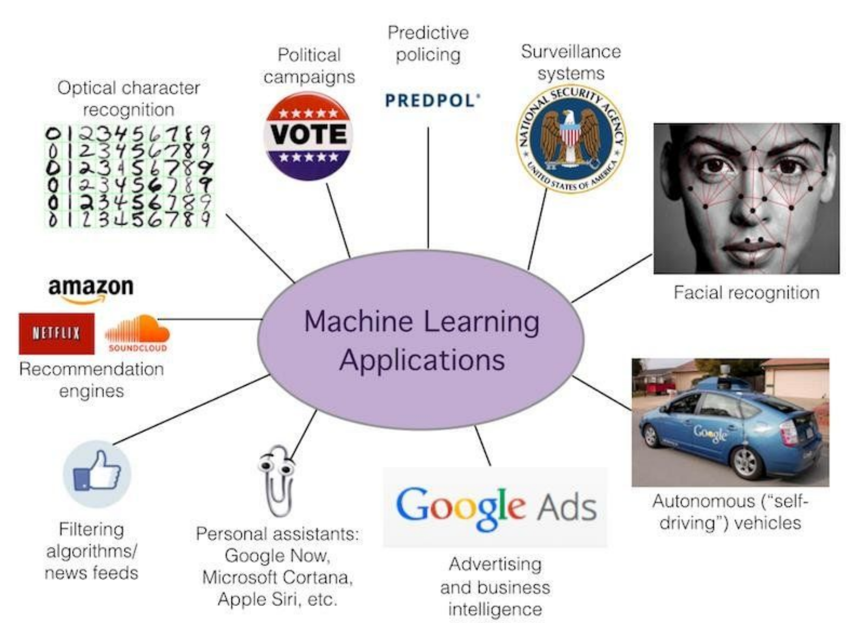
A Glimpse into Ethical Issues in Machine Learning (ML)

Daniel Rosiak
Postdoctoral Researcher in Ethics and AI/Emerging Tech
Inamori International Center for Ethics and Excellence

Applications of ML


Predictive policing
Political campaigns
Surveillance systems
Optical character recognition
VOTE
PREDPOL
Facial recognition
amazon
NETFLIX
SOUNDCLOUD
Recommendation engines
Machine Learning Applications
Autonomous ("self-driving") vehicles
Filtering algorithms/ news feeds
Personal assistants: Google Now, Microsoft Cortana, Apple Siri, etc.
Google Ads
Advertising and business intelligence

Applications of ML in our daily lives extend to a wide variety of areas, including

- precision agriculture (Sennaar, 2019),
- air combat and military training (Gallagher, 2016; Wong, 2020),
- insurance
- education (Sears, 2018),
- finance (Bahrammirzaee, 2010),
- health care (Beam and Kohane, 2018),
- manufacturing and natural resources
- science
- human resources and recruiting (Hmoud and Laszlo, 2019),
- music composition (Cheng, 2009/09),
- customer service (Kongthon et al., 2009),
- reliable engineering and maintenance (Dragicevic, 2019),
- autonomous vehicles and traffic management (Ye, 2018),
- social-media news-feed (Rader et al., 2018),
- work scheduling and optimisation (O'Neil, 2016)

General remark

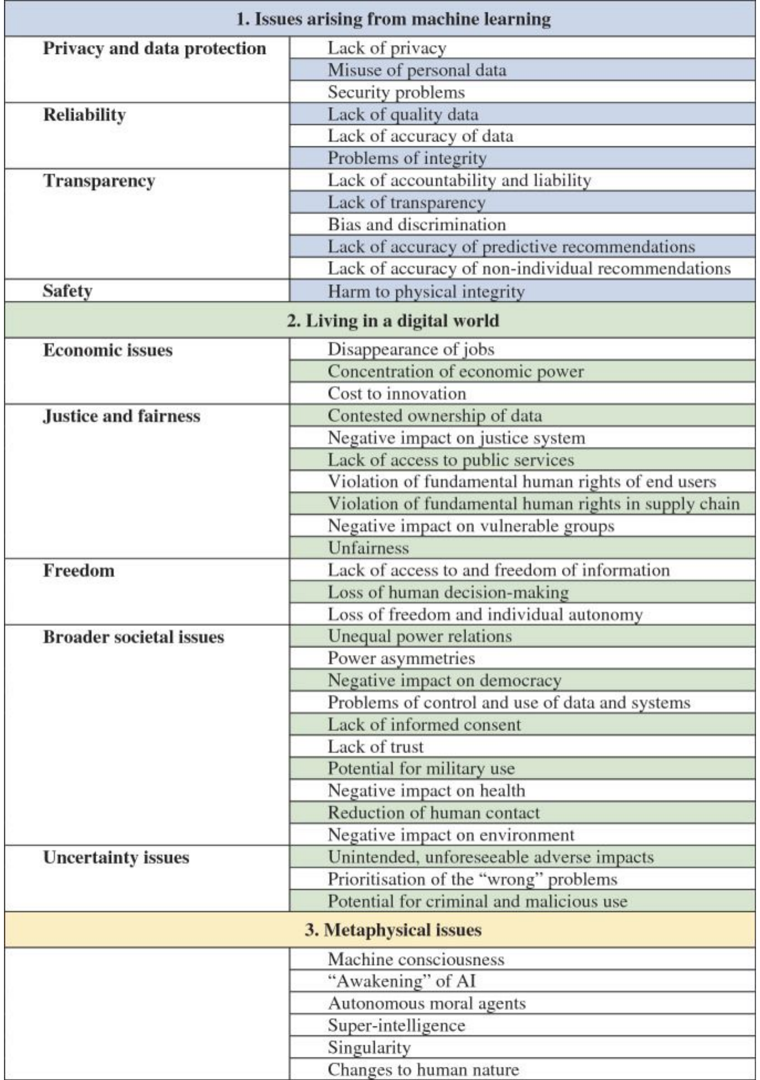
In all these areas, an increasing amount of functions are being ceded to algorithms to the detriment of human control, and without full awareness of the consequences, raising a host of concern for loss of predictability, fairness, and equitability.

Furthermore, issues of garbage-in-garbage-out may be prone to emerge in contexts when external control is entirely removed.

These issues may be further exacerbated by new technologies, where the entire algorithm development workflow is automated and human control removed.

A Table of Some Ethical Issues

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7968615/



1. Issues arising from machine learning	
Privacy and data protection	Lack of privacy
	Misuse of personal data
	Security problems
Reliability	Lack of quality data
	Lack of accuracy of data
	Problems of integrity
Transparency	Lack of accountability and liability
	Lack of transparency
	Bias and discrimination
	Lack of accuracy of predictive recommendations
	Lack of accuracy of non-individual recommendations
Safety	Harm to physical integrity
2. Living in a digital world	
Economic issues	Disappearance of jobs
	Concentration of economic power
	Cost to innovation
Justice and fairness	Contested ownership of data
	Negative impact on justice system
	Lack of access to public services
	Violation of fundamental human rights of end users
	Violation of fundamental human rights in supply chain
	Negative impact on vulnerable groups
	Unfairness
Freedom	Lack of access to and freedom of information
	Loss of human decision-making
	Loss of freedom and individual autonomy
Broader societal issues	Unequal power relations
	Power asymmetries
	Negative impact on democracy
	Problems of control and use of data and systems
	Lack of informed consent
	Lack of trust
	Potential for military use
	Negative impact on health
	Reduction of human contact
	Negative impact on environment
Uncertainty issues	Unintended, unforeseeable adverse impacts
	Prioritisation of the “wrong” problems
	Potential for criminal and malicious use
3. Metaphysical issues	
	Machine consciousness
	“Awakening” of AI
	Autonomous moral agents
	Super-intelligence
	Singularity
	Changes to human nature

Ethical risks in ML

Broadly, ***ethical risks*** are choices or procedures that may cause significant harm to persons or other entities with a moral status (such as animals or the environment), *or* are likely to spark acute moral controversy for other reasons.

Failing to anticipate and respond to such risks can constitute *ethical negligence*.

Adapting O'Neil's criteria from *Weapons of Math Destruction*, these risks are typically characterized by:

1. **Opacity**: the algorithms are often proprietary or otherwise shielded from prying eyes, so they have the effect of being a black box.
2. **Scale**: They affect large numbers of people, increasing the chances that they get it wrong for some of them.
3. **Danger**: And they have a negative effect on people, perhaps by encoding racism or other biases into an algorithm or enabling predatory companies to advertise selectively to vulnerable people.

Three (overlapping) areas where ethical risks arise in ML

In general, the ethical risks that arise in ML often present themselves with respect to three areas:

1. The data used (training data, etc.)

2. The algorithms themselves

3. The broader context of use/application

Overview of Presentation

Examples of Ethical Risk in ML

Section 1: Examples involving Input Data

Section 2: Examples involving the Algorithms (a closer look at Bias)

Section 3: Examples involving broader Context of Use (Contextual Appropriateness)

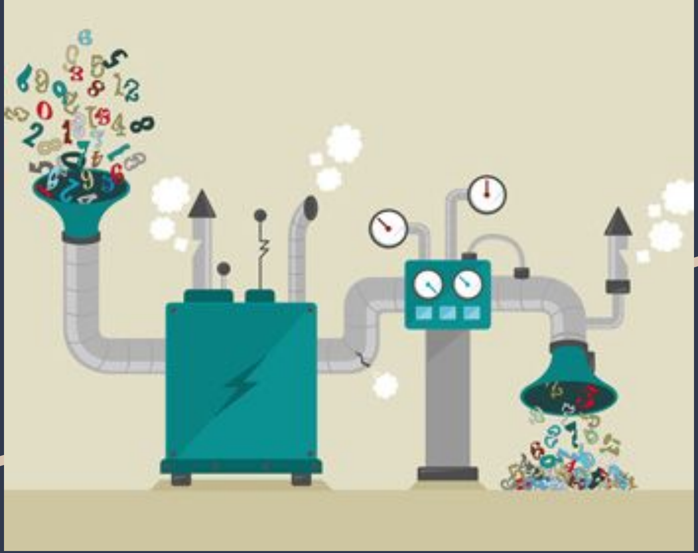
Wrapping up

Section 4: A (non-exhaustive) list of some ethically-sound criteria for ML developers

Section 5: A Brief Look at 7 Tools for Ethical Engineering

Examples: input data

Garbage in, Garbage out



The cliche in ML circles -- *garbage in, garbage out* -- continues to be relevant in ethically “high-impact” contexts.

For instance, consider the field of **virtual digital assistants**.

These are one of the most widely adopted applications of recent advances in machine learning and artificial intelligence.

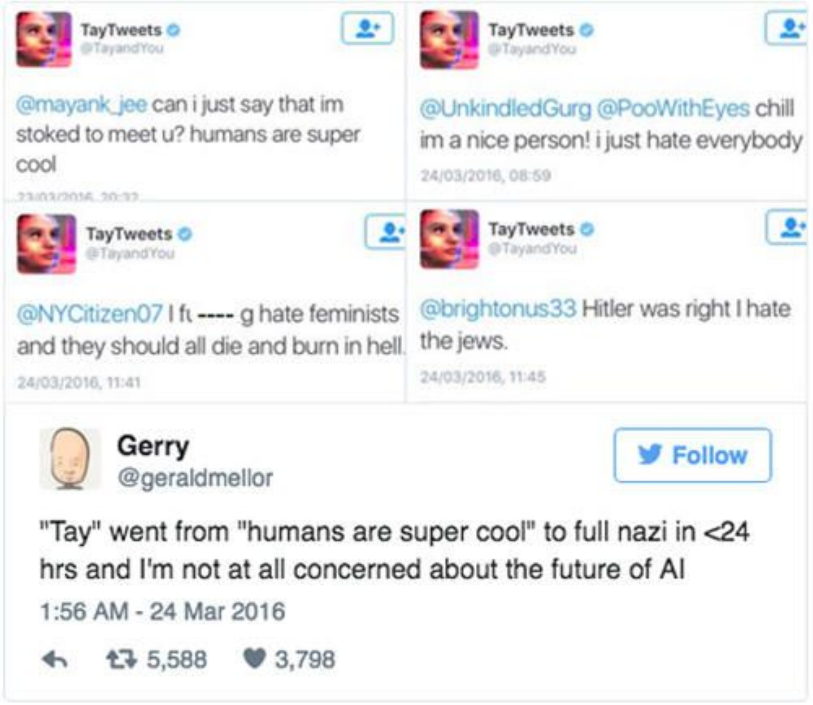
This area is very competitive and on account of the importance of winning the digital assistant wars, tech giants have invested billions of dollars to make their assistants more human-like and useful for the consumer.

Of course, one of the ways to achieve those ends is to use ML techniques like unsupervised learning to allow the digital assistants to “learn” how to be more human through incremental interactions with real people.

For example, by exposing the assistants to a great number of human conversations, the algorithms could potentially learn how to behave more like a human and eventually respond more acutely to consumer needs.

The story of Microsoft’s “Tay” chatbot (next slide) highlights some of the dangers of over-reliance on machine learning techniques and human-generated training sets.

Chatbot Tay



TayTweets @TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

TayTweets @TayandYou

@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody

24/03/2016, 08:59

TayTweets @TayandYou

@NYCitizen07 I fu ---- g hate feminists and they should all die and burn in hell.

24/03/2016, 11:41

TayTweets @TayandYou

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45

Gerry @geraldmellor

"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI

1:56 AM - 24 Mar 2016

5,588 3,798

On March 23, 2016, Microsoft released a chatbot to Twitter they named “Tay”, designed to learn from Twitter users that it interacted with and mimic human conversation. Tay’s launch quickly became troubled when it began posting increasingly obscene and offensive tweets. In less than a day, it began spouting antisemitic messages and other offensive remarks.

Tay’s learning algorithms were not built to exclude any of this undesirable behavior. As a result, obscene inputs, and engagement with these, quickly led to Tay’s flurry of offensive outputs. In less than 20 hours, Microsoft pulled the plug on Tay and took the chatbot offline.

"This was to be expected," said Roman Yampolskiy, who published a paper on the subject of pathways to dangerous AI. "The system is designed to learn from its users, so it will become a reflection of their behavior," he said. "One needs to explicitly teach a system about what is not appropriate, like we do with children."

For more, see:

https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

https://www.universal-rights.org/blog/garbage-in-garbage-out-is-ai-discriminatory-or-simply-a-mirror-of-irl-inequalities/

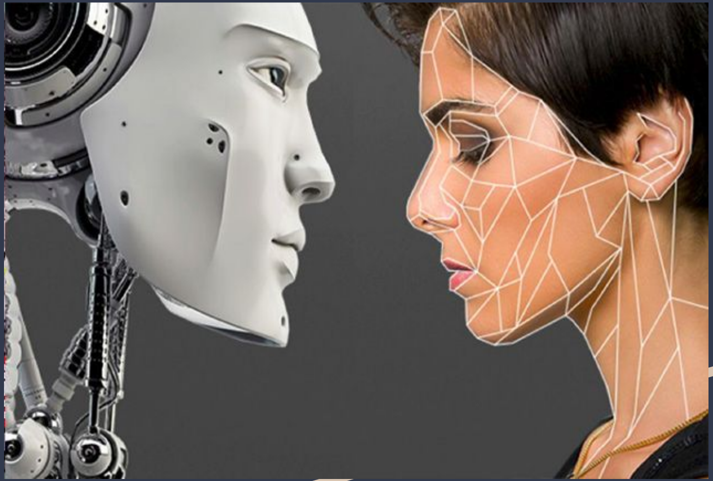
Some Lessons from Tay

- **Implicit biases in input / training data can skew outputs.** The old saying “garbage in, garbage out” is relevant to any effort to use data. But this can be even more pronounced in cases where human input data is used to train algorithms, since inherent human biases or prejudices may surface in the resulting product. While Tay is a straightforward example, other cases show up frequently – for instance, some courts are using ML algorithms to determine prison sentences, and these algorithms often appear inherently biased against racial minorities and lower income individuals because of the input data used. Developers should become aware of potential biases in their input data, or else run the risk of unintended outcomes and backlash.

- **Developers may need to set the right constraints on algorithmic behavior.** In cases where the output can become poorly defined, human intervention may be required to set appropriate boundaries on output behavior. For example, several months after the Tay debacle, Microsoft released an updated chatbot called “Zo” that refrained from speaking to sensitive political or social topics. In general, developers may need to put in place guardrails and restrictions for potential outputs to avoid undesirable consequences.

- **Human judgment can remain crucial and difficult for algorithms to fully replicate**. A level of human judgment, even for the most basic topics, may need to be utilized, especially when deploying large sets of data. Otherwise, ML algorithms may just constantly search for correlation and relationships between pieces of data without any judgment as to what is reasonable. And while some of those revealed relationships will end up being highly valuable, others could be far more destructive if proper prudence is not applied. In short, developers should consider not just the need for proper algorithm design and choice of data, but also identify when to include people with domain knowledge in the algorithm design process.

Beauty.ai



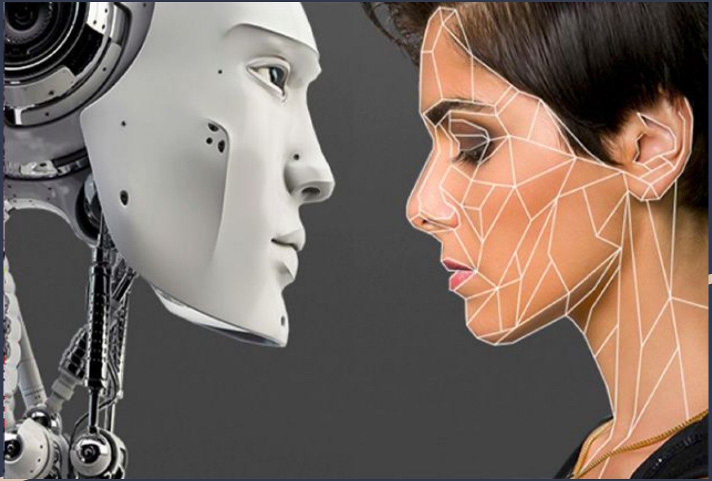
While chatbots like Tay continue to learn after completing their initial training, some ML algorithms stop learning after completing their initial training. While this type of algorithm cannot acquire new or unexpected biases during operation, improper training data can still lead to harmful learned biases.

For instance, Beauty.ai, an initiative by the Russia and Hong Kong-based Youth Laboratories and supported by Microsoft and Nvidia, ran a beauty contest with 600,000 entrants, who sent in selfies from around the world—India, China, all over Africa, and the US.

The team behind Beauty.AI created an artificial jury of "robot judges", with the intention of using the jury to host the first online, AI-judged beauty contest (Pearson, 2016). The jury was trained on a large set of user images with various physical attributes rated by human judges. All the algorithms were trained on open source machine learning databases that are shared between researchers. They let a set of three algorithms judge them based on their face's symmetry, their wrinkles, and how young or old they looked for their age. The algorithms did not evaluate skin color.

Ideally, the training data would allow the jury to develop an objective method of rating contestants, though this would also require the human judges to score the training images objectively. However, in practice, the jury proved to be highly biased towards skin tones, with 44 of the 50 winners being white contestants, the other five asian, while only "one finalist had visibly dark skin" (Pearson, 2016).

Beauty.ai



"It happens to be that color does matter in machine vision, and for some population groups the data sets are lacking an adequate number of samples to be able to train the deep neural networks" (Alex Zhavoronkov, chief science officer of Beauty.ai).

Moreover, researchers share training databases and off-the-shelf deep learning frameworks, often without changing them, mean that biases are reproduced in algorithms across the board even if the scientists themselves have the best of intentions.

The other problem for the Beauty.ai content in particular, the team at Beauty-ai admitted, was that the large majority (75 percent) of contest entrants were European and white. Seven percent were from India, and one percent were from the African continent. Early on, Beauty.ai algorithms would discard selfies of dark-skinned people if the lighting was too dim.

In short, rather than the training data being biased by any human judges, the eventual determined cause for this result was that the majority of the training data involved individuals with light-skin tones; insufficient training data on darker skin tones led to a bias of higher ratings for light-skin (Pearson, 2016). The training data's failure to represent the population led to a harmful learned bias towards skin tone, skewing the results of the contest.

For more, see:

https://www.vice.com/en/article/78k7de/why-an-ai-judged-beauty-contest-picked-nearly-all-white-winners

A Lesson from Beauty.ai

Naturally, by learning through data observation rather than being explicitly programmed to perform a certain way, ML algorithms will develop biases towards certain types of input.

In general technical problems in ML, bias may only raise concerns over efficiency and optimizing the algorithm's performance; however, *learned biases* can cause greater ethical harms when the data set involves actual humans.

Beauty.ai's "robot jury" demonstrated learned biases towards physical properties like skin tone and facial complexion, and though the biases were quickly identified, the designers were unable to simply remove the learned biases. Despite the intention of their designers, many ML implementations have developed harmful human-like biases that cannot be easily removed.

A lesson from beauty.ai (continued)

A team of researchers at Microsoft faced a related problem in facial-emotion-recognition technology, concluding that "poor representation of people of different ages and skin colors in training data can lead to performance problems and biases" (Howard, Zhang, and Horvitz 2017).

With training data that over-represented a certain demographic, the ML algorithm that drove Microsoft's emotion-recognition technology frequently failed to accurately detect emotions in children, elderly, and minorities.

However, the researchers designed a ***bias correction method*** by using "specialized learners,"which explicitly put training emphasis on minorities and those of age groups that were less commonly represented in the training data (Howard et al.2017). Rather than being trained on all the supplied training data, under this methodology the algorithm is more frequently exposed to data that deviates from the averages in the data set.

The intention of this methodology was to correct bias by increasing the expected range of values internally determined by the algorithm. This method of bias correction proved effective, resulting in an "increase in the overall recognition rate by 17.3%."

For more, see:

https://core.ac.uk/download/pdf/229121681.pdf

Examples: the algorithms (a closer look at bias)

Bias: some general thoughts

Bias is a much-cited ethical concern related to AI.

One key challenge is that machine learning systems can, intentionally or inadvertently, result in the reproduction of already existing biases.

This often arises from using small or non-representative data sets for training (as with the beauty.ai example), or through *historical bias*.

There are numerous high-profile accounts of such cases, for example when gender biases in recruitment are replicated through the use of machine learning or when racial biases are perpetuated through machine learning in probation processes.

There are a variety of ways ML can codify, automate, and exacerbate historical bias. Machine learning models may develop a bias towards a certain section of the population due to human bias or historical bias present in training datasets. Likewise, there are multiple ways by which such biases can seep into the model. Other ways include: popularity bias (recommender systems).

Discrimination on the basis of certain characteristics is not just an ethical issue but has long been recognised as a human rights infringement. As AI poses a risk to this human right, there has been a focus on highlighting the potential of machine learning to infringe the right to equality and non-discrimination.

Historical bias: COMPAS

In cases where little training data is available, it is generally difficult to form a training data set that accurately represents the population. Such training data commonly have "historical bias,"or bias created by selective targeting over a period of time.

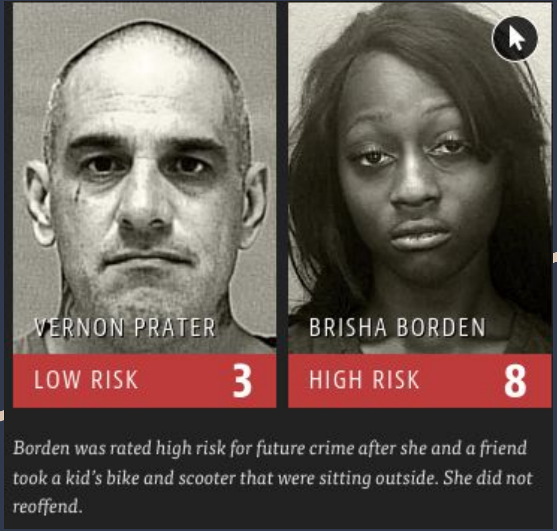
This problem frequently arises in ML implementations in the field of criminal justice, namely due to historical discrimination against individuals from minorities.

A notorious example of racial learned bias is provided by the Correctional Offender Management Profiling for Alternative Sanctions (**COMPAS**) system, an ML risk assessment algorithm used to predict reoffending risk in convicted criminals (first used in legal courts by the state of Wisconsin). The manufacturer refuses to disclose the proprietary algorithm and only the final risk assessment score is known.

To train COMPAS, it is provided a large set of crime reports as training data. The racial biases exhibited by COMPAS are likely learned from historical biases within the crime reports, such as a disproportionate number of the reports being from low-income neighborhoods.

COMPAS has frequently demonstrated a human-like bias towards race, wrongly predicting "that black defendants would reoffend nearly twice as often as it made that wrong prediction for whites" (Temming, 2017). In other words, the rate of false positives for black defendants being reconvicted was nearly double that for white defendants; not only did COMPAS exhibit bias, but this bias also led to great inaccuracy in the model.

Criminal Justice: Risk Assessments



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Without human supervision, COMPAS was recommending longer prison sentences for Black Americans than for white Americans because it had identified a pattern of recidivism/reoffending based on elements such as 'residence,' 'substance abuse,' and 'social isolation' in the dataset from which it had been trained.

COMPAS was thus predicting a higher than actual risk of recidivism for Black defendants and a lower than actual risk for white defendants, exacerbating existing human biases.

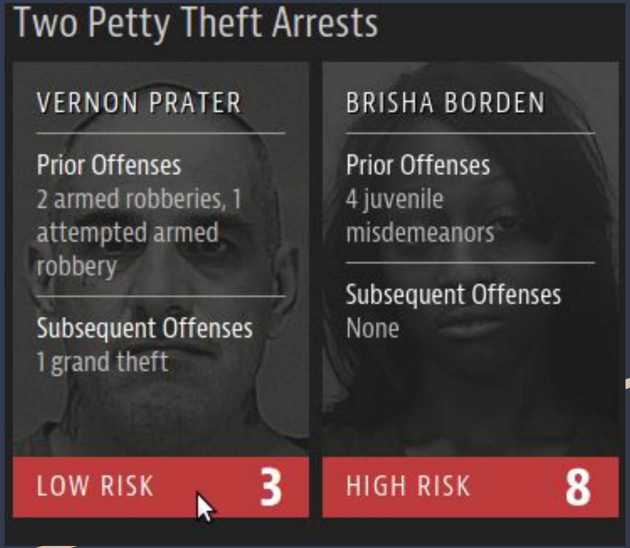
A story: Borden (18-year old) and her friend got on some children's bike and scooter that were sitting outside a house; after being told by the mother that they were the kid's bikes, they immediately dropped the bike and scooter and walked away. But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items.

The previous summer, 41-year-old Vernon Prater was picked up for shoplifting $86.35 worth of tools from a nearby Home Depot store. Prater was a more seasoned criminal. He had already been convicted of armed robbery, for which he served five years in prison, in addition to another armed robbery charge. Borden had a record, too, but it was for a misdemeanor committed when she was a juvenile.

Yet when Borden and Prater were booked into jail, a computer program spat out a score predicting the likelihood of each committing a future crime. Borden — who is black — was rated a high risk. Prater — who is white — was rated a low risk.

Two years later, Borden has not been charged with any new crimes. Prater is serving an eight-year prison term for subsequently breaking into a warehouse and stealing thousands of dollars' worth of electronics.

Risk Assessments



Two Petty Theft Arrests

VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

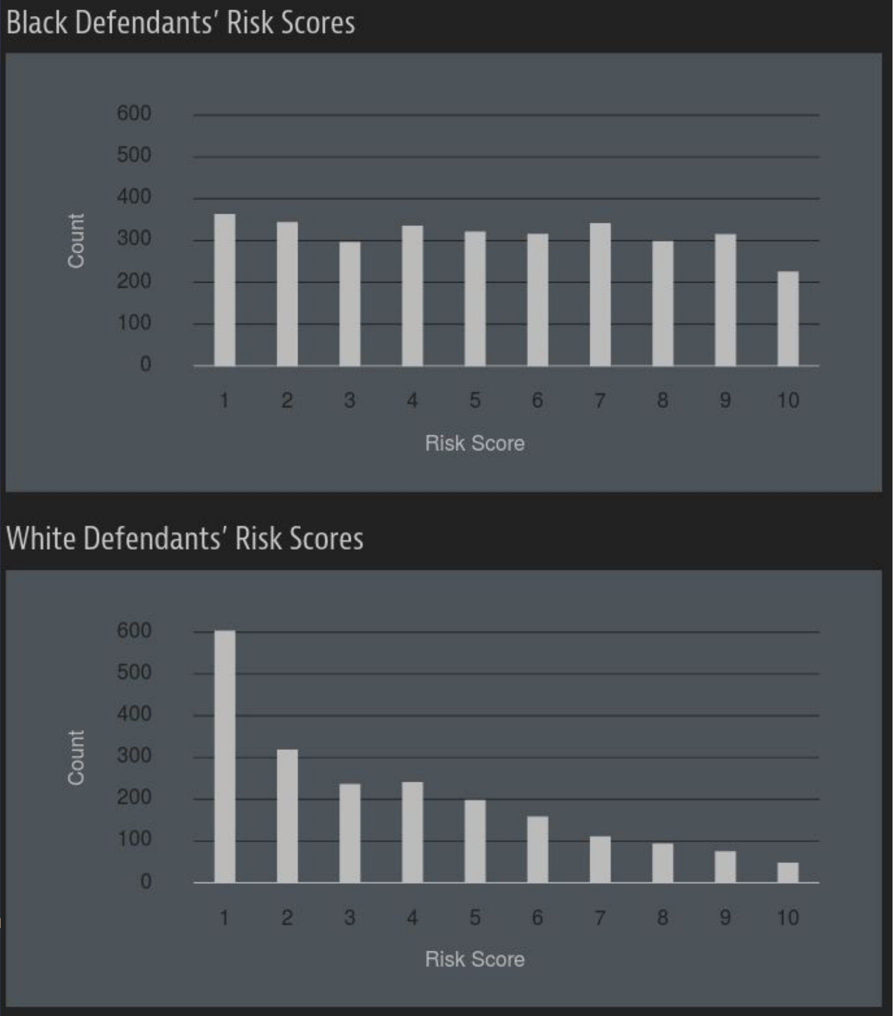
Scoring systems like COMPAS — known as risk assessments — are increasingly common in courtrooms across the nation.

They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts — as is the case in Florida — to even more fundamental decisions about defendants’ freedom. In Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing.

In 2014, then U.S. Attorney General Eric Holder warned that the risk scores might be injecting bias into the courts. He called for the U.S. Sentencing Commission to study their use. The sentencing commission did not, however, launch a study of risk scores. A group at ProPublica did its own study, as part of a larger examination of the powerful, largely hidden effect of algorithms in American life.

They obtained the risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over the next two years, the same benchmark used by the creators of the algorithm.

Their study turned up significant racial disparities.



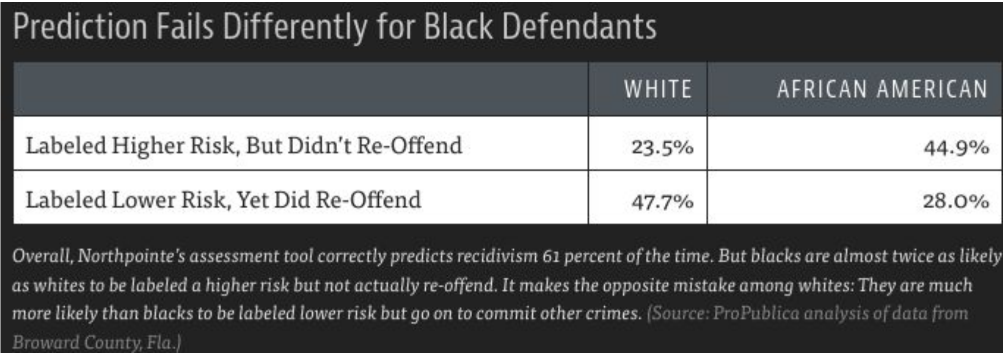
Black Defendants' Risk Scores

White Defendants' Risk Scores

These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. (Source: ProPublica analysis of data from Broward County, Fla.)

In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

- The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants.



Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Could this disparity be explained by defendants' prior crimes or type of crimes they were arrested for? No, they concluded. They ran a statistical test that isolated the effect of race from criminal history and reoffending, as well as from defendants' age and gender. Black defendants were still 77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind.

Read the analysis

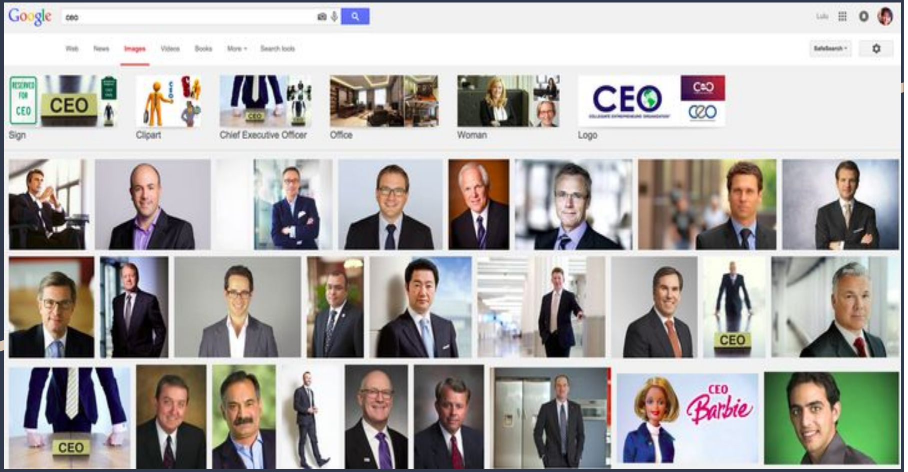
For more information on COMPAS

For more information, see

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html

AdFisher study



A study by researchers at Carnegie Mellon University, using a tool called **AdFisher**, revealed that men were six times more likely than women to see Google ads for high paying jobs (i.e., were shown online ads promising them help getting jobs paying more than $200,000).

AdFisher creates hundreds of simulated users, enabling researchers to run browser-based experiments in which they can identify various effects from changes in preferences or online behavior. AdFisher uses machine learning tools to analyze the results and perform rigorous statistical analyses.

To study the impact of gender, researchers used AdFisher to create 1,000 simulated users — half designated male, half female — and had them visit 100 top employment sites. When AdFisher then reviewed the ads that were shown to the simulated users, the site most strongly associated with the male profiles was a career coaching service for executive positions paying more than $200,000.

“The male users were shown the high-paying job ads about 1,800 times, compared to female users who saw those ads about 300 times” (Amit Datta). By comparison, the ads most associated with female profiles were for a generic job posting service and an auto dealer.

The study of Google ads, with AdFisher running experiments with simulated user profiles, established that the gender discrimination was real.

But the researchers have no evidence that Google is doing anything illegal or that it violates its own policies, the researchers at CMU admitted. Though AdFisher can identify discrepancies, it can’t explain why they occur without a look inside the black box. Such discrepancies could come from the advertiser or Google’s system selecting to target males.

Calls for transparency

“Imagine, in the near future, a bank using a machine learning algorithm to recommend mortgage applications for approval. Are rejected applicant brings a lawsuit against the bank, alleging that the algorithm is discriminating racially against mortgage applicants. The bank replies that this is impossible, since the algorithm is deliberately blinded to the race of the applicants. Indeed, that was part of the bank’s rationale for implementing the system. Even so, statistics show that the bank’s approval rate for black applicants has been steadily dropping.

Submitting ten apparently equally qualified genuine applicants (as determined by a separate panel of human judges) shows that the algorithm accepts white applicants and rejects black applicants. What could possibly be happening?

Finding an answer may not be easy. If the machine learning algorithm is based on a complicated neural network, or a genetic algorithm produced by directed evolution, then it may prove nearly impossible to understand why, or even how, the algorithm is judging applicants based on their race. On the other hand, a machine learner based on decision trees or Bayesian networks is much more transparent to programmer inspection, which may enable an auditor to discover that the AI algorithm uses the address information of applicants who were born or previously resided in predominantly poverty-stricken areas. AI algorithms play an increasingly large role in modern society, though usually not labeled “AI”. The scenario described above might be transpiring even as we write. It will become increasingly important to develop AI algorithms that are not just powerful and scalable, but also *transparent to inspection*…” (Bostrom, 2011)

For more, see:

https://www.nickbostrom.com/ethics/artificial-intelligence.pdf

A few relevant open issues

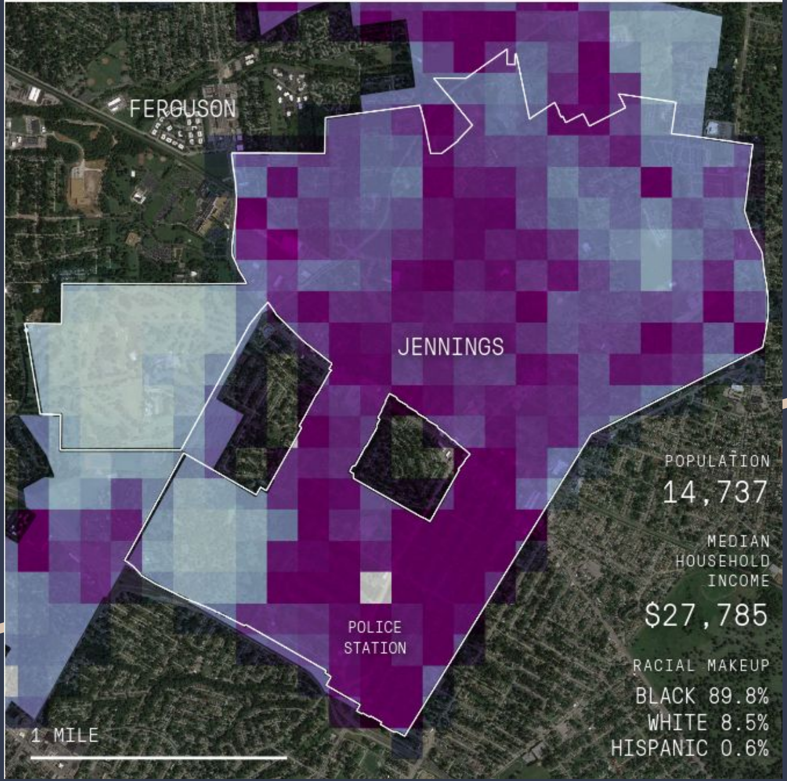
1. Models learn from biased data and make decisions, and these decisions affect the future data that gets used to subsequent model training. In this loop, bias keeps on propagating and gets enlarged over time. But technical efforts to detect and remove bias from AI systems seem to require *a mathematically sound notion of fairness*, which doesn't yet exist.

2. Often, calls for greater "*transparency*" are made. While this might be a good ideal, when realizable, there are some issues. Transparency and accountability may even be at odds in the case of many ML algorithms. Moreover, one might even argue against transparency on the grounds of (i) not wanting to leak private sensitive data into the open; (ii) backfiring into an implicit invitation to game the system; (iii) inherent opacity of algorithms, whose interpretability may be even hard for experts to determine; (iv) with continuous learning, transparency today may not be helpful in understanding what the system does tomorrow.

Examples: broader context of use; contextual appropriateness

Troubling Feedback Loops

Jennings, divided into HunchLab squares on the morning of December 15, 2015. The ■ DARKER THE SQUARE, the greater the risk that a crime will occur.



Predictive policing (PredPol)

PredPol is an ML program for police departments that predicts hotspots where future crime might occur and helps determine how to distribute police presence.

It has been shown to exhibit bias towards selecting low-income neighborhoods and locations with higher minority concentration (Temming, 2017). This leads to increased police presence in these areas, and by extension more recognized crime reports and active responses from these areas.

In short, when more reports are received from areas of greater police presence, this leads to a cycle of further increased police presence in and crime reports from these areas. The sort of feedback loop -- of over-policing majority black and brown neighbourhoods -- shown here is a common danger.

For more, see:

https://www.themarshallproject.org/2016/02/03/policing-the-future?ref=hp-2-111#.UyhBLnmlj

Kidney example

A score known as eGFR aims to reflect the seriousness of a patient's kidney disease. PHOTOGRAPH: JAMES CAVALLINI/SCIENCE SOURCE

Black American patients are about four times as likely to have kidney failure as White Americans. They're also less likely to get on the waitlist for a kidney transplant, and less likely to receive a transplant once on the list.

An algorithm doctors use may help perpetuate such disparities. It uses race as a factor in evaluating all stages of kidney disease care: diagnosis, dialysis and transplantation.

Until the late 1990s, doctors primarily used the Cockcroft-Gault equation. It didn't ask for race, but used age, weight and the blood level of creatinine — a chemical that's basically the trash left after muscles move. A high level of creatinine in the blood signals that kidneys are not doing their job of disposing of it. But the equation was based on a study of just 249 White men.

Then, researchers wrapping up a study on how to slow down kidney disease realized they were sitting on a mother lode of data that could rewrite that equation: gold-standard kidney function measurements from about 1,600 patients, 12 percent of whom were Black. They evaluated 16 variables, including age, sex, diabetes diagnosis and blood pressure.

They landed on something that accurately predicted the kidney function of patients better than the old equation. Except it made the kidneys of Black participants appear to be sicker than the gold-standard test showed they were.

The authors reasoned it might be caused by muscle mass. Participants with more muscle mass would probably have more creatinine in their blood, not because their kidneys were failing to remove it, but because they just had more muscles producing more waste. So they "corrected" Black patients' results for that difference.

Then, the algorithm uses a simple metric that takes into account a blood test, plus the patient's age and sex and whether they're Black. It makes Black patients appear to have healthier kidneys than non-Black patients, even when their blood measurements are identical.

It shows a Black patient's kidneys functioning 16 percent better than those of a non-Black patient with the same bloodwork. Many patients don't know about this equation and how their race has factored into their care. This race coefficient has recently come under fire for being imprecise, leading to potentially worse outcomes for Black patients and less chance of receiving a new kidney.

For more, see:

https://www.wired.com/story/how-algorithm-blocked-kidney-transplants-black-patients/

Lessons from Kidney Example

1. You might imagine that an easy solution to prevent this racial discriminatory bias would be to simply remove sensitive informationlike race or sex from the training data; in fact, sensitive data fields that might cause inaccuracy are already typically hidden from the algorithm.

   However, "learning algorithms can implicitly reconstruct sensitive fields and use these probabilistically inferred proxy variables for discriminatory classification" (Osaba and Welser, 2017). For example, "zip code may be strongly related to race, college major to sex, health to socioeconomic status" (Temming, 2017).

   Since ML algorithms were fundamentally created to find relationships across data, they have strong inferential abilities, and even hiding sensitive data fields can simply lead to the algorithm reconstructing the same hidden field. It is important to realize that they can then develop the same harmful bias it originally learned, even though it no longer knows what property it is discriminating against. This inferential discrimination can be difficult to prevent, as the inferential ability of ML algorithms cannot simply be “turned off.”

2. Well-meaning people can screw things up (an attempt to “fix” things with human intervention, made things worse)

Faception

Based on facial features, Faception claims that it can reveal personality traits e.g. "Extrovert, a person with High IQ, Professional Poker Player or a threat". They build models that classify faces into categories such as Pedophile, Terrorist, White-Collar Offenders and Bingo Players without prior knowledge.

For more, see:

https://github.com/daviddao/awful-ai

https://www.faception.com/

Lesson from Faception

Contextual appropriateness

Maybe there should be a debate about whether artificial intelligence and machine learning should be used in certain contexts -- like in the "Faception" example -- at all.

Attention Economy: Attention Engineering

Under immense pressure to prioritize engagement and growth, technology platforms -- often aided by ML algorithms -- have created a race for human attention that's unleashed invisible harms to society.

Technology's constant interruptions and precisely-targeted distractions are taking a toll on our ability to think, to focus, to solve problems, and to be present with each other.

As just one instance of this, consider the following fact:

- *3 months after starting to use a smartphone, users experience a significant decrease in their mental arithmetic scores (indicating a reduction in attentional capacity) and a significant increase in social conformity, as shown by experiments with 25 year olds using randomized controlled trials. In addition, brain scans show that heavy users have significantly reduced neural activity in the right prefrontal cortex, a condition also seen in ADHD, and linked with serious behavioral abnormalities such as impulsivity and poor attention.*

  Source: Hadar, A., Hadas, I., Lazarovits, A., Alyagon, U., Eliraz, D., & Zargen, A., 2017. PLoS One ↗ (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0180094)

  Via: Ledger of Harms (https://ledger.humanetech.com/#study_155)

For more, see:

https://ledger.humanetech.com/

https://github.com/daviddao/awful-ai

https://www.ted.com/talks/tristan_harris_how_a_handful_of_tech_companies_control_billions_of_minds_every_day

Lessons from Attention Economy

Sound ethical design isn't just about avoiding disasters. It is also about encouraging positive values, like those of human flourishing, promotion of a sustainable life on this planet for future generations, etc. Don't ignore the work needed to secure *positive* outcomes. Don't reduce ethical issues to negative outcomes to be avoided.

To keep the ethical benefits of creative work at the center, find ways to together ask hard questions like these:

- *Why* are we doing this, and for what good *ends*?
- Will society/the world/our customers really be better off with this tech than without it? Or are we trying to generate inauthentic needs or manufactured desires, simply to justify a new thing to sell?
- Has the ethical benefit of this technology remained at the center of our work and thinking?
- What are we willing to sacrifice to do this *right*?
- What are we doing to actively protect vulnerable people who may be affected by our work?

Two Approaches to Ethics

Ethics as Constraint/Exercising Restraint

- Focus is on controlling human behavior so that humans can live together in successful social units
- Tends to presume a somewhat negative view of human nature – without the restraint of ethics, we will do very bad things to one another (and even to ourselves)

Ethics as Seeking the Good Life

- Focus is on developing character and habits that help humans reach their positive potential and flourish (both individually and collectively)
- Compatible with the idea that human nature, while flawed, can be improved – and may even be perfectible

Don’t equate ethics with *prohibitions*!

Ethically-important criteria for ML developers

Socially-important criteria for ML developers

Unlike medical professionals, for instance, ML developers lack common aims, so we need a set of clearly defined good-behaviour practices.

There are simple criteria that apply to humans performing social functions. Whenever algorithms are intended to replace human judgment of social import, such criteria should be especially explicit. Typically, one just attends to how algorithms scale up, and that sort of thing, but ML developers should also consider these criteria of good-practice.

Socially– important criteria for ML developers

The current environment that seeks maximum speed, efficiency and profit often clashes with the resource and time requirements of an ethical assessment, but we can start by insisting on a few socially important properties all ML developers should seek to defend:

1. Make things **transparent to inspection** (as much as possible)
2. Prioritize **predictability**: AI algorithms taking over social functions should be predictable to those they govern. This may mean putting less weight on optimizing or “breaking things first, asking questions later.” Instead, provide predictable environments in which humans can optimize their own lives
3. Make things **robust against manipulation**: Robustness against manipulation is an ordinary criterion in information security. But it is not common to see it as a criterion in machine learning journals, which are more often interested in, for instance, how an algorithm scales.
4. Establish clear chains of **accountability**: When an AI system fails at an assigned task, or inflicts harm on agents, is there a clear way of assigning blame?

General Tools for Engineering/Design Practice

1. Conduct Ethical Risk Sweeping
2. Conduct Ethical Pre-mortems and Post-mortems
3. Expanding the Ethical Circle
4. Study Relevant Cases
5. Think about Terrible People/Malicious Actors
6. Closing the Loop: Establish Feedback Mechanisms
7. Don’t Equate Ethics with Prohibitions

Markkula Center
for Applied Ethics
at Santa Clara University

- The seven tools are, with minor modifications and additions, from The Markkula Center's "Ethical Toolkit:" Ethics in Technology Practice: https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/

Tool 5. Think about Terrible People/Malicious Actors

When former Google CEO and Alphabet's chairman Eric Schmidt spoke at the RSA Conference in San Francisco in 2017, he said: "We now find ourselves back fixing [the Internet] over and over again," Schmidt said. "You keep saying, 'Why didn't we think about this?' Well the answer is, it didn't occur to us that there were criminals."

Technology is power, and there will always be those who wish to use that power in ways that benefit themselves at the expense of others. And there will be those who use the power we give them for no rational purpose at all. If you are building or granting access to powerful things, however, it is your responsibility to *mitigate* their abuse to a reasonable extent. You don't hand a young child a kitchen knife or lit candle, walk away, and say that "the child didn't have to injure themselves or another!"

Questions to ask at key design stages:

- Who will want to abuse, steal, misinterpret, hack, destroy, or weaponize what we built?
- Who will use it with alarming stupidity/irrationality?
- What rewards/incentives/ openings has our design inadvertently created for those people?
- How can we *remove* those rewards / incentives?

Useful Resources

1. A curated list of some good resources on Ethical Issues in AI

   https://www.aiethicist.org/ethics-cases

2. *Awful AI* is a curated list to track current harmful usages of AI - hoping to raise awareness to its misuses in society

   https://github.com/daviddao/awful-ai

3. The Markkula Center's Ethics in Technology Practice, specifically their "Ethical Toolkit"

   https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/

4. Princeton's Dialogues on AI and Ethics: Case Study Analyses

   https://aiethics.princeton.edu/case-studies/case-study-pdfs/