I had a hard time picking just one algorithm-gone-wrong, so here's two!

Number One:

In 2017, a professor at Stanford published this paper, in which he demonstrated that an AI could predict the sexual orientation of a person (or at least the subset of people in the training set of data for the study, sourced from dating sites, I believe) at much higher rates than a human.

This obviously raises all kinds of questions about how such technology might be misused--unlike most of the algorithms here, this one hasn't, to my knowledge, been abused so far (although, who knows, since so many governments around the world employ different forms of facial recognition software) but there was an outcry from LGBTQ organizations about this algorithm's potential for misuse. (One critic argued the algorithm is the equivalent of a "13-year-old bully.")

Interestingly, according to the New York Times, the professor behind this work wanted to draw attention to the very dangers of such algorithms: he "wanted to call attention to privacy risks", the article states, by showing that AI can detect something that "people should have full rights to keep private."

This raises the troubling question of how best to prevent algorithms such as this one from causing harm: is it better to draw attention to untapped potential for algorithmic harm, the way this researcher apparently hoped to? Or is drawing attention to the harmful applications of technology itself a preamble to the very abuse it aims to prevent?

Indeed, an AI that can identify queer folks on sight raises troubling echoes of the Holocaust, when queer folks were forced to wear pink triangles. And yet--as the professor behind the research argues, the capability for this abuse already exists, and we are fooling ourselves if we think someone won't eventually enact it. (Sadly, he didn't offer any clear next steps on how to prevent this sort of abuse.)

At the same time, the study *does* offer evidence--and some LGBTQ groups quoted by the *Times* celebrated this fact--that being queer is genetic, as the facial differences identified by the AI would have to be genetic. (So, even though the algorithm presents potential to single out queer folks for abuse, it simultaneously acknowledges that queer folks are born queer, underscoring that being queer has a biological basis, and is not a "choice.")

Number Two:

In the 1990s, a system called COMPAS went online, which promised to use computing to accurately predict recidivism rates, and thereby assist in the sentencing of criminal defendants.

According to a recent article, over a million folks have been processed by COMPAS, which predicts recidivism rates based on factors that are now recognized as biased (i.e., race, etc.)

The upshot is that plenty of folks (mostly all Black) have been given inappropriately long sentences--one example that ProPublica highlighted in a 2016 investigation was a Black man arrested at a protest who was deemed at high risk of committing another crime, while a white man who dealt drugs was given a low risk. (in fact, their criminal records since show the opposite was true.)

On the one hand, you can understand why such a system was developed--most major criminal courts have terrible backlogs, which itself imposes a cost on defendants, some of whom are imprisoned and wait years for a trial while still not convicted of a crime. (Like Kalief Browder, the young man accused of stealing a backpack who spent years in Rikers for a crime he wasn't even convicted of, an experience that so deeply scarred him that he later took his own life.)

In other words, the way the algorithm was designed seems to have replicated the very bias we see in policing, where Black folks are systematically disadvantaged. Rather than accurately predict the likelihood that someone will commit a crime, thereby improving the fairness of criminal sentencing, the algorithm just reproduces the caste system of American society. Even worse, it gives such sentences the veneer of impartiality, because the algorithm is not human, and therefore might seem less prone to error. (Of course, it was designed by humans.)