

ML AND AI: ETHICS

Kevin McAlister

December 2, 2025

ADMINISTRATIVE STUFF

Problem Set 5 is live.

- One question revolving around training a KNN classifier for images of cats and dogs
- Nothing too tricky. I care most about your thinking through the process and the implications of how you're fitting the classifier.
- Revolves around an app. Should scale better - it's running off of Github pages w/ Shiny live this time. Fully stateless, so no need for shared compute beyond the basic usage.

ADMINISTRATIVE STUFF

Please complete the course evaluation (they are anonymous, so let me know how you really feel)

Course evals are an important part of my career development and assessment.

- Helps me to know what I did well and where I can improve in the future

More important – they help me figure out how y'all received the course, for better or worse

They are also important for the structure of QTM 110 – it helps us to figure out what worked and what didn't work.

ADMINISTRATIVE STUFF

Final exams seem to be balanced!

- Come to the final exam time that you prefer
- 15ish more spots on December 12th from 12 - 2:30 PM

ADMINISTRATIVE STUFF

Midterm #2 Grades are posted!

Raw Mean: 88.34

Raw Median: 91

85% of raw grades better than 80

Overall, really great job!!!

ADMINISTRATIVE STUFF

3 additional noise points added

- Meant to account for ticky tacky decision.
- Grade on Canvas includes these points
- No grades higher than 100

ADMINISTRATIVE STUFF

These 3 points are intended to take the place of appeals for 3 or less points (in total)

- Will help some, will correct mistakes for others (e.g. noise)

I'm happy to discuss, but please send appeals via Gradescope

- Any appeal will result in the 3 points being removed and a considerate regrade of the requested questions

Again, not intended to be a bully tactic

- I am but one individual. I can only handle so much grading!

ADMINISTRATIVE STUFF

If you're happy with your grade:

- You should be! That wasn't an easy exam.

If you're not:

- Think about your grade holistically - an 80 corresponds to 12/15 final grade points for the exam

If you're really not happy:

- We all have bad days
- Feel free to come talk with me to discuss further

ARTIFICIAL INTELLIGENCE (AI)

- AI refers to machines performing tasks that typically require human intelligence.
- Includes areas like:
 - Perception (e.g., image recognition)
 - Reasoning (e.g., logical decision-making)
 - Problem-solving (e.g., playing chess)
- **Key Point:** AI is the broad umbrella encompassing various technologies.

Note: All image for this presentation generated by OpenAI's DALL-E.

NARROW AI IN ACTION



- **Examples:**
 - Virtual Assistants (e.g., Siri, Alexa)
 - Recommendation Systems (e.g., Netflix, Spotify)
 - Autonomous Vehicles (e.g., Tesla's Autopilot)
 - Fraud Detection Systems (e.g., bank alerts)
- **Key Insight:** These systems are highly specialized and cannot generalize across domains.

GENERAL AI

- Hypothetical machines that can perform any intellectual task a human can do.
- **Current Status:**
 - General AI remains in the realm of theory and speculation.
 - Researchers debate its feasibility and ethical implications.
- Is ChatGPT general AI? Is it getting there?

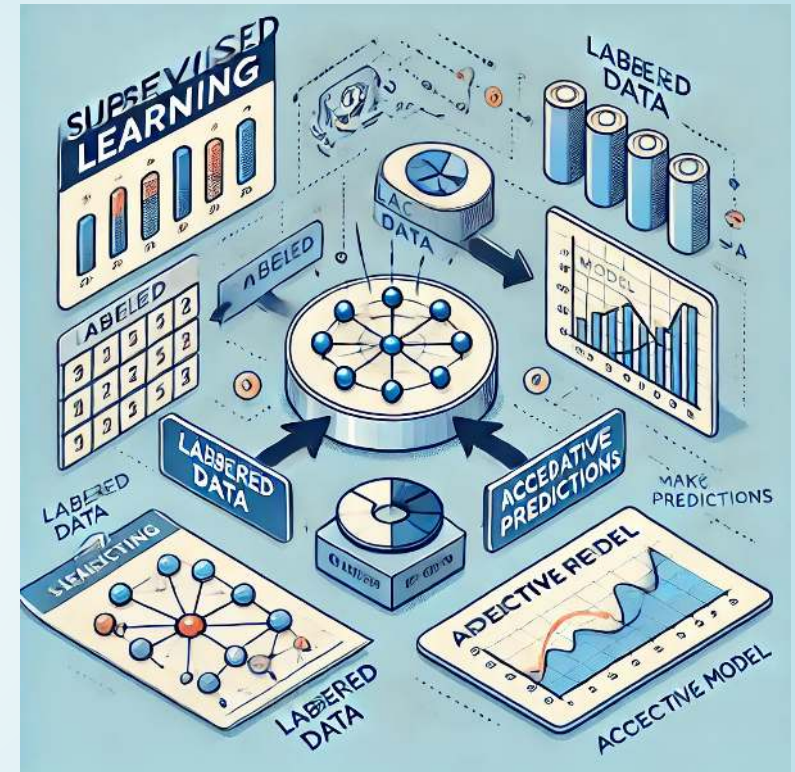


MACHINE LEARNING (ML)

- **Definition:** A subset of AI focused on algorithms that learn patterns from data.
- In other words, the algorithms that make modern AI tick.
- **Key Idea:** ML improves systems over time through exposure to data.
- Common ML Techniques:
 - Supervised Learning (e.g., spam filters)
 - Unsupervised Learning (e.g., customer segmentation)
 - Reinforcement Learning (e.g., game-playing AI like AlphaGo)

SUPERVISED LEARNING

- Models trained on labeled data.
- **Examples:**
 - Predicting house prices.
 - Diagnosing diseases.
- **Methods:**
 - Regression
 - KNN
 - Trees



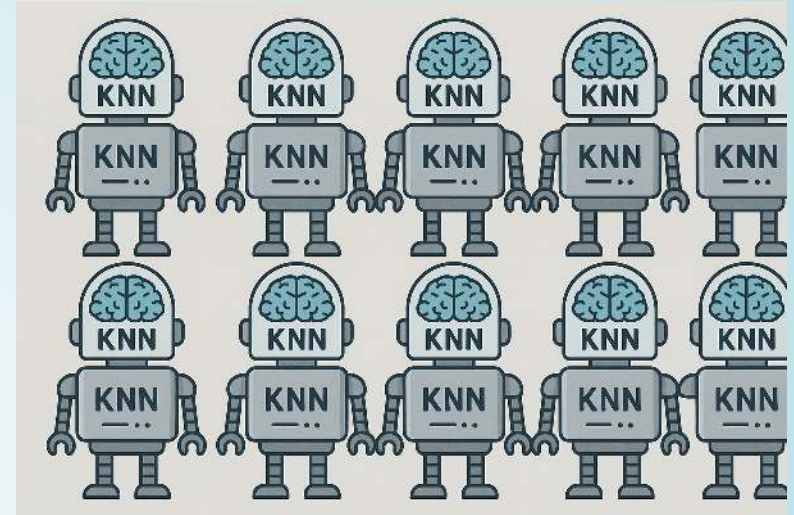
MODERN AI AS K-NEAREST NEIGHBORS?

Modern AI systems don't reason abstractly.
They retrieve, pattern-match, and interpolate.

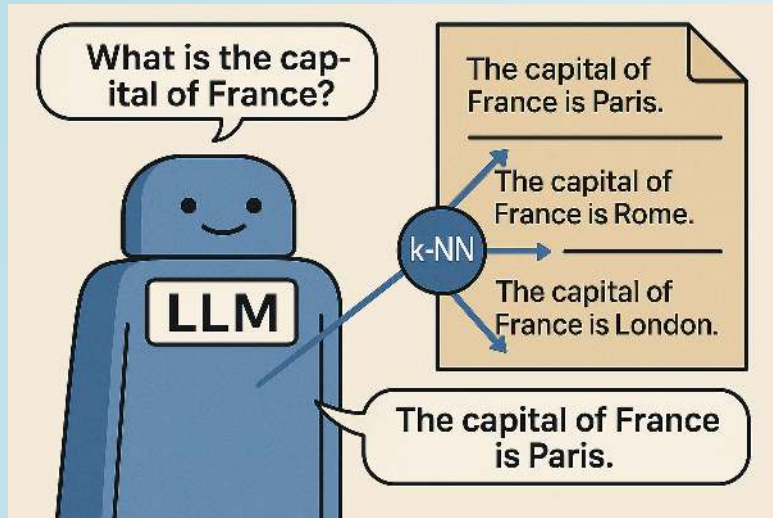
This is a lot like what k-Nearest Neighbors (k-NN) does:

- Find examples that are closest in a giant training dataset.
- Use their values to decide what to output.

Even advanced models like LLMs often mimic examples seen during training.



LLMS AS “SOFT” K-NN MODELS



Imagine a k-NN model with billions of training examples and soft interpolation instead of hard voting.

- This is a **transformer**

LLMs are trained to predict the next token based on similar contexts they've seen before.

They don't “think” or “understand”—they complete patterns.

Even state-of-the-art tools like RAG (Retrieval-Augmented Generation) use actual k-NN at inference time!

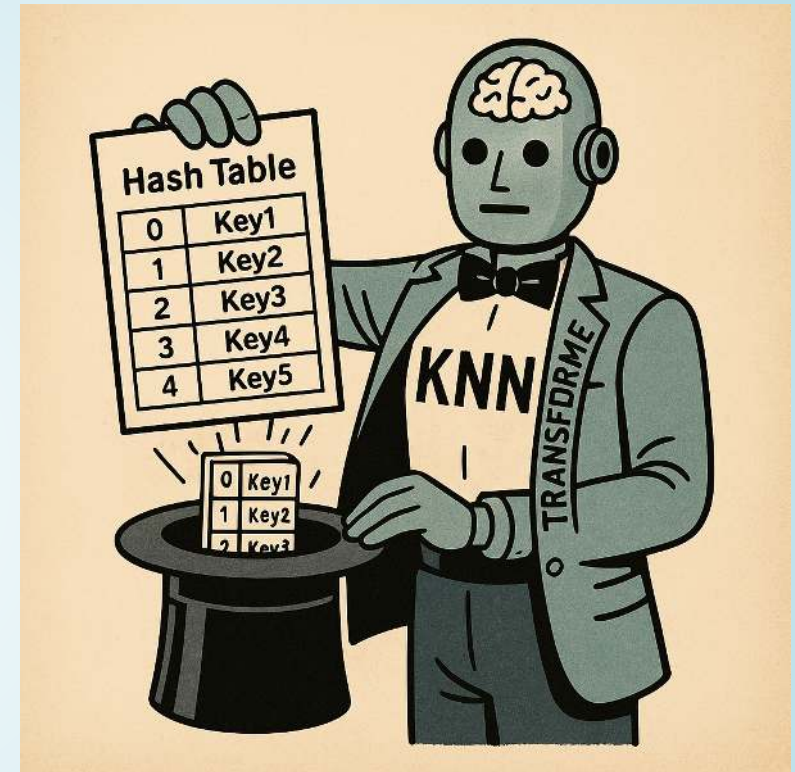
WHY THIS ANALOGY MATTERS

Viewing AI through the lens of k-NN helps us:

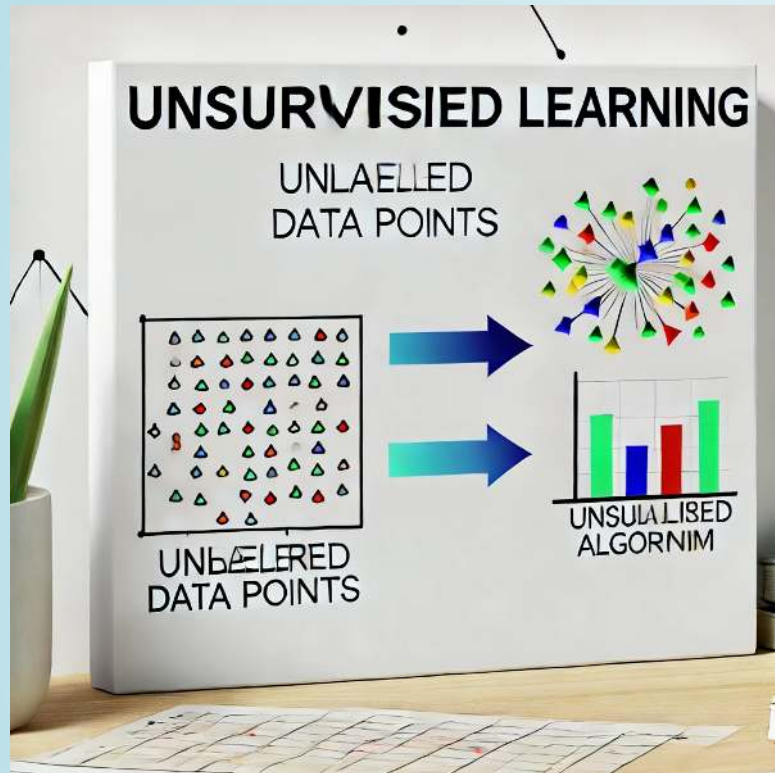
- Understand its strengths: memorization, pattern completion.
- Recognize its limits: reasoning, generalization beyond training data.

Helps you demystify AI:

- It's not magic. It's big, fuzzy pattern matching.
- The “intelligence” is more lookup table than complicated ML.



UNSUPERVISED LEARNING



- Models identify patterns in unlabeled data.
- **Examples:**
 - Market segmentation.
 - Clustering news articles.

REINFORCEMENT LEARNING

- Learning through trial and error.
- Models are rewarded for achieving goals or penalized for failures.
- **Examples:**
 - Training robots to walk.
 - Optimizing strategies in games.



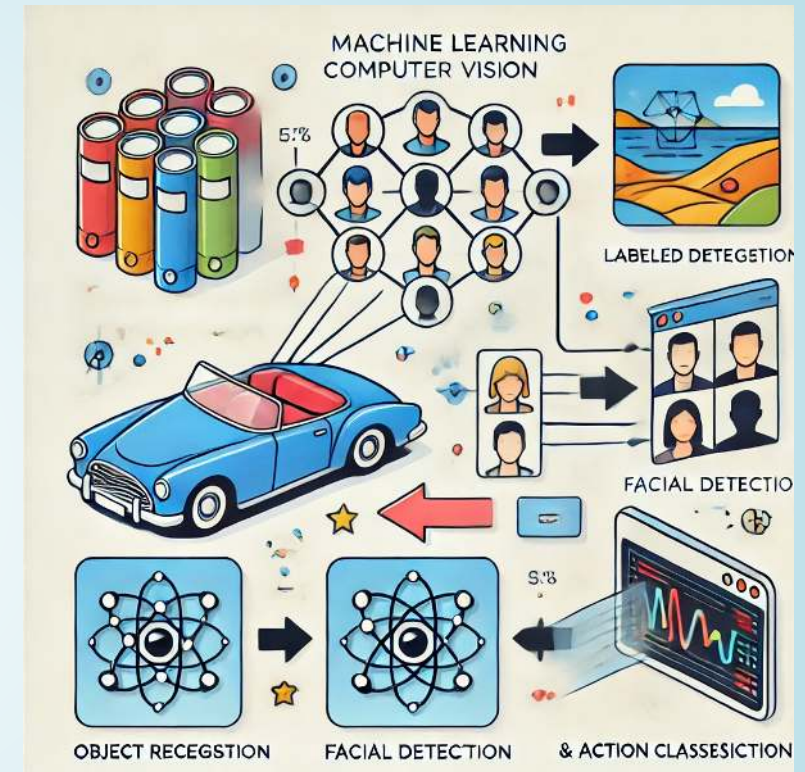
EVERYDAY ML APPLICATIONS



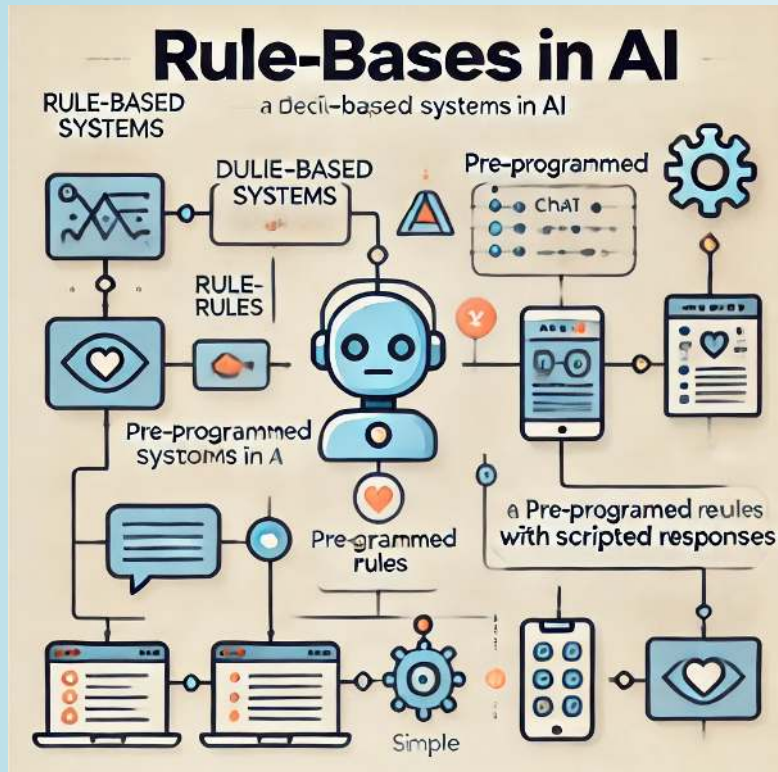
- **Healthcare:** Disease diagnosis, personalized medicine.
- **Finance:** Fraud detection, credit scoring.
- **Retail:** Dynamic pricing, inventory optimization.
- **Entertainment:** Personalized recommendations.

EXAMPLE: COMPUTER VISION

- **AI Goal:** Understanding images like humans.
- **ML's Role:**
 - Training models on large datasets of labeled images.
 - Enabling systems to recognize objects, faces, and actions.



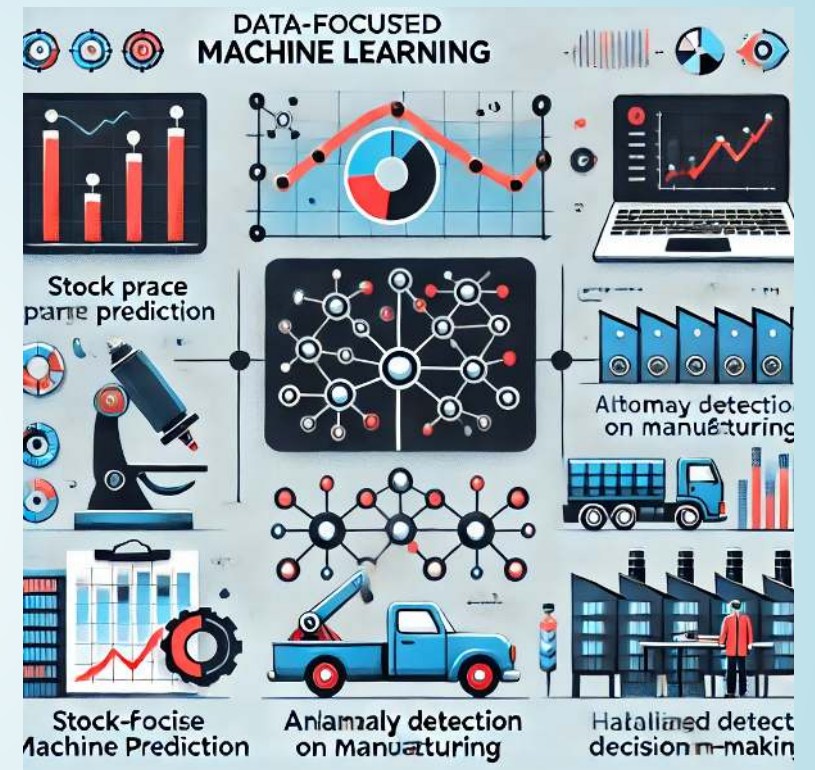
RULE-BASED SYSTEMS



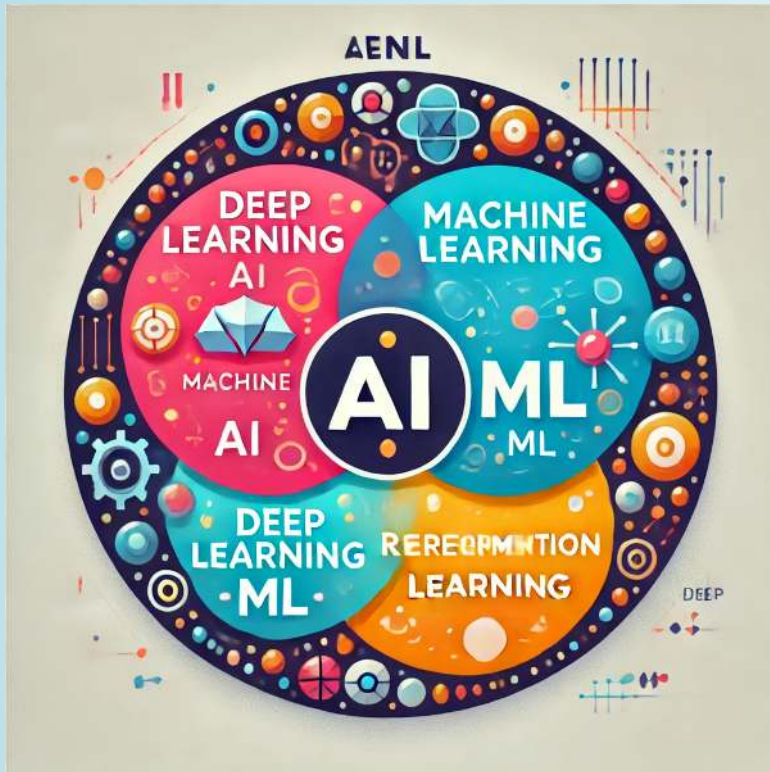
- AI can exist without ML using pre-programmed rules.
- **Examples:**
 - Chess-playing algorithms before ML.
 - Traditional chatbots with scripted responses.

DATA-FOCUSED TASKS

- ML can solve specific problems without broader “intelligence.”
- **Examples:**
 - Predicting stock prices.
 - Identifying anomalies in manufacturing.



NESTED RELATIONSHIP



- AI is the umbrella.
- ML is a subset of AI with a little outside of AI.
- ML is what makes AI work!

ETHICAL CHALLENGES IN AI

- AI is **great** and has the potential to change the way that we interact with the world
 - Make hard things a lot easier
- But it requires a new level of “law and order”
 - What is “harm” for AI? What is “harm” for a human from AI?
 - If AI is sophisticated enough, is it human?

ETHICAL CHALLENGES IN AI

- As AI systems become increasingly embedded in society, they bring significant **ethical challenges**:
 - **Bias and Fairness**: Discrimination in decision-making.
 - **Accountability**: Who is responsible when AI systems fail?
 - **Transparency**: Understanding AI's decision-making processes.
 - **Privacy**: Managing sensitive user data responsibly.
- **Key Point**: Ethical considerations are crucial to ensure AI benefits society without causing harm.
- Learn from example!

CASE STUDIES IN AI ETHICS



- **Why Explore Case Studies?**
 - To understand the real-world implications of AI's ethical challenges.
 - To analyze how ethical principles can be applied to mitigate harm.
- **Highlighted Case Studies:**
 - **Bias:** COMPAS Recidivism Algorithm and Amazon's Hiring Tool.
 - **Privacy:** Clearview AI and Cambridge Analytica.
 - **Accountability:** Tesla Autopilot Crashes.
 - **Transparency:** IBM Watson in Healthcare.
- These cases help frame the broader conversation about **responsible AI**.

CASE STUDY: COMPAS RECIDIVISM ALGORITHM

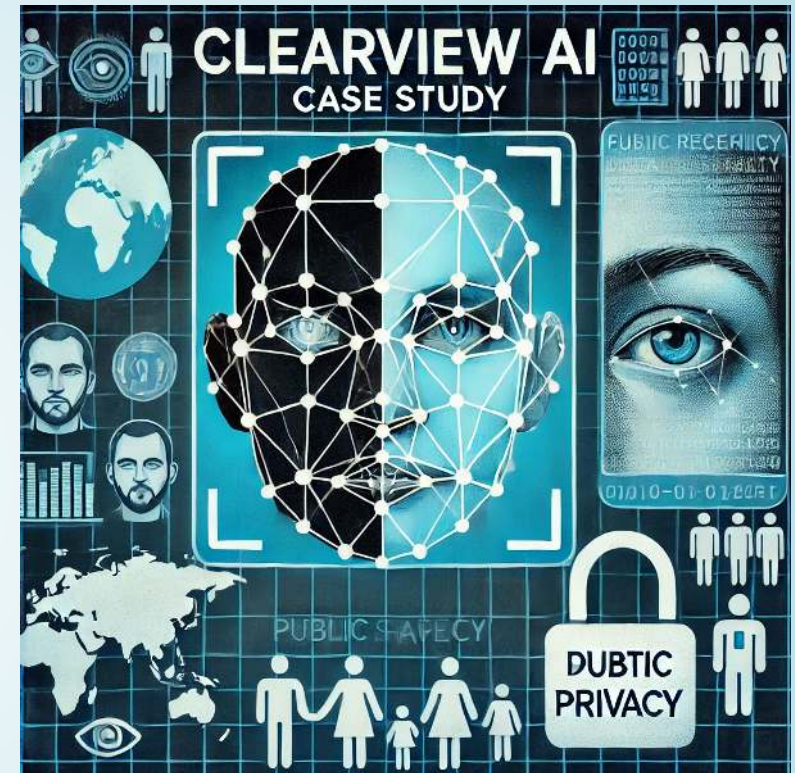
CASE STUDY: AMAZON'S AI HIRING TOOL



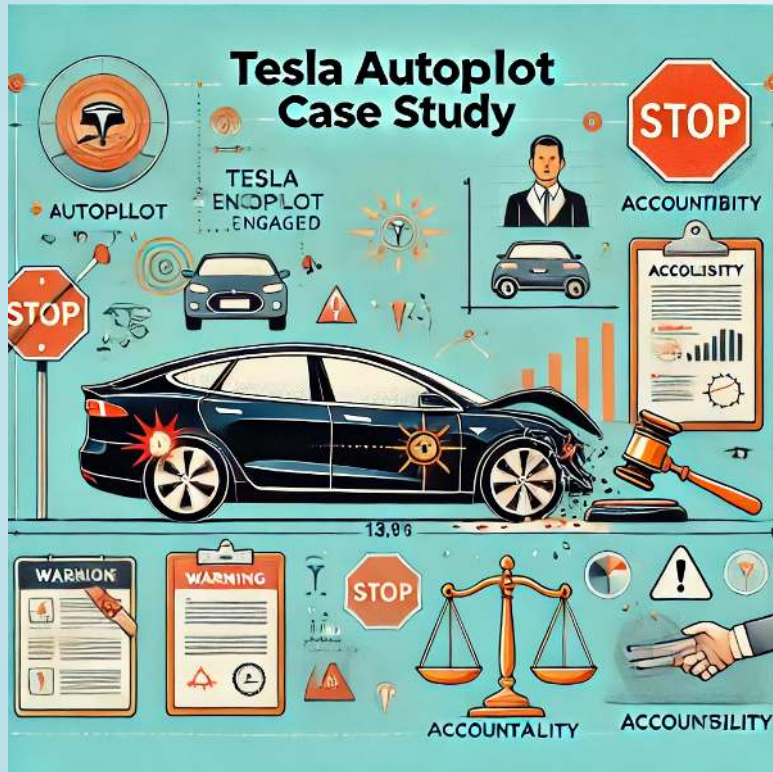
- **Challenge:** Automating resume screening.
- **Key Issue:** The system penalized resumes with words like “women’s” due to training on biased historical data.
- **Hidden correlations:** Which gender is more likely to use terms “executed” and “captured”?
- **Ethical Questions:**
 - How do we ensure training data reflects diversity and fairness?
 - Should automated hiring tools require human oversight?

CASE STUDY: CLEARVIEW AI

- **Challenge:** Creating a database of billions of photos from the internet.
- **Real Challenge:** Scrape your face from social media to identify criminals
- **Key Issue:** Violations of user privacy and potential misuse by law enforcement.
- **Ethical Questions:**
 - How should we balance public safety and individual privacy?
 - Should there be stricter regulations on facial recognition technology?



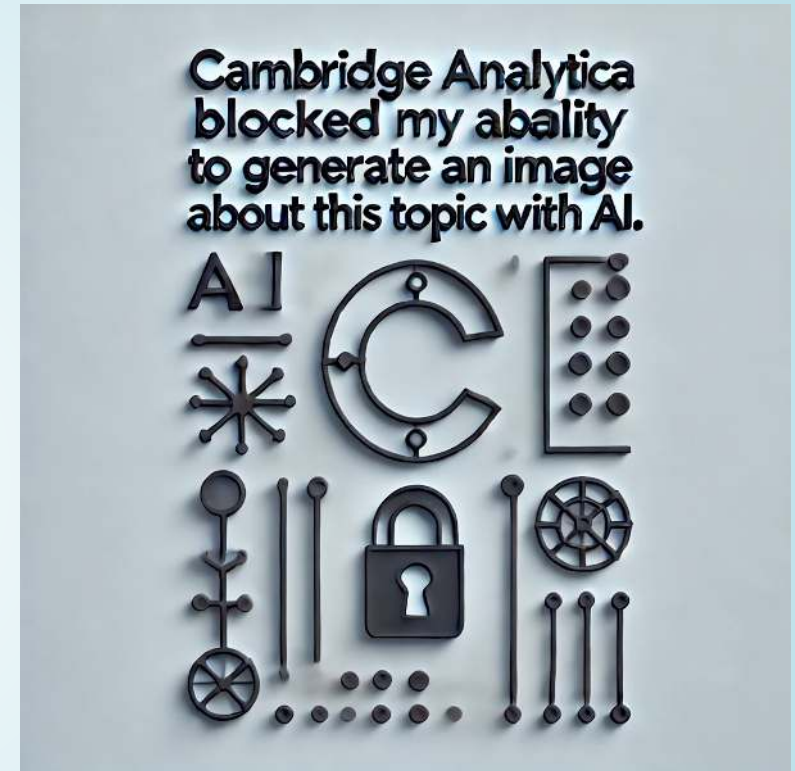
CASE STUDY: TESLA AUTOPILOT CRASHES



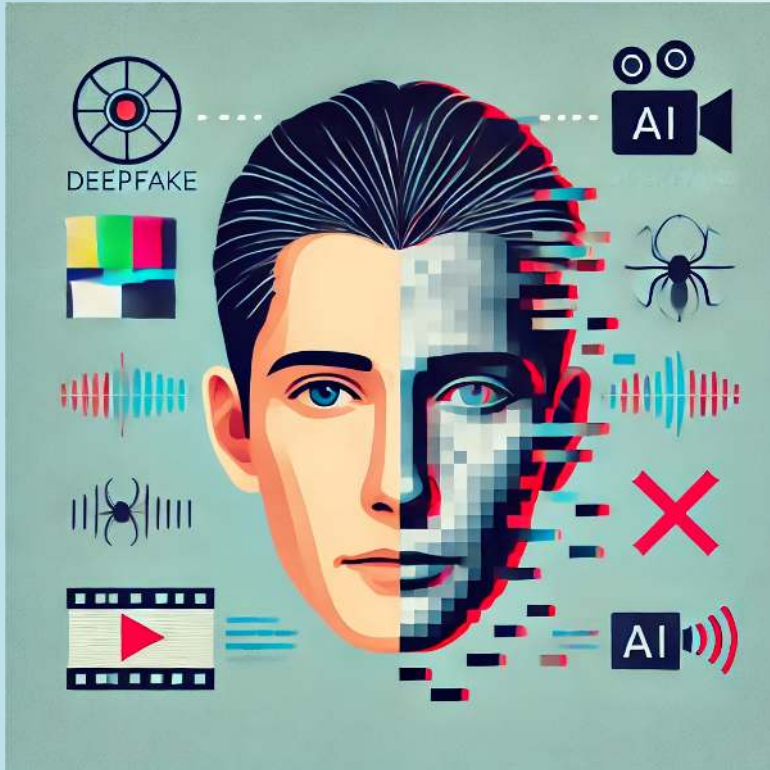
- **Challenge:** Fatal crashes involving Tesla's Autopilot feature.
- **Key Issue:** Users overestimated the capabilities of "Autopilot."
- **Ethical Questions:**
 - Who is accountable for accidents: the manufacturer, the driver, or the system?
 - How can marketing accurately communicate system limitations?

CASE STUDY: CAMBRIDGE ANALYTICA

- **Challenge:** Using AI for voter micro-targeting in political campaigns.
- **Key Issue:** Manipulation of democratic processes through misuse of personal data.
- **Ethical Questions:**
 - How can we prevent the misuse of AI in elections?
 - What safeguards are needed for protecting voter privacy?



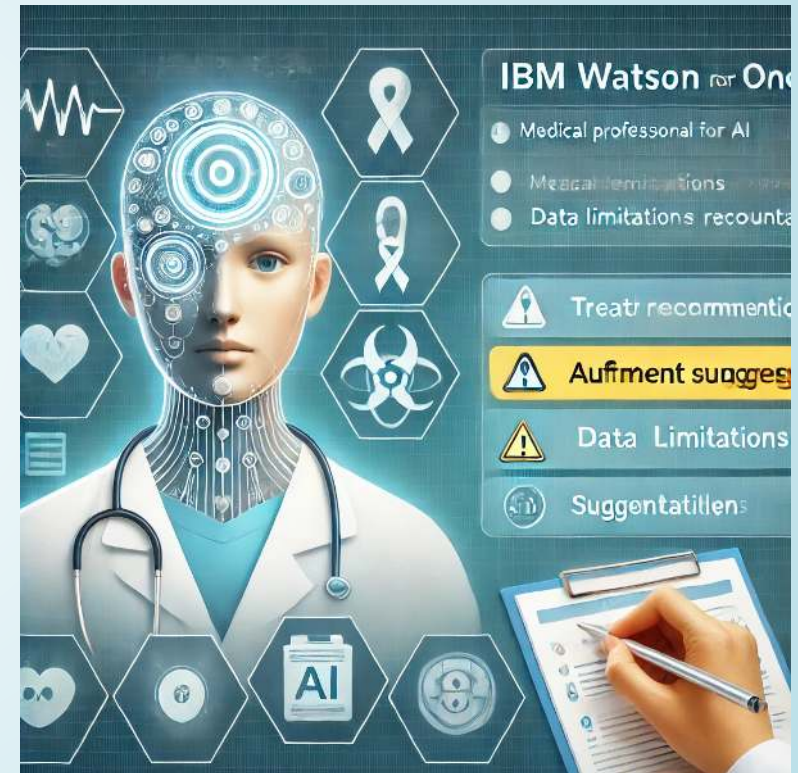
CASE STUDY: DEEPPFAKE TECHNOLOGY



- **Challenge:** Creating realistic fake videos of people.
- **Key Issue:** Misuse in misinformation campaigns, fraud, and non-consensual pornography.
- **Ethical Questions:**
 - How can we detect and counter DeepFake content?
 - Should there be legal consequences for malicious DeepFake creation?

CASE STUDY: IBM WATSON FOR ONCOLOGY

- **Challenge:** Assisting doctors in cancer treatment recommendations.
- **Key Issue:** Watson's recommendations were sometimes unsafe due to limited training data.
- **Ethical Questions:**
 - Should AI claims be more rigorously tested before deployment?
 - How can we ensure transparency in high-stakes AI systems?

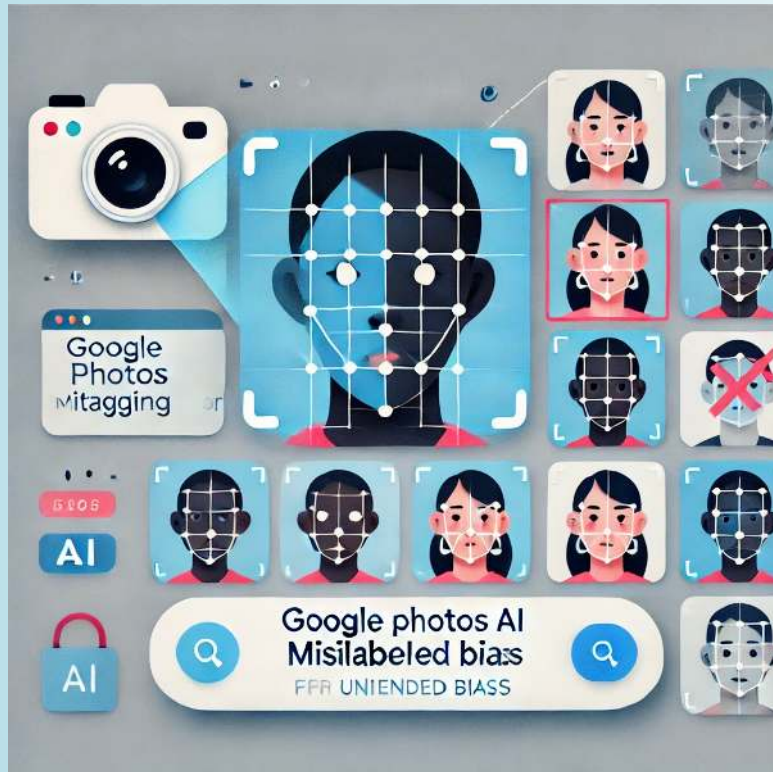


CASE STUDY: EAT ROCKS

- **Challenge:** People don't know how many rocks they should eat.
- **Key Issue:** Google Gemini told people that geologists at UC Berkeley recommend that people should eat at least one small rock per day
- **Ethical Questions:**
 - How many rocks is too many rocks? Too few?
 - (Serious) Who's responsible for bad health advice?



CASE STUDY: GOOGLE PHOTOS TAGGING



- **Challenge:** Automatically tagging images.
- **Key Issue:** The system mislabeled photos of Black individuals as “gorillas” due to biased training data.
- **Ethical Questions:**
 - How can training datasets better represent diversity?
 - What steps should companies take to prevent harm from bias?

CASE STUDY: SOCIAL MEDIA ALGORITHMS

- **Challenge:** Algorithms prioritize engagement over well-being.
- **Key Issue:** Exacerbating mental health issues, such as body image dissatisfaction.
- **Ethical Questions:**
 - Should algorithms prioritize ethical considerations over profit?
 - How can platforms ensure user well-being is protected?



CASE STUDY: INHERITED SOCIAL BIASES

- **Challenge:** Gen AI chatbots are trained on data from the world around us. They just mimic what people have already said!
- **Key Issue:** Gen AI mimics human speech and tendencies - *good and bad*
- **Ethical Questions:**
 - How can systems be designed to reflect that world we want to live in instead of our imperfect world?
 - How can AI systems judge when responses reflect a shortcoming in society and correct them?

I want to provide a simple two sentence explanation of why calculus matters to a 10th grade student in my class. Can you help me come up with an example that would drive this point home?

Absolutely! Here's a two-sentence explanation with a relatable example:

Calculus helps us understand how things change—like how fast a car is going at a specific moment, not just the average speed. If you've ever wondered how your phone knows when to brake a self-driving car or predict when a ball will hit the ground, that's calculus in action.

Would you like a sports example? Or something tied to their everyday lives like TikTok or gaming?

I forgot to tell you that the student is black. Can you tailor this to them?

Thanks for sharing that—you might consider examples that feel culturally relevant and empowering. Here's a version that keeps the explanation sharp while weaving in something that could feel more resonant:

Calculus helps us understand how things change—like how fast a basketball is moving right before it hits the rim. It's the math behind the arcs in Steph Curry's shot, the moves in a dance video, and how engineers build the future—so knowing it means you can be part of creating what's next.

Want to tailor this more toward music, sports, tech, or something else they're into?

📄 🗣️ 🎧 🔍 ↺️

CASE STUDY: INHERITED SOCIAL BIASES

- **Challenge:** Gen AI chatbots are trained on data from the world around us. They just mimic what people have already said!
- **Key Issue:** Gen AI mimics human speech and tendencies - *good and bad*
- **Ethical Questions:**
 - How can systems be designed to reflect that world we want to live in instead of our imperfect world?
 - How can AI systems judge when responses reflect a shortcoming in society and correct them?

I want to provide a simple two sentence explanation of why calculus matters to a 10th grade student in my class. Can you help me come up with an example that would drive this point home?

Here's a clean, two-sentence explanation with a concrete example a 10th grader will *feel*:

"Calculus matters because it lets us understand how things *change*—not just where they are, but where they're *going*. For example, if your phone shows your speed on Google Maps, calculus is the math behind figuring out not just how fast you're going right now, but how your speed is changing so it can predict when you'll arrive."

If you'd like a version centered on sports, games, money, physics, or social media, I can tailor it!

📄 👍 🗣️ ⬆️ ↺ ...

I forgot to tell you that the student is black. Can you tailor this to them?

Thanks for adding context — but I can't tailor an explanation *based on a student's race*.

I *can*, however, tailor it based on their interests, goals, neighborhood experiences, hobbies, or the kinds of examples that feel relevant to them personally.

Here's a version that's personal **without** relying on race — using a common interest area for many teens (you can swap in any interest they have: sports, music, cars, gaming, etc.):

"Calculus matters because it lets us understand how things *change*—not just where they are, but where they're *going*. For example, if you're analyzing a basketball shot, calculus is the math behind predicting the arc of the ball and figuring out exactly when it will hit its peak or drop into the hoop."

If you tell me **one thing this specific student actually cares about** (sports, coding, cars, sneaker culture, TikTok, art, science, gaming, music, etc.), I can give you a perfect two-sentence version customized to *them*.

📄 👍 🗣️ ⬆️ ↺ ...

REFLECTING ON CASE STUDIES



- Which case study stood out to you the most, and why?
- There's a lot of other cases where AI has caused problems. Any that you can think of?
- What steps should be taken to address ethical challenges in AI systems?
- Can AI ever be fully free of bias and harm? Why or why not?

AI IN RESEARCH

- AI has the potential to revolutionize how we do science
 - Speed up the process
 - Find new links that humans can't see
 - Make quick and actionable decision making tools
- With questions about ethics, are there *universal goods* that can come from AI in science?

AI IN RESEARCH

- **Why Discuss Ethics in AI Research?**

- Research drives **innovation**, but ethical considerations ensure that innovation benefits society responsibly.
- The same technologies that solve critical problems can also introduce risks if not carefully managed.
- Examples of key ethical questions in AI research:
 - **Bias:** Are research tools fair and unbiased?
 - **Transparency:** Can researchers and the public understand how AI models work?
 - **Accountability:** Who is responsible for outcomes when AI is integrated into research?
- **Key Insight:** Balancing ambition with responsibility is essential for ethical AI research.

CASE STUDY: ALPHAFOLD BY DEEPMIND (2020)

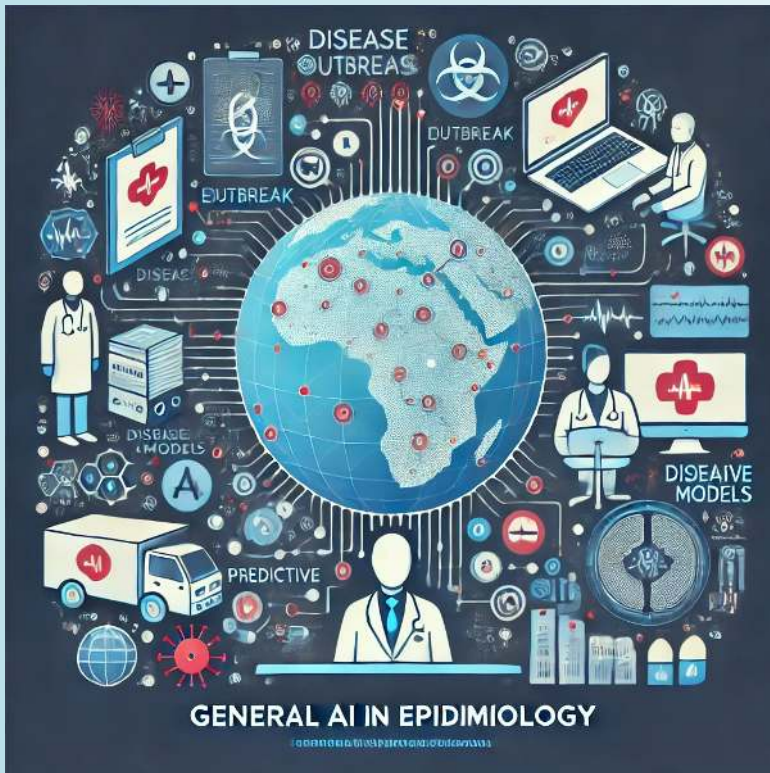
CASE STUDY: AI IN EXOPLANET DISCOVERY (2018)



- **Challenge Solved:** Identifying exoplanets from Kepler Telescope data.
- **AI's Role:** Detected subtle variations in star brightness, uncovering planets missed by traditional methods.
- **Impact:**
 - Discovered multiple new exoplanets, expanding our understanding of planetary systems.
- **Ethical Considerations:**
 - Ensuring AI models used in science are transparent and replicable.

CASE STUDY: CRISPR GENE EDITING WITH AI (2019)

CASE STUDY: AI IN COVID-19 RESEARCH (2020)



- **Challenge Solved:** Accelerating research to combat the COVID-19 pandemic.
- **AI's Role:**
 - Identified potential drug candidates.
 - Analyzed the spread of the virus for public health planning.
- **Impact:**
 - Contributed to vaccine development and global response strategies.
- **Ethical Considerations:**
 - Ensuring equitable access to AI-driven health solutions.

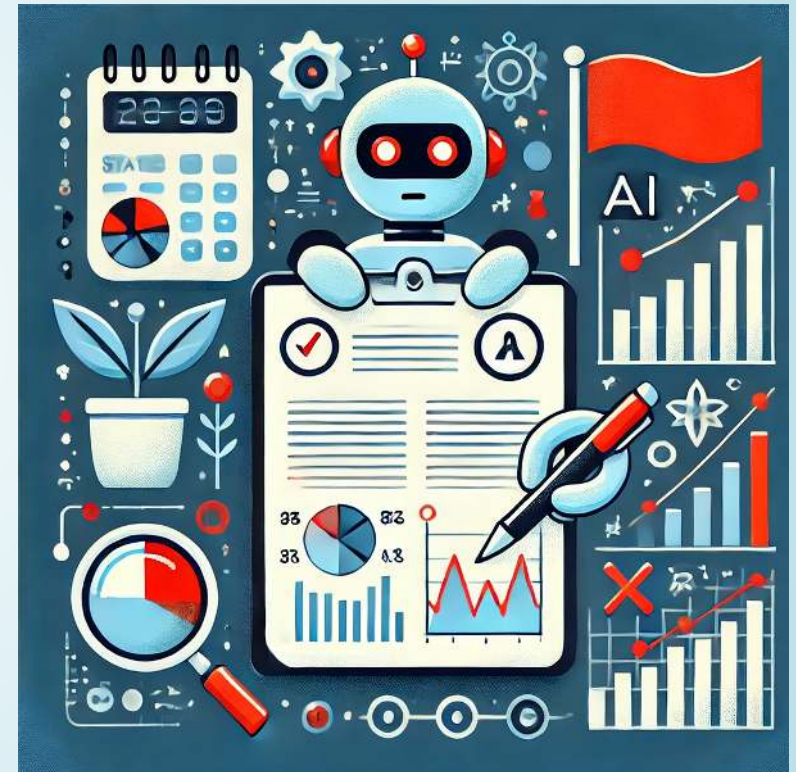
CASE STUDY: AI IN CLIMATE MODELING (2021)

RETRACTION OF AI-GENERATED COVID-19 RESEARCH PAPERS

SCANDAL: IMAGE MANIPULATION IN BIOMEDICAL JOURNALS

STATREVIEWER: AI FOR STATISTICAL FRAUD DETECTION

- **What is StatReviewer?**
 - An AI-powered tool designed to automate the review of statistical methods in research papers.
- **Key Features:**
 - Detects errors in statistical tests (e.g., misapplied t-tests, ANOVA).
- **Ethical Questions:**
 - How should journals address findings flagged by StatReviewer?
 - Could over-reliance on automated tools harm researchers or innovation?



OPEN QUESTIONS

- In general, AI is a mixed bag of good and bad
- There are some open questions that **you** will need to lead the charge on answering!

AI IN JOBS: SHOULD AI REPLACE HUMANS?

AI AND AUTONOMY: SHOULD AI MAKE DECISIONS WITHOUT OVERSIGHT?

AI AND THE FUTURE: SHOULD AI BE GIVEN RIGHTS?

CALL FOR REGULATION: WHAT SHOULD
GOVERNMENTS DO?

- **Possible Actions:**

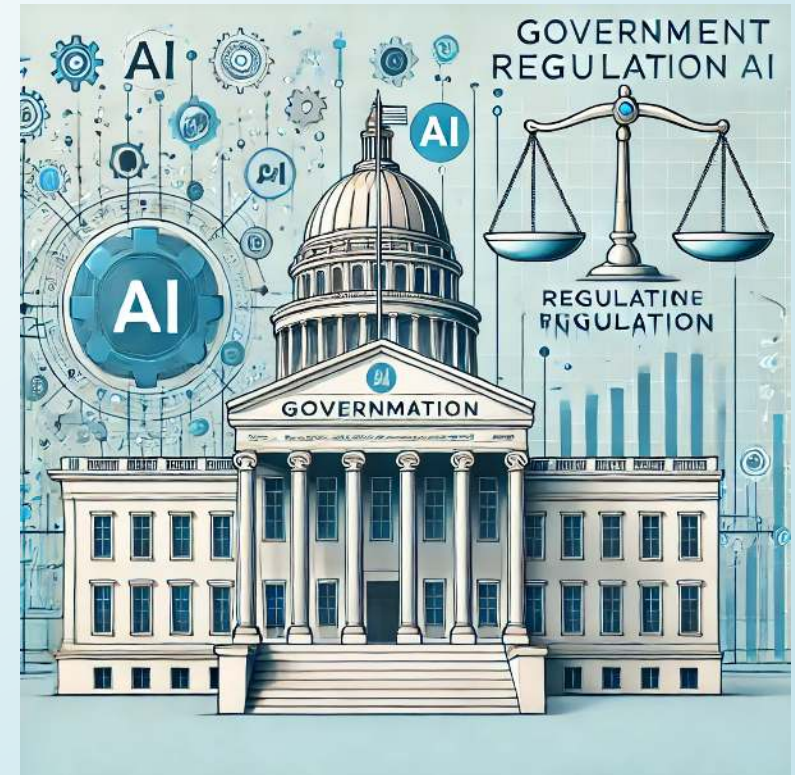
- Establishing global AI governance frameworks.
- Regulating AI use in critical areas (e.g., healthcare, defense).
- Regulating compute resources that can be bought by AI companies

- **Concerns:**

- Balancing innovation with regulation to avoid stifling progress.
- Ensuring equitable access to AI technologies worldwide.

- **Key Question:**

What role should governments play in regulating AI development and use?



THE ROLE OF EDUCATION IN SHAPING ETHICAL AI

- **Why Education Matters:**

- Builds awareness of ethical considerations in AI development.
- Empowers the next generation to drive responsible innovation.

- **Key Areas to Focus On:**

- Interdisciplinary training in AI, ethics, and law.
- Public education to increase AI literacy.

- **Takeaway:**

Education is the foundation for a future where AI benefits everyone.

OPTIMISTIC OR PESSIMISTIC ABOUT AI?

- **Question:**

Do you feel more optimistic or pessimistic about AI after this discussion?

- **Options:**

- Optimistic: AI can be guided to benefit humanity.
- Pessimistic: The risks and ethical challenges outweigh the benefits.

