DOI:10.1145/3241367

Frank Pasquale

▸ **Stefan Bechtold,** Column Editor

Law and Technology
When Machine Learning Is Facially Invalid

Observations on the use of machine learning and facial inferences to classify people using inexplicable data.

MACHINE LEARNING RESEARCHERS have stirred controversy by claiming our faces may reveal our sexual orientation and intelligence.[a] Using a database of prisoners' faces, some have even developed stereotyped images of criminal features.[b] A start-up now claims it can deploy facial recognition to identify pedophiles and terrorists.[c] Facial inferences via machine learning are deeply troubling. When such methods of pattern recognition are used to classify persons, they overstep a fundamental boundary between objective analysis and moral judgment. When such moral judgments are made, people deserve a chance to understand and contest them.

The machine learning community must decide whether to improve such facial inference work or shun it. This column explores what each approach would entail. Better, more representative data could save the facial inference project from its worst tendencies. However, there are some scientific research programs best not pursued—and this might be one of them.

IMAGE BY METAMORWORKS

a See https://bit.ly/2wmgZ3L
b See https://bit.ly/2g6m0qW
c See https://wapo.st/2mnuJbO. Faception also presumes to classify persons in more positive ways; https://www.faception.com/our-technology.

Internal Critiques of the Facial Inference Project

Critics have faulted facial inference studies for inadequate datasets and overinterpretation of findings. For example, a study purporting to infer criminality from a dataset based on existing prisoners' and non-prisoners' faces has serious endogeneity problems. Prison itself may advance aging, or affect routine expressions, or even lead to disproportionate risk of facial damage (such as a broken nose or missing teeth). It would be odd to trumpet an inference of a broken bone from a cast. It is similarly off-putting to see machine learning researchers promote prediction of a status as a major achievement when key data supporting the inference of a certain status may be epiphenomenal of the status itself.

Another problem is representative-

ness, either in time, or across space. Training data consisting of prisoners' faces is not representative of crime, but rather, represents which criminals have been caught, jailed, and photographed. A high percentage of criminals are never caught or punished, so the dataset cannot adequately represent crime itself.

To the extent they motivate or rationalize additional surveillance for those identified as resembling typical faces in the training set, such identifications simply encourage society to double down on its existing priorities in crime detection. Thus facial inferences of criminality risk becoming self-fulfilling prophecies in exactly the same way as predictive policing (which has been extensively critiqued). In predictive policing, administrators use past arrest data (among other information sources) to decide where to allocate future police resources. However, that past data reflects long-standing biases: police were often assigned to minority neighborhoods, because political appointees assumed those locations would have the most crime.

More crime will naturally be found where there is more policing. To the extent above-normal crime in the neighborhood simply reflected past patterns of intense policing that reflected racism, then the "science of society" promised by big data morphs into a subjugation of certain parts of society. The algorithms behind such judgments become less objective arbiters of risk than ways of laundering subjective, biased decisions into ostensibly objective, fair scoring. Those affected lose a chance at individualized treatment and understanding, as technical systems treat people as a mere collection of data points.

When a dataset is not even close to representing the putative class it is used to classify, any results based on it should be qualified clearly. For example, a machine learning classifier may properly be said to succeed in classifying some percentage of faces *in its dataset* as criminal or non-criminal. But it should not be trumpeted as a potential classifier for all persons, unless and until we have some sense of how the training set maps to the full set of persons it ostensibly classifies.

There are other slippages in the facial inference studies. For example, a study that purported to identify "gay faces" may have merely picked up on certain patterns of self-presentation of persons who use the dating sites that were the source of the "gay" and "non-gay" images uses to train the classifier. Gay and lesbian persons of a certain time and place may be more or less likely to wear eyeglasses, have particular patterns of facial hair, or present themselves smiling or more serious. While the authors of a prominent "gay face" study tried to link their results to less-mutable aspects of physical development, such as hormone levels, their work may be time-bound to a certain pattern of make-up use, expressions, and other transient factors. A study on inferring sexual orientation from facial features also included a dataset based entirely on white faces—yet another limitation on its classifier's extrapolability.

The Importance of Explainability

All of these shortcomings support a larger critique of many opaque forms of artificial intelligence and machine learning: a lack of explainability as to how a classifier operates leaves it vulnerable to a critique of the representativeness of its training data. Such classifiers can mislead. For example, imagine an overwhelmed court that uses natural language processing to determine which of its present petitions are most like petitions that succeeded in the past, and then prioritizes those petitions as it triages its workflow. To the extent the past petitions reflect past conditions that no longer hold, they cannot be a good guide to which of the present petitions are actually meritorious.[3] However, a more explainable system, which identified why it isolated certain words or phrases as indicating a particularly grave or valid claim, could be more useful.

When it comes to criminal law, extreme caution should be exercised with respect to the new physiognomy.

Of course, these concerns would not arise if AI or machine learning mysteriously parsed the surface of a mountain in order to determine if it contains diamonds. Whatever claims nature may have on our conscience, they do not include the right to a well-reasoned account of suspicion. "Whatever works," a popular ethic of big data and probabilistic reasoning since Chris Anderson's 2009 article "The End of Theory,"[1] may end up ruling swathes of the natural world. However, Judea Pearl has cautioned that even in contexts far removed from human classification, machine learning will be a much more scientific enterprise to the extent it develops more robust accounts of causation.[4]

Internal critiques of the facial inference project counsel in favor of greater surveillance, more representative datasets, and a revamped physiognomy that traces correlations between facial features and intelligence, criminality, and sexuality to some common genetic or environmental factors. Perhaps a universal database of the faces of all criminals, or all gay persons, coupled with even more intimate assessments of health and education data, environmental exposures, and more, would advance sciences of sexuality or criminology. But this vision rests on a naively scientistic perspective on social affairs, where reflexivity (the effect of social science on the social reality it purports to merely understand) compromises any effort (however well intended) to straightforwardly apply natural science methods to human beings.

When it comes to criminal law, extreme caution should be exercised with respect to the new physiognomy. To the extent it "works" to videntify potential criminals, it will also work to reinforce the same structures of oppression and exclusion that generated the predictive power of certain facial features. The allocation of surveillance based on facial characteristics also ignores a fundamental ethical commitment in criminal

justice—that individuals are not to be held responsible or harmed based on features they cannot control. For the most part, we cannot control how our faces appear. To the extent dieting or plastic surgery could permit someone to escape or reduce the burden of face-driven investigations, an incentive to engage in such behavior is bizarre, because there is no evidence that changing one's facial appearance reduces one's propensity to criminality.

This lack of causal connection points up another troubling aspect of face-based surveillance. To the extent we lack any evidence that face shape actually affects criminality, basing policy on mere correlation is creepy. It is effectively the elevation of an alien, non-human intelligence in a system where human meaning and communication are crucial to legitimacy.[5] But to the extent we can actually tell people that there is some way in which changing their faces can reduce their likelihood of being criminal, the big data intervention presages exceptionally intense and granular social control. This double bind—between black-box manipulation and intimate control—counsels against further work in the area.

Given the tragic history of projecting criminality from facial or cranial features, false positives due to appearance (including the mere chance of increased surveillance burden) are intolerable. Moreover, even if the system were by some miracle 100% accurate, its methods are not consistent with the rule of law. As Kiel Brennan-Marquez has explained, a jurisprudence of well-founded suspicion (largely arising out of Fourth-Amendment law in the U.S.) demands that authorities give a plausible, and not merely probabilistic, statistical, or artificially intelligent, account of the reasons why they investigate suspects.[2] This is a limit on the power of the state, which may be all too tempted to use advanced surveillance technology to achieve complete control over citizens. Given the repugnance of "general warrants" (which give the police a general right to search persons), a big-data-driven risk assessment of vall persons' threat profile for crime would undermine important privacy protections. Black-box predictive analytics could easily give a police force an excuse to investigate nearly anyone, since we are all likely to have engaged in some behavior with *some* correlation with that of potential criminals.

Even if the system were by some miracle 100% accurate, its methods are not consistent with the rule of law.

Brennan-Marquez's concept of plausibility should resonate with regulators in Europe, who have developed the concept of a "right of explanation"[d] under the General Data Protection Regulation, to enable citizens to understand how they are being profiled by automated systems in the private sector. Some machine learning researchers claim such a right will impede research by retarding the adoption of ML systems in credit, insurance, banking, and beyond. But explainability is a key component of science, and should be recognized as such. Efforts to strangle the emergent right to explanation in the cradle should be resisted not just on human rights grounds, but also because of basic commitments to science and justice. The computational power and algorithmic acumen that would be directed at classifying people, in the absence of a right of explanation, is better redirected to research on how to make the natural world tractable to human ends. C

d See https://arxiv.org/abs/1606.08813

References

1. Anderson, C. The end of theory: The data deluge makes the scientific method obsolete. *Wired* (June 23, 2008).
2. Brennan-Marquez, K. Plausible cause: Explanatory standards in the age of powerful machines. *Vanderbilt Law Review* (2017).
3. Pasquale, F. and Cashwell, G. Prediction, persuasion, and the jurisprudence of behaviourism. *U. Toronto Law Journal 68*, 1 (Jan. 2018), 63–81.
4. Pearl, J. and McKenzie, D. *The Book of Why* (2018).
5. Sloan, R.H. and Wagner, R. Avoiding alien intelligence: A plea for caution in the use of predictive systems; https://bit.ly/2Lh1vK2

Frank Pasquale (fpasquale@law.umaryland.edu) is a Professor of Law at the University of Maryland Carey School of Law, Baltimore, MD, USA.

Calendar of Events

August 28–31
ASONAM '18: International Conference on Advances in Social Networks Analysis and Mining,
Barcelona, Spain,
Contact: Jon Rokne,
Email: rokne@ucalgary.ca

August 28–31
DocEng '18: ACM Symposium on Document Engineering 2018,
Halifax, Nova Scotia, Canada
Sponsored: ACM/SIG,
Contact: Evangelos Milios,
Email: eem@cs.dal.ca

September 3–6
MobileHCI '18: The 20th International Conference on Human-Computer Interaction with Mobile Devices and Services,
Barcelona, Spain,
Sponsored: ACM/SIG,
Contact: Lynne Baillie,
Email: l.baillie@hw.ac.uk

September 3–7
ASE '18: ACM/IEEE International Conference on Automated Software Engineering,
Montpellier, France,
Contact: Marianne Huchard,
Email: marianne.huchard@lirmm.fr

September 5–7
NANOCOM '18: The 5th ACM Annual International Conference on Nanoscale Computing and Communication,
Reykjavik, Iceland,
Sponsored: ACM/SIG,
Contact: Falko Dressler,
Email: dressler@ccs-labs.org

September 14–17
ICTIR '18: ACM SIGIR International Conference on the Theory of Information Retrieval,
Tianjin, China,
Sponsored: ACM/SIG,
Contact: Dawei Song,
Email: dawei.song@open.ac.uk

September 19–22
TAPIA '18: Richard Tapia Celebration of Diversity in Computing Conference,
Orlando, FL,
Sponsored: ACM/SIG,
Contact: Valerie E. Taylor
Email: taylor@cse.tamu.edu