Data&Society

**This case study was utilized at an AI and Human Rights workshop, held at the Data & Society Research Institute on April 26-27, 2018.*

AI Systems and Research Revealing Sexual Orientation
Case Study

Background

In September 2017, Michal Kosinski and Yilun Wang pre-printed a study from Stanford University entitled, 'Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images.' The study, published in the February 2018 edition of the Journal of Personality and Social Psychology, claims that a deep neural network can be trained to detect sexual orientation from photographs with a higher level of accuracy than humans can.

The academic paper details how Kosinski and Wang trained an algorithm on a sample of 35,326 facial images of self-identified gay and straight individuals, which were obtained from an undisclosed dating platform on which users self-identify as seeking homosexual or heterosexual partners.[1] The algorithm composed of publicly-available facial recognition software compared different facial features and found that gay men and women tended to have "gender atypical" faces. The results showed that the faces of gay men were more feminine and the faces of lesbians were more masculine. When presented with two side by side photos, one gay and one straight, the algorithm could purportedly distinguish between them 91% of the time for men and 83% of the time for women. The authors say that in contrast, the computer algorithm was far more accurate than human judges who were only correct 61% of the time for men and 54% for women (with 50% success representing a random guess). The paper cited this finding as yet another example of artificial intelligence outperforming humans.[2]

Ethical dilemmas

Academia and Research

The research had significant limitations, some of which were addressed in an author's note accompanying their peer-reviewed paper. For example, the study only looked at white men and women who self-reported as being gay or straight. The reason cited was because the dating site they were using as their data set "had served up too few faces of color to provide for meaningful analysis."[3]

The study and its findings were immediately challenged by human rights organizations, data scientists, and researchers, among others. Critics questioned the accuracy of the model utilized in the study, the limited data pool, and the binary classification of sexual orientation. They also raised concerns about the study's conclusions and the ethics of how it was conducted. Kate Crawford, co-founder of the AI Now Institute, said the study was "AI phrenology, and it's very, very dangerous."[4]
The authors defended the study citing that it had been approved by Stanford University's Institutional Review Board (IRB). However, this approval does not mean that the study was ethical.[5] The goal of IRBs

---

[1] Michal Kosinski, Yilun Wang. "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images." Journal of Personality and Social Psychology. February 2018, Vol. 114, Issue 2, Pages 246-257.

[2] Ibid., note 1

[3] Alan Burdick. "The A.I. 'Gaydar' Study and the Real Dangers of Big Data." September 15, 2017. https://www.newyorker.com/news/daily-comment/the-ai-gaydar-study-and-the-real-dangers-of-big-data

[4] @katecrawford. "This is AI phrenology, and it's very, very dangerous." Twitter, November 18, 2016, 8:29 AM. https://twitter.com/katecrawford/status/799650861370720258

[5] Jacob Metcalf. "'The study has been approved by the IRB': Gayface AI, research hype and the pervasive data ethics gap." November 30, 2017. https://medium.com/pervade-team/the-study-has-been-approved-by-the-irb-gayface-ai-research-hype-and-the-pervasive-data-ethics-ed76171b882c

is to protect human subjects "from the potential harms caused to them by the research methodologies" and are "legislatively forbidden to consider downstream consequences for people outside of the study." While IRBs are often necessary, they may not be sufficient or applicable in cases using big data 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."[6]

Business

Although this example comes from academia, the ethical dilemmas stemming from the combination of big data and artificial intelligence extend to the private sector as well and can be exacerbated by questionable data sharing practices. For instance, it was recently revealed that Grindr, a dating app for gay, bisexual, and transgender men, shared the HIV data of its users to a third-party vendor, without their informed consent. This data included personally identifiable information, such as HIV status, geolocation, sexuality, relationship status, and ethnicity.[7] While the app shares users' profile information to optimize its services, such sensitive health information could be misused if appropriate safeguards are not in place.[8]

In this case, the sharing of someone's HIV status could lead to discrimination and stigmatization, which can further marginalize HIV-positive individuals and make them more vulnerable. This can manifest itself in ill treatment, an erosion of human rights, psychological damage, and limited or denial of healthcare services for people living with HIV.[9] Furthermore, in countries where homosexuality is banned, governments could use this information as a means to discriminate against and penalize LGBTQ individuals, which could result in physical abuse.

Companies have an obligation to protect their users information. The role of Facebook in the Cambridge Analytica controversy is a prime recent example. In addition to a breach of user's trust, the carelessness of Facebook's actions reveals a clear disregard for any sort of ethical considerations, which can lead to unforeseen consequences. Furthermore, there is added layer of risk for online users when their information is shared with outside parties over unencrypted connections, making them more vulnerable to hacks and data breaches.

Human rights implications

Discrimination

Kosinski said he began his study to demonstrate how readily available data and technologies could facilitate discrimination. As human biases can be transferred into algorithms, the risks of using them to categorize individuals can lead to harmful consequences and embed societal discrimination into decision-making processes. This can include racially biased risk assessment software to predict recidivism,[10] compromised automated systems to gauge eligibility for welfare and housing,[11] and targeted ads for predatory payday loans for low income households. This sensitive information could be further revealed without a person's knowledge or consent.

---

[6] Ibid., note 5

[7] "Congressional letter from Senator Edward Markey and Senator Richard Blumenthal to Zhou Yahui." April 3, 2018. https://www.markey.senate.gov/imo/media/doc/grindr%20letter.pdf

[8] Ibid., note 7

[9] Avert. "HIV Stigma and Discrimination." April 9, 2018. https://www.avert.org/professionals/hiv-social-issues/stigma-discrimination

[10] Matthias Spielkamp. "Inspecting Algorithms for Bias." June 12, 2017. https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/

[11] Tanvi Misra. "When Welfare Decisions Are Left to Algorithms." February 15, 2018. https://www.theatlantic.com/business/archive/2018/02/virginia-eubanks-automating-inequality/553460/

Privacy

As the influx of big data continues to grow, it offers machines more meaningful and contextual information[12] from a variety of sources. When integrated, such datasets can create a very detailed picture of a person's life and their relationships. This kind of data layering can also reveal new information about an individual and serve as a form of algorithmic surveillance. The implications of Kosinski and Wang's research are that, in the wrong hands, the tool they created could pose a real threat to the privacy and safety of the LGBTQ community,[13] especially for those living under repressive regimes. For example, it is technologically trivial to add a "gaydar" plugin using such an algorithm to a closed-circuit surveillance system to automatically detect "gender atypical" faces.

Freedom of expression and association

The ability to safely communicate online is extremely important for vulnerable and marginalized communities. This is compromised when confidential details are disclosed to an unknown party. In the case of Grindr, a consequence of sharing sensitive health information can lead users to self-censor for fear of discrimination or other repercussions. Such disclosures can hamper the visibility of individuals who self-identify as LGBTQ. The irresponsible release of this personally identifiable information can also prompt the unjust targeting of LGBTQ communities or organizations.

Discussion questions

1. The study raises concerns around research methodology and the ethics of deploying AI tools. Do university IRBs need to be reconsidered? What safeguards are necessary for industry R&D? Would a human rights impact assessment be an effective guide to evaluate risks? How would these protections be enforced?
2. The authors created an algorithm run by a neural network to carry out their study. Given that there are many hidden layers in a deep learning process, how can AI developers be held accountable if machine learning technologies are, by nature, opaque?
3. In many instances regulation has not caught up with technology. For example, some might argue that images posted on websites are "public information" with no expectation of privacy. What role should governments have in regulating the development of AI based technologies? How can this be approached without stifling innovation?
4. Kosinski specifically did not release the algorithm that he trained to detect sexual orientation as an open-source tool. He refrained from doing so because he recognized the possible dangers that could come from its misuse. In considering the importance of algorithmic accountability, to what extent do the researchers have a responsibility to share this algorithm for independent review? With whom would they share it?
5. How can human rights organizations effectively engage companies to uphold human rights obligations? What are some key strategies and tactics?

[12] Steven Hansen. "How Big Data Is Empowering AI and Machine Learning?" Nov 24, 2017. https://hackernoon.com/how-big-data-is-empowering-ai-and-machine-learning-4e93a1004c8f

[13] JD Schramm. "AI 'gaydar' could compromise LGBTQ people's privacy — and safety." February 19, 2018. https://www.washingtonpost.com/opinions/ai-gaydar-could-compromise-lgbtq-peoples-privacy--and-safety/2018/02/19/172156bc-126d-11e8-9065-e55346f6de81_story.html?utm_term=.c007ad0c65d4