DATA ETHICS

DATA & AI ETHICS

MAVEN® ANALYTICS

COURSE OUTLINE

1 **Data Ethics 101**

Introduce the fundamental concepts of ethics, and discuss how data ethics differs from legal frameworks regulating data

2 **Ethical Data Stewardship**

Understand how to be an effective steward of data, and review the ethical implications of collecting and managing sensitive data

3 **Data & Algorithmic Bias**

Learn how to detect and mitigate common forms of bias, including sampling, selection, algorithmic and confirmation bias

4 **AI Ethics & Impact**

Dive into the world of AI and explore unique ethical challenges in terms of data collection, bias, and societal harm

SETTING EXPECTATIONS

✓ This is a **high-level introduction** to the principles of ethics in data & AI

- *We'll introduce key concepts and terminology, but won't dive into advanced or specialized topics*

✓ Our goal is to provide **foundational knowledge** of core ethical concepts

- *This course is for data leaders, individual contributors, or anyone who wants to better understand the ethical implications of an increasingly data and AI-driven world*

✓ The goal of this course is to generate **awareness** rather than solutions

- *Cultivating awareness is the first step in being able to identify and mitigate potential ethical lapses*

✓ This is **NOT** a prescriptive guide to navigating specific ethical issues

- *We will highlight real-world ethical issues and use thought exercises to help cultivate an ethical mindset, but do not have all the answers when it comes to navigating complex ethical scenarios*

DATA & AI ETHICS 101

WHAT IS ETHICS?

ETHICS

Moral principles that govern a person's behavior or the conducting of an activity

DATA ETHICS

The system of moral principles that guide how data is collected, shared, and used

AI ETHICS

The field of study concerned with the moral implications and guidelines for the development, deployment, and use of AI systems

CASE STUDY: CAMBRIDGE ANALYTICA

THE **SITUATION**

In 2014, **Cambridge Analytica** used a Facebook app called "This is Your Digital Life" to harvest personal information from 87 million global users, many of whom were unaware.

This was used to create detailed voter profiles and target users with political ads and misinformation, impacting key votes in the US and UK.

THE **FALLOUT**

Because it failed to protect user privacy and subjected users to psychological manipulation without their consent, **Facebook was fined $5B dollars by the US government**.

The EU and Great Britain passed much more stringent data privacy laws, including GDPR, the EU's landmark data privacy law *(more in that later!)*

THE **ETHICS**

- **Consent**:
  - Most were unaware their data could be harvested by a third party
  - Many people who didn't use the app had their data harvested
- **Privacy**:
  - Facebook violated user privacy by allowing a third party to access personal info
- **Societal Impact**:
  - Voters were targeted with misinformation, causing immeasurable societal impact

ETHICS VS. LAW

Ethics are the moral principles that govern a person or society's behavior

Laws are systems of rules that govern a society's conduct, often enforced by a controlling authority

- Laws are often informed by ethics, **but not all ethical behaviors are codified in the law**
- The ethical effects of technology (like social media or AI) can take decades to understand, let alone translate to law

Unethical Behaviors Enforced by Law

- Murder
- Theft
- Fraud
- Reckless Driving

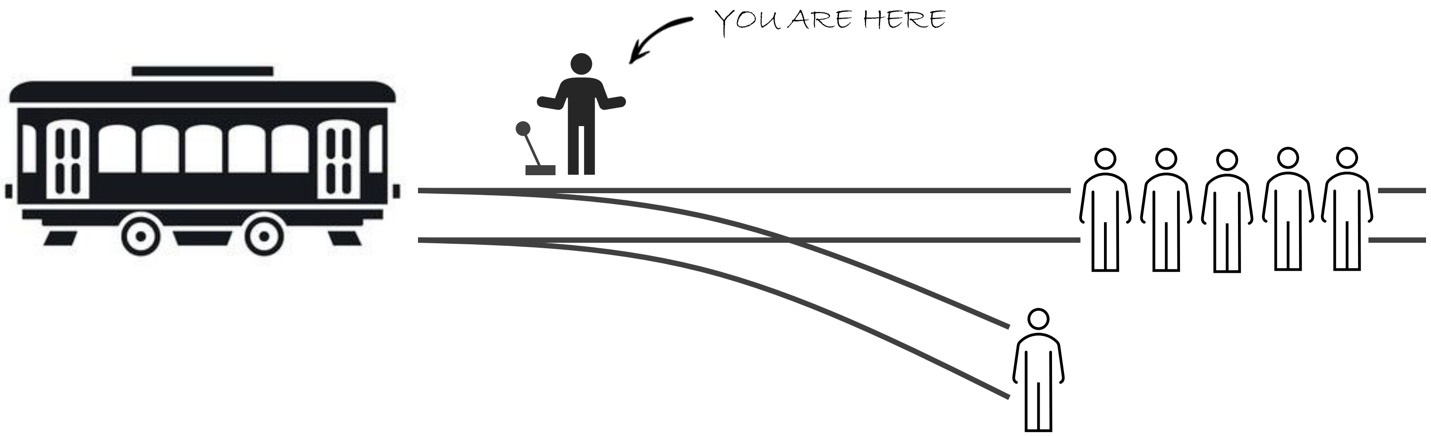
Unethical Behaviors NOT Enforced by Law

- Lying
- Being unfair to employees or children
- Taking credit for others' work
- Eavesdropping or reading private messages

THOUGHT EXERCISE: THE TROLLEY PROBLEM

THE SCENARIO

A runaway trolley speeds down a track. If it stays on course, 5 people on the track will be hit and killed. **By pulling a lever, you can divert the trolley to another track where a single bystander will be hit**.



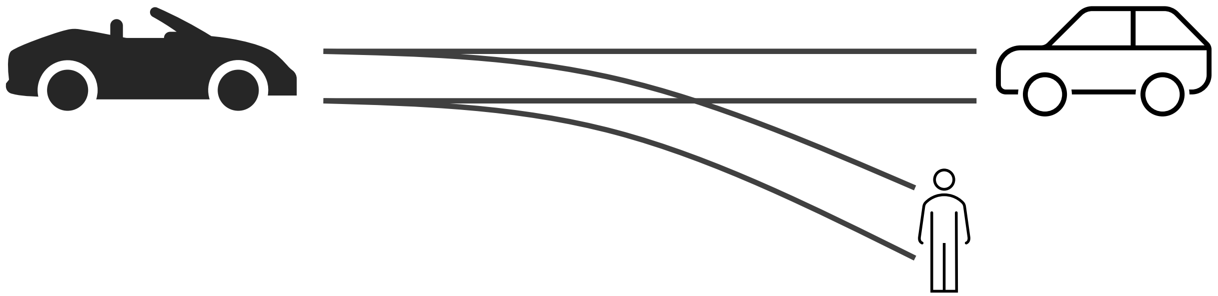
Do you pull the lever?

- 90% of people say they would pull the lever, to save the most lives
- But what if the one person was your spouse or best friend? What if it's one child and five senior citizens?

THOUGHT EXERCISE: SELF-DRIVING CAR

THE SCENARIO

A self-driving car is facing a head-on collision with a reckless driver. If it stays on course, it estimates that the passenger and 4 people in the other car will die. **If it veers off course, it will hit one person on the sidewalk**.



How should the self-driving car be programmed to respond?

- What type of algorithm would determine the "best" course of action?
- Does the company that programmed the car decide how to solve this problem, or should there be laws that regulate how AI companies handle decisions like this?

KEY TAKEAWAYS

✓ **Data & AI Ethics** are increasingly important in a data driven world

- *Data practitioners and organizations alike should foster systems and create guidelines for the ethical handling of data and use of AI algorithms*

✓ Ethics and the law **are not the same thing**

- *While the law often reinforces a society's ethical values, when it comes to tech, it can be years or decades before the law catches up*

✓ **There are many grey areas** when it comes to issues in data and AI ethics

- *While we don't have a prescription for every potential issue, cultivating an ethical mindset is critical for identifying areas of concern & safeguarding against ethical lapses*

ETHICAL DATA STEWARDSHIP

DATA STEWARDSHIP

Data stewardship is the practice of managing and using data ethically

- Stewardship ultimately comes down to the golden rule: **treat user data how you'd like your own to be treated**
- Ethical data stewardship requires a thoughtful approach to **user consent**, **security**, **privacy** and **confidentiality**

User Consent

- Getting permission to collect and store user data
- Making sure users are aware of how their data will be used
- Allowing users to opt out or withdraw consent at any time

Security

- Taking reasonable measures to protect user data once you've collected it
- This can include anonymizing data, running employee awareness training, and implementing strong network security practices

Privacy & Confidentiality

- Collecting the minimum amount of data necessary for your business needs
- Restricting access to personal data to only those employees who need to use it
- Getting consent when sharing user data with third parties

CASE STUDY: OKCUPID DATASET

THE
SITUATION

In 2016 researchers in Denmark scraped 70k profiles from dating app **OkCupid**, including usernames, locations, ages, religions, and answers to personal questions.

Critics pointed out that it would be easy to to identify individuals based on clues provided by the data, representing a huge breach in user privacy and trust.

THE
FALLOUT

Reputations were significantly damaged for the researchers, their university, and OkCupid. The event may have also **violated the US Computer Fraud and Abuse Act** and would have been punishable under GDPR (which was passed a few years later).

While OkCupid was not directly involved, questions were raised about its ability to protect user privacy, and their commitment to ethical data stewardship.

THE
ETHICS

- **Consent**:
  - OkCupid users did not consent to their data being used for public datasets
- **Privacy**:
  - The data was not sufficiently anonymized, which meant things like the political views and deeply personal dating preferences were now in the public domain

CONSENT

Consent is the permission for something to happen or an agreement to do something

- In the world of data, this can relate to **what data is collected**, **how it is used**, and **who it is shared with**
- There are common practices for collecting, using, and sharing personal information that are **legal** but **unethical**

Meaningful Consent

Many users will check a box agreeing to complex legal agreements or terms of service without reading or fully understanding them, to get access to a product.

While this fulfills legal obligations, it isn't considered *meaningful* consent.

☑ A better approach:

- Ask users whether they consent to specific use cases of their data
- Remind them periodically about what data is collected, how it is used, and who it is shared with

Opting out

Users should have a way to **opt out of data collection** or **non-essential uses**, including:

- ✓ Opting out of sharing their information with third parties
- ✓ Withdrawing consent at any time, including data storage

SECURITY

Data security requires taking reasonable steps to shield data from unauthorized access & leaks

- Risks can be mitigated through proper **data collection & storage**, strong **network security**, and **employee training**

Collection & Storage

- Collecting the minimum amount of data necessary
- Encrypting sensitive data
- Anonymizing personal records

Network Access

- Firewalls
- Virtual private networks (VPN)
- Threat detection & continuous monitoring

Employee Training

- Role-based access to data
- Security awareness training

Cybersecurity is a complex and rapidly evolving field, and is not just an ethical issue. Numerous data breaches occur every year, and can cause **reputational damages**, **revenue loss**, and in some cases **costly lawsuits or government sanctions**

PRIVACY & CONFIDENTIALITY

Users should have a reasonable expectation of **privacy & confidentiality** when they share data

- Sharing data without their permission or collecting/tracking data without their consent and awareness is unethical

Data Minimization Only the minimum amount of data should be collected to operate the product or service

Transparency Users should be clearly informed about what data is being collected, how it will be used, and who it will be shared with

Confidentiality Personal information should never be shared with others without express user consent

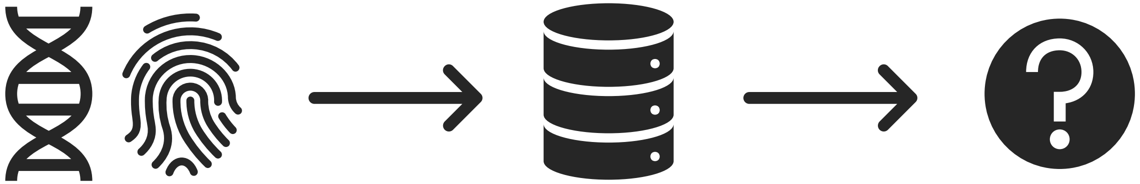
Anonymization When possible, data should be anonymized to minimize risks of a security breach, and identities should be shielded from both internal and third-party users

Birthdays are an example of a data collection ethical grey area. They may be required for things like loan approval, car rentals or other services requiring age verification, but should only be required inputs by customers in cases where it's absolutely required

THOUGHT EXERCISE: GENETIC DATABASES

THE SCENARIO

You submit your DNA to a company that analyzes it to determine your ancestry and risk for certain diseases.
What should that company do with your data after it has been analyzed?



How would you feel if…

- The company shared your data with medical researchers?
- The company shared your data with law enforcement agencies?
- They sold your data to pharmaceutical companies to use for targeted marketing campaigns?
- They sold your data to foreign governments with unknown motives?
- They told you that your data was breached in a hack, but they didn't know by whom?

THOUGHT EXERCISE: SOCIAL MEDIA DATABASE

THE SCENARIO

You work as an analyst for a social media company that your several of your friends and family members use.
How should you think about the way you access and analyze the data?

Which of the following feels like crossing an ethical line?

- You perform an analysis on aggregated data, reporting on overall trends like time spent on the platform
- You analyze messaging behavior on anonymized user account data
- You look up your friend’s account in the database to see whose photos they’ve liked
- You see that your friend sent a message to their ex, and tell others about it

DATA ETHICS REGULATIONS

Governments have increasingly sought **regulate the ethical storage and use of data**

- This gives them the ability to fine and punish those who violate ethical standards of data stewardship

UNITED STATES (HIPAA)

The **Health Insurance Portability and Accountability Act** (HIPAA), passed in 1996, set strict rules for the storage and use of individuals' medical information

- **The Privacy Rule** established standards for the use of Protected Health Information (PHI)
- **The Security Rule** set standards for the storage of health information
- Patients are also given control of personal health data, including the right to access and request corrections to their records
- Data professionals who work with health data in the US are expected to pass HIPAA training prior to being granted access

DATA ETHICS REGULATIONS

Governments have increasingly sought **regulate the ethical storage and use of data**

- This gives them the ability to fine and punish those who violate ethical standards of data stewardship

EUROPEAN UNION (GDPR)

The **General Data Protection Regulation** (GDPR), passed in 2019, is a landmark set of regulations governing how organizations collect, store, and process the personal data of individuals in the EU

- It demands informed consent for the collection of personal information and mandates that organizations ensure transparency, security, and accountability when handling it
- The **"right to be forgotten"** mandates that individuals have the right to request their data be deleted
- Failure to comply with GDPR can result in significant fines, up to 4% of a company's global revenue

Critics argue that some regulations limit innovation and make it difficult for small startups to compete against larger companies

KEY TAKEAWAYS

✓ Ethical **data stewardship** is a core pillar of data ethics

- *Consent, security, privacy and confidentiality are all important responsibilities of ethical data stewards*

✓ **Consent** is arguably the single most important aspect of ethical data stewardship

- *Users should know what data is collected and how it is being used, and be able to opt out at any time*

✓ **Data security** is the practice of taking reasonable steps to keep user data safe

- *Best practices include anonymizing user data, collecting only necessary data, and conducting employee training*

✓ **Privacy & confidentiality** are closely tied to consent – sharing or tracking user data without an individual's knowledge is unethical

- *When users consent to data collection, that doesn't mean they consent to you sharing that data with others*

DATA & ALGORITHMIC BIAS

WHAT IS BIAS?

BIAS

Prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair

DATA BIAS

Systematic distortion or skewing of data that causes it to inaccurately represent the true characteristics of the population or phenomena it is intended to describe

- This can lead to inaccurate or unrepresentative outcomes in analysis, modeling and decision-making

Not all bias is bad, unethical, or fixable, but being aware of what types of bias exist and their root causes can help you prevent bias and minimize potential harm

TYPES OF BIAS

There are several **types of bias** that data professionals should be aware of

- The presence of bias **isn't always an ethical issue**, but ethical issues often accompany bias in the data
- Ethics aside, being aware of bias can help improve insights and encourage better data practices

Sampling Bias

Data collected is not representative of a population, leading to skewed results

- ***Example***: *A company only surveys its Californian (CA) users on product quality*

Selection Bias

Data is not randomly selected, or the population self-selects

- ***Example***: *Customers volunteer to review your product and detractors don't volunteer*

Algorithmic Bias

Machine learning models perpetuate biases present in training data

- ***Example***: *A model only targets CA customers which reinforces historical targeting, at the expense of new markets*

Confirmation Bias

Interpreting data in a way that confirms existing hypotheses

- ***Example***: *An analyst concludes the product won't sell outside of California due to low sales, ignoring market research*

CASE STUDY: THE EIGENFACE DATASET

THE
SITUATION

The **Eigenface Algorithm**, developed in the early 1990s, was a breakthrough in facial recognition, using a clever application of linear algebra to identify and match faces.

For convenience, the original dataset was likely collected by taking photos of colleagues, coworkers, and students at the institution, who were predominately white, male and middle-aged.

THE
FALLOUT

The algorithm was effective at detecting faces in images with strong accuracy, and there wasn't anything unethical about the way the data was collected for the initial research project.

However, the algorithm was then applied more broadly and used to build new facial recognition datasets by identifying faces in photos on the web, **leading to decades of algorithms that struggled to identify dark skinned or female faces**.

THE
ETHICS

- **Data Bias**:
  - The lack of diversity in the original dataset led to compounding issues of data bias in future datasets collected for facial recognition algorithms
- **Algorithmic bias**:
  - Biased facial datasets led to biased facial detection algorithms - they were much less likely to correctly identify dark skinned and female faces

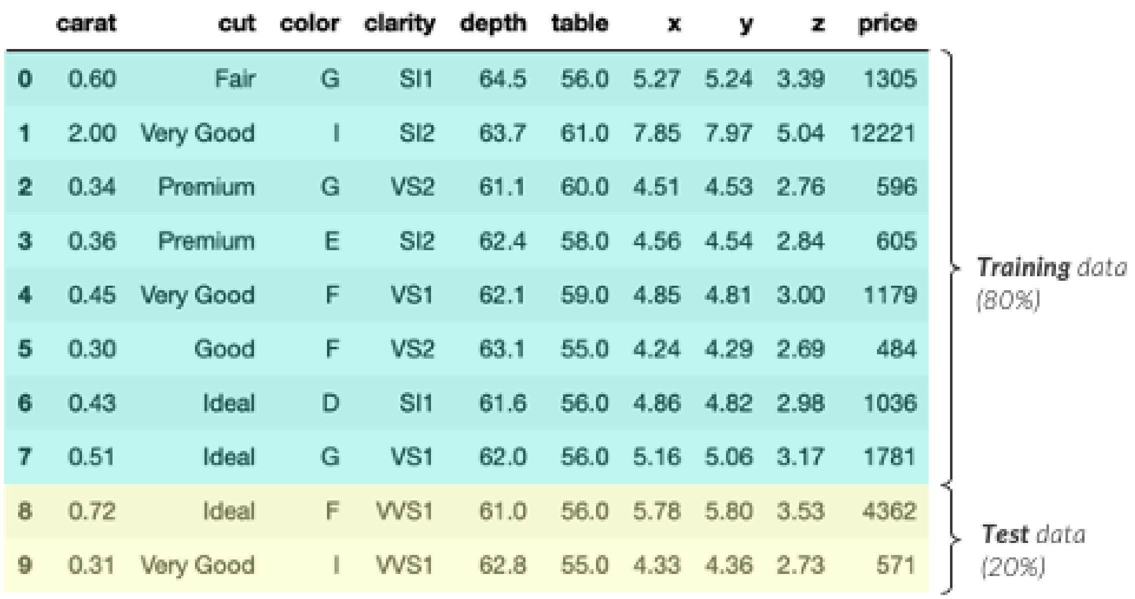
THE MODEL TRAINING PROCESS

ML models learn from **training** data, and are evaluated by their performance on **test** data

- If both the training and test data sets are biased, or not representative of the population the model will be applied to, **the model will likely underperform on underrepresented groups**

Training and test data is often drawn from the **same larger dataset**, so if bias exists in one it will likely exist in the other

This means that **poor performance on underrepresented groups can often go undetected** unless you evaluate the model on subgroups of the data



	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.60	Fair	G	SI1	64.5	56.0	5.27	5.24	3.39	1305
1	2.00	Very Good	I	SI2	63.7	61.0	7.85	7.97	5.04	12221
2	0.34	Premium	G	VS2	61.1	60.0	4.51	4.53	2.76	596
3	0.36	Premium	E	SI2	62.4	58.0	4.56	4.54	2.84	605
4	0.45	Very Good	F	VS1	62.1	59.0	4.85	4.81	3.00	1179
5	0.30	Good	F	VS2	63.1	55.0	4.24	4.29	2.69	484
6	0.43	Ideal	D	SI1	61.6	56.0	4.86	4.82	2.98	1036
7	0.51	Ideal	G	VS1	62.0	56.0	5.16	5.06	3.17	1781
8	0.72	Ideal	F	VVS1	61.0	56.0	5.78	5.80	3.53	4362
9	0.31	Very Good	I	VVS1	62.8	55.0	4.33	4.36	2.73	571

Training *data (80%)*

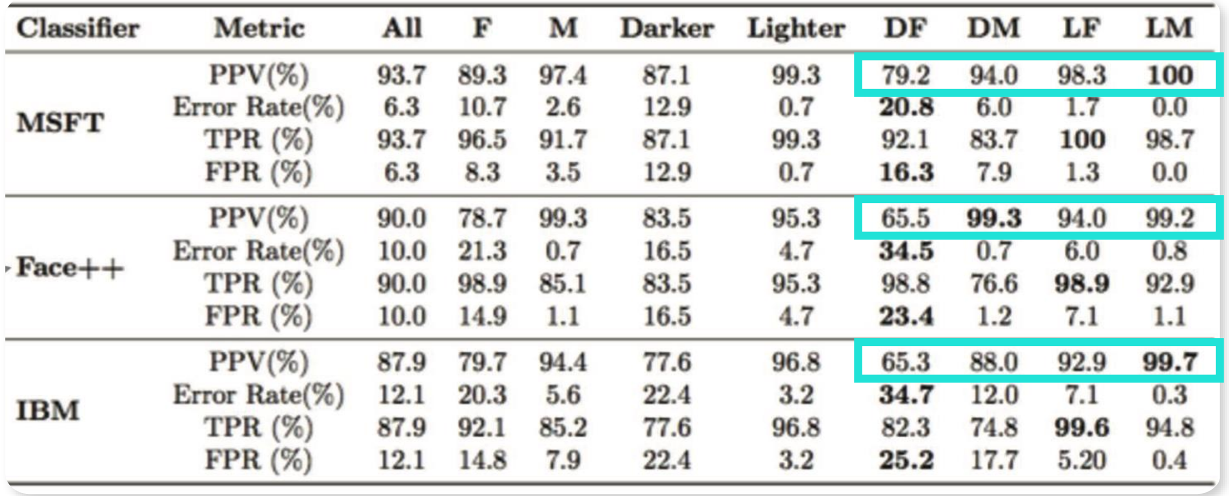
Test *data (20%)*

THE MODEL TRAINING PROCESS

ML models learn from **training** data, and are evaluated by their performance on **test** data

- If both the training and test data sets are biased, or not representative of the population the model will be applied to, **the model will likely underperform on underrepresented groups**

EXAMPLE | FACIAL RECOGNITION ACCURACY



Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Dark skinned females

Dark skinned males

Light skinned females

Light skinned males

Do you notice anything about these results?

- All models significantly underperform when it comes to detecting **dark skinned female faces**
- This population was underrepresented in the training data, leading to **algorithmic bias** in future models

EFFECTS OF DATA BIAS

Data bias is often difficult to avoid, and in many cases may be harmless

- However, when data bias does cause harm, **the impact can be substantial and may take several forms**

Poor Decision-Making

- Conclusions drawn from biased data may be incorrect or suboptimal, leading to poor decisions or missed opportunities

***Example**: You create an optional survey on your website, and respondents tend to be younger and more digitally inclined than average. Based on the survey results, you heavily promote select digital products and miss larger opportunities.*

Discrimination

- Data that is biased along demographic lines can lead to discrimination, where certain groups are disproportionately harmed or overlooked

***Example**: Your survey asks whether customers find value in your call center customer service. Because the survey skews young, you decide to shut down your call center. leaving many of your older customers without a viable contact option.*

Algorithmic Bias

- Biased data can lead to biased algorithms, which can discriminate and cause societal harm at a much larger scale *(more on that later!)*

***Example**: You fit a model on your survey results to predict if people will like your product. Because your model only included a few older respondents, it performs poorly on that audience and drives marketing decisions that bias towards younger consumers.*

THOUGHT EXERCISE: CUSTOMER SURVEY

THE SCENARIO

You're conducting market research for a consumer goods company, and plan to survey customers to ask how they feel about the brand. **How should you collect your sample to get representative results?**

Which of the following sampling strategies might lead to biased data?

- Randomly sampling the company's existing customers
- Randomly sampling existing customers, stratified by key demographic groups
- Sending a text to all customers asking them to fill out the survey
- Sampling shoppers at the company's retail shop on an iconic shopping street at noon on a Tuesday

ALGORITHMIC BIAS

Algorithmic bias is when models perform poorly on or cause disproportionate harm to certain groups

- If your data is not representative, or represents past inequities, that bias can be perpetuated by a model
- Because models automate decision making at scale, the negative impacts can be massive

Biased Data

- If the training data represents past human biases the algorithm will replicate them
- If groups are underrepresented in your data, the model may perform poorly on them

Proxy Variable Bias

- Variables that serve as a proxy for demographics (like geographic regions) can indirectly introduce bias into your model

Feedback Loops

- When biased decisions feed back into your data, algorithms continue to learn from and reinforce biased patterns

CASE STUDY: AMAZON HIRING ALGORITHM

THE
SITUATION

Amazon, a global e-commerce giant, sees tens of thousands of job applicants annually.

To streamline the hiring process, they **built an algorithm to identify top candidates by scanning resumes and LinkedIn profiles** and trained it using information from past hires and top performers.

THE
FALLOUT

Despite removing gender from the model, **the algorithm was shown to discriminate against women** by picking up on other things – whether the person played predominately women's sports, went to a women's college, or included terms more likely to be used by men like "executed" or "dominated" – which served as proxies for gender.

In the face of algorithmic discrimination, Amazon scrapped the initiative.

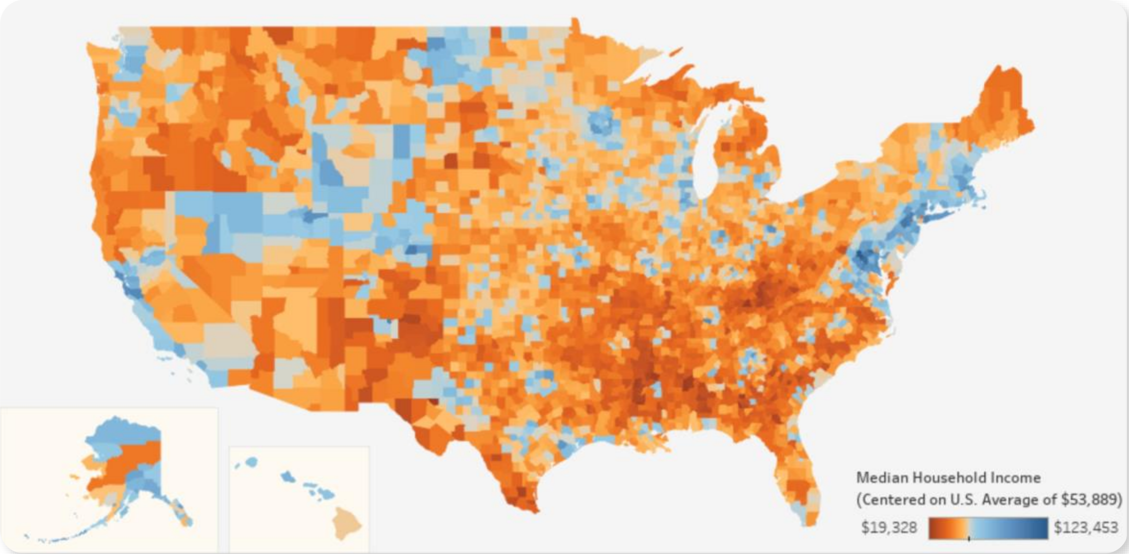
THE
ETHICS

- **Data Bias**:
  - Past hires for tech roles were predominately male, so the data wasn't representative
- **Proxy Variable Bias**:
  - Despite not using gender directly, other features in the model (like the college the applicant attended, their hobbies, and vocabulary) served as an indicator for gender
- **Algorithmic Unfairness**:
  - The algorithm discriminated against women in the hiring process, rejecting or overlooking qualified female candidates as a result

PROXY VARIABLES

Proxy variables are often used in place of variables that are difficult to measure directly

- In the context of data ethics, proxy variables often **unintentionally represent demographics like race or gender**
- This can be difficult to detect, but can lead to harmful discrimination and bias



For example, a geographic variable like **Postal Code** can be a proxy for things like:

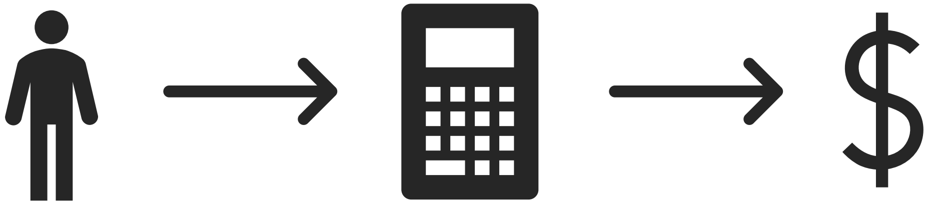
- Ethnicity
- Income Level
- Education Level
- Health Outcomes
- Religion
- Environmental Conditions

Geographic variables can be extremely powerful tools for analysis, precisely because they summarize so much information. Our advice is not to avoid them entirely, **but to be aware of potential bias and how it may impact the decisions you make**.

THOUGHT EXERCISE: PROXY VARIABLES

THE SCENARIO

You work for a health insurance company and have been asked to build a pricing model that doesn't discriminate on age, which traditionally an important factor. **Is excluding age from the model enough?**



Which of the following variables might be a proxy for age?

- Years of work experience
- Hours per week spent on social media
- Homeownership status
- Doctor visits per year
- Preferred communication method (text vs. phone call)

ALGORITHMIC HARM POTENTIAL

The **effects of algorithmic bias can be massive**, and the potential for harm is influenced by factors that include the model's domain, transparency, oversight and scale

- The higher the risk and potential for harm, the more care should be taken to build and train the model

Domain

Algorithms used in areas like criminal justice, healthcare, lending or hiring have a greater potential for harm than those used for things like outbound sales or customer retention

Scale

Algorithms that operate on a large scale and reach broader populations have a greater potential to cause harm than smaller, more targeted models

Transparency

"Black box" algorithms are complex and difficult for humans to interpret, which increases the likelihood that unethical or harmful decisions will be made

Oversight

When algorithms make decisions without human oversight or intervention, they can cause significant harm before any ethical issues are detected

MODEL TRANSPARENCY

Model transparency measures how easy it is to understand why a model makes certain predictions

- In general, more transparency means less risk of unethical decision making
- Complex models that are difficult or impossible to interpret are known as **"black box"** models, while highly transparent models are sometimes referred to as **"white box"** models

White Box Models

- Simple Model Structure
- Clear relationships between variables
- Emphasis on interpretation
- Linear Regression, Simple Decision Trees

Black Box Models

- Complex Model Structure
- Complex relationships between variables
- Emphasis on predictive accuracy
- Ensemble Models, Neural Networks

In some US industries like insurance, **government regulators require white box model structures for pricing**, to ensure that they don't discriminate against certain demographics. However, most industries are not subject to these types of regulations.

COMBATTING BIAS

There are several ways to help **combat bias**, but fostering an ethics-driven mindset is the first step

Identify Sources of Bias

- Understand how data has been collected and whether it is susceptible to bias
- Determine if bias has been introduced or amplified by algorithms, or by human decisions made at any point in the process

Audit Your Data

- Profile the data to confirm that it's representative of the population of interest
- Augment your data or tweak the data collection process to ensure that populations are fairly and accurately represented

Practice Mindful Modeling

- Screen training data for bias and use sampling techniques when needed
- Avoid using demographic variables and be aware of potential proxies
- Use transparent models when the potential for harm is high, and review outputs regularly for bias

Welcome Diverse Perspectives

- Invite and encourage diverse perspectives during the project planning, data collection and modelling process
- Ask for feedback to minimize confirmation bias and avoid ethical blind spots

THOUGHT EXERCISE: DISASTER RESPONSE

THE SCENARIO

You've been hired by the government's disaster management agency to build an algorithm that helps deploy aid to where it's needed most. **How do you build a model that distributes aid fairly?**

Which of the following model biases could lead to unfairness?

- The model prioritizes areas with the highest values of economic damage
- The model prioritizes areas with the best infrastructure to ensure timeliness of aid
- The model prioritizes areas where the populations were already vulnerable (high poverty, elderly)
- The model prioritizes areas where weather conditions are safe for aid workers
- The model deprioritizes areas with high quantities of resources (food, shelter, medical supplies)

CASE STUDY: RECIDIVISM ALGORITHM

THE **SITUATION**

In the United States, judges are turning to algorithms to assess which criminals are likely to commit another crime after being released from jail.

One such algorithm, known as **COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)**, was developed to serve such a purpose.

THE **FALLOUT**

Analysis of the model showed that “low risk” White defendants were **twice as likely to reoffend** as “low risk” Black defendants, and that Black defendants **were 45% more likely than White defendants to be labeled “high risk”**, even after controlling for age, prior crime, gender and other factors.

Recent court rulings have restricted, but not fully eliminated, the use these algorithms.

THE **ETHICS**

- **Transparency:**
    - The company behind the model will not disclose details, making it difficult to assess how the algorithm actually makes its predictions, and if it was trained on biased data
- **Algorithmic Unfairness:**
    - The risk scores generated by the model do not align with actual outcomes, and disproportionately harm Black defendants

ALGORITHMIC MORAL JUDGEMENTS

Many decisions once made by humans are now made by algorithms, raising questions about **when human judgement should take priority**

- Factors like accountability, transparency and context all come into play

Accountability

Who is responsible if an algorithm makes an unfair or unjust decision?

- ***"Sorry, that's what the model says, it's out of our hands"***

Transparency

People impacted by a model's decision should have a right to understand why

- ***"Sorry, we can't tell you how the algorithm works, it's proprietary"***

Context

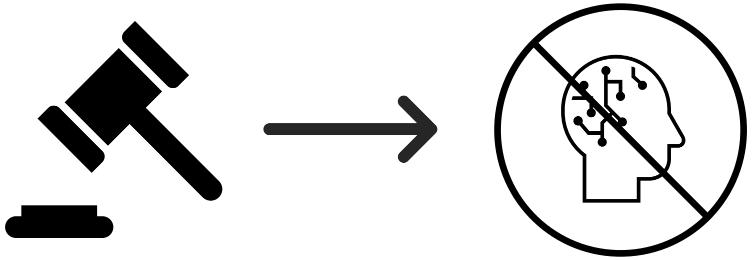
Models can't always capture the same context or nuance that humans can

- ***"The model doesn't have the data to understand the full context, but it's the best we can do"***

THOUGHT EXERCISE: LEAVING IT TO HUMANS

THE SCENARIO

You are the the government’s top AI Ethicist, and you have the authority to help shape new legislation.
Are there areas where you would outlaw the use of algorithmic or AI-based decision making?



Which decisions would you NOT want AI to make?

- Which customers should get a discount coupon mailed to their house
- Whether or not a person is guilty of a crime
- Which job a person is best suited for, regardless of their interests
- Would your perspective change if it could be proven that models were more accurate than humans?
- Would you be open to AI “assistance” if a human used model outputs as one factor in their decision?

KEY TAKEAWAYS

✓ **Data bias** exists when data doesn't represent the population of interest

- *It is not always harmful or unethical, but is generally tied to data collection and in some cases can lead to poor decision making, discrimination and algorithmic bias*

✓ **Algorithmic bias** is when models perform poorly on subsets of the population

- *Algorithmic bias is often caused by biased training data, and can lead to discrimination at a massive scale*

✓ **Algorithms aren't always the answer** when it comes to decision making

- *Models can help us make better decisions, but they aren't perfect; in some cases, offloading high-impact decisions to models can cause significant harm and lack of accountability*

AI ETHICS

AI TODAY AND TOMORROW

AI is rapidly becoming part of our daily lives and creating a positive impact in many ways, but there are still **significant ethical tradeoffs and implications to be considered**

Self-Driving Cars

Self-driving cars now legally operate in many US cities and are generally cheaper and safer than human-operated vehicles. **But how many jobs will they displace**?

AI for Students

Students are increasingly using AI to help study, research new topics, and write papers. **Will they still learn basic skills if they rely on these tools too much**?

AI for Work

Professionals are using tools like ChatGPT to boost their efficiency at work, from writing emails to writing code. **When AI makes a mistake, who is responsible**?

AI Content

AI can produce credible text, images, videos and voice samples, which can potentially be used maliciously. **Will we be able to distinguish what is real and what is AI**?

We do not intend to be alarmist, and we are still extremely optimistic about the benefits of AI despite its inherent risks. That said, one thing is clear: **understanding the ethical implications of AI has never been more important.**

AI ETHICS

While AI is subject to familiar challenges like bias and data stewardship, **its unique scale and complexity have raised some brand new and unfamiliar ethical questions**

Societal Impact

- From drivers to artists to white collar workers, millions of people could find their jobs replaced by AI
- Will AI prevent students from learning skills like writing and critical thinking?
- With the ability to mimic human words, can AI-generated misinformation or propaganda shift human behavior and opinion at scale?

Data Stewardship

- The creative works of countless writers and artists were used to train AI, often without their consent
- User chats are often captured as model inputs, which many users aren't aware of
- Private and proprietary information can be ingested by these models and potentially returned as outputs to other users

Algorithmic Bias

- The complexity of modern AI algorithms is unprecedented, and it may be impossible to understand why they produce certain outputs
- Because AI is largely trained on data collected from the internet, is it overrepresenting the views and images of developed countries or those which speak certain languages?

Hallucinations & Fraud

- Who is responsible when AI makes a mistake, and how can users validate outputs and make sure they aren't being mislead?
- AI can generate amazing new images and entertaining voice samples, but what happens when these capabilities are used to deceive or to defraud?

CASE STUDY: ARTIST LAWSUITS

THE
SITUATION

Many writers and artists are outraged that their creative works have been ingested by AI models, which can be used to produce similar, AI-generated content in mere seconds.

While these creators may have posted their work on the internet for others to view, **they didn't consent to it being used to train algorithms designed to mimic their work**.

THE
FALLOUT

Numerous lawsuits have been filed in the US against OpenAI, the creator of ChatGPT, and Stable Diffusion, a generative image algorithm.

While currently pending, there are serious implications in terms of copyright law in the age of AI. These might affect how models are trained, how creators are compensated, whether AI-generated work should be eligible for copyright protection, and the role AI plays in the creative realm in general.

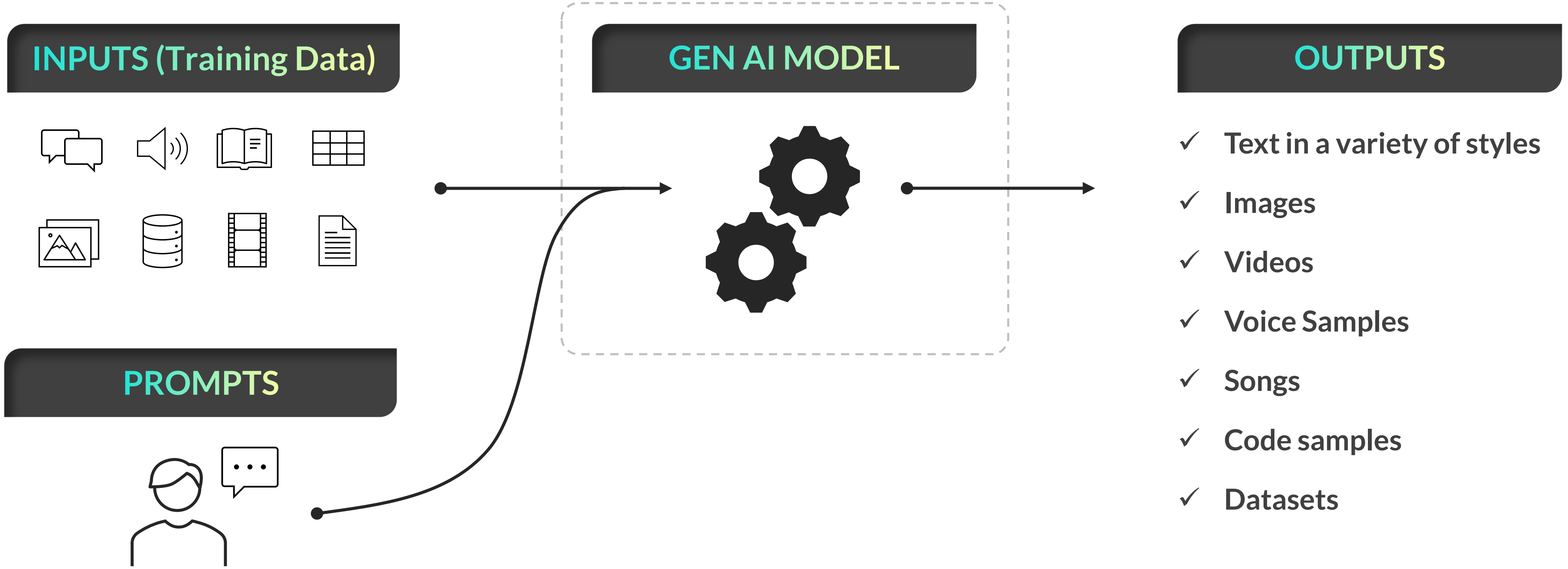
THE
ETHICS

- **Intellectual Property (IP) Theft**:
  - Countless images, paintings, and written works were used to train generative AI models, without asking for consent or compensating their original creators
- **Societal Impact - Employment**:
  - Gen AI tools couldn't exist without these creators, but ironically, they make it much easier to perform tasks that used to be labor intensive. This could mean lower compensation and fewer opportunities for the people whose work helped create these tools in the first place.

GENERATIVE AI

One of the most widely used forms of artificial intelligence is **Generative AI** – models designed to produce brand new content (text, images, video, audio, code, etc.) based on user prompts

- Gen AI tools like ChatGPT are already commonly used, but the ethical implications are still largely unknown



INPUTS (Training Data)

PROMPTS

GEN AI MODEL

OUTPUTS

- ✓ Text in a variety of styles
- ✓ Images
- ✓ Videos
- ✓ Voice Samples
- ✓ Songs
- ✓ Code samples
- ✓ Datasets

AI TRAINING DATA

Popular Generative AI algorithms, like OpenAI's GPT or Stable Diffusion for image generation, were primarily trained on **data scraped from the web**, without the consent of creators

- Ethically, AI training data faces criticisms about **IP Theft**, **bias**, **privacy** and **confidentiality**

IP Theft & Consent

- Images, articles, videos and books posted online have been used to train AI models, but does publishing your work mean that you consent to it being used for model training?
- Prompts and user inputs for Gen AI tools can also be used for training, which isn't always clear to end users

Bias

- Since models are trained on public internet data, the training dataset could be biased towards nationalities and ethnicities with larger or wealthier populations
- This can lead to biased outputs in terms of societal values, harmful stereotypes, or failure to represent diverse ethnicities

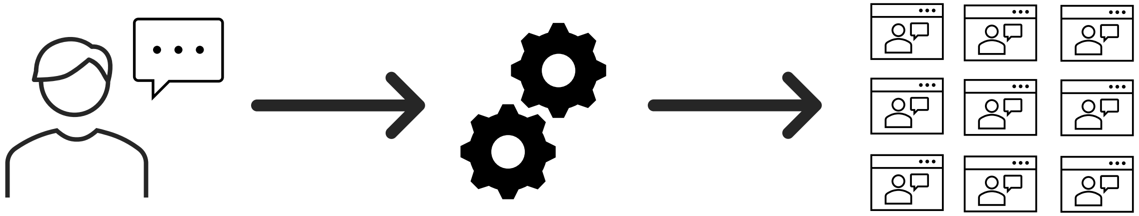
Privacy & Confidentiality

- Conversations on sites like Reddit, where users post anonymous and personal information, are often used as inputs for Gen AI models
- User chats with AI tools may also be used as training data, increasing the risk of proprietary business or personal information being inadvertently shared

THOUGHT EXERCISE: YOUTUBE CREATOR

THE SCENARIO

A popular YouTuber spent years perfecting her distinct and engaging style. When a new Gen AI video tool launches, she notices her unique style is now instantly reproduced with a simple prompt. **The company behind the algorithm profits from their new tool, while her own revenue suffers as copycats flood the platform.**



Has YouTube or the Gen AI tool crossed ethical lines?

- Does this feel different, ethically, from a human imitating her style (vs. an algorithm)?
- Should the artist be able to opt out of her stylistic influence on the algorithm?
- Is she entitled to compensation or a revenue share for the copycat videos or use of the algorithm?
- Did YouTube violate the principle of consent if it permitted her videos to be used by a third party?

CASE STUDY: AI PRODUCT REVIEWS

THE
SITUATION

Online shoppers are noticing that an **increasing number of product reviews are written by Gen AI tools** like ChatGPT. In some cases, they even begin with phrases like *"As an AI Language Model…"*

THE
FALLOUT

Consumers are losing trust in product reviews, and in some cases avoiding products that feature AI-generated reviews entirely. AI-generated reviews may also falsely advertise or promote low quality products, leaving shoppers unhappy and leading to an increase in product returns or refunds.

While companies like Amazon are looking into solutions, online shoppers will become increasingly skeptical and continue to lose trust in online ratings and reviews if the issue isn't addressed.

THE
ETHICS

- **Fraud:**
  - Using Generative AI to write product reviews is misleading and manipulates people into spending money on products they might otherwise avoid
- **Societal Impact**:
  - If consumers lose faith in in online reviews, which are a powerful signal of product quality and seller reputation, both buyers and sellers alike may be harmed and conduct fewer online transactions

GENERATIVE AI OUTPUTS

As models become more advanced, **Generative AI outputs** are becoming increasingly difficult to distinguish from human, expert-generated content

- Consider the following examples, and think about the **benefits and ethical risks** associated with them

Text Chat messages, social media posts, student essays, article-length works, etc.

Image Fantastical artwork, profile photos, pictures of friends and family, etc.

Video Sci-fi landscapes, AI avatars, fake crimes, scenes featuring politicians or public figures, etc.

Audio AI-generated music, celebrity vocals, real-time voice cloning, etc.

Code Practice SQL questions, code comments, debugging, production-level app code, etc.

HALLUCINATIONS & FRAUD

Gen AI models occasionally invent false information (known as **hallucinations**), and have the ability to produce hyper-realistic outputs that can be leveraged for **malicious or fraudulent purposes**

Text Incorrect diagnosis or potentially dangerous medical advice provided to a user

Image Fake images of real individuals used to damage their reputation or blackmail them

Video Videos featuring characters from other people’s IP, or actors in roles they did not consent to

Audio Voice samples used to manipulate family members or gain access to financial accounts

Code Code deployed to production with security vulnerabilities that put user data at risk

HALLUCINATIONS & FRAUD

Gen AI models occasionally invent false information (known as **hallucinations**), and have the ability to produce hyper-realistic outputs that can be leveraged for **malicious or fraudulent purposes**

Text Incorrect diagnosis or potentially dangerous medical advice provided to a user

Image Fake images of real individuals used to damage their reputation or blackmail them

Video Videos featuring characters from other people's IP, or actors in roles they did not consent to

Audio Voice samples used to manipulate family members or gain access to financial accounts

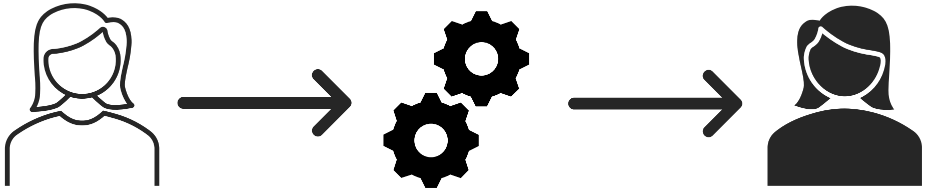
Code Code deployed to production with security vulnerabilities that put user data at risk

In cases like these, who should take responsibility?

THOUGHT EXERCISE: CELEBRITY LIKENESS

THE SCENARIO

You work for an entertainment company that produces TV shows and movies, and your boss has asked you to make an AI-generated clone of a star actor, including their voice and likeness, without telling you any other details. **Do you make the digital clone, not knowing what it will be used for?**



Which of the following use cases would you be comfortable with?

- It's for an internal research project, and will not be seen outside your team
- It will be used to help with reshoots for a film the actor is starring in
- It will be used for the next film in the series, which the actor refused to participate in
- Does your opinion change if you know the actor has given their consent?
- Do you think it matters if viewers know that the actor is AI-generated?

MITIGATING RISKS OF AI

Mitigating AI risk is a complex and challenging task; in some cases, government policy may be the only way to curb some of the potential negative effects

- That said, individual consumers and producers do have ways to reduce the adverse impact of AI

AI Producers

- Use ethically sourced and trustworthy data for model training
- Provide transparency into the model training process
- Conduct rigorous testing of model outputs for bias
- Continuously monitor and outputs and model usage

AI Consumers

- Use critical thinking and fact-check model outputs whenever possible
- Use AI tools for ethical and responsible purposes
- Be aware of and compliant with workplace AI policies
- Don't share proprietary data or sensitive personal information

Governments

- Create regulatory programs and standards to ensure the safe and ethical use of AI
- Allocate funding to research the societal impacts of AI
- Create public awareness campaigns to educate people on AI and its associated risks
- Promote global cooperation

CURRENT LEGAL STATUS

As AI usage becomes more widespread, **national and state governments are beginning to develop regulatory frameworks** to establish standards for AI safety, transparency and accountability

United States

In the US, the Biden administration released the **AI Bill of Rights** in 2022 and issued an executive order in 2023 emphasizing the need for safe and secure AI development. Several states are working on their own set of AI regulations.

European Union

The EU passed the **AI Act**, which is one of the world's most comprehensive AI laws. It aims to regulate AI use based on risk categories (from minimal to high risk), with stringent requirements for transparency and accountability.

International Efforts

- In 2023, global stakeholders (including the US, EU and China) signed the **Bletchley Declaration**, advocating for trustworthy AI development through international cooperation

KEY TAKEAWAYS

✓ The **impact of AI** will be significant and difficult to predict in the coming years

- *AI is already changing the way that we work and live, but the exact role it will play – and the ethical issues it will continue to raise – are yet to be determined*

✓ The **responsible use of AI** is critical – especially for data professionals

- *AI models aren't perfect – validate outputs whenever possible and remember that you are ultimately responsible for how that information is used*

✓ The **benefits of AI** will ultimately be maximized through a combination of responsible model building, informed use, and government regulation

- *Maintaining an ethics-driven mindset is key for ensuring the safe and responsible development of AI tools, and minimizing potential risk*