IRIE
International Review of Information Ethics Vol. 34 (10/2024)

Author(s): Peter Andes

Can Machine Learning Identify Criminals Just by Looking at Their Faces?

Agenda:

Author(s):

Dr. Peter Andes

- University of Alberta
- ✉ andes@ualberta.ca

ISSN 1614-1687

Case Description

In 2016, AI researchers Xiaolin Wu and Xi Zhang posted a paper to a non-peer reviewed depository that explained their plan to use machine learning to recognize criminals just by looking at their faces. "These two populations should be among the easiest to differentiate," the researchers stated, "…because being a criminal requires a host of abnormal (outlier) personal traits." They claimed to discover that "some discriminating structural features for predicting criminality have been found by machine learning." There was an immediate backlash. According to Wu and Zhang, critics called the project racist. The researchers defended their stance, admitting that some of the language they used did have negative connotations due to oversights in translating their research into English, but overall standing by the project. "Some of our critics seemed to suggest that machine learning tools cannot be used in social computing simply because no one can prevent the garbage of human biases from creeping in," Wu and Zhang stated in a follow-up article[1]. "We do not share their pessimism," they declared. The researchers believed that even as biases can enter algorithms, algorithms can also be used to detect human biases. In their research they claimed to be "playing devil's advocate" and not to be endorsing the use of the algorithm by law enforcement.

Questions

1. Can criminality be recognized simply from facial features? How likely is it that our perceptions of criminality indicate some deeper biological foundation for criminal behavior? Are our perceptions simply culturally influenced and unlikely to indicate the presence of actual criminal tendencies?

2. Should some areas of research be ethically off-limits, even if the research is meant to be purely academic?

3. Are algorithms simply neutral tools, available to be used in helpful or harmful ways, as the researchers in this case argue? Should they be avoided entirely in "social computing" because there is no way to eliminate biases? Even if there is no way to eliminate biases in algorithms, is there a responsible way to carefully use biased algorithms?

4. Are some technological applications just too potentially dangerous for it to be morally permissible to engage with them, even when "playing devil's advocate"? What might they be?

5. How might this research be put to use outside of academia in ways that lead to negative consequences? Are there any positive consequences that could result from the research?

6. Suppose there are some media reports that have misrepresented the intentions and aims of the researchers. Are there also dangers that journalists and commentators will in see bias everywhere, and look for the least charitable interpretation of a project, just to get a click-worthy, AI-related headline?

Exercises

1. Imagine you are serving on a committee that determines which AI projects receive grant funding. Develop a set of guidelines specifying what sort of ethical restrictions, if any, should be placed on the projects that get funded.

[1] Wu, X., & Zhang, X.. (2016). Responses to Critiques on Machine Learning of Criminality Perceptions (Addendum of arXiv:1611.04135).

2. Suppose that you are tasked with taking over this project from the researchers. How might you redirect and revise the research toward a positive social impact? Or is it simply impossible to do so, the only ethically appropriate course of action being to shut the project down?

Applying the Principles of AI Ethics

Using the chart provided, identify which principles of AI ethics are at issue in this case and, if principles conflict, which seems to be the weightiest and so the one that should override other principles.

Principle	Application (If Any)
Nonmaleficence	
Beneficence	
Respect for Autonomy	
Justice	
Explicability	
Accountability	

Normative Theories

Apply two of the normative theories explained in the introductory section at the beginning of this volume to bring out issues in the case that might have been overlooked. How would a utilitarian approach this case? A deontologist? A virtue ethicist? Do either of the two approaches you pick accord with your own moral judgments or not? Do they arrive at the same verdict as the principles of AI ethics?

Expert Analysis by Peter Andes

Although they use new technology, the researchers in this case study are part of a long history of efforts to detect criminality by examining a person's appearance. One of the most famous efforts was the work of the 19th century Italian criminologist and phrenologist Cesare Lombroso. (Phrenology is a defunct pseudoscience which held that the shape of the skull could reveal information about a person's intelligence and personality.) Lombroso proposed the idea that some people were just born criminals. They could be identified by certain physical traits such as the shape of the forehead and nose. Much like the present-day researchers in this case study, another 19th century figure, the eugenicist Francis Galton, used a series of photographs to compare criminals with law-abiding citizens to detect differences between their faces. Galton coined the term "eugenics" (meaning good birth or creation) and his ideas influenced the Nazis as well as the forced sterilization movement in North America.

Given this disturbing lineage, I argue we need to be far more careful than these researchers have been in venturing into the realm of predicting criminality using facial features. In what follows, I will identify the many ethical problems that arise for me in this case. To do so I will apply the principles of AI ethics using a Rossian prima facie methodology. Based on the application of these principles, I argue that this research should not have been conducted in the form it was, but a revised form could be responsibly carried out. I conclude by showing how, on any of the major normative theories, this research should not have been conducted.

This case raises issues involving beneficence, nonmaleficence, justice, and accountability. To begin with, the benefits of the research are unclear. Who exactly will benefit from this technology? Presumably the researchers would answer that the study offers an interesting addition to our existing knowledge. However, whether it does is doubtful. And, even if it does, this benefit, and so the principle of beneficence, can be outweighed by other principles.

Turning to nonmaleficence, the research could potentially harm people by leading some people with no criminal tendencies to be thought of as criminals simply because of their facial features. The potential for harm is a concern when applying the principle of nonmaleficence.

Promoting this kind of research could also lead to discrimination against those deemed to have "criminal" traits and tendencies by leading to stereotyping and prejudice. Discrimination is in violation of the principle of justice. People deserve to be treated equally unless there is a good reason for treating them differently. Given the unlikelihood that criminality is indeed linked to facial features, and the fact that even if it were this would not mean that all people with "criminal faces" would commit crimes, treating some people as criminals is not justifiable. As a failure to treat people in a way they deserve, it is a violation of the principle of justice.

By putting this research out into the public sphere but essentially washing their hands of the matter in saying they are "playing devil's advocate" and not advocating its use by law enforcement, the researchers seem to be distancing themselves from any responsibility for what might be done with their research. This seems like an unacceptable avoidance of accountability. This claim depends on an understanding of when it is appropriate to play devil's advocate.

The term devil's advocate comes from the process of canonization in the Catholic Church. The office was established in 1587. When the Church is deciding whether a particular person is to become a saint or not, historically someone takes the side of arguing against the canonization. It is not that this person really believes the case that he is making. Rather, he takes up the role to make sure the arguments are heard and so all evidence is presented and assessed to make the determination accurately. This is indeed an important part of a comprehensive inquiry which is important for the formation of knowledge. As the 19th century Victorian philosopher John Stuart Mill famously argued,

> He who knows only his own side of the case knows little of that. His reasons may be good, and no one may have been able to refute them. But if he is equally unable to refute the reasons on the opposite side, if he does not so much as know what they are, he has no ground for preferring either opinion (Mill 1901, 67).[2]

There are indeed times where controversial research may be morally permissible to conduct in the spirit of playing devil's advocate. However, I argue that this is only permissible in cases where playing devil's advocate can offer substantial benefits to the formation of knowledge that are not outweighed by potential harms. We would not accept as morally defensible a position that sought to challenge the idea that men and women are equally cognitively capable simply for the sake of playing devil's advocate when there are no benefits to be gained from doing so. And, as I have already argued, no such benefits are on offer here, or at the very least only weak ones. At the same time, the potential harms are serious.

The current scientific consensus would militate against any supposition that facial features are linked to criminal behavior. There is nothing to be gained from hearing arguments against that consensus, since these arguments won't help anyone, but only lead to harm, prejudice, discrimination, and injustice. Rather, it seems here that

[2] Mill went on to write that it is no good to hear an opinion defended merely in the manner of a devil's advocate. Rather, one must hear it from sincere proponents. But surely if there are no sincere proponents to be found, or none whose arguments are any good, it seems fine to settle for a devil's advocate.

the researchers are using the idea of playing devil's advocate to escape responsibility for the potential social impact of their work. This is in violation of the principle of accountability.

Of course, each of the principles of AI ethics is here being treated as having prima facie weight. That is, each principle is overridable by a stronger principle. It is conceivable that a controversial study could yield such great benefits that this would justify conducting it despite other principles counting against it. However, the benefits of this study are weak at best, and perhaps nonexistent. As a result, the principle of beneficence is outweighed and the other principles which count against conducting the study should be followed.

A similar study could be morally permissible. If the research were revised so that it was aimed more clearly at showing that there is no correlation between physical appearance and criminality then this would be a positive revision. That is not to say that ethical worries don't still remain. Perhaps there is so little benefit to be gained from this sort of research that even the more clearly positively position project would not be worth carrying forward.

Even if someone were to reject the principles of AI ethics as an appropriate basis for moral judgment, I argue that any of the main normative theories would arrive at the same conclusion. To begin with, consider utilitarianism. In this case, there is little benefit to the research and a great potential for harm. If in the dissemination of these results the study is taken to imply that some people are more likely to be criminals simply because of their facial features, then this could lead them to be harmed through discrimination and prejudice when they have never and might never commit any criminal act. Although one can always run a cost/benefit analysis differently to arrive at a different conclusion, there is a strong case to be made for thinking that the utilitarian would prohibit this research.

Next, consider deontology. We already looked at one form of deontology above in applying the principles of AI ethics. But suppose someone subscribed to a different deontological tradition, such as the Kantian view. We can invoke Kant's Second Formulation of the Categorical Imperative here. Kant tells us not to use people merely as means and to respect them as ends in themselves. Seeking to identify someone as a criminal just based on that person's facial features fails to respect that person's rational dignity. As rational agents, people are in charge of their actions. They legislate their next course of action, if all goes well, using their rational capacities. At the very least, each of us is capable of acting this way and ought to be acting this way. To suppose that some people with certain facial features will have a tendency to commit crimes ignores the degree to which agents are capable of controlling their actions through reason. Even if agents did have this tendency, the Kantian would exhort them to master their instincts and act on reason alone. And the Kantian would say it was in the capacity of everyone to do so as a rational being.

Finally, a virtue ethics perspective would also be opposed to conducting this research. Virtue theory tells us to do what the virtuous person would do. To know what the virtuous person would do, we have to consider what virtues the virtuous person would act from. Even as virtue theorists differ on which character traits are virtues, I argue all lists should contain the virtue of being just. Even as virtue theorists can also differ in their characterization of the virtue of justice and all that it involves, in general any reasonable account of justice will involve treating others fairly. The virtuous person is just, and so would seek not to potentially encourage prejudgment of the innocent. Thus the virtuous person would not conduct the research.

In conclusion, then, our best ethical reasoning tells us that this research should not have been conducted. Whether it is based on the principles of AI ethics, or the major normative theories, we see that the ethical concerns outweigh any benefits that might arise from this research.

Student Reflection

Did you touch on everything this discussion identifies in your own analysis of the case? What did you miss? What did you think of that could be added to the discussion?

References

Arcas, B. A. y. (2017, May 20). *Physiognomy's new clothes*. Medium. from https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a.

Bailey, K. (2016, November 29). *Put away your machine learning hammer, criminality is not a nail*. Wired. https://www.wired.com/2016/11/put-away-your-machine-learning-hammer-criminality-is-not-a-nail/.

Mill, John Stuart. (1901). *On Liberty*. https://www.gutenberg.org/files/34901/34901-h/34901-h.htm.

Van Noorden, R. (2020). The ethical questions that haunt facial-recognition research. Nature, 587(7834), 354–358. https://doi.org/10.1038/d41586-020-03187-3.

The dark past of algorithms that associate appearance and criminality. (2020, December 17) American Scientist. https://www.americanscientist.org/article/the-dark-past-of-algorithms-that-associate-appearance-and-criminality.

Wu, X., & Zhang, X. (2016). Responses to critiques on machine learning of criminality perceptions (Addendum of arXiv: 1611.04135). *arXiv preprint arXiv:1611.04135*. https://arxiv.org/abs/1611.04135.