ETHICS AND A.I.

CIVIC INNOVATION LAB
UNAM

R3D
Red en Defensa de los Derechos Digitales

EPIC Queen

This class is based on:

- CMU's Human-AI Interaction by Chinmay Kulkarni and Mary Beth Kery.
- Industry A.I. courses and standards

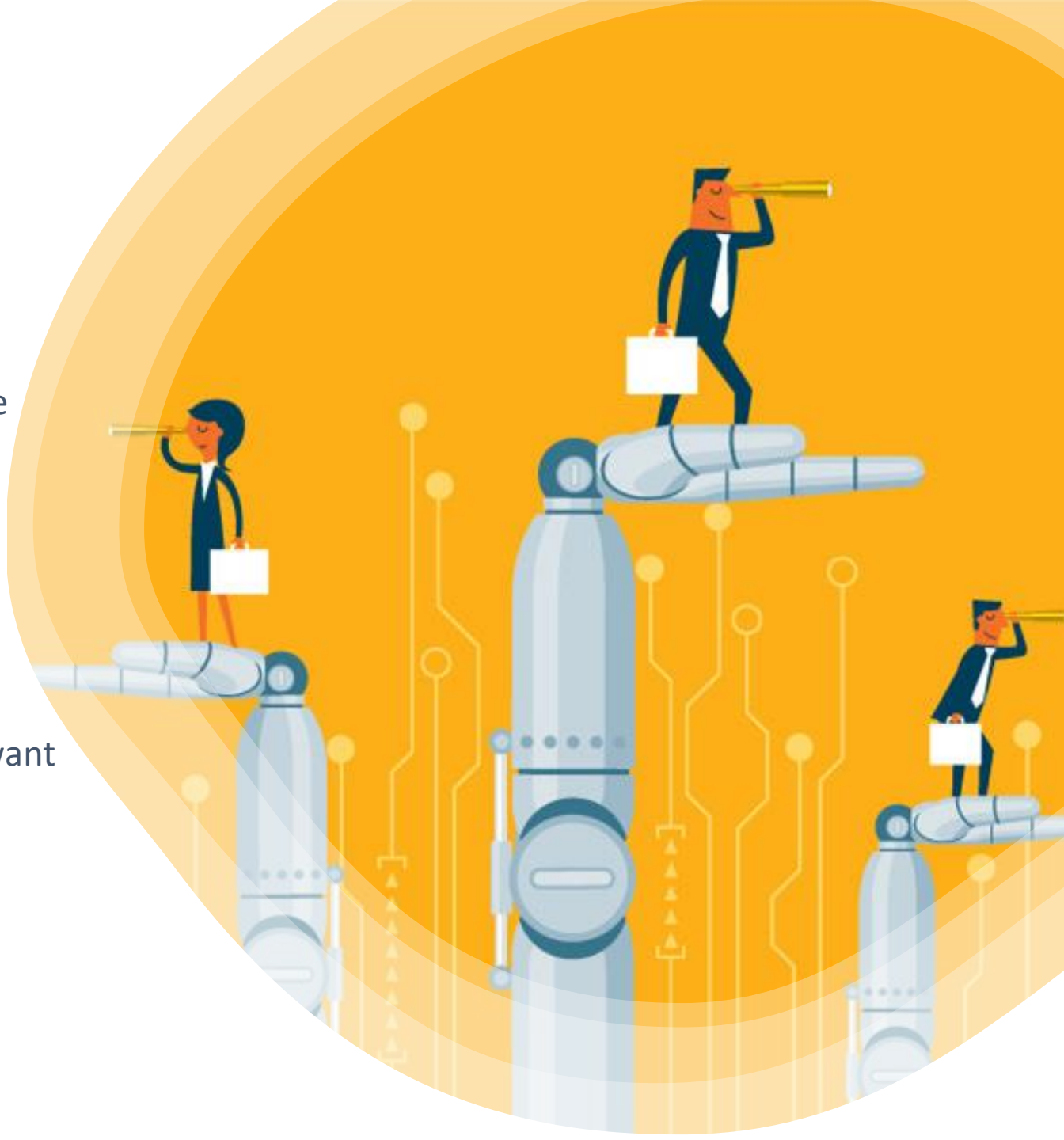
Goals

- After today's micro-couse , you should be able to:
  - Understand main concepts around ethics & A.I.:
    - Why Ethics & A.I.?
    - Understand basic A.I. model
    - Identify basic principles for thinking about A.I. and Ethics
      - Fairness
      - Transparency
      - Inclusivity
  - Understand technical concepts around ethics & A.I.:
    - Determine where the Standard A.I. models fails and ethical problems can emerge
      - Describe what are the overlooked factors that can bring ethical problems
    - Define common Terminology used in talking about A.I. and Ethics
  - Have practical experience around A.I. and Ethics
    - Coding: Deep Fakes and Disinformation

Main Concepts in AI & Ethics

Why Ethics & A.I.?

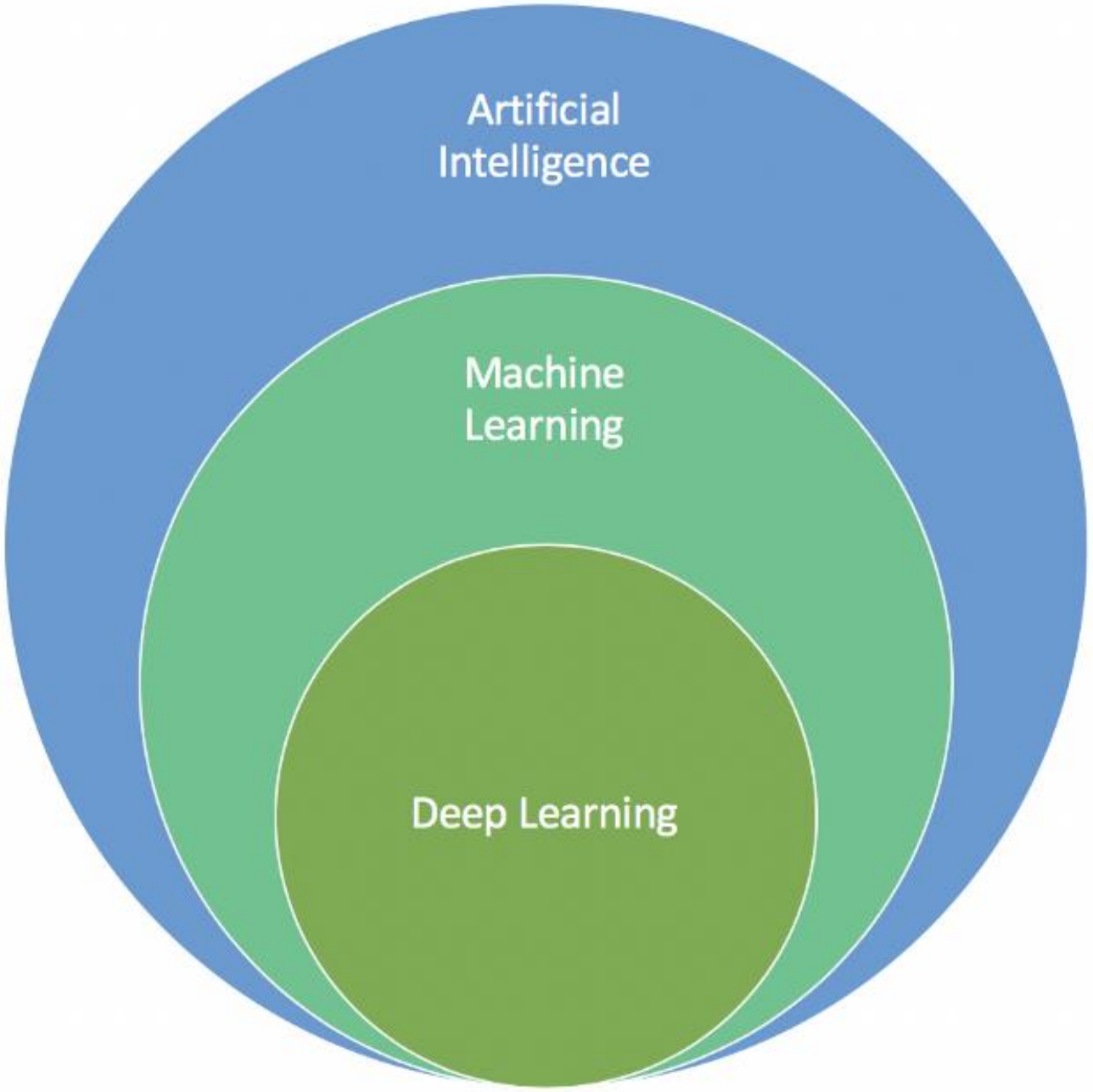
- We are the first generation to have machines that do the decisions that humans have traditionally done.
- It is important we get this right.
- We need to enable machines that think ethically.
- The public no longer just wants great technology, they want people who can design technology in a responsible way.



What is Artificial Intelligence?

- **Artificial intelligence** — "It *is the science and engineering of making computers behave in ways that, until recently, we thought required human intelligence.”* --- Andrew Moore

- **Machine Learning** — It is an application of artificial intelligence that provides the AI System with the ability to automatically learn from the environment and applies that learning to make better decisions



Machine Learning

- A “machine” that is able to improve based on past experience without explicit human programming on how to improve each time.
  - *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."* -- Tom M. Mitchell

Standard Machine Learning Pattern

Raw Data

1

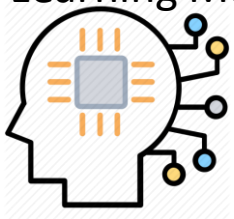
Data Cleaning

2

Feature Extraction

3

Machine Learning Model



4

- Machine Learning Training Model

Standard Machine Learning Pattern

Raw Data

1

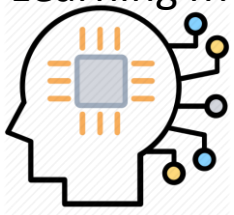
Data Cleaning

2

Feature Extraction

3

Machine Learning Model



4

- Machine Learning Training Model

New Data

Data Cleaning

Feature Extraction

Trained Machine Learning Model

Predicted Label

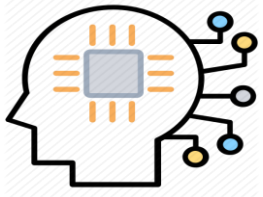
Example: App that learns to predict whether a person has Coronavirus

Medical Data with labels on who has COVID19

Clean Medical Data on COVID19

Feature Extraction

Machine Learning Model



- Machine Learning Training Model

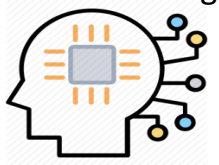
Example: App that learns to predict whether a person has Coronavirus

Medical Data with labels on who has COVID19

Clean Medical Data on COVID19

Feature Extraction

Machine Learning Model



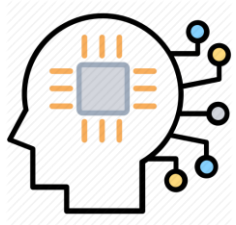
- Machine Learning Training Model

Medical Data of a Bob Jones

Data Cleaning

Feature Extraction

Trained Machine Learning Model



Bob Jones likely has COVID19

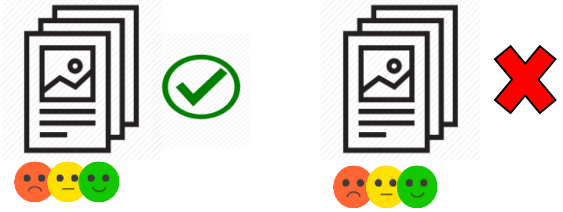
Example Machine Learning Pipeline:

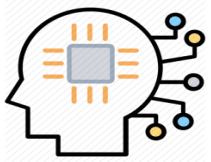
Learn to detect when someone posts negative content about Government and take it down

1) Social Media Posts

2) Data Cleaning: Removal of stopwords, label with crowd workers whether the Social Media Posts attack the government

3) Feature Extraction: Sentiment analysis over the content; ngrams





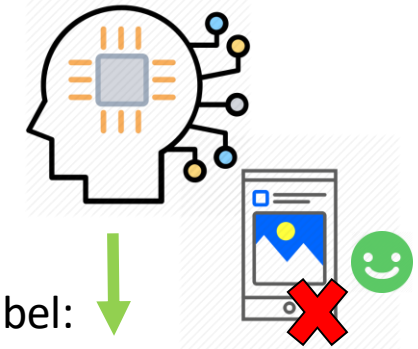
Machine Learning Model learns to detect when a post attacks the government

New Data:
New social media post

Data Cleaning:
removal of Stopwords

Feature Extraction:
sentiment analysis over the social media post; ngrams

Trained Machine Learning Model



Predicted Label:
Post is not friendly towards the government! Take it down.

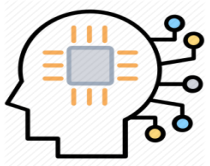
Example Machine Learning Pipeline:

Learn to predict the number of ventilators a hospital will need

1) Data from Hospitals

2) Data Cleaning: Removal of stopwords, ensure same dates in all entries, ensure similar naming patterns

3) Feature Extraction: Represent each hospital as a vector denoting number of doctors available, average patient age, number of patients with cronic illness, overview of the population size



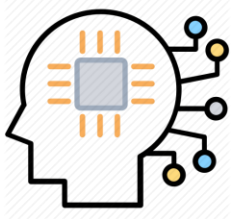
4) Machine Learning Model learns to predict the number of ventilators that are needed for a hospital with given characteristics.

New Data:
New hospital data

Data Cleaning:
Removal of stopwords, ensure same dates in all entries, ensure similar naming patterns

Feature Extraction:
Represent hospital data as a vector

Trained Machine Learning Model



Predict number of ventilators that will be needed for that hopsital.

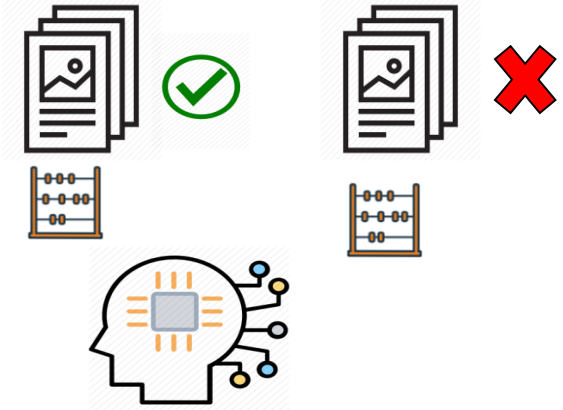
Example Machine Learning Pipeline:

Learn to detect when someone posts disinformation and take it down

1) Social Media Posts

2) Data Cleaning: Removal of stopwords, label with crowd workers whether the Post is sharing disinformation

3) Feature Extraction: represent each post as a vector holding the topics, public figures mentioned



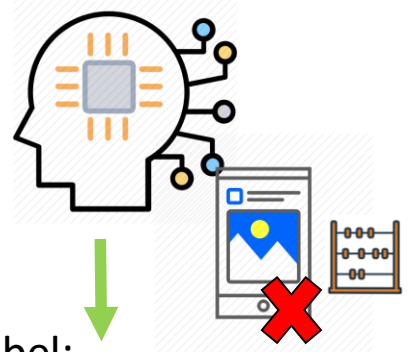
Machine Learning Model
learns to detect when a post is about disinformation

New Data:
New social media post

Data Cleaning:
removal of Stopwords

Feature Extraction:
represent each post as a vector holding the topics, public figures mentioned

Pre-Trained Machine Learning Model



Predicted Label:
Post holds disinformation! Take it down.

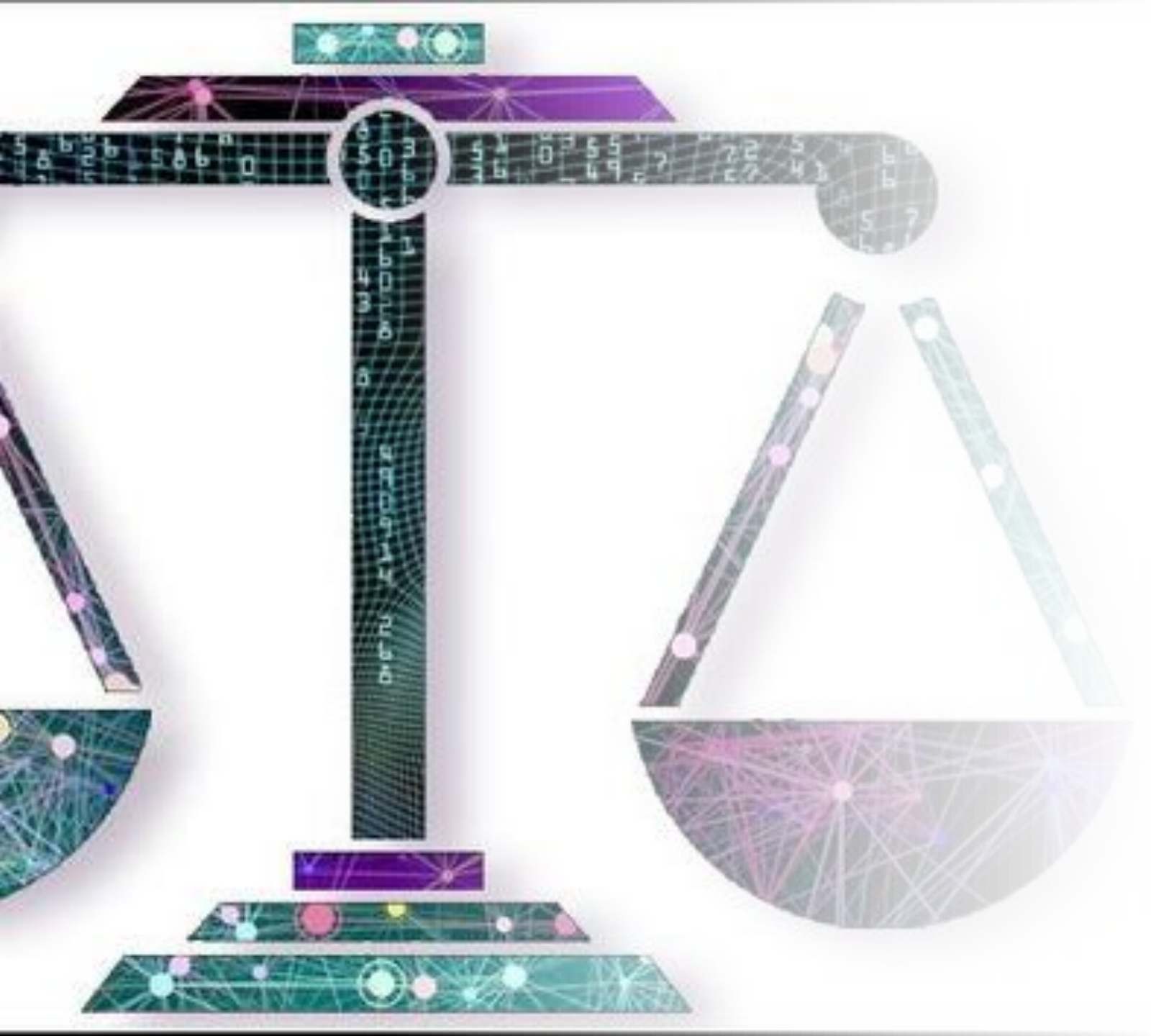
How do we enable machines that think ethically?

Ethics & A.I. Principales

- The literature has defined principles to help us navigate how we create ethical A.I. to be applied on every instance of the AI Life Cycle:
    - Human centerdeness
    - Fairness and Non discrimination
    - Transparency and Explainability
    - Inclusivity
    - Privacy
    - Necessity and proportionality
    - Multi-stakeholder and interdisciplinary governance
    - Safety and Security

Fairness

Avoid techno-chauvinism



Fairness in Ethics & Artificial Intelligence (A.I.)

- We need to create A.I. that either reduces unfairness in our society or at least keeps it the same way.
- Efforts must be conducted to avoid creating reinforcing or perpetuating human biases throughout the full lifecycle of the AI system.

Fairness in Ethics & Artificial Intelligence (A.I.)

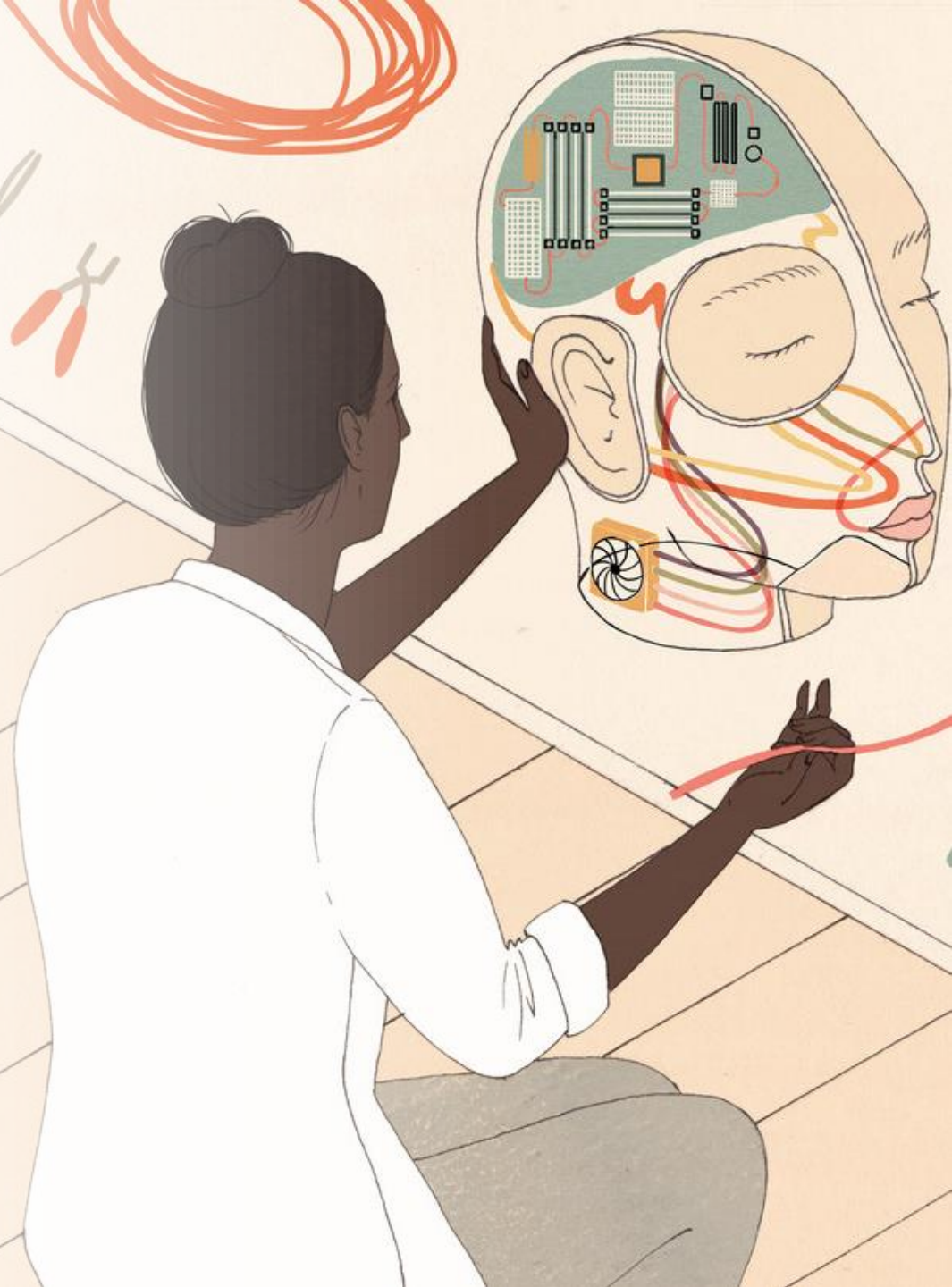
- It is important we think about **fairness** because A.I. is used in:
  - Criminal justice
  - Criminal investigation
  - Employment and hiring
  - Finance and credit

Fairness in Ethics & Artificial Intelligence (A.I.)

- AI can reinforce existing
  - Societal stereotypes
  - Cultural denigration
  - Over and under representation

  Example: Facial Recognition or using a biased data set for recruiting employees

Fairness in Ethics & Artificial Intelligence (A.I.)



- Fairness in A.I. relates to not just the technical component, but also the societal context in which the system is deployed.

- Fairness in A.I. is a sociotechnical challenge.
    - We need a greater diversity of people developing and deploying A.I. systems

    - The assumptions and decisions made by teams at each stage of the A.I. development and deployment lifecycle can introduce biases.
    - Avoid "Trash in, Trash out"

        - We can't delegate this to one or two people.
        - Everyone needs to be thinking about this actively.

Transparency

Transparency in A.I. & Ethics

Transparency in A.I. & Ethics

- Transparency in A.I. is important because it can help us:
    - Mitigate unfairness
    - Help developers debug their A.I. systems
    - Gain more trust from end-users who now can understand the platform better

Accountability in A.I.

- We are accountable for how our A.I. impacts the world.
- We need to help end-users also be accountable for their decisions with our A.I.
- Humans must always remain responsible for decisions aided or mediated by AI systems
- Ethical and legal responsibility for the research, design, development, deployment, funding, acquisition and use of AI systems, must be assigned to a physical person, notwithstanding the responsibility of legal entities.
- There must always exist human oversight along the life cycle of systems involving AI. Additionally, there must be processes in place to allow oversight from outside stakeholders, including evaluation, transparency and auditability mechanisms, as well as accountability.

Accountability in A.I. invovles setting principles that guide:

Transparency in A.I.: How and Why You Are Using A.I.

- Need to bring transparency to each part of the A.I. pipeline as the A.I. is affected by the pipeline choices we make :

  - **Data:**
    - Every dataset can have a datasheet explaining its:
      - Motivation
      - Creation
      - Maintenace
      - Intended use

    - Datasheets can help:
      - Data creators understand and uncover :
        - Potential biases in their data.
        - Unintentional assumptions we are making
      - Data consumers:
        - Determine if a dataset is right for their needs

Raw Data

1

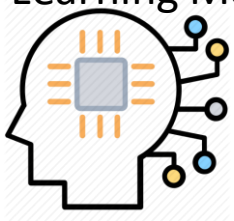
Data Cleaning

2

Feature Extraction

3

Machine Learning Model



4

Transparency in A.I.: How and Why You Are Using A.I.

- Need to bring transparency to each part of the A.I. pipeline as the A.I. is affected by the choices made:
  - For Feature Extraction and the Machine Learning Models we can create similar information sheets.
    - This could also extend to the related APIs we build.

Raw Data

1

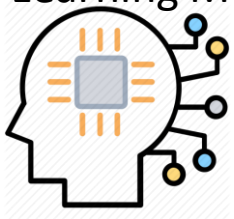
Data Cleaning

2

Feature Extraction

3

Machine Learning Model



4

Inclusivity

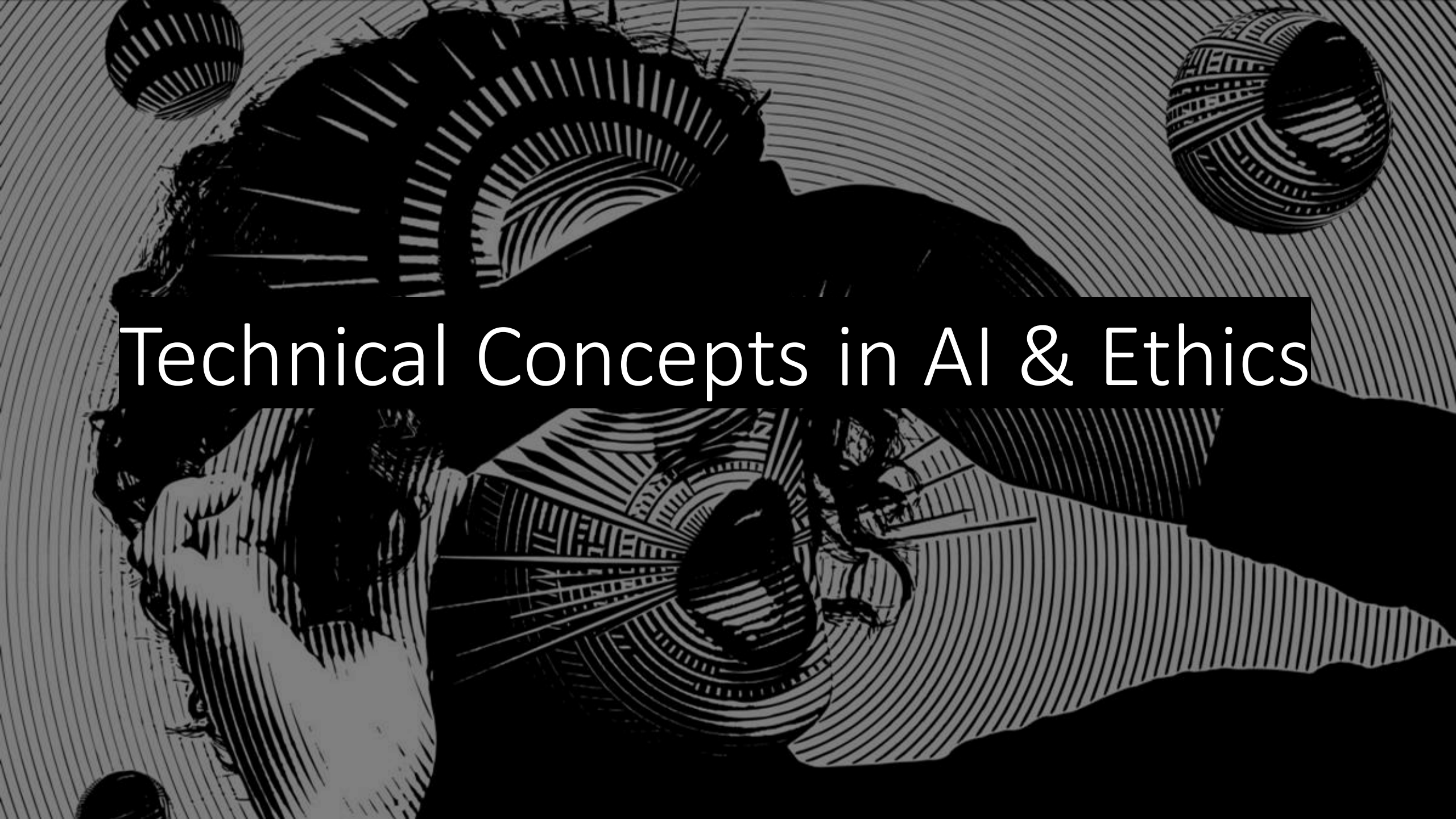
in A.I.

Inclusiveness in A.I.:

- Create A.I. that can empower and engage communities around the world.
- Equality of opportunity, of access and of benefits for all groups and communities, including vulnerable groups.

- Ensure that nobody is left out in the A.I. we design.

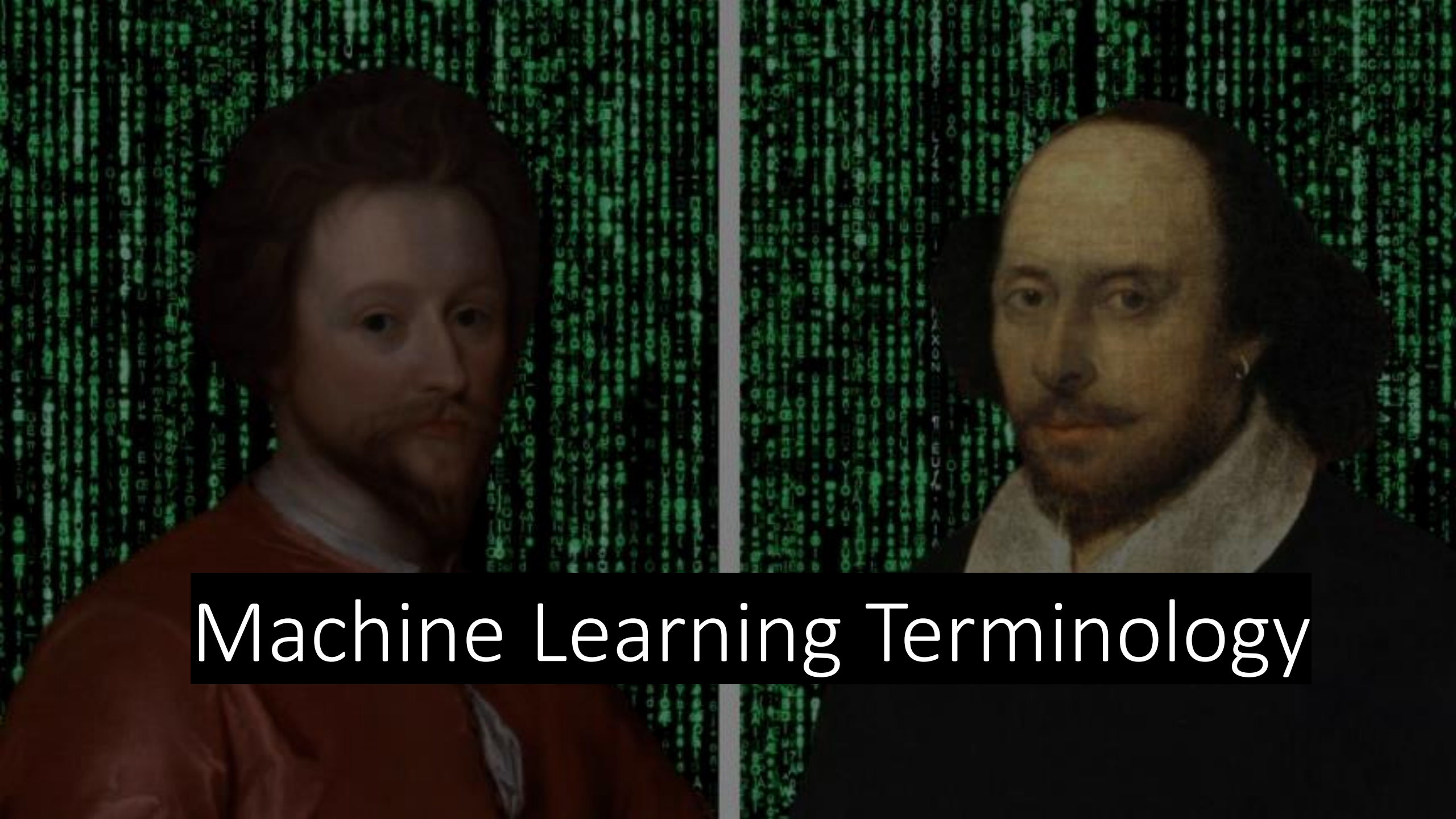
- Making sure we are intentionally inclusive and intentionally diverse in how we design A.I.
- Make sure that the full spectrum of communities are covered.
- Design side-by-side individuals to ensure we don't take ableist perspective



Technical Concepts in AI & Ethics

Why dig deep in Technical Concepts?

- It helps you to understand the technical terms that are used to raise your voice and be able to participate in conversations around A.I. and Ethics.



Machine Learning Terminology

Terminology: Type of Machine Learning Model

- Classification
  - Is this a dog or a cat?
- Regression
  - How warm will it be tomorrow?
- Clustering
  - Here are a few articles. Organize them into topics, and predict whether a given article is of a certain topic
- Representation
  - What is the best way to represent words? (e.g. so representation of “autumn” and “fall” are similar)
  - What is the color of happiness?
- “Dear algorithm: figure it out, and if you’re right, you get a reward”
  - Drive a car
  - Recommend movies to people

Terminology: Type of Machine Learning Model

- Classification
  - Is this a dog or a cat?
- Regression
  - How warm will it be tomorrow?

Supervised Learning

- Clustering
  - Here are a few articles. Organize them into topics, and then predict the topic of a new article.
- Representation
  - What is the best way to represent words? (e.g. so representation of “autumn” and “fall” are similar)
  - What is the color of happiness?

Unsupervised Learning

- “Dear algorithm: figure it out, and if you’re right, you get a reward”
  - Drive a car
  - Recommend movies to people

Reinforcement Learning

Terminology: Evaluation of Machine Learning

- What metrics do we use to discuss performance?
- How do we set up a meaningful evaluation?

Terminology: Evaluation of Machine Learning

- We're generally interested in the following:
  - How often is the prediction wrong?
  - How is the prediction wrong?
  - What is the cost of wrong predictions?
  - How does the cost vary by the type of prediction that was wrong?

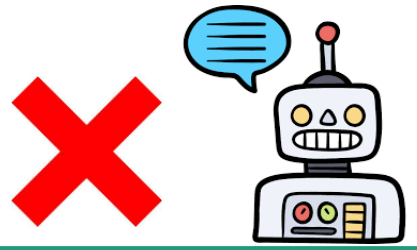
Terminology: Specific Evaluation Metrics

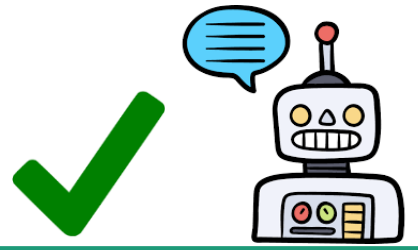
- Accuracy. Correct predictions/total predictions

Terminology: Metrics

- What can happen when your model classifies something wrong?

Imagine a machine learning model that detects when an elder has taken her medication corectly.





	Machine Identifies that wrong mediaction has been taken	Machine Identifies that correct medication has been taken
Real world: Incorrect medication taken	**true positive** (They take wrong medication and the machine said it was wrong)	**false negative** (They took wrong medication, but the machine says they took correct one.)
Real world: Correct medication taken	**false positive** (They took correct medication, but machine said it was wrong)	**true negative** (They took correct medicine, and the machine says it is correct)

Terminology: Metrics

- False Positive, False Positive Rate
- False Negative, False Negative Rate

- Sensitivity == True Positive Rate
  - ability of the test to identify those elders who have NOT taken their medication

- Specificity == True Negative Rate
  - ability of the test to identify those elders who have taken their medication

Terminology: Types of Errors

- Not all errors are created equal.
  - Depending on your task, different errors have different costs.

  - Cost of “false negative” in pregnancy detection?
  - Cost of “false positive” in pregnancy detection?

  - In law enforcement?

  - In detecting the “Alexa” command?
  - In detecting person in the road?

FALSE POSITIVES: SELF-DRIVING CARS AND THE AGONY OF KNOWING WHAT MATTERS

According to a preliminary report released by the National Transportation Safety Board last week, Uber's system detected pedestrian Elaine Herzberg six seconds before striking and killing her. It identified her as an unknown object, then a vehicle, then finally a bicycle. (She was pushing a bike, so close enough.) About a second before the crash, the system determined it needed to slam on the brakes. But Uber hadn't set up its system to act on that decision, the NTSB explained in the report. The engineers prevented their car from making that call on its own "to reduce the potential for erratic vehicle behavior." (The company relied on the car's human operator to avoid crashes, which is a whole separate problem.)

Uber's engineers decided not to let the car auto-brake because they were worried the system would overreact to things that were unimportant or not there at all. They were, in other words, very worried about false positives.

RN MORE

Terminology: Overfitting

- Many machine learning model learns every detail of its own training data, which makes it impossible for the model to be able to make accurate decisions (predict) outside the training data.
  - The model can't correctly predict anything outside the training data.

Community Exercise: Types of Errors

- What type of error is this?

Community Exercise: Types of Errors

Johnson

In the world of voice-recognition, not all accents are equal

But you can train your gadgets to understand what you're saying

- What type of error is this?

Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots

By Jacob Snow, Technology & Civil Liberties Attorney, ACLU of Northern California
JULY 26, 2018 | 8:00 AM

TAGS: Face Recognition Technology, Surveillance Technologies, Privacy & Technology

Amazon's face surveillance technology is the target of growing opposition nationwide, and today, there are 28 more causes for concern. In a test the ACLU recently conducted of the facial recognition tool, called "Rekognition," the software incorrectly matched 28 members of Congress, identifying them as other people who have been arrested for a crime.

The members of Congress who were falsely matched with the mugshot database we used in the test include Republicans and Democrats, men and

The false matches were disproportionately of people of color, including six members of the Congressional Black Caucus, among them civil rights legend Rep. John Lewis (D-Ga.). These results demonstrate why Congress should join the ACLU in calling for a moratorium on law enforcement use of face surveillance.

Types of Errors

What type of error is this?

The Unquestioned Textbook Assumption

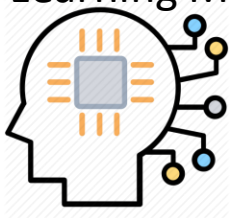
Training Data

This not yet fixed!

Data Cleaning

Feature Extraction

Machine Learning Model

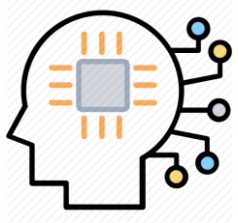


New Data

Data Cleaning

Feature Extraction

Machine Learning Model



How was this selected and why?

Predicted Label

The Unquestioned Textbook Assumption

- everything can and should be iterated on, including the problem itself … what are you trying to solve?

Summary

- Building Human-AI Systems is Iterative
- Terminology and Way of Thinking About Performance/Errors
  - these apply regardless of what you stick in the black box
- HCI for AI is fundamentally about dealing w/ uncertainty and errors
- Deploying Intelligent Systems globally bring new type of challenges.

Practical experience around A.I. and Ethics

The Malicious Use of AI + Deep Fakes Example

Properties of AI

- AI is a dual-use area of technology: it can be used toward beneficial and harmful ends. AI is dual-use in the same sense that human intelligence is.

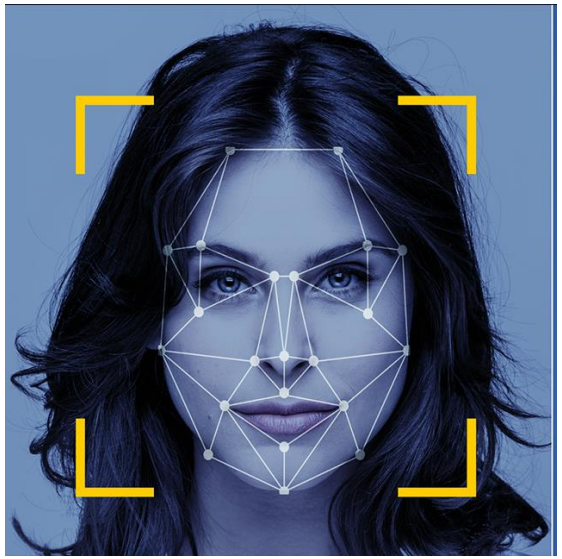
- AI systems can exceed human capabilities: they can perform certain tasks better than any human could (i.e. play. chess)

Properties of AI

- AI systems are commonly both efficient and scalable:

"efficient": once trained and deployed, it can complete a certain task more quickly or cheaply than a human could.

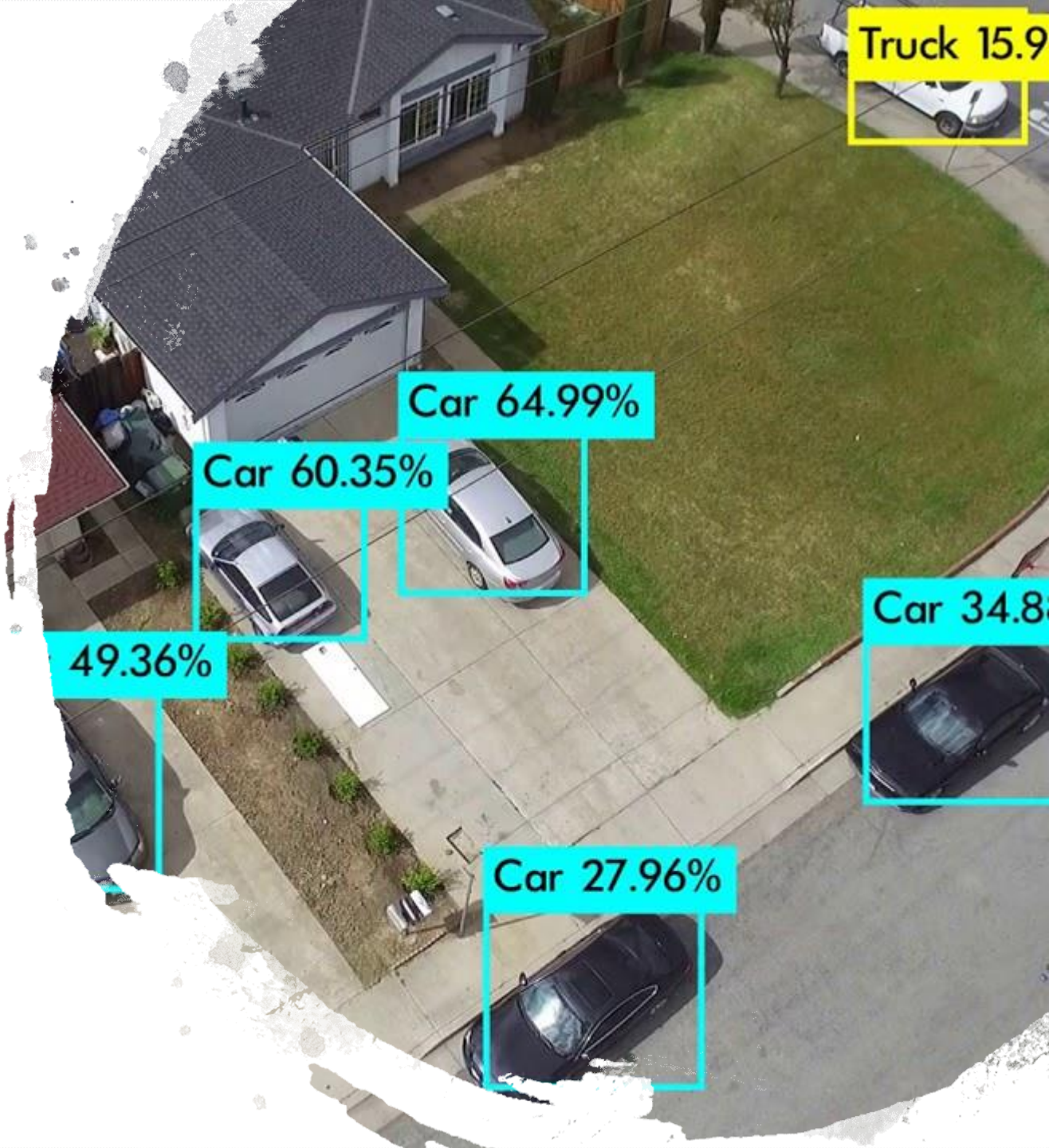
"scalable" increasing the computing power it complete many more instances of the task.



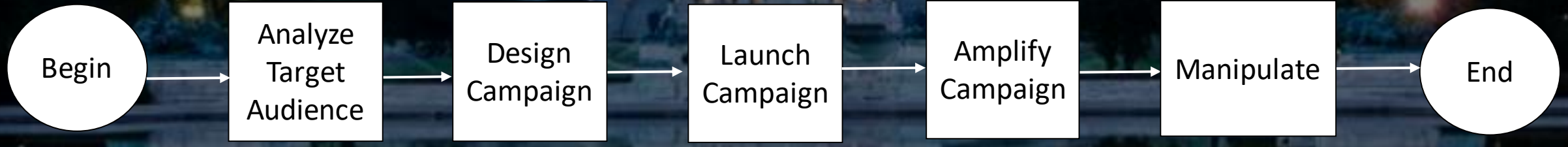
- A facial recognition system is both efficient and scalable; once it is developed and trained, it can be applied to many different cameras.

Properties of AI

- AI systems can increase anonymity and psychological distance



AI + Disinformation

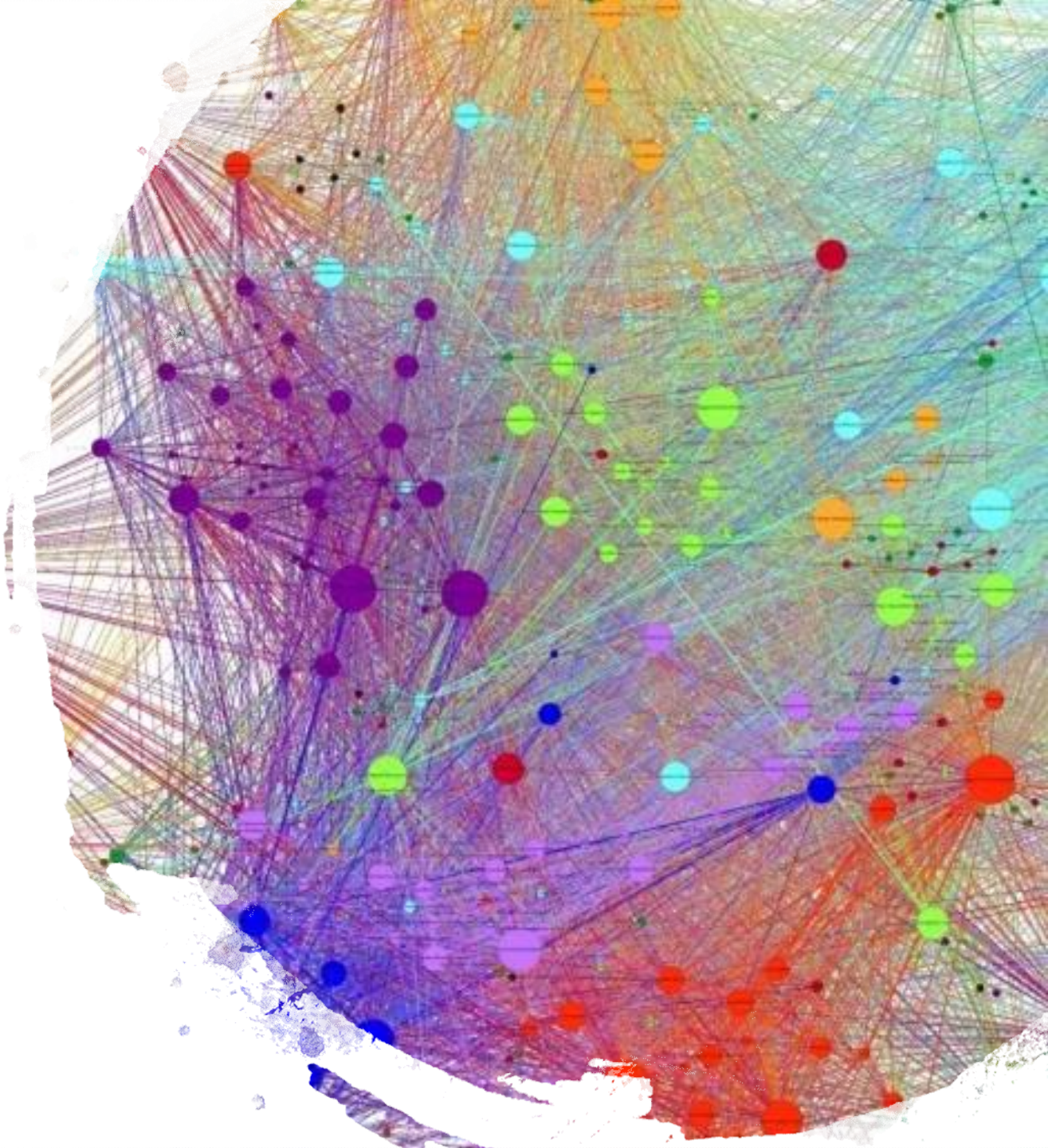


The Malicious Use of AI

- Disinformation.

<u>Analyze Target Audience.</u>

To identify key influencers, who can then be approached with (malicious) offers or targeted with disinformation.

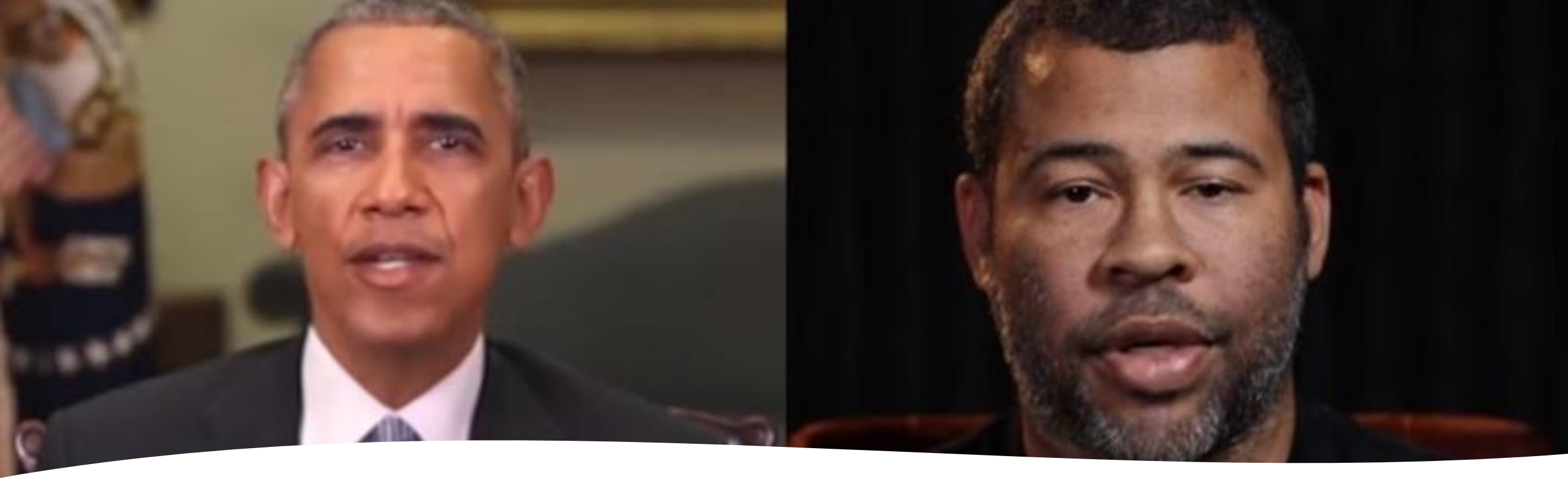


The Malicious Use of AI

- Disinformation.

<u>Design, hyper-personalised disinformation campaigns.</u>

Create personalised messages in order to affect their voting behavior.



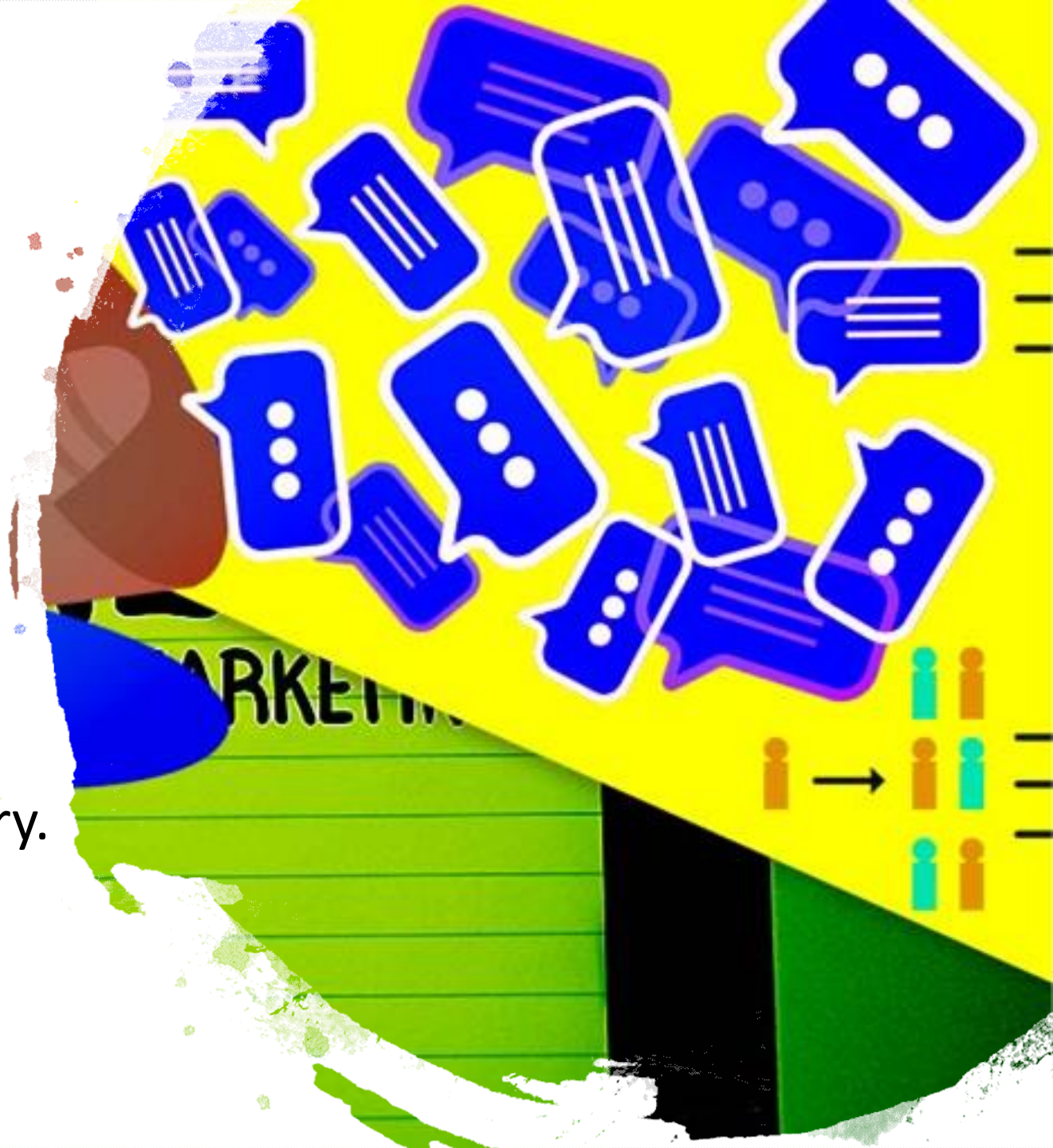
- Fake news reports with realistic fabricated video and audio.
- Highly realistic videos are made of state leaders seeming to make inflammatory comments they never actually made.

The Malicious Use of AI

- Disinformation.

<u>Launch Campaign.</u>

Deliver content using different accounts to create the illusion that there are multiple sources of a story.

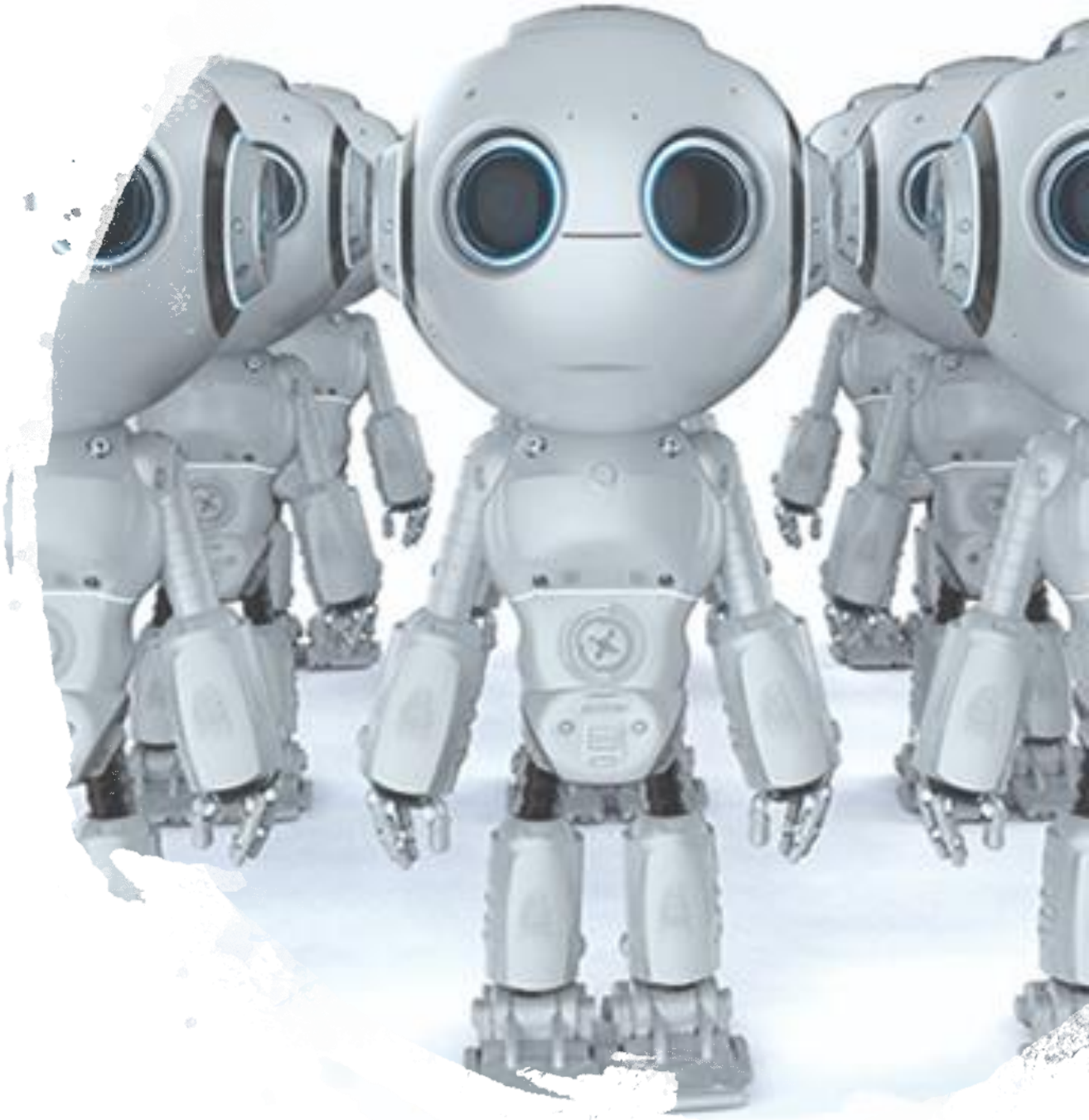


The Malicious Use of AI

- Disinformation.

<u>Amplify.</u>

Amplify content by pushing the story using fake accounts, massive likes, shares, etc.



The Malicious Use of AI

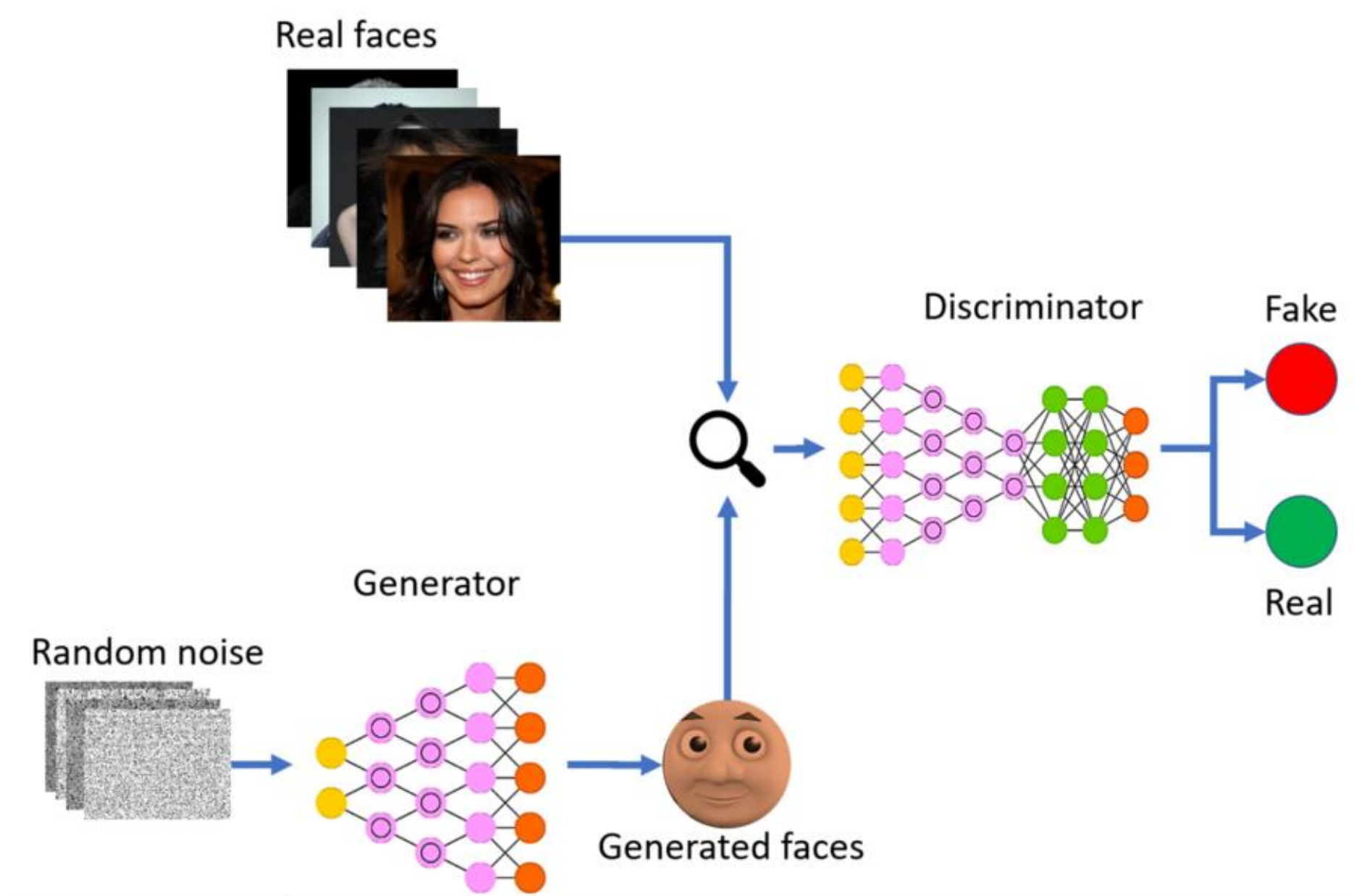
- Disinformation.

<u>Manipulate.</u>

Manipulate the target's reaction by infiltrating conversations about the content. Incite conflict and or streghthen the ilusion of consensus by trolling comment sections of online posts.

JULIO ASTILLERO
PERIODISMO CON CREDIBILIDAD

Deep Fake Demo


Real faces
Discriminator
Fake
Real
Generator
Random noise
Generated faces

https://github.com/saviaga/riiaa_mx_workshop

What do we need?

Driving Video

Image

DeepFake
https://youtu.be/6jWSjaaX404

Link Google Colab: https://bit.ly/3gnUHph

Thank you!

Saiph Savage

Twitter: @saiphcita

www.saiph.org

Claudia Flores-Saviaga

Twitter: @saviaga

www.saviaga.com

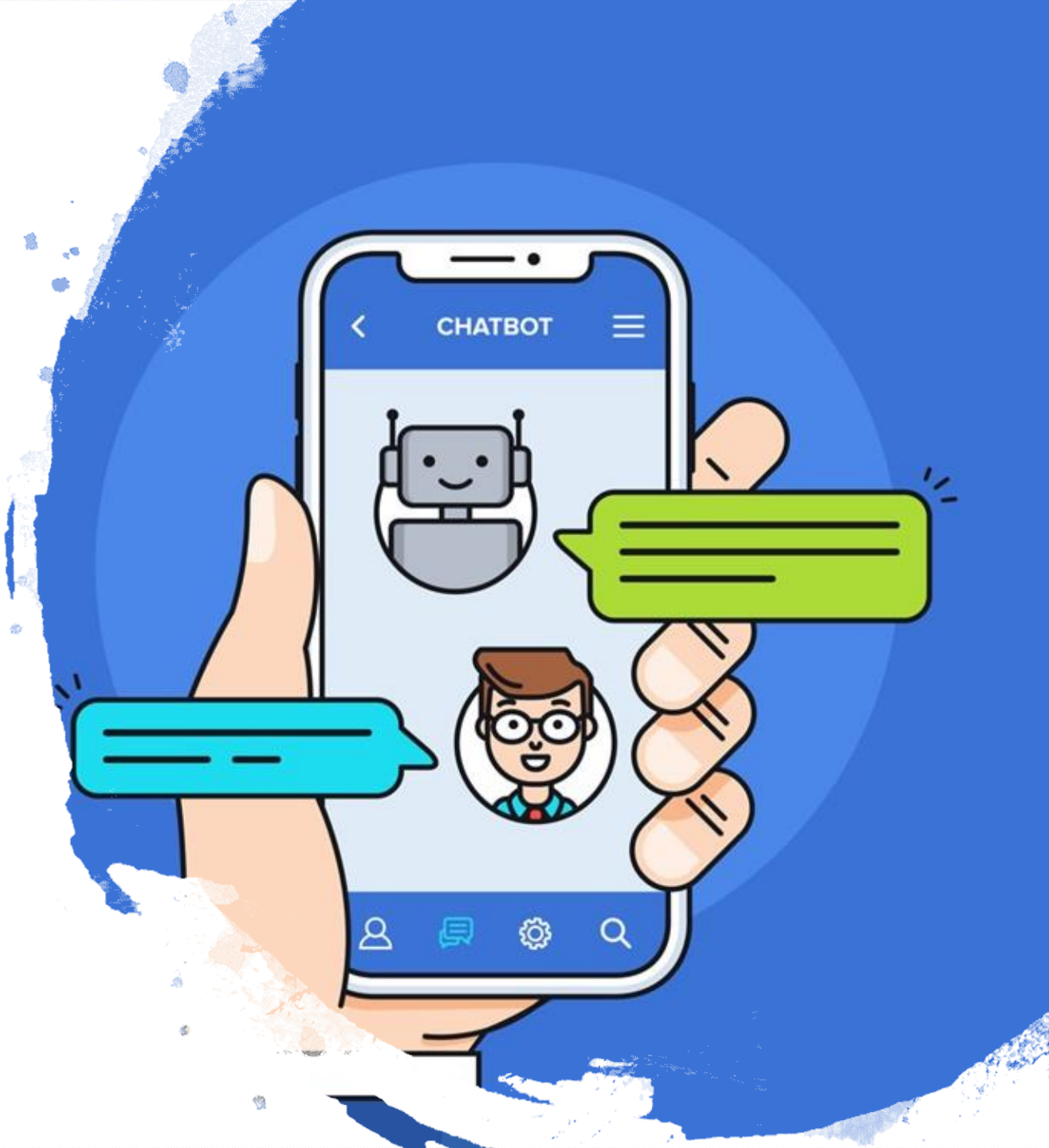
Backup Slides

The Malicious Use of AI

- Digital Security:

<u>Automation of social engineering attacks.</u>

Convincing chatbots may elicit human trust by engaging people in longer dialogues, and perhaps eventually masquerade visually as another person in a video chat.



The Malicious Use of AI

- Digital Security:

<u>Prioritising targets for cyber attacks using machine learning.</u>

Large datasets are used to identify victims more efficiently, e.g. by estimating personal wealth and willingness to pay based on online behavior.