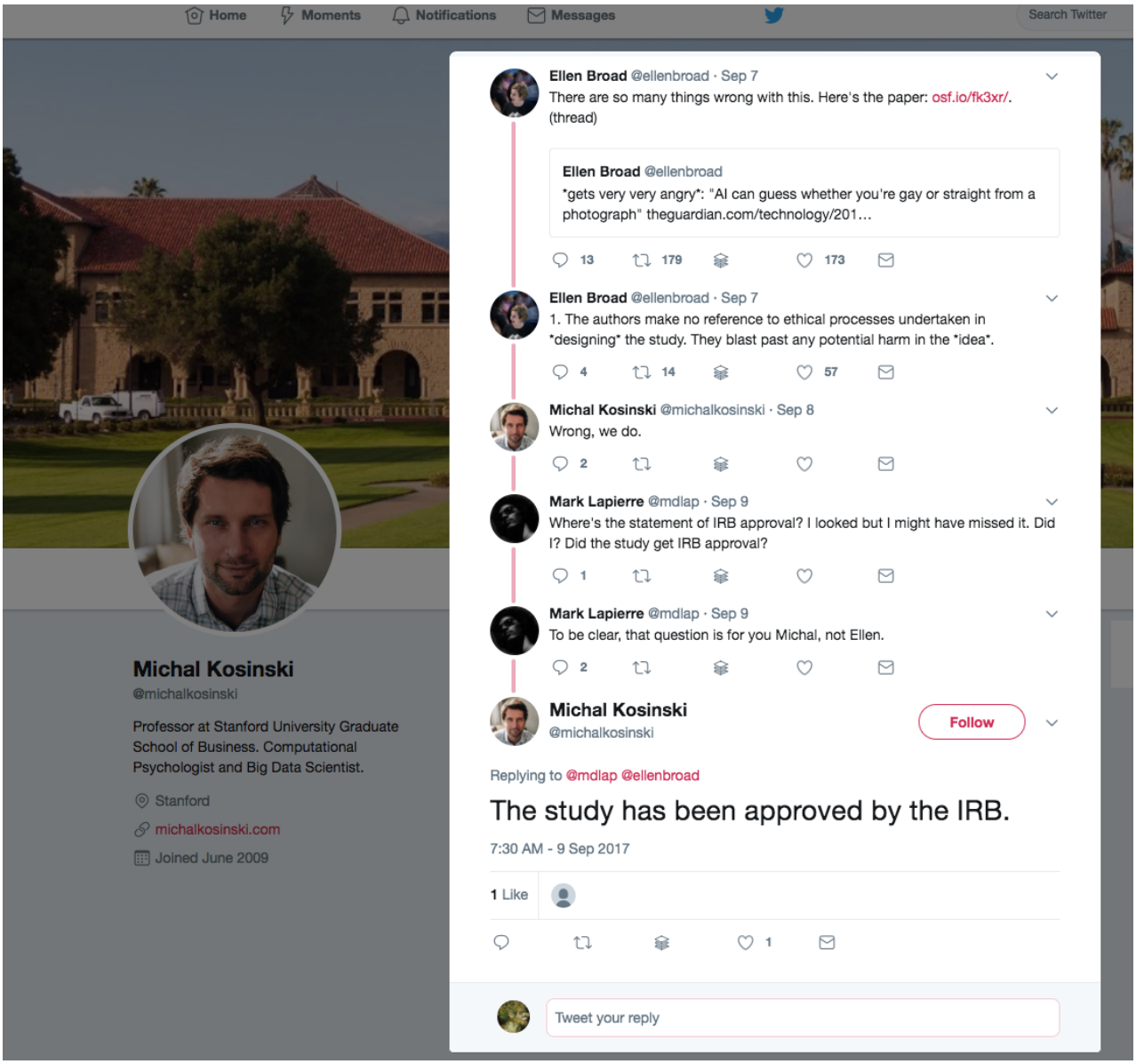
By ANDRADA BĂLEANU, from https://www.huffingtonpost.com/entry/a-more-queer-bucharest_us_57f50b9be4b0b7215072caef

For an example of how such a tool could go wrong, consider the ease with which a gayface plugin could be incorporated into the new AI-facilitated customs procedures at Dubai's airports, a country where consensual homosexual acts are punishable by prison sentences. Or how the government of Chechnya could use a gayface algorithm to entrap and purge homosexual men using surveillance cameras. And as Mattson says, there's plenty of reason to be concerned about the "bathroom police" closer to home using AI to humiliate and persecute transgendered people.

Kosinski appears to recognize this threat, and to his credit did not release the gayface algorithm as open-source tool (unlike his previous brushes with public controversy). Indeed, one of the most interesting aspects of the gayface controversy is that the authors and their critics agree on one central point: most plausible use cases for the tool that they built are ethically terrible. Yet they disagree about whether that means the study functions as an effective warning about possible future harms or is an ethical lapse in itself.

So is their suggestion that this research is necessary because it functions as a warning call actually ethically justifiable? Here's where the matter of research methodology and ethics regulation becomes paramount.

Noted prominently on the cover page of the preprint article, the authors state "The study has been approved by the IRB [Institutional Review Board] at Stanford University." In the below exchange about ethical justifications of the study on Twitter, Kosinski again refers to the Stanford IRB's approval:



https://twitter.com/michalkosinski/status/906525285394403328

Does that mean the study is ethical? Not at all. As I have argued in the past (1, 2, 3), research ethics regulation in the form of university IRBs is poorly suited to interrogating the methods of data science.

The core mission of IRBs is to protect individual research subjects from the potential harms caused to them by the research methodologies. The Common Rule—the U.S. federal law governing how IRBs should regulate human subjects research—only kicks in when the research methodology meets two conditions: 1) the research creates generalizable knowledge from datasets containing new and non-public data, and 2) acquiring that data requires intervening in a person's life (interview or psychological experiment) or body (blood draw or medicine) in a way that poses more than normal daily risks. These assumptions make sense using traditional research methods: you don't need ethics supervision if you are studying anonymized public census data, but you probably do need supervision if you are using deceptive tactics in a psychology study about sensitive personality traits. Furthermore, IRBs are legislatively forbidden to consider downstream consequences for people outside of the study. They are strictly tasked with controlling risk to individual study participants posed by the proposed research methods.

> Pervasive datasets dramatically change the research ethics landscape. The research methods and risks have changed, but the regulations have not.

Pervasive datasets dramatically change the research ethics landscape. The research methods and risks have changed, but the regulations have not. The vast majority of research that uses “big data” in one way or another does not fall under the purview of IRBs because 1) it does not create new data, it uses existing data as a learning set; 2) the data it uses is considered public, which includes data that can be purchased, lent, or gleaned from an Internet service like Facebook or OkCupid; and 3) it does not require any contact (“intervention”) with the individuals whose data is being used.

Although we can’t know for sure without Stanford releasing their IRB application (which are typically never viewed by the public), what Kosinski means when he says that “the study has been approved by the IRB” is likely just that the IRB decided his research does not create new data in such a way that poses risk to individual research subjects. *Which is technically correct* because there is no additional risk to the people whose facial images were anonymously used in the study. After all, they already outed themselves in “public” and put their pictures on a dating site.

IRBs are specifically mandated to avoid even considering the types of harms this research poses, which is downstream consequences to groups of people or society overall. Pervasive data of the type they draw upon is distinct from the type historically familiar to IRBs. Machine learning tools are designed to leverage general knowledge about patterns in a population in order to have an effect on individuals at a later point. This is an inverse of the traditional pattern of potential risks and benefits in human subjects research, wherein studying individuals leads to potential effects on populations. Machine learning can be weaponized in ways that traditional psychological or sociological research simply cannot.

As such, data scientists need to be aware that when their research is “approved by an IRB,” it does not mean “the research is ethical.” Rather, it means that whatever harms your research may pose are quite possibly invisible to the IRB’s review process.

> As such, data scientists need to be aware that when their research is “approved by an IRB,” it does not mean “the research is ethical.” Rather, it means that whatever harms your research may pose are quite possibly invisible to the IRB’s review process.

In my reading, those data science risks are greatly increased when reported results reach beyond the parameters of the research methods. Kosinski and Wang significantly increase the possibility that their work will be used against individuals by tying their work to claims about biological roots of sexual behavior. It’s actually not clear from their paper why PHT is a necessary component of the project at all. It certainly is not a component of their machine learning experiment as they do not actually measure any phenomena that could be used to empirically confirm or refute PHT.

Whether empirically justified or not, claims about the origin of sexual behavior in biology are consequential to many people's lives, and it is certainly possible that their over-reaching findings will be leveraged against individuals in automated decision-making. As Kosinski stated, the proper site for protecting individual rights is through politics, not through technologies. But that does not mean the risks of the research are within acceptable limits. And it certainly does not warrant leaning on the judgment of an IRB designed specifically to not consider these types of harms.

Tips for data scientists wanting to approach their work ethically:

- IRBs are often ill-suited to judge the most significant consequences of data science work.
- IRBs are nonetheless often necessary (but not sufficient).
- Data science research needs to be interrogated about downstream consequences because that is the type of harm most likely caused by the methods.
- If there are possible harmful consequences that are going to receive public attention, provide advice about avoiding them. Don't just throw your hands in the air and claim we live in a post-privacy society and individuals are responsible for protecting themselves from malicious data use. Ethical research and design is always a distributed responsibility.
- If your work involves sorting people by sensitive demographic categories, discuss the research with those communities and their advocates and listen to what they have to say. Asking for their input implies a responsibility to take it seriously and alter or possibly drop your research agenda to protect them in the manner they ask.
- If your work makes or implies empirical claims about other domains of expertise, include collaborators from those domains.

*All citations come from the version published on the OSF preprint repository on 09/10/2017. This is the version available when the press first started discussing the study. Updated versions have been published at later dates (most recent version available here), but none that substantially changes any claims made herein.

More from PERVADE: Pervasive Data Ethics

NSF-Funded Pervasive Data Ethics for Computational Research: a multi-disciplinary project examining the reuse of personal/social data in computational research