Queen Mary
University of London

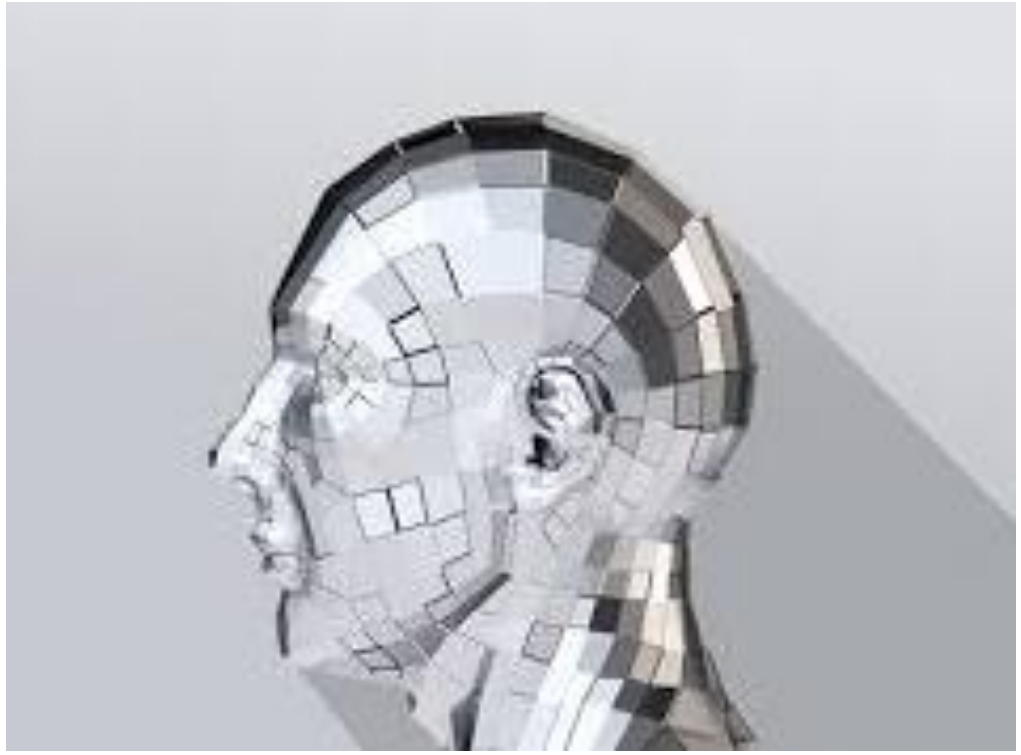
Biases in AI
Lecture 5

Learning outcomes of the session

- **Understand the inherent nature of human bias**
- **Understand the different types of biases**
- **Understand algorithmic bias**
- **Be familiar with the Alan Turing guidelines and research on counterfactual fairness**
- **Can the world ever be fair, and can machines be ever fair???**

Why should a machine decide my future !!

- **Algorithms are increasingly assisting in life-changing decisions, such as in parole hearings, loan applications, and university admissions. However, if the data used to train an algorithm contains societal biases against certain races, genders, or other demographic groups, then the algorithm will too. (The Alan Turing Institute)**



Human beings are consistently, routinely, and profoundly biased.

- People in supermarkets buy more French wine when French music is playing in the background, and more German wine when the music is German?
- White National Basketball Association (NBA) referees have been found to call more fouls on black players, and black referees call more fouls on white players?
- Scientists have been found to rate potential lab technicians lower, and plan to pay them less, if the potential technicians are women?
- Doctors treat patients differently when the patients are overweight, and that patients treat doctors differently when the doctors are overweight?

Human Bias

All of these behaviors, and many more, happen without people realizing they are happening, and that these behaviors are demonstrations of biases, biases people don’t even know they possess.

Even more worryingly these Biases occur without people knowing why they occur.

Therefore, it is almost impossible to prevent some of these biases influencing everything we do, without us even realizing that they are doing so.

Examples Of Human Bias

Selective Attention, also known as *inattentional blindness*, is a mental process through which we selectively see some things but not others. Selective attention is helpful to us, and important to maintaining our ability to function in a world in which we are constantly being bombarded with stimuli.

https://youtu.be/UtKt8YF7dgQ

Discuss:

1. **Could this kind of bias be transferred to an AI ? If so, how ?**
2. **What are the dangers posed by such biases ?**

- **Trafton Drew is an attention researcher at Harvard University Medical School who took the gorilla experiment to a whole new level.**

- **Drew decided to check if the inattentional blindness effect would be true when it involved people who were extremely intelligent and highly trained in observation.**

- **He developed a research protocol in which he superimposed a small picture of somebody in a gorilla suit shaking its fist into a lung X-ray (see figure). The gorilla was roughly the size of a book of matches or forty-eight times larger than an average cancer nodule!**

- **He then asked a team of highly trained radiologists and a group of unskilled observers to examine the image for nodules**

- **More than 80% of radiologists and 100% of unskilled observers, said they had seen nothing**

How do we define bias?

- Bias is a disproportionate weight *in favor of* or *against* an idea or thing, usually in a way that is disproportionate or unfair. Biases can be innate or learned.
- People may develop biases for or against an individual, a group, or a belief.

- In science and engineering, a bias is a systematic error. Statistical bias results from unfair sampling of a population or from an estimation process that does not give accurate results on average.
- People are often **biased** against others outside of their own social group, showing **prejudice** (emotional **bias**), stereotypes (cognitive **bias**), and discrimination (behavioral **bias**).

Bias

- In the past, people used to be more explicit with their biases, but during the 20th century, when it became less socially acceptable to exhibit bias, such things like prejudice, stereotypes, and discrimination became more subtle (automatic, ambiguous, and ambivalent).
- Even in one's own family, everyone wants to be seen for who they are, not as "just another typical X." But still, people put other people into groups, using that label to inform their evaluation of the person as a whole—a process that can result in serious consequences. (Dovidio & Gaertner, 2010; Fiske, 1998).
- **For example, sometimes people have a negative, emotional reaction to a social group (prejudice) without knowing even the most superficial reasons to dislike them (stereotypes).**

Social dominance orientation

- This is the belief that group hierarchies are inevitable in all societies and are even a good idea to maintain order and stability (Sidanius & Pratto, 1999).
- Those who score high on SDO believe that some groups are inherently better than others, and because of this, there is no such thing as group "equality."
- At the same time, though, SDO is not just about being personally dominant and controlling of others; SDO describes a preferred arrangement of groups with some on top (preferably one's own group) and some on the bottom.
- For example, someone high in SDO would likely be upset if someone from an outgroup moved into his or her neighborhood. It's not that the person high in SDO wants to "control" what this outgroup member does; it's that moving into this "nice neighborhood" disrupts the social hierarchy the person high in SDO believes in (i.e. living in a nice neighborhood denotes one's place in the social hierarchy—a place reserved for one's in-group members).

SDO

- For example, researchers have found that those who score higher on SDO are usually lower than average on tolerance, empathy, altruism, and community orientation. In general, those high in SDO have a strong belief in work ethic—that hard work always pays off and leisure is a waste of time.
- People higher on SDO tend to choose and thrive in occupations that maintain existing group hierarchies (police, prosecutors, business), compared to those lower in SDO, who tend to pick more equalizing occupations (social work, public defense, psychology).

Automatic Biases

- Most people like themselves well enough, and most people identify themselves as members of certain groups but not others.
- Logic suggests, then, that because we like ourselves, we therefore like the groups we associate with more, whether those groups are our hometown, school, religion, gender, or ethnicity.
- Liking yourself and your groups is human nature. The larger issue, however, is that own-group preference often results in liking other groups less. And whether you recognize this "favoritism" as wrong, this trade-off **is relatively automatic that is, unintended, immediate, and irresistible.**
- Social psychologists have developed several ways to measure this relatively automatic own-group preference, the most famous being the **Implicit Association Test** (IAT;Greenwald, Banaji, Rudman, Farnham, Nosek, & Mellott, 2002; Greenwald, McGhee, & Schwartz, 1998).

Own Group preference bias

- As the IAT indicates, people's biases often stem from the spontaneous tendency to favor their own, at the expense of the other. Social identity theory (Tajfel, Billig, Bundy, & Flament, 1971) describes this tendency to favor one's own in-group over another's outgroup. And as a result, outgroup disliking stems from this in-group liking (Brewer & Brown, 1998).

Self categorization theory Example ……

- For example, if two classes of children want to play on the same soccer field, the classes will come to dislike each other not because of any real, objectionable traits about the other group.
- The dislike originates from each class's favoritism toward itself and the fact that only one group can play on the soccer field at a time. With this preferential perspective for one's own group, people are not punishing the other one so much as neglecting it in favor of their own.
- However, to justify this preferential treatment, people will often exaggerate the differences between their in-group and the outgroup. In turn, people see the outgroup as more similar in personality than they are.
- This results in the perception that "they" really differ from us, and "they" are all alike. Spontaneously, people categorize people into groups just as we categorize furniture or food into one type or another. The difference is that we people inhabit categories ourselves, as **self-categorization theory** points out (Turner, 1975).

More human Biases leading to discrimation

Diagnosis Bias, the propensity to label people, ideas, or things based on our initial opinions.

University of Pittsburgh reported that African American patients were almost three times as likely to be diagnosed as schizophrenic because the clinicians subjectively determined that the patients were not responding as honestly as white patients. Studies in England have shown similar results in the diagnoses of Afro-Caribbean people by white psychiatrists, often based on the dubious notion that the patients are “strange, undesirable, bizarre, aggressive, and dangerous.”

Stereotype Threat, or *internalized bias*, is the experience of anxiety or concern in a situation where a person has the potential to confirm a negative stereotype about their social group.

In more recent studies, Claude Steele found that simply asking African American students to answer one additional question before taking their SAT tests significantly lowered their scores. The question was: “What is your race?” For many of the African American students, being reminded of being black seemed to internalize negative performance bias.

How the Best Bosses Interrupt Bias on Their Teams

- https://hbr.org/2019/11/how-the-best-bosses-interrupt-bias-on-their-teams

Anchoring bias?

- Anchoring describes the propensity to rely on the first piece of information encountered when making a decision.
- According to this individuals begin with an implicitly suggested reference point (the "anchor") and make adjustments to it to reach their estimate.
- For example, the initial price offered for a used car sets the standard for the rest of the negotiations, so that prices lower than the initial price seem more reasonable even if they are still higher than what the car is worth.

What is attribution bias

- An attribution bias can happen when individuals assess or attempt to discover explanations behind their own and others' behaviors.
- People make attributions about the causes of their own and others' behaviors; but these attributions don't necessarily precisely reflect reality.
- Rather than operating as objective perceivers, individuals are inclined to perceptual slips that prompt biased understandings of their social world.
- When judging others, we tend to assume their actions are the result of internal factors such as personality , whereas we tend to assume our own actions arise because of the necessity of external circumstances.

Confirmation bias

- This is the tendency to search for, interpret , favor, and recall information in a way that confirms one’s beliefs while giving disproportionately less attention to information that contradicts it.

Statistical biases

- Statistical bias is a systematic tendency in the process of data collection, which results in lopsided, misleading results.
- This can occur in any of a number of ways, in the way the sample is selected, or in the way data are collected.It is a property of a statistical technique or of its results whereby the expected value of the results differs from the true underlying quantitative parameters being estimated .

Selection bias

- This is the conscious or unconscious bias introduced into a study by the way individuals, groups or data are selected for analysis, if such a way means that true randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population intended to be analyzed.
- This results in a sample that may be significantly different from the overall population.

Algorithmic bias

- Algorithms are generally understood as lists of instructions that determine how programs read, collect, process, and analyze data to generate output.
- **Algorithmic bias** describes systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others.
- Bias can emerge due to many factors, including but not limited to the design of the algorithm or the unintended or unanticipated use or decisions relating to the way data is coded, collected, selected or used to train the algorithm.
- The study of algorithmic bias is most concerned with algorithms that reflect "systematic and unfair" discrimination (Alan Turing Guidelines) .

Danger of algorithmic biases ….Alan Turing Guidelines

- There is a risk to individual rights and freedoms in systems where algorithms determine who may get a mortgage, how much an insurance policy costs, and whether certain behaviours or places people visit can be flagged as ‘suspicious’ activities.
- For example, if someone does not have an algorithmic profile, they may be excluded from systems where diverse voices are needed for decision-making. However, to have such a profile can mean we are further monitored and tracked to make algorithmic predictions more accurate.
- Source - https://www.turing.ac.uk/sites/default/files/2019-07/response_of_the_alan_turing_institute_to_the_centre_for_data_ethics_and_innovations_review_on_bias_in_algorithmic_decision-making_0.pdf

Bias which afflict AI

Bias can creep in long before the data is collected, at many other stages of the deep-learning process.

- ***Framing the problem.*** The first thing computer scientists do when they create a deep-learning model is decide what they actually want it to achieve. Eg: Creditworthiness algorithm and the business decision which influence it

Discuss what type of biases would influence this decision

- ***Collecting the data.*** There are two main ways that bias shows up in training data: either the data you collect is unrepresentative of reality, or it reflects existing prejudices.

Discuss possible situations this would occur

Almost all AI's learn from data, unlike traditional code, with any AI system, the data is in every practical sense the code which powers the AI

Training Data is it reliable ?

Bias which afflict AI

A lack of diversity in the data science field as a whole is one factor to consider when trying to identify why most of algorithmic bias is centered around race and gender.

Source : https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

However, the consensus is mostly focused on Biased Data = Biased Algorithms

- For any system that learns, the output is determined by the data it receives. This nothing new, however its something we tend to forget when we look at systems driven by millions of examples.
- The hope has been that the sheer volume of examples will overwhelm any human bias.
- But if the training set itself is skewed; there is a real possibility that the results will also be skewed.

Source: https://www.forbes.com/sites/cognitiveworld/2020/02/07/biased-algorithms/?sh=484e68f776fc

Algorithmic decision-making bias -other causes

- Bias can arise from the lack of transparency in black-box methods, and the modelling done by data scientists which feeds into machine-learning methods.
- Such modelling also requires transparency, as some data scientists might embed bias or not be sufficiently conscious of potential biases.
- On the other hand, the data itself is a significant potential source of bias, and more of a focus needs to be put on its quality and limitations.

Important question to ponder

- Some checks can be automated, such as verifying whether minority groups are underrepresented in the training data, but selection biases are not always evident from the information available. What are your views on these?

Short Case – The Prejudiced Computer

- For one British university, what began as a time-saving exercise ended in disgrace when a computer model set up to streamline its admissions process exposed - and then exacerbated - gender and racial discrimination.
- As detailed here in the British Medical Journal, staff at St George's Hospital Medical School decided to write an algorithm that would automate the first round of its admissions process. The formulae used historical patterns in the characteristics of candidates whose applications were traditionally rejected to filter out new candidates whose profiles matched those of the least successful applicants.
- By 1979 the list of candidates selected by the algorithms was a 90-95% match for those chosen by the selection panel, and in 1982 it was decided that the whole initial stage of the admissions process would be handled by the model. Candidates were assigned a score without their applications having passed a single human pair of eyes, and this score was used to determine whether or not they would be interviewed.
- Quite aside from the obvious concerns that a student would have upon finding out a computer was rejecting their application; a more disturbing discovery was made. The admissions data that was used to define the model's outputs showed bias against females and people with non-European-looking names.
- The truth was discovered by two professors at St George's, and the university co-operated fully with an inquiry by the Commission for Racial Equality, both taking steps to ensure the same would not happen again and contacting applicants who had been unfairly screened out, in some cases even offering them a place.

- Source theguardian,com

Discussion questions

What could have caused this problem?

Would going back to a total human driven selection process solve the issue?

Biases With AI

- **Bias Through Interaction**
- While some systems learn by looking at a set of examples in bulk, other sorts of systems learn through interaction. Bias arises based on the biases of the users driving the interaction.
- A clear example of this bias is Microsoft Tay , a Twitter-based chatbot designed to learn from its interactions with users. Unfortunately, Tay was influenced by a user community that taught Tay to be racist and misogynistic. In essence, the community repeatedly tweeted offensive statements at Tay and the system used those statements as grist for later responses.

Biases With AI

- **Latent Bias**
- In latent bias, an algorithm may incorrectly identify something based on historical data or because of a stereotype that already exists in society.
- An AI may recognize a doctor to be male and not female, because the data tells it that doctors are primarily men. This could also be called prejudice bias. In it, the data characterizes our preconceived notions or historical predispositions and that skews the dataset
- **Famously, Amazon had to scrap an automated recruiting tool that was favoring male candidates over female candidates because the algorithm was trained on historical patterns in which men are primarily those hired.**

Biases With AI

Selection Bias

- A dataset overrepresents one certain group and underrepresents another.
- “Selection bias occurs when a data set contains vastly more information on one subgroup and not another,” says White.
- On the data collection side biases can emerge when the data you collect is unrepresentative of reality or when existing human biases are amplified during the collection process.
- For instance, many machine learning algorithms are taught by scraping the Internet for information. Major search engines and their algorithms were developed in the West. Therefore, algorithms are often more likely to recognize a bride and a groom in a western-style wedding but fail to do so in an African wedding.

Critical points …..

- Assuming this bias is occurring or at risk of occurring in the future, what is being done to mitigate it?
- And who should be leading efforts to do this? If the provenance of data used for training algorithms is not known or trustworthy, definitions are arising in the research community to force predictions to respect categorical constraints. For example, predictions should not be sensitive to gender, race, and so on. More generally, prediction shouldn’t be sensitive to so-called protected attributes (as well as proxy attributes that might indirectly reveal the protected attribute).

So, what do?

- Four researchers, a mix of Turing Fellows and Research Fellows who met at the Institute, have developed a framework that aims to ensure fairness in algorithm-led decision-making systems by taking into account different social biases and compensating for them effectively. The research stems from the idea that **a decision is fair towards an individual if the outcome is the same in the actual world as it would be in a 'counterfactual' world, in which the individual belongs to a different demographic.**
- The framework calls for the need for decision-making algorithms to be designed with the input of expert knowledge about the situations the algorithms are being used in. Knowledge about the relationships between key factors and attributes that relate to the people and processes involved.
- The work is providing practitioners with customised techniques for solving a wide array of problems, in applications from policy-making to policing.

Counterfactual fairness

- Does counterfactual fairness equal actual fairness?

- Source - https://www.turing.ac.uk/research/impact-stories/fairer-algorithm-led-decisions
- Counterfactual Fairness: Unidentification, Bound and Algorithm
- https://www.ijcai.org/Proceedings/2019/0199.pdf

- Centre for Data Ethics and Innovation's review on Bias in Algorithmic Decision-Making
- https://www.turing.ac.uk/research/publications/centre-data-ethics-and-innovations-review-bias-algorithmic-decision-making

Counterfactual fairness

- Counterfactual fairness is a notion of fairness derived from Pearl's causal model, which considers a model is fair if for a particular individual or group its prediction in the real world is the same as that in the counterfactual world where the individual(s) had belonged to a different demographic group. However, an inherent limitation of counterfactual fairness is that it cannot be uniquely quantified from the observational data in certain situations, due to the unidentifiability of the counterfactual quantity. (Yongkai Wu , Lu Zhang and Xintao Wu, 2009)
- https://www.ijcai.org/Proceedings/2019/0199.pdf

Counterfactual Fairness

Statistics > Machine Learning

[Submitted on 20 Mar 2017 (v1), last revised 8 Mar 2018 (this version, v3)]

Counterfactual Fairness

Matt J. Kusner, Joshua R. Loftus, Chris Russell, Ricardo Silva

Machine learning can impact people with legal or ethical consequences when it is used to automate decisions in areas such as insurance, lending predictive policing. In many of these scenarios, previous decisions have been made that are unfairly biased against certain subpopulations, for exa of a particular race, gender, or sexual orientation. Since this past data may be biased, machine learning predictors must account for this to avoid p or creating discriminatory practices. In this paper, we develop a framework for modeling fairness using tools from causal inference. Our definition c counterfactual fairness captures the intuition that a decision is fair towards an individual if it is the same in (a) the actual world and (b) a counterfac where the individual belonged to a different demographic group. We demonstrate our framework on a real-world problem of fair prediction of succe school.

Subjects: **Machine Learning (stat.ML)**; Computers and Society (cs.CY); Machine Learning (cs.LG)
Cite as: arXiv:1703.06856 **[stat.ML]**
(or arXiv:1703.06856v3 **[stat.ML]** for this version)

Essential Reading

Landscape Summary: Bias in Algorithmic Decision-Making

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819055/Landscape_Summary_-_Bias_in_Algorithmic_Decision-Making.pdf

Fairer algorithm-led decisions

- https://www.turing.ac.uk/research/impact-stories/fairer-algorithm-led-decisions

- Next few slides are not related directly to bias but for your reading pleasure for those curious minds!

CYBERCRIME, SECURITY & ETHICS

Cybercrime, Security & Ethics

The last two decades has seen the rapid growth of the Internet, mobile technology and the Correspondingly rapid growth of online crimes, or cybercrimes.

FIRST GEN OF CYBER CRIMINALS/CRIMES

- At the tail end of the 1980s, computers had become smaller and with the introduction of the IBM PC, became very widely used.
- Schools were filling with students wanting to be computer programmers as were the newspapers with stories of cybercrimes. These Wiz kid criminals were mostly thrill seekers after bragging rights, their operations lacked of organizational structure

SECOND GEN OF CYBER CRIMINALS/CRIMES

- From around 1990 through to 2000. This period was characterized by serious, often devastating, and widespread virus attacks.
- Interconnected interdependent networks were an attractive target for virus attacks by criminals who were mostly Wiz Kids from the 1980's

Cybercrime, Security Ethics

THIRD GENERATION OF CYBER CRIMES

- From 2000 onwards saw denial of service attacks being replaced by unauthorized access as there were thousands of hacking tools available on the internet by then.
- Unlike before, the targets were preselected to maximize personal gains, usually these were financial.

FOURTH GENERATION OF CYBER CRIMES

- With the increasing sophistication of communication technology, the 4th generation of cybercrime started around 2010.
- Two major trends were; the cyber criminals were organizing themselves more into criminal enterprise cartels, and, there is more state sponsored hacking activities than ever before.

EFFORTS TO COMBAT AND TARGET CYBER CRIMINALS

Formation of public-private partnerships

- The United Kingdom's Cyber Crime Reduction Partnership (CCRP). This effort is to provide a forum in which government, law enforcement, industry and academia can regularly come together to tackle cybercrime more than before.

- DHS collaborates with financial and other critical infrastructure sectors to improve network security. additionally, DHS components, such as the U.S. Secret Service and U.S. Immigrations and Customs Enforcement (ICE), have special divisions dedicated to fighting cybercrime.

- The FBI has the following cybercrime partnerships and initiatives. National Cyber Investigative Joint Task Force – a focal point for all U.S. government agencies to coordinate, integrate, and share information related to domestic cyber threat investigations

- Setting up publicly funded agencies to go after cyber criminals. Representative examples include:

The Secret Service maintains Electronic Crimes Task Forces (ECTFs), which focus on identifying and locating international cyber criminals connected to cyber intrusions, bank fraud, data breaches, and other computer-related crimes.

ICE's Cyber Crimes Center (C3) works to prevent cybercrime and solve cyber incidents. From the C3 Cyber Crime Section, ICE identifies sources for fraudulent identity and immigration documents on the Internet.

C3's Child Exploitation Section investigates large scale producers and distributors of child pornography and well as individuals who travel abroad for the purpose of having sex with minors

NEED FOR A MORAL CODE

When you ask people what kind of life they like most, the most popular answer is always going to be a life full of freedoms - they will say that freedom is doing what they want to do, when they want to do it, and in the way that they want to do it. In other words, a life without restraints.

But this kind of freedom will result in anarchy. This raises the point about a Moral code. Humans have the capacity to reason, freedom within reason is the bedrock of morality. The moral code, therefore, is essential for humanity to attain and keep the freedoms they need.

Moral standards

- A moral standard is a moral norm, a standard to which we compare human actions to determine their goodness or badness

ETHICS

Robert C. Solomon, in Morality and the Good Life, defines ethics as a set of "theories of value, virtue, or of right (valuable) action." O.J. Johnson, on the other hand, defines ethics as a set of theories "that provide general rules or principles to be used in making moral decisions and, unlike ordinary intuitions

WHAT DOES ETHICS HAVE TO DO WITH TECHNOLOGY?

- Technologies are not ethically ‘neutral’, for they reflect the values that we ‘bake in’ to them with our design choices, as well as the values which guide our distribution and use of them
- Many lawmakers lack the technical expertise needed to guide effective technology policy. This means technology experts get involved in creating this policy, but are they not biased as they are in the business of making profits?

Eg: face- and voice-recognition algorithms can now be used to track and create a lasting digital record of your movements and actions in public, even in places where previously you would have felt anonymous.

- There is no consistent legal framework governing this kind of data collection, even though such data could potentially be used to expose a person’s medical history, their religiosity, their status as a victim of violence or other sensitive information, up to and including the content of their personal conversations in the street

This goes to show the that a person given access to all that data, or tasked with keeping it secure, needs to have a good understanding about its ethical significance and power to affect a person’s life.

Another factor driving the recent explosion of interest in technology ethics is the way new technologies are reshaping the global distribution of power, justice, and responsibility. Companies like Facebook, Google, Amazon, Apple, and Microsoft are now seen as having levels of global political influence comparable to, even greater than, that of states and nations.

WHAT DOES ETHICS HAVE TO DO WITH CYBERSECURITY?

- Cybersecurity practices have as their aim the securing of data, computer systems and networks (software and hardware). What cybersecurity practices primarily protect are the integrity, functionality, and reliability of human institutions/practices that rely upon such data, systems, and networks. In turn are protecting the lives and happiness of the human beings who depend upon them.
- If you are a cybersecurity professional tasked with securing a hospital's network and critical data from invasion and attack, you are intimately involved in protecting sick patients, even if you have no medical training. Patients' privacy, health, even their survival can hinge upon your success or failure.
- Hence ethical issues are at the core of cybersecurity practices, because these practices are increasingly required to secure and shield the ability of human individuals and groups to live well.

WHICH OF THESE LIFE-IMPACTING EVENTS MIGHT RESULT FROM CYBERSECURITY PRACTICES?

- Kent, a hard-working first-generation college senior, has just requested that copies of his university transcript be sent to the ten graduate schools to which he has applied. Kent does not know that he was recently the victim of a malicious and undetected intruder into his 5 university's network; as a prank, the intruder changed a random selection of students' course grades to an 'F.'

- Dev and Katia, a pair of talented freelance hackers, identify a previously unknown but easily fixed vulnerability in the current operating system of a particular manufacturer's mobile phones, which allows the remote injection and execution of malicious code. As they discuss what they should do next—contact the affected the manufacturer via a backchannel, notify a popular tech media news site, or expose the vulnerability on their own cybersecurity blog— Dev and Katia are approached by a friend, who works for the phone manufacturer's primary competitor. The friend offers them both lucrative jobs, on the condition that they remain silent about the exploit they have found.

In each of these examples, one or more unsuspecting persons' chances of living good lives are profoundly impacted by what cybersecurity professionals and other actors in the information security space have or have not done—or by what they will or will not do.

COMMON ETHICAL CHALLENGES FOR CYBERSECURITY PROFESSIONALS

It is important to note that even when a cybersecurity practice is legal, it may not be ethical. Unethical or ethically dubious cybersecurity practices can result in significant harm and reputational damage to network users, clients, companies, the public, and cybersecurity professionals themselves.

ETHICAL CHALLENGES IN BALANCING SECURITY WITH OTHER VALUES:

- Have we struck an ethically acceptable balance – Device usability, Stakeholder resource needs
- Have all the stakeholders been consulted on how this balance has been struck?
- Is there an enough threat to justify resource allocation?
- Do our actions display consistency, sincerety and transparency and align with our value commitments?

ETHICAL CHALLENGES IN THREAT/INCIDENT RESPONSE:

- Do we have a response plan for eacj type of incident we anticipate?
- Do we have the resources in lace to implement this plan?
- Are there any ethically grey areas in the plan? Eg: Collateral damage

COMMON ETHICAL CHALLENGES FOR CYBERSECURITY PROFESSIONALS

ETHICAL CHALLENGES IN SECURITY BREACH/VULNERABILITY:

- Do we have a plan on how to inform users of any breaches?
- How do we meed the need for accurate and timily reporting
- Have we considered other stakeholders perspectives on the action plan?
- Have we identified the ethically appropriate balance between overreaction to breaches and vulnerabilities and underreaction?

ETHICAL CHALLENGES IN NETWORK MONITORING AND USER PRIVACY:

- How can our network be monitored effectively without unjustifiable intrusions upon users and their privacy?
- To what extent should users of the network be made aware of our security monitoring activities?

COMMON ETHICAL CHALLENGES FOR CYBERSECURITY PROFESSIONALS

ETHICAL CHALLENGES WITH COMPETING INTERESTS & OBLIGATIONS:

- Have we adequately considered the ethical harms that may be done by a security breach, both in the short-term and long-term, and to whom?
- What should we do if asked by an employer or client to grant someone a level of system access privileges that we have good reason to think are inappropriate
- What should we do if asked by an employer or client to put off disclosure of a serious or critical system vulnerability, or to delay a patch
- What will we do if we are asked to violate a professional duty of cybersecurity practice in the interests of national security, or some other non-professional ethical interest?

ETHICAL CHALLENGES IN DATA STORAGE AND ENCRYPTION:

- How can we responsibly and safely store and transmit sensitive information?
- How will we respond to requests from law-enforcement or intelligence agencies to weaken our encryption practices or decrypt specific devices?
- What are the ethical risks of long-term data storage? How long is long enough?

COMMON ETHICAL CHALLENGES FOR CYBERSECURITY PROFESSIONALS

ETHICAL CHALLENGES IN PRODUCT DESIGN:

- Have we been sufficiently imaginative in conceiving how a network, product or feature might be abused, exploited, or misused?
- For networked utilities and wireless devices that could cause significant harm if exploited, have we employed added/upgraded security measures, including end-user training

ETHICAL CHALLENGES WITH ACCOUNTABILITY FOR CYBERSECURITY:

- Who should and will be held accountable for various risks or harms that might be imposed on others by our cybersecurity practice?
- What organizational/team policies and routines must we establish in advance and enforce, in order to safeguard and promote ethical cybersecurity practice? Eg Audits
- Are the team members adequately competent
- Is investigative process for breaches

COMMON ETHICAL CHALLENGES FOR CYBERSECURITY PROFESSIONALS

ETHICAL CHALLENGES IN SECURITY RESEARCH AND TESTING:

- When is it ethical to publish information about new security techniques, tools, or vulnerabilities for the benefit of security researchers
- What are the ethical implications of developing and releasing automated security tools, especially those intended to spread ‘in the wild’? Eg: to block the spread of a worm
- How should we balance the different ethical interests and timescales of security
- What are the ethical implications of participating in the private market for ‘zero-day’ exploits, either as an exploit seller, buyer, or developer?

UNDERSTANDING BROADER IMPACTS OF CYBERSECURITY PRACTICE

- Overall, have we fully considered how our cybersecurity practices today may impact others, now and well down the road? Is our cybersecurity team sufficiently diverse to understand and anticipate these effects?
- Do our cybersecurity practices violate anyone’s legal or moral rights,
- Would information about our cybersecurity practices be morally or socially controversial, or damaging to the professional reputations of those involved

Resources

- Altemeyer, B. (2004). Highly dominating, highly authoritarian personalities. *The Journal of Social Psychology, 144(4)*, 421-447. doi:10.3200/SOCP.144.4.421-448
- Altemeyer, B. (1988). *Enemies of freedom: Understanding right-wing authoritarianism*. San Francisco: Jossey-Bass.
- Bodenhausen, G. V., & Peery, D. (2009). Social categorization and stereotyping in vivo: The VUCA challenge. *Social and Personality Psychology Compass, 3(2)*, 133-151. doi:10.1111/j.1751-9004.2009.00167.x
- Brewer, M. B., & Brown, R. J. (1998). Intergroup relations. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology, Vols. 1 and 2* (4th ed.) (pp. 554-594). New York: McGraw-Hill.
- Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56(1)*, 5-18. doi:10.1037/0022-3514.56.1.5
- Dovidio, J. F., & Gaertner, S. L. (2010). Intergroup bias. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology, Vol. 2* (5th ed.) (pp. 1084-1121). Hoboken, NJ: John Wiley.
- Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology, Vols. 1 and 2* (4th ed.) (pp. 357-411). New York: McGraw-Hill.
- Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11(2)*, 77-83. doi:10.1016/j.tics.2006.11.005
- Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109(1)*, 3-25. doi:10.1037/0033-295X.109.1.3

Activity 1

- https://youtu.be/IGQmdoK_ZfY

Back in 1962, a famous person was about to catch a flight when they made an interesting observation. Here's how they described the moment in their memoirs:

> *"As I was boarding the plane I saw that the pilot was black. I had never seen a black pilot before, and the instant I did I had to quell my panic. How could a black man fly an airplane?"*

> *"...But a moment later I caught myself: I had fallen into the apartheid mind-set, thinking Africans were inferior and that flying was a white man's job. I sat back in my seat, and chided myself for such thoughts."*


Jackie Kennedy


Nelson Mandela


Enoch Powell


Audrey Hepburn


Harold MacMillan