DS 517– Lecture 1

Ethics and Bias in AI

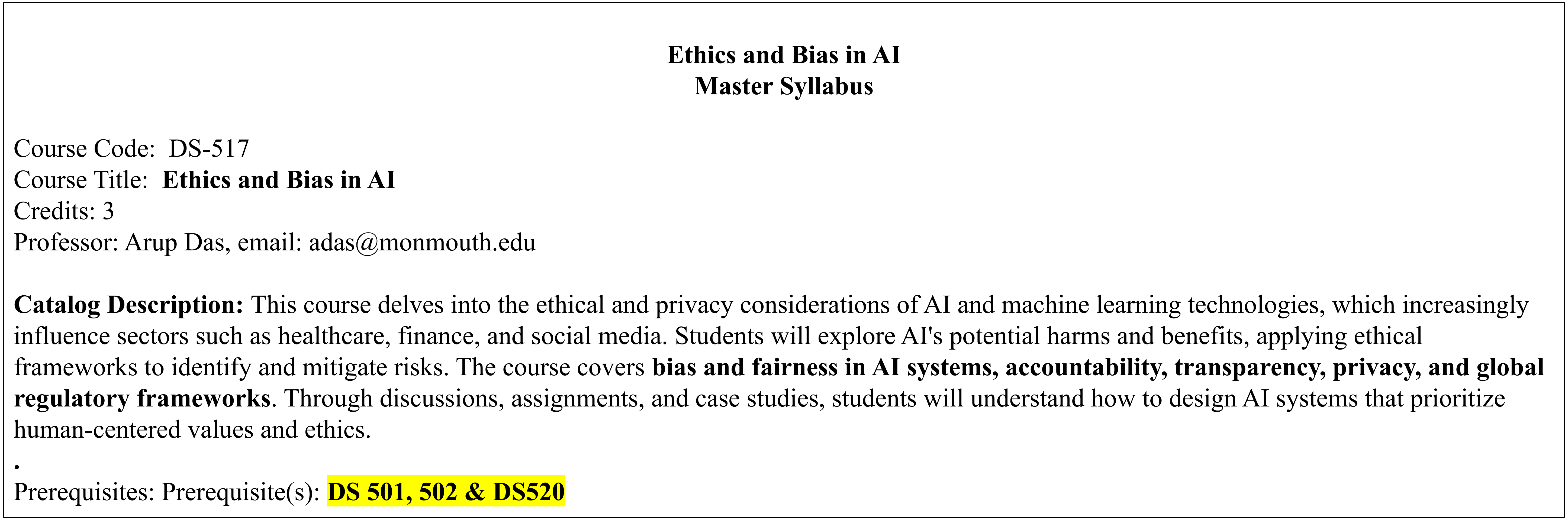
DS-517-50: Ethics and Bias in AI
2024 Fall
MONMOUTH CAMPUS
M 7:30 PM - 10:20 PM
9/3/2024 - 12/9/2024
Howard Hall, 309 LECTURE

Arup Das

adas@Monmouth.edu

Disclaimer:

- The views expressed are solely those of the presenter and not affiliated with any other party.
- This presentation is free of copyright violations, and external sources have been appropriately credited.
- T**he content within this presentation is legally protected; unauthorized reproduction, including photography, will result in legal action.**
- **This material is not intended for distribution and must remain solely within the confines of this class. Do not distribute slide or assignments to other students**
- Using cameras to take screenshots or photographs of the slides is strictly prohibited.



Ethics and Bias in AI
Master Syllabus

Course Code: DS-517
Course Title: **Ethics and Bias in AI**
Credits: 3
Professor: Arup Das, email: adas@monmouth.edu

Catalog Description: This course delves into the ethical and privacy considerations of AI and machine learning technologies, which increasingly influence sectors such as healthcare, finance, and social media. Students will explore AI's potential harms and benefits, applying ethical frameworks to identify and mitigate risks. The course covers **bias and fairness in AI systems, accountability, transparency, privacy, and global regulatory frameworks**. Through discussions, assignments, and case studies, students will understand how to design AI systems that prioritize human-centered values and ethics.

.

Prerequisites: Prerequisite(s): **DS 501, 502 & DS520**

DS 517- Ethics and Bias in AI

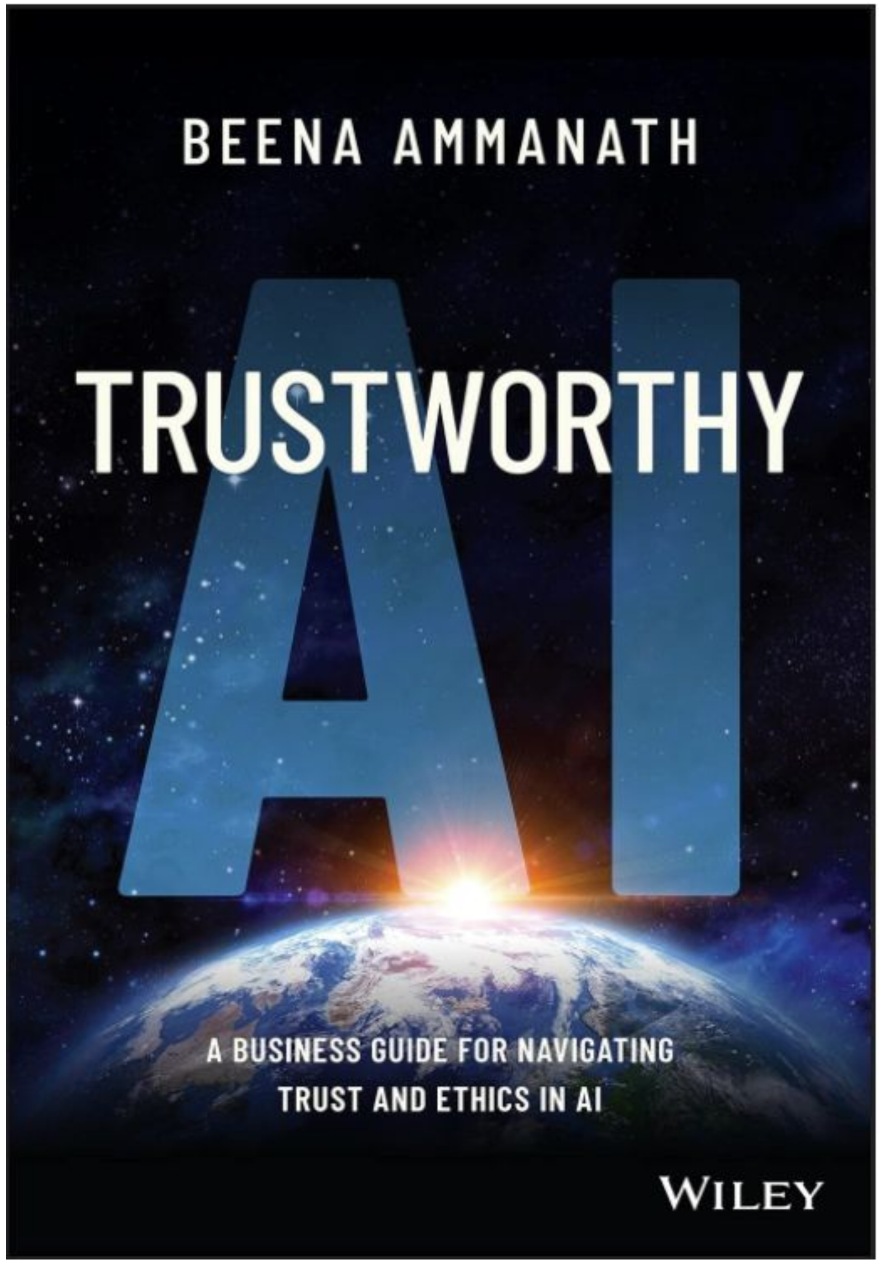
Course Objectives: By the end of this course, students will have a solid understanding of AI ethics and the ability to design human-centered AI systems that prioritize accountability, transparency, and fairness. They will also learn to detect and mitigate bias, navigate global regulatory frameworks, and implement explainable AI techniques for interpreting complex models.

Assessable Learning Outcomes:

- Apply ethical principles to the design and development of AI systems.
- Critically assess AI systems for accountability and transparency.
- Evaluate AI systems for privacy, security, and inclusion compliance.
- Detect and mitigate bias in AI/ML systems.
- Navigate and apply global regulatory frameworks related to AI.
- Implement and evaluate methods for AI model interpretability and explainability.
- Articulate ethical considerations in AI through essays and case studies.

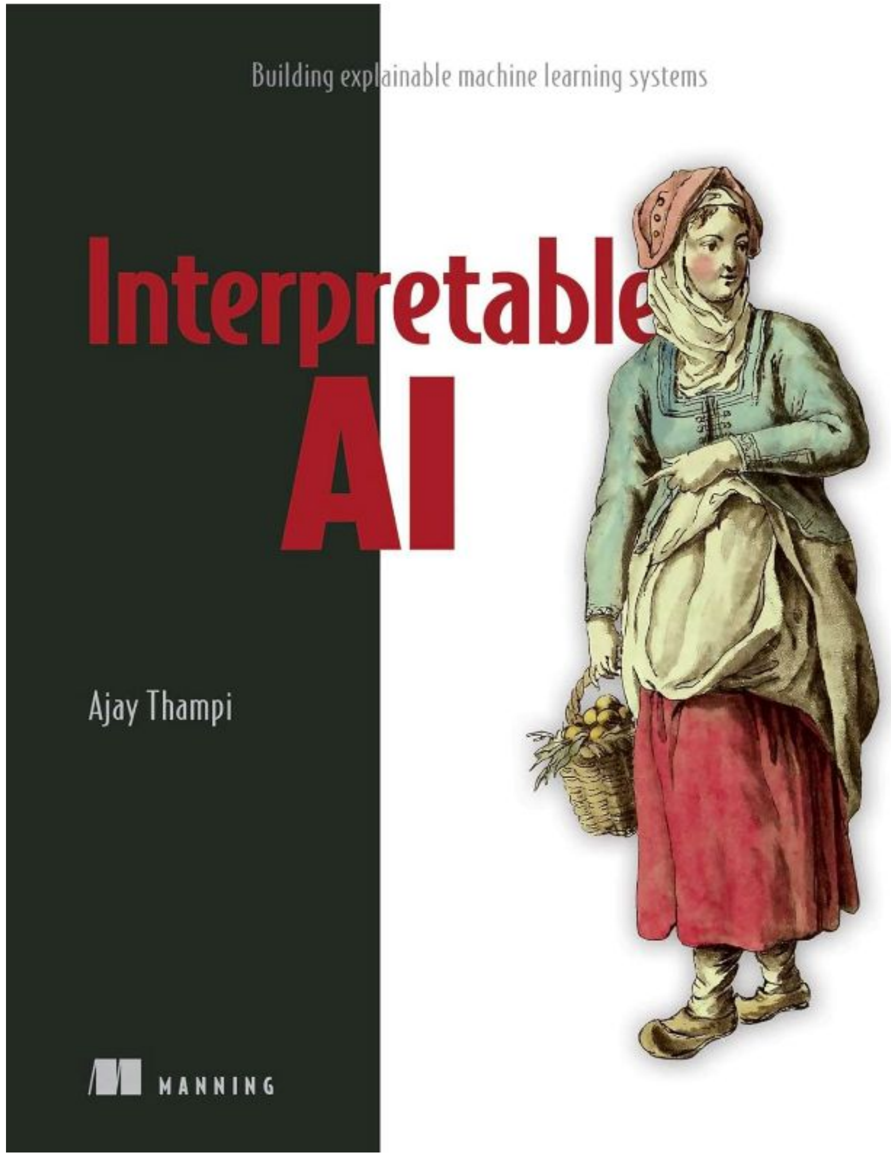
Book1 –

Trustworthy AI: A Business Guide for Navigating Trust and Ethics in AI- ISBN#978-1119867920



Book 2 –

Interpretable AI: Building explainable machine learning systems– ISBN#-978-161729 7649



Week 1: AI Ethics & Human-Centered Design

- Introduction to AI Ethics
- Designing AI systems to serve human needs

Week 2: Algorithms and Accountability

- What is accountability in AI?
- Ethical implications of AI accountability

Week 3: Transparency in AI

- Importance of transparency in AI systems
- Risks associated with transparency

Week 4: Privacy, Security, and Inclusion

- Human rights and AI
- The intersection of privacy, security, and inclusion in AI

Week 5: AI Fairness & Bias

- Analyzing AI bias in society
- Methods to mitigate bias in AI/ML systems

Week 6: AI Fairness & Bias

- Case studies and hands-on exercises on bias detection and mitigation

Week 7: AI Regulatory Frameworks (US and Europe)

- Overview of proposed AI regulatory frameworks
- Comparing US and European AI regulations

Week 8: AI Regulatory Frameworks (Continued)

- Case studies and in-depth analysis of global AI regulations

Week 9: Explainable AI - Introduction to Model Interpretability

- Fundamental concepts of AI model interpretability
- Taxonomy of interpretability methods

Week 10: Explainable AI - Advanced Topics

- Evaluation and properties of explanations
- Challenges of interpreting black-box models

Week 11: Case Studies in Explainable AI

- Real-world examples of AI model interpretability
- Hands-on exercises with interpretability tools

Week 12: AI Ethics in Practice

- Applying ethical principles in real-world AI projects
- Group discussions and reflections

Week 13: AI Ethics Case Essay Preparation

- Guidance on writing the ethics case essay

Week 14: Final Exam and Course Wrap-Up

- Final exam covering all course topics
- Review and discussion of key concepts

Fall 2024 Academic Calendar

September
Tuesday, Sept. 3 – Classes Begin
Tuesday, Sept. 3 – Tuesday, Sept. 10 – Late Registration / Drop – Add / Leave of Absence
Friday, Sept. 27 – “W” Deadline for “A” Session

October
Saturday, Oct. 12 – Tuesday, Oct. 15 – **Fall Holiday (non-weekend students)**
Tuesday, Oct. 22 – Undergraduate Midterm Grades Due
Tuesday, Oct. 22 – Session “A” Classes End
Wednesday, Oct. 23 – Session “B” Classes Begin
Wednesday, Oct. 30 – Pattern B add/drop
Thursday, Oct. 31 – “W” Deadline

November
Monday, Nov. 18 – “W” Deadline for “B” Session Classes
Wednesday, Nov. 27 – Sunday, Dec. 1 – **Thanksgiving Holiday**

December
Monday, Dec. 9 – Thirteenth Week Ends
Tuesday, Dec. 10 – Reading Day
Wednesday, Dec. 11 – Tuesday, Dec. 17 – Fourteenth Week Adjusted Schedule
Friday, Dec. 20 – End of Final Grading Period

Date	Week	Class Format/Location/Time	Topics	Readings Required (Due before class)	Assignment/Quiz
September 9, 2024	Week_1	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	AI Ethics & Human-Centered Design		
September 16,2024	Week_2	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	Algorithms and Accountability	**Book 1 – Chapter 1, 3, 9**	**Assignment 1- Presentation on the professor assigned reading Due Sep 23,2024**
September 23,2024	Week_3	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	Transparency in AI	Book 1 – Chapter 4	
September 30, 2024	Week_4	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	Privacy, Security, and Inclusion	Book 1- Chapter 6, 7,8	**Assignment 2- Presentation on the professor-assigned reading Due Oct 7, 2024**
October 7, 2024	Week_5	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	AI Fairness & Bias	Book 1- Chapter 2 Book 2 – Chapter 8	
October 14, 2024	Week_6	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	AI Fairness & Bias	Book 2 – Chapter 8	**Assignment 3- Presentation on the professor-assigned reading Due Oct 21, 2024**
October 21, 2024	Week_7	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	AI Regulatory Frameworks (US and Europe)	Professor Handout	
October 28,2024	Week_8	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	AI Regulatory Frameworks (Continued)	Professor Handout	**Assignment 4 - Presentation on the professor-assigned reading Due Nov4,2024**
November 4, 2024	Week_9	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	Explainable AI - Introduction to Model Interpretability	Book 2 – Chapter 1 ,2	
November 11, 2024	Week_10	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	Explainable AI - Advanced Topics	Book 2 – Chapter 3, 4 ,5	**Coding Assignment – Due Nov 22, 2024**
November 18,2024	Week_11	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	Case Studies in Explainable AI	Book2 – Chapter 6,7	
November 25, 2024	Week_12	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	AI Ethics in Practice	Book 2 – Chapter 9	
December 2, 2024 (Last class)	Week_13 &14	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20 PM	AI Ethics Case Essay Preparation/Final Exam and Course Wrap-Up		**Final Essay, Final Exam Due Dec 8, 2024, before midnight EST**

Course Logistics

1. OneDrive link for professor notes and assignments/quiz -
2. Check your Monmouth email for announcements
3. Check your Monmouth calendar for Zoom links for office hours and remote lectures
4. My contact information: adas@monmouth.edu, Cell # 917-523-7683
4. Office hours (zoom only) – Friday (EST)
5. Assignment submission to professoraruprdas@gmail.com (Notation for files: Assignment_1_Name_of_Student), Colab notebooks ipynb file and html file, all presentation in ppt format.
6. Quiz submission to professoraruprdas@gmail.com (Notation for file : Quiz_1_Name_of_Student. doc , Quiz_2_Name_of_Student.doc)

Methods of Evaluation

- **Assignments – 20% (4 assignments - each 5%)**
- **Final Exam – 30%**
- **Essay – 30%**
- **Coding Assignment – 20%**

Letter Grade	Percentage Points
A	100-93
A-	92-90
B+	89-87
B	86-83
B-	82-80
C+	79-77
C	76-73
C-	72-70
F	69-0

Academic Honesty

Everything you turn in for grading must be your work. Academic dishonesty subverts the University's mission and undermines the student's intellectual growth. Therefore, we will not tolerate violations of the code of academic honesty. Penalties for such violations include suspension or dismissal and are elaborated upon in the Student Handbook.

A guide to plagiarism can be found at http://www.plagiarism.org/.

Topics

1. AI Ethics News & Industry Certifications
2. Impact of AI on society and why it has to be regulated
3. What is Ethics, and how is AI ethics a sub-field
4. AI Ethics
5. AI Ethics framework
6. Human Centered Design

Course Outline

How to start thinking about AI from an ethical point of view

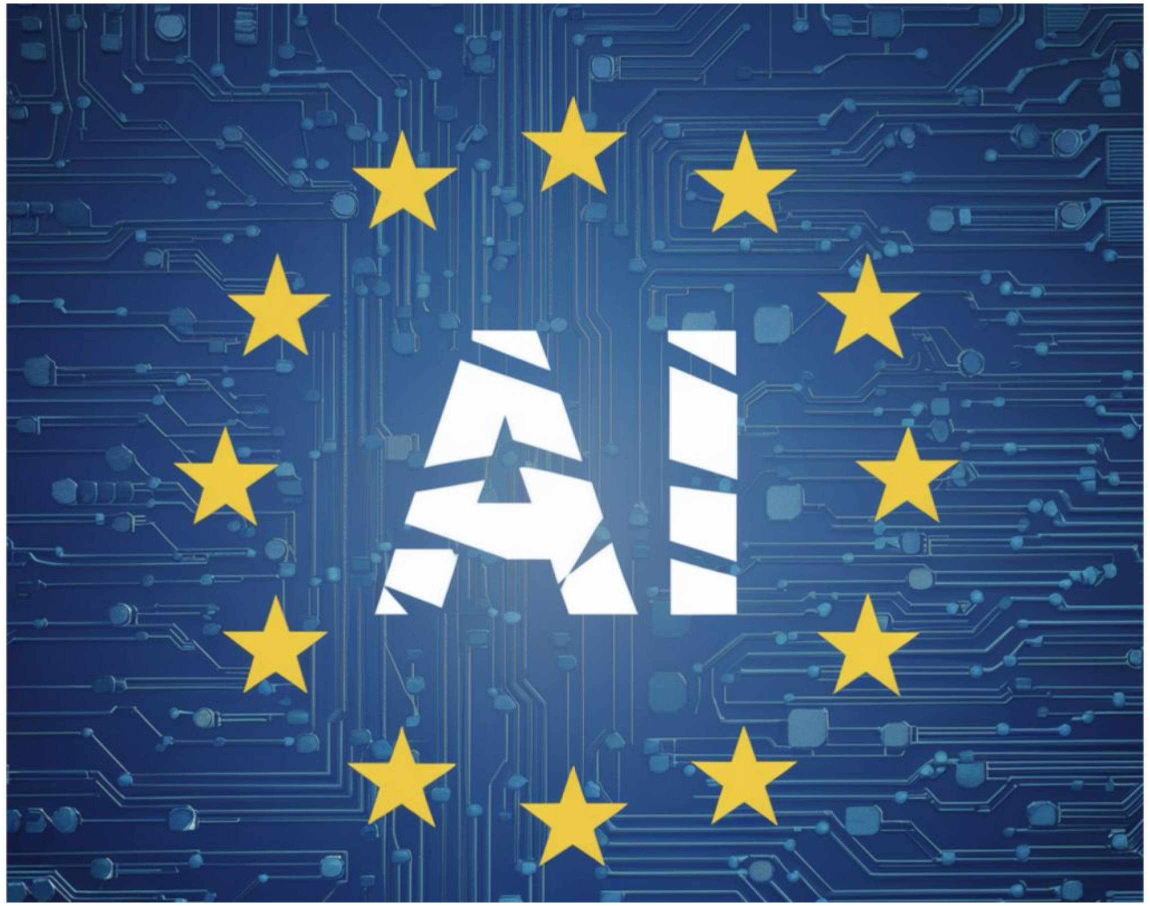
AI increasingly has an impact on everything from social media to healthcare. AI makes loan decisions, detect diseases, and recruits' personnel. These technologies can potentially harm or help the people they serve.

By applying an ethical lens, we can work toward identifying the harms that these technologies can cause. Additionally, we can design and build them to reduce these harms—or decide not to build them.

This course examines the full range of issues around BIAS and Ethics implications in AI/ML-enabled systems. Additionally, it will cover a wide range of proposed global AI regulatory frameworks.

AI Ethics News & Industry Certifications

AI Ethics in the News



The law was passed in the European Parliament on 13 March 2024, with 523 votes in favor, 46 against, and 49 abstaining. The EU Council approved it on 21 May 2024.

The European Union's (EU) Artificial Intelligence (AI) Act imposes fines for non-compliance with the law:

- Prohibited AI systems
- Fines can be up to €35 million or 7% of a company's worldwide annual turnover, whichever is higher.

- General-purpose AI models
- Fines can be up to €15 million or 3% of a company's worldwide annual turnover, whichever is higher.

- Providing false information
- Fines can be up to €7.5 million or 1.5% of a company's worldwide annual turnover

AI Ethics in the News

California legislature passes sweeping AI safety bill

Illustration by Cath Virginia / The Verge | Photos from Getty Images

SB 1047 has passed the California Senate.
The Senate was widely expected to pass the bill, which has now officially cleared every hurdle except a final signature from Governor Gavin Newsom. Newsom has until the end of September to make his call.

The bill, a flashpoint for debate in Silicon Valley and beyond, would obligate AI companies operating in California to implement several precautions before they train a sophisticated foundation model. Those include making it possible to quickly and fully shut the model down, ensuring it is protected against “unsafe post-training modifications,” and maintaining a testing procedure to evaluate whether a model or its derivatives is especially at risk of “causing or enabling a critical harm.”

AI Ethics News

A previous lawsuit filed against UnitedHealthcare by TeamHealth last year alleged the insurer used an algorithm to routinely deny claims based on diagnostic codes. The Humana lawsuit targets the AI algorithm, nH Predict, designed to predict how long patients will stay in skilled nursing facilities.

The Importance of AI Ethicists & How to Become One

Key Responsibilities

AI ethicists carry out a myriad of responsibilities that play a vital role in guiding an organization’s ethical compass:

- **Developing ethical guidelines and policies for AI projects:** AI Ethicists craft comprehensive guidelines and policies that direct responsible AI development.

- **Establishing standards for responsible AI development:** They set forth standards to ensure AI development adheres to ethical principles and societal norms.

- **Conducting ethics reviews of AI projects:** Before an AI project goes live, AI ethicists conduct a review to identify potential ethical issues and suggest remedies.

- **Assessing potential ethical risks and ensuring compliance:** They actively evaluate potential risks associated with AI projects and ensure these projects comply with existing ethical guidelines and regulations.

- **Collaborating with cross-functional teams:** AI ethicists work closely with developers, product managers, legal experts, and other team members to integrate ethical considerations into every stage of AI development.

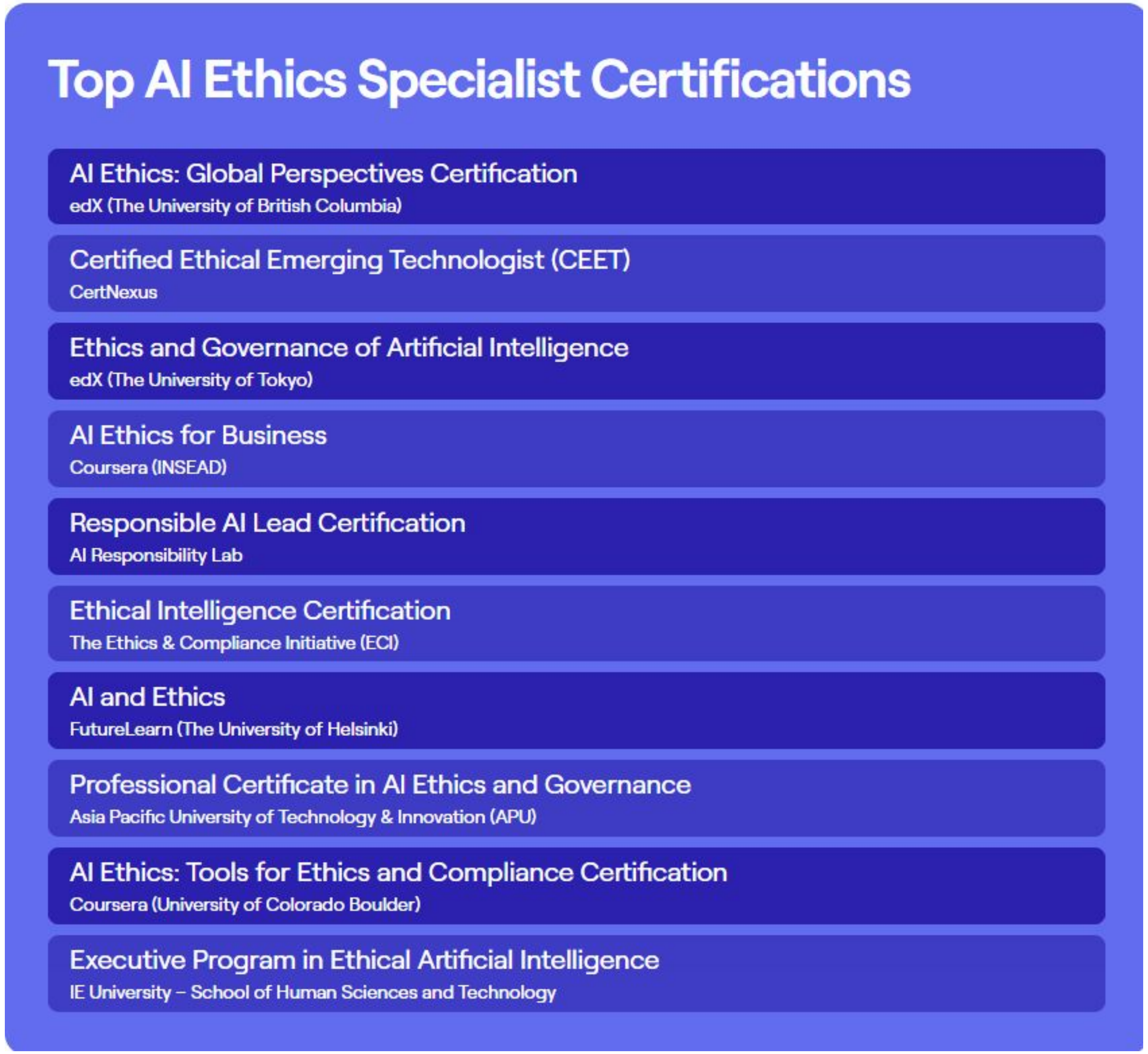
- **Educating and raising awareness of ethical AI practices:** They educate team members about ethical AI practices, fostering a culture of responsibility and ethical consciousness within the organization.

https://onlinedegrees.sandiego.edu/ai-ethicist-career/

New Careers in Responsible AI This Week!

https://alltechishuman.org/all-tech-is-human-blog/new-careers-in-responsible-ai-this-week-feb-16

AI Ethics Industry Certifications



Top AI Ethics Specialist Certifications

AI Ethics: Global Perspectives Certification
edX (The University of British Columbia)

Certified Ethical Emerging Technologist (CEET)
CertNexus

Ethics and Governance of Artificial Intelligence
edX (The University of Tokyo)

AI Ethics for Business
Coursera (INSEAD)

Responsible AI Lead Certification
AI Responsibility Lab

Ethical Intelligence Certification
The Ethics & Compliance Initiative (ECI)

AI and Ethics
FutureLearn (The University of Helsinki)

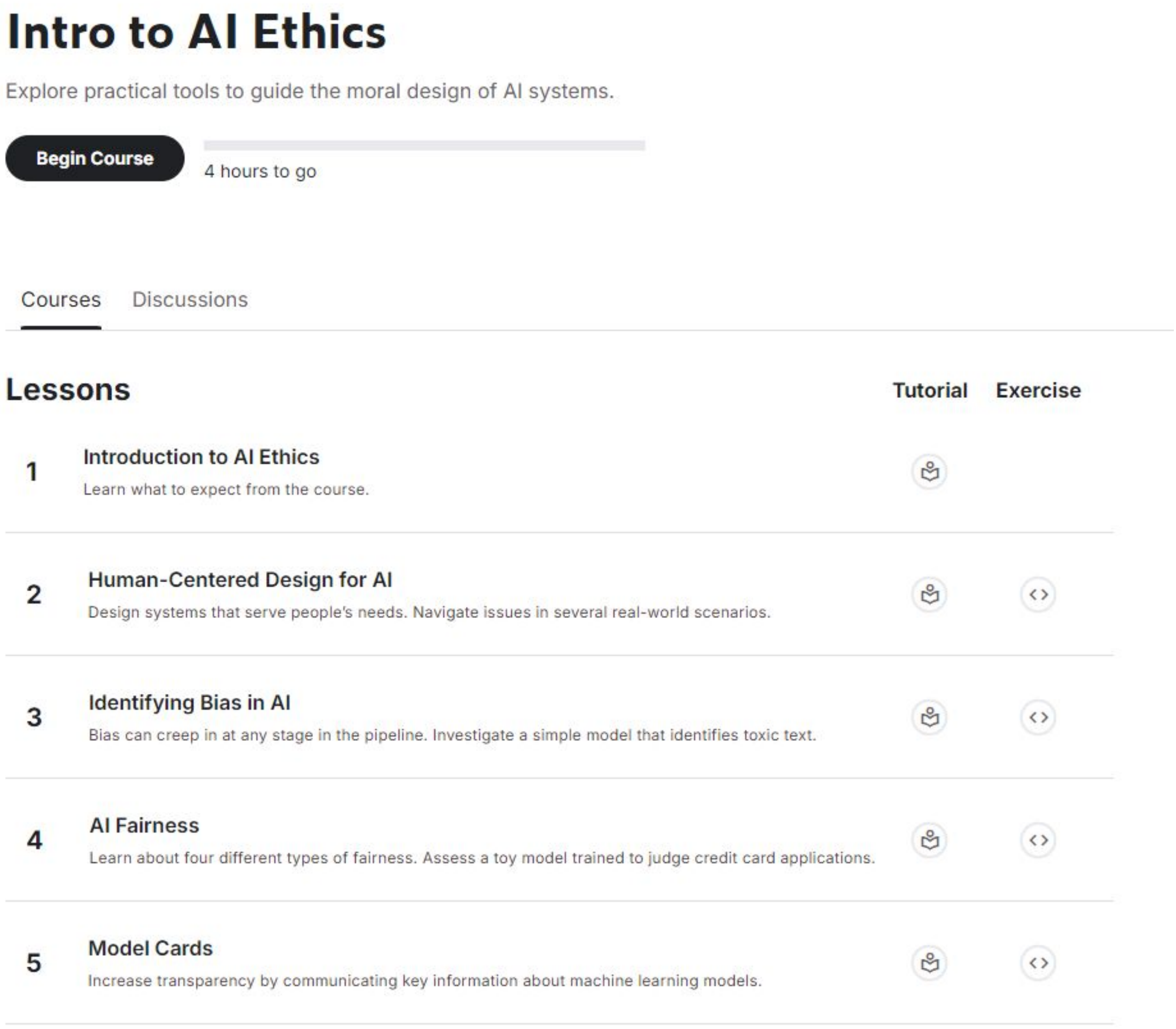
Professional Certificate in AI Ethics and Governance
Asia Pacific University of Technology & Innovation (APU)

AI Ethics: Tools for Ethics and Compliance Certification
Coursera (University of Colorado Boulder)

Executive Program in Ethical Artificial Intelligence
IE University – School of Human Sciences and Technology

https://www.tealhq.com/certifications/ai-ethics-specialist

Kaggle Badge – Add to LinkedIn



Intro to AI Ethics

Explore practical tools to guide the moral design of AI systems.

Lessons

Tutorial Exercise

1 **Introduction to AI Ethics**
Learn what to expect from the course.

2 **Human-Centered Design for AI**
Design systems that serve people's needs. Navigate issues in several real-world scenarios.

3 **Identifying Bias in AI**
Bias can creep in at any stage in the pipeline. Investigate a simple model that identifies toxic text.

4 **AI Fairness**
Learn about four different types of fairness. Assess a toy model trained to judge credit card applications.

5 **Model Cards**
Increase transparency by communicating key information about machine learning models.

https://www.kaggle.com/learn/intro-to-ai-ethics

Impact of AI on Society and why it needs to be regulated

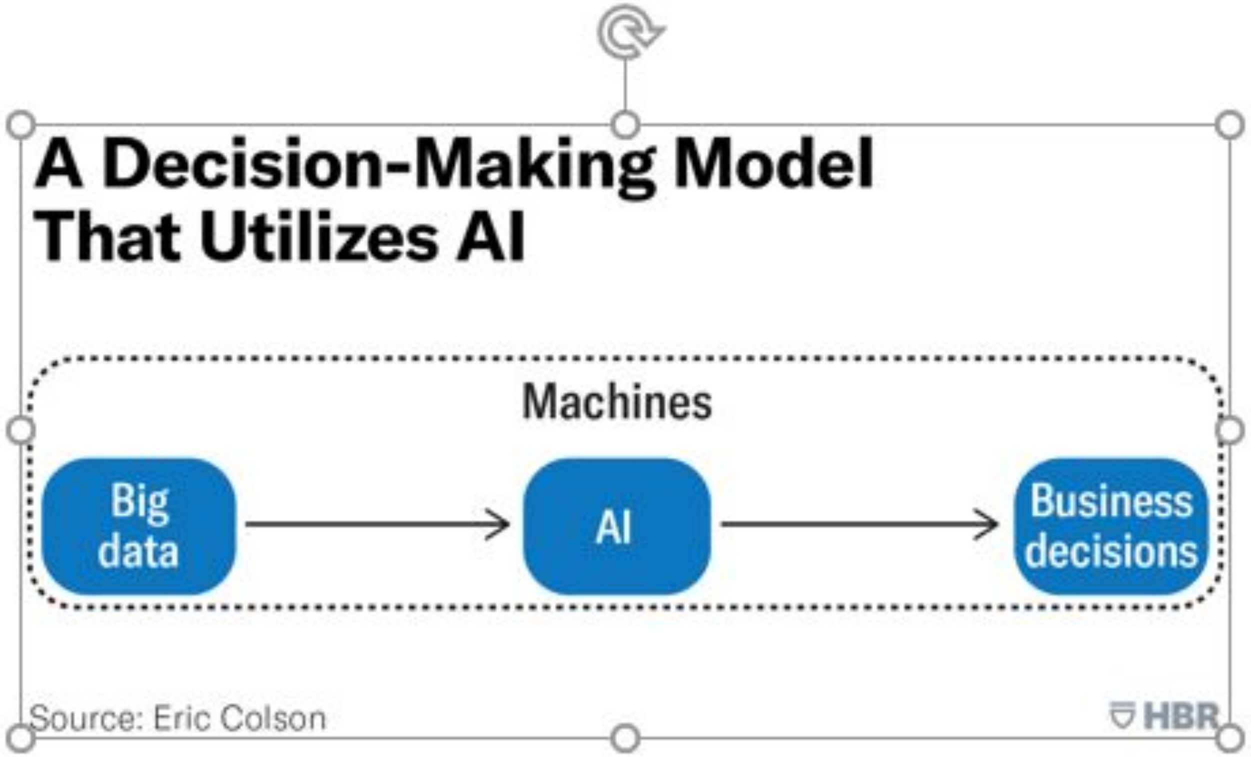
Impact of AI in Society today

- AI represents the most significant economic opportunity of our lifetime – estimated to contribute $15.7 trillion to the global economy by 2030, according to PwC research

- AI is increasingly employed to make decisions affecting most aspects of our lives, mainly as digital transformation accelerates in the face of COVID-19, both before and after.

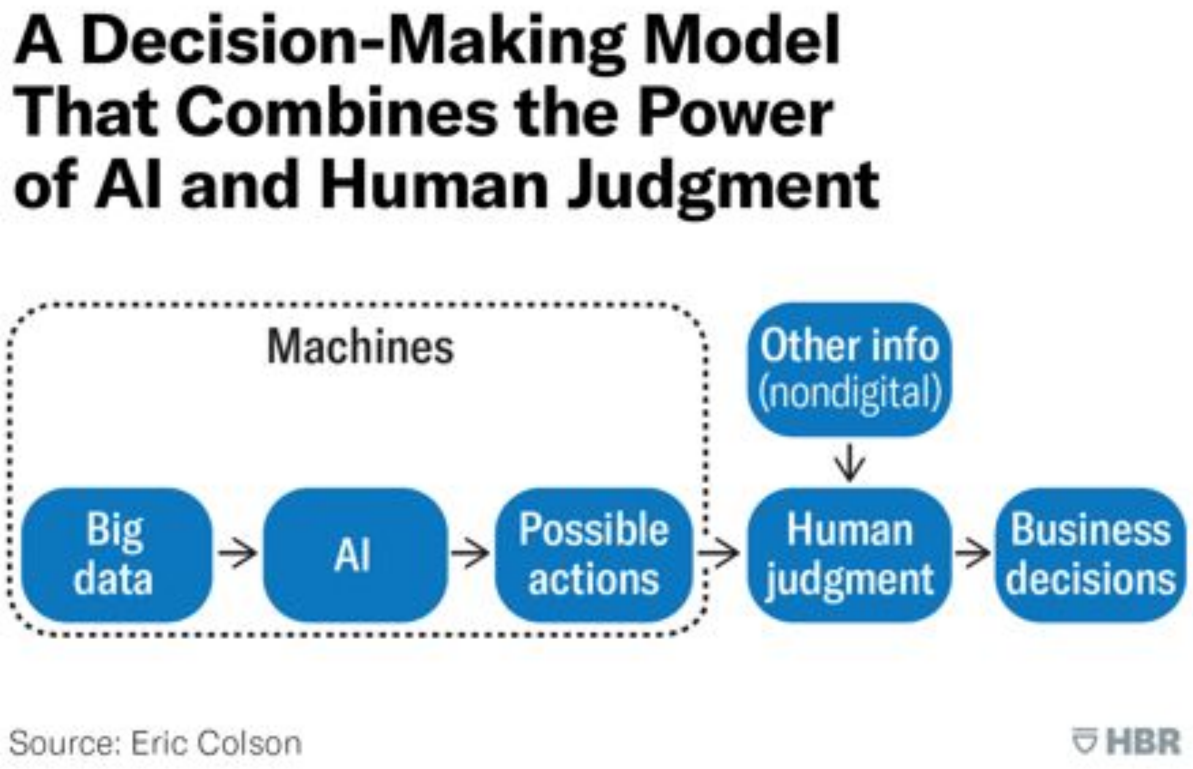
- **AI is automating decision-making in our everyday life:**
  - Inform who will receive an interview.
  - Who gets credit or who does not?
  - Which product gets advertised to which consumers?
  - Who gets welfare?
  - Who gets parole?
  - Which neighborhoods are targeted at high risk?
  - What is a home priced at? What level of mortgage is a home buyer qualified for?
  - Do Patients have disease symptoms or not ?

Impact of AI in Society today – AI Based Decision Making



A Decision-Making Model That Utilizes AI

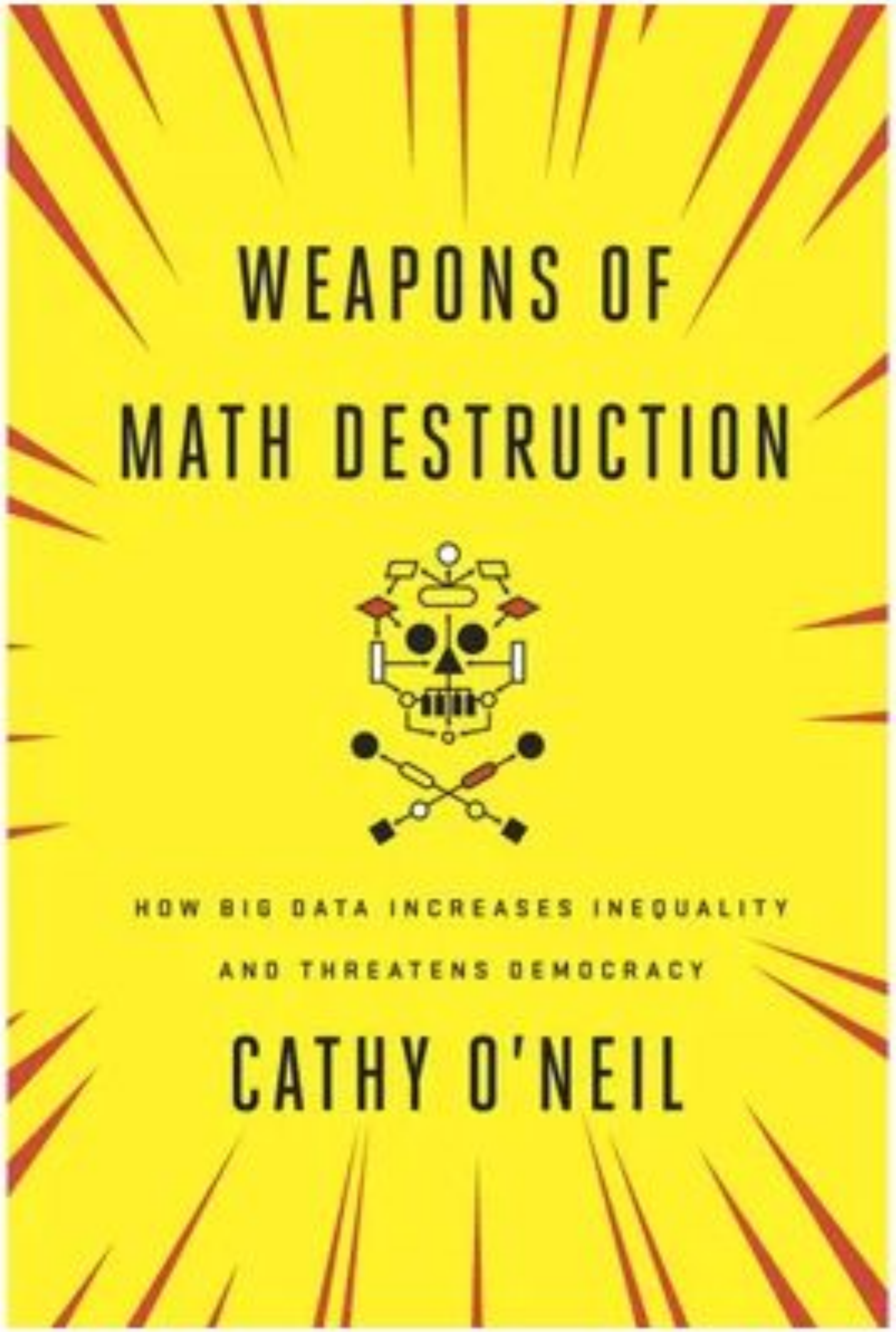
Source: Eric Colson

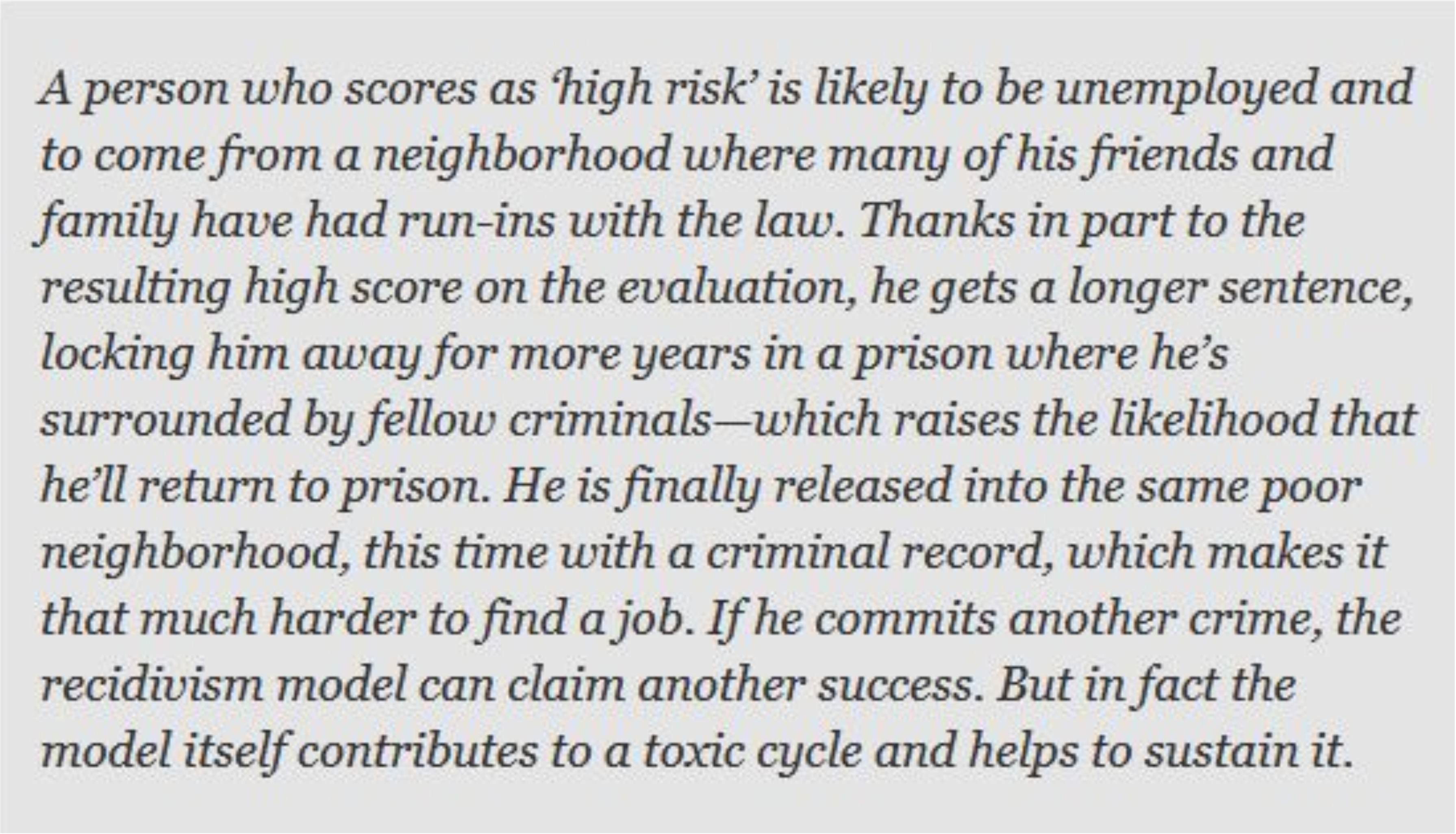


A Decision-Making Model That Combines the Power of AI and Human Judgment

Source: Eric Colson

Why AI Needs to be regulated ?



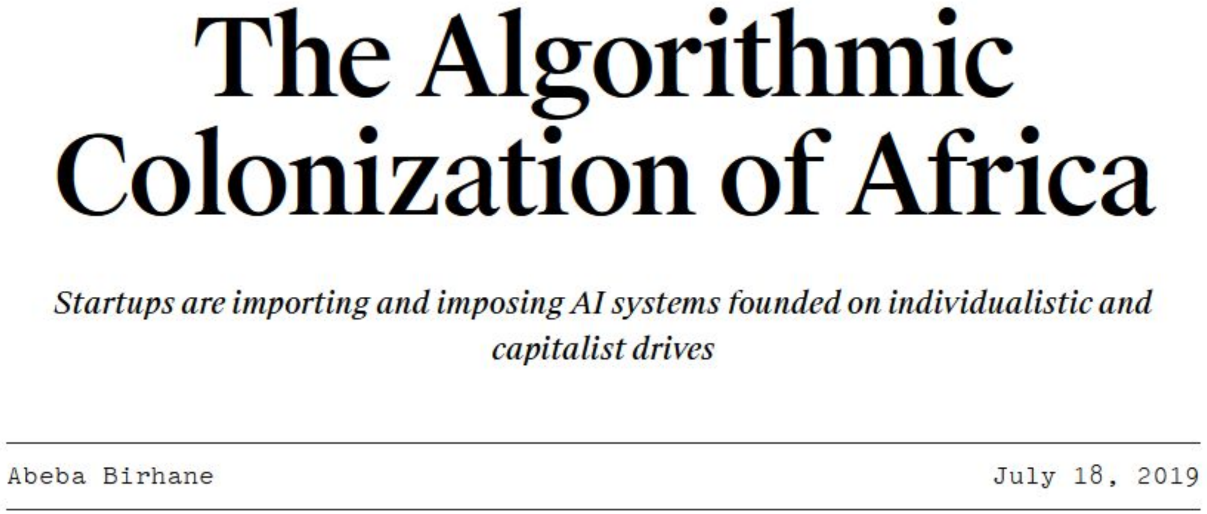


A person who scores as 'high risk' is likely to be unemployed and to come from a neighborhood where many of his friends and family have had run-ins with the law. Thanks in part to the resulting high score on the evaluation, he gets a longer sentence, locking him away for more years in a prison where he's surrounded by fellow criminals—which raises the likelihood that he'll return to prison. He is finally released into the same poor neighborhood, this time with a criminal record, which makes it that much harder to find a job. If he commits another crime, the recidivism model can claim another success. But in fact the model itself contributes to a toxic cycle and helps to sustain it.

Machines serve the poor and humans are catering to the rich leading to society inequality

https://blogs.scientificamerican.com/roots-of-unity/review-weapons-of-math-destruction/

Why AI Needs to be regulated ?

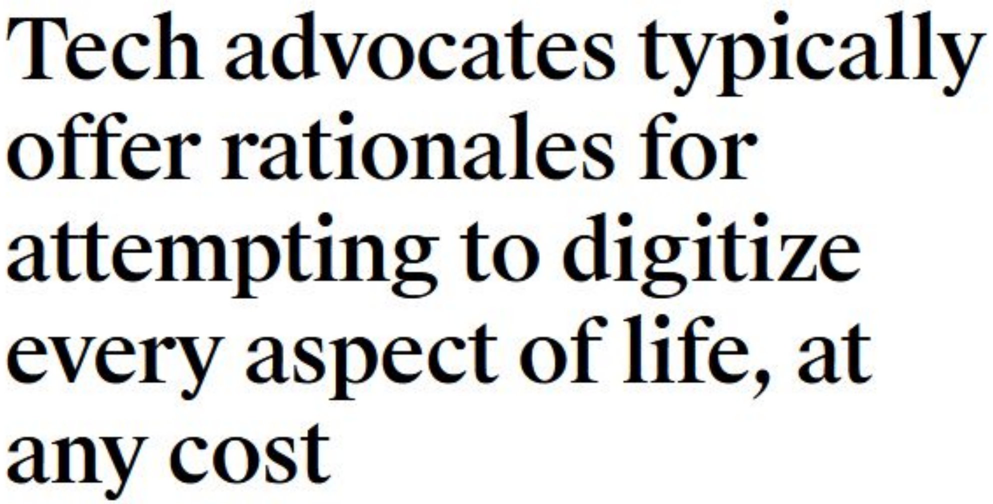


The Algorithmic Colonization of Africa

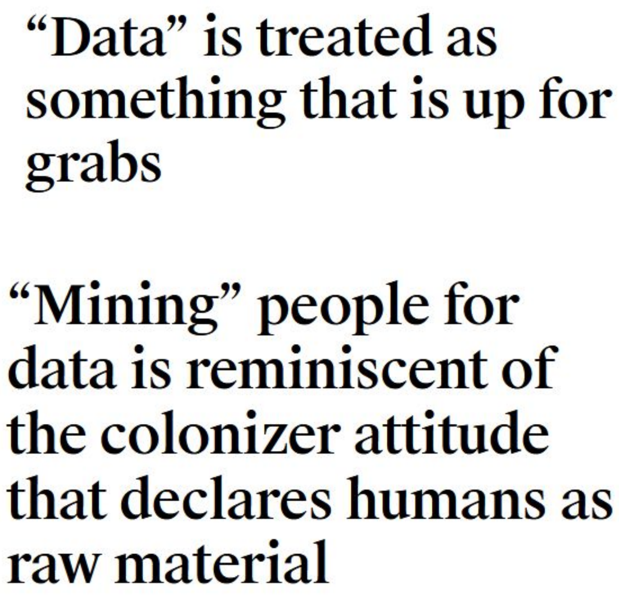
Startups are importing and imposing AI systems founded on individualistic and capitalist drives

Abeba Birhane　　July 18, 2019

https://reallifemag.com/the-algorithmic-colonization-of-africa/



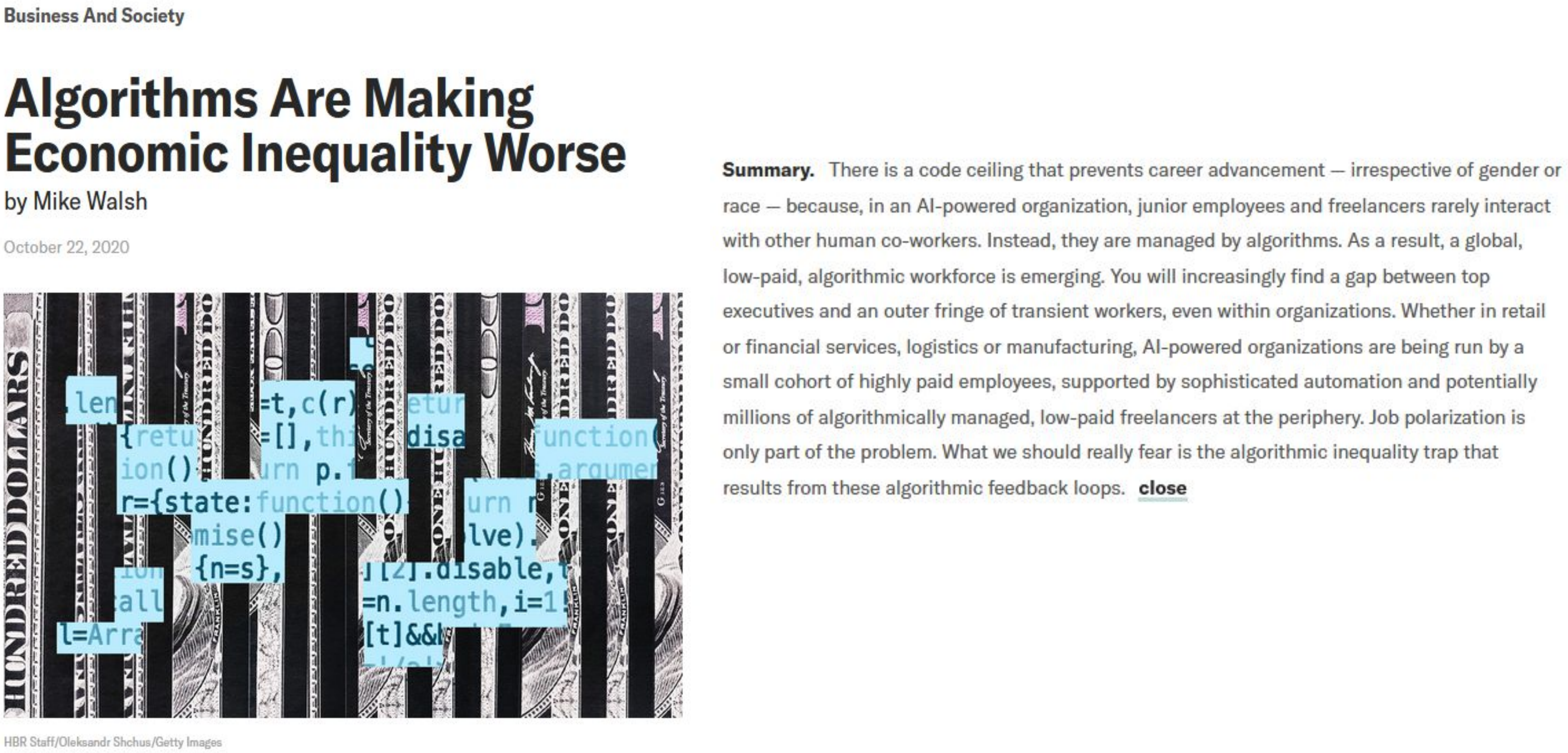
Tech advocates typically offer rationales for attempting to digitize every aspect of life, at any cost



“Data” is treated as something that is up for grabs

“Mining” people for data is reminiscent of the colonizer attitude that declares humans as raw material

Why AI Needs to be regulated ?



Business And Society

Algorithms Are Making Economic Inequality Worse

by Mike Walsh

October 22, 2020

HBR Staff/Oleksandr Shchus/Getty Images

Summary. There is a code ceiling that prevents career advancement — irrespective of gender or race — because, in an AI-powered organization, junior employees and freelancers rarely interact with other human co-workers. Instead, they are managed by algorithms. As a result, a global, low-paid, algorithmic workforce is emerging. You will increasingly find a gap between top executives and an outer fringe of transient workers, even within organizations. Whether in retail or financial services, logistics or manufacturing, AI-powered organizations are being run by a small cohort of highly paid employees, supported by sophisticated automation and potentially millions of algorithmically managed, low-paid freelancers at the periphery. Job polarization is only part of the problem. What we should really fear is the algorithmic inequality trap that results from these algorithmic feedback loops. **close**

https://hbr.org/2020/10/algorithms-are-making-economic-inequality-worse

AI Ethics Why ?

- Need to explore the ethical, social, and legal aspects of AI systems given the impact of AI decisions on society

- **Ethics of AI: How can we develop and use this technology in an ethically acceptable and sustainable way?**

What is Ethics and how is AI ethics a sub-field

Ethics

Ethics – Answer questions Good or bad, What is right or wrong, what is justice, well-being, equality

- **Subfields of Ethics:**
  - **Meta-ethics**
  - **Normative ethics**
  - **Applied ethics -** ***concerns what a moral agent (someone who can judge right and wrong and be held accountable) is obligated or permitted to do in a specific situation or a particular domain of action.***

Ethics & AI Ethics

Ethics is a set of moral principles that help us discern right from wrong. AI ethics is a set of guidelines that advise on the design and outcomes of artificial intelligence

Since data is the foundation for all machine learning algorithms, it’s important to structure experiments and algorithms with this in mind. Artificial intelligence has the potential to amplify and scale these human biases at an unprecedented rate.

AI Ethics

AI Ethics is a sub-field of Applied Ethics.

- **Part of the ethics of technology specific to robots and other artificially intelligent entities**
- **Concerns question how developers, manufacturers, and corporations should behave to minimize the ethical risks that can arise from AI in society**
- **Minimize ethical risk concerns:**
  - **Design**
  - **Inappropriate application**
  - **Misuse of technology**

ML Algorithm (Software)

ML Applications

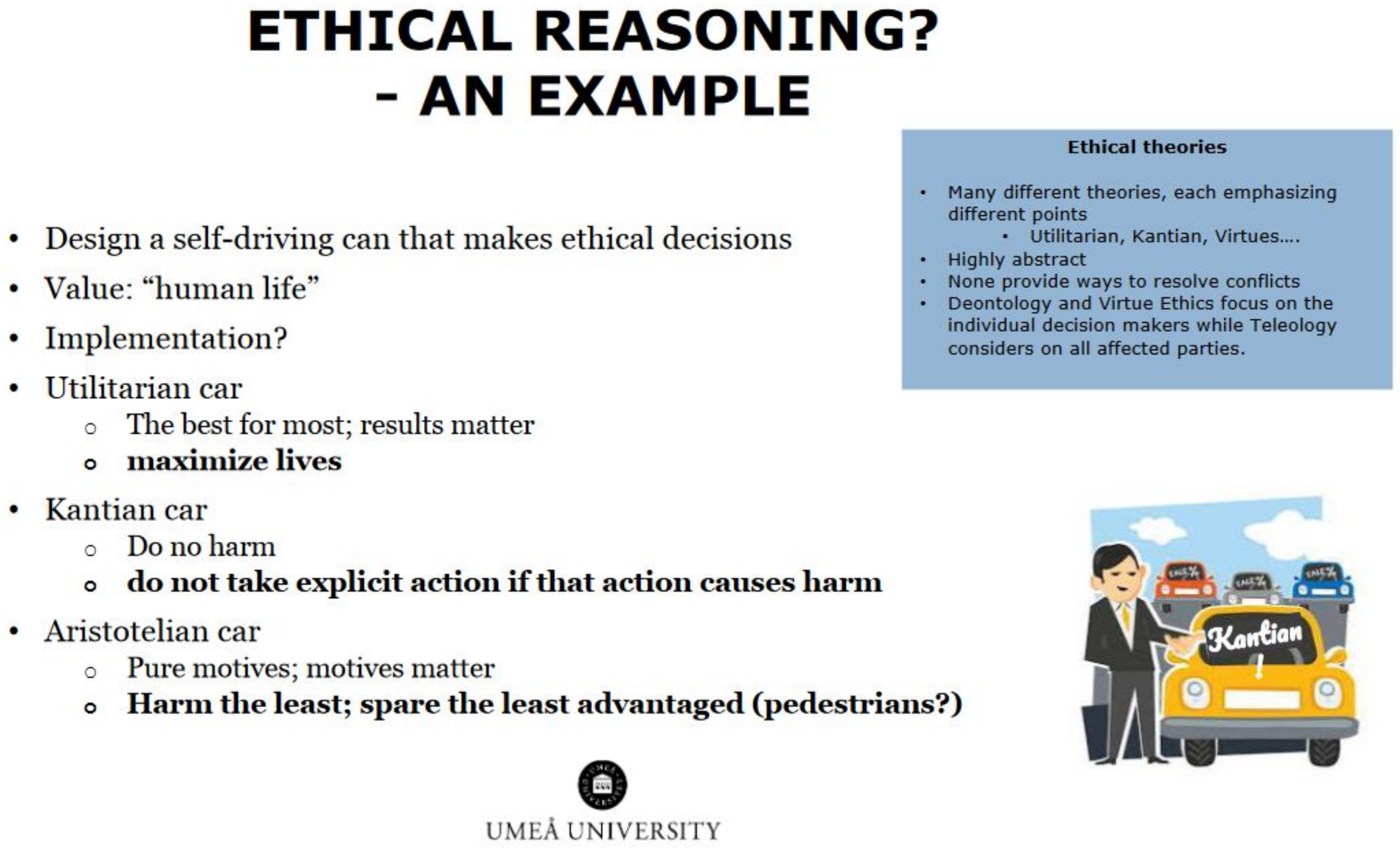
ML Robots

Ethical Reasoning - Five Ethical Lenses (AI Ethics is a sub-field of Ethics)

- **Right approach** – Which option best respects the rights of all who have stakes
- **Justice approach** – Which option treats people equally or proportionately
- **Utilitarian approach** – Which option will produce the best and do the least harm
- **Common good approach** – Which option best serves the community as a whole and not just members
- **Virtue approach** – Which option leads me to act as the sort of person I want to be

https://www.scu.edu/media/ethics-center/technology-ethics/Tech_and_Engineering_Practice-Ethical_Lenses-2022.pdf

Five Ethical Lenses in designing a self-driving car



ETHICAL REASONING?
- AN EXAMPLE

- Design a self-driving can that makes ethical decisions
- Value: “human life”
- Implementation?
- Utilitarian car
  - The best for most; results matter
  - **maximize lives**
- Kantian car
  - Do no harm
  - **do not take explicit action if that action causes harm**
- Aristotelian car
  - Pure motives; motives matter
  - **Harm the least; spare the least advantaged (pedestrians?)**

Ethical theories

- Many different theories, each emphasizing different points
  - Utilitarian, Kantian, Virtues….
- Highly abstract
- None provide ways to resolve conflicts
- Deontology and Virtue Ethics focus on the individual decision makers while Teleology considers on all affected parties.

UMEÅ UNIVERSITY

IEAI Institute of Ethics in Artificial Intelligence
Technical University of Munich

AI Ethics

Autonomous Driving

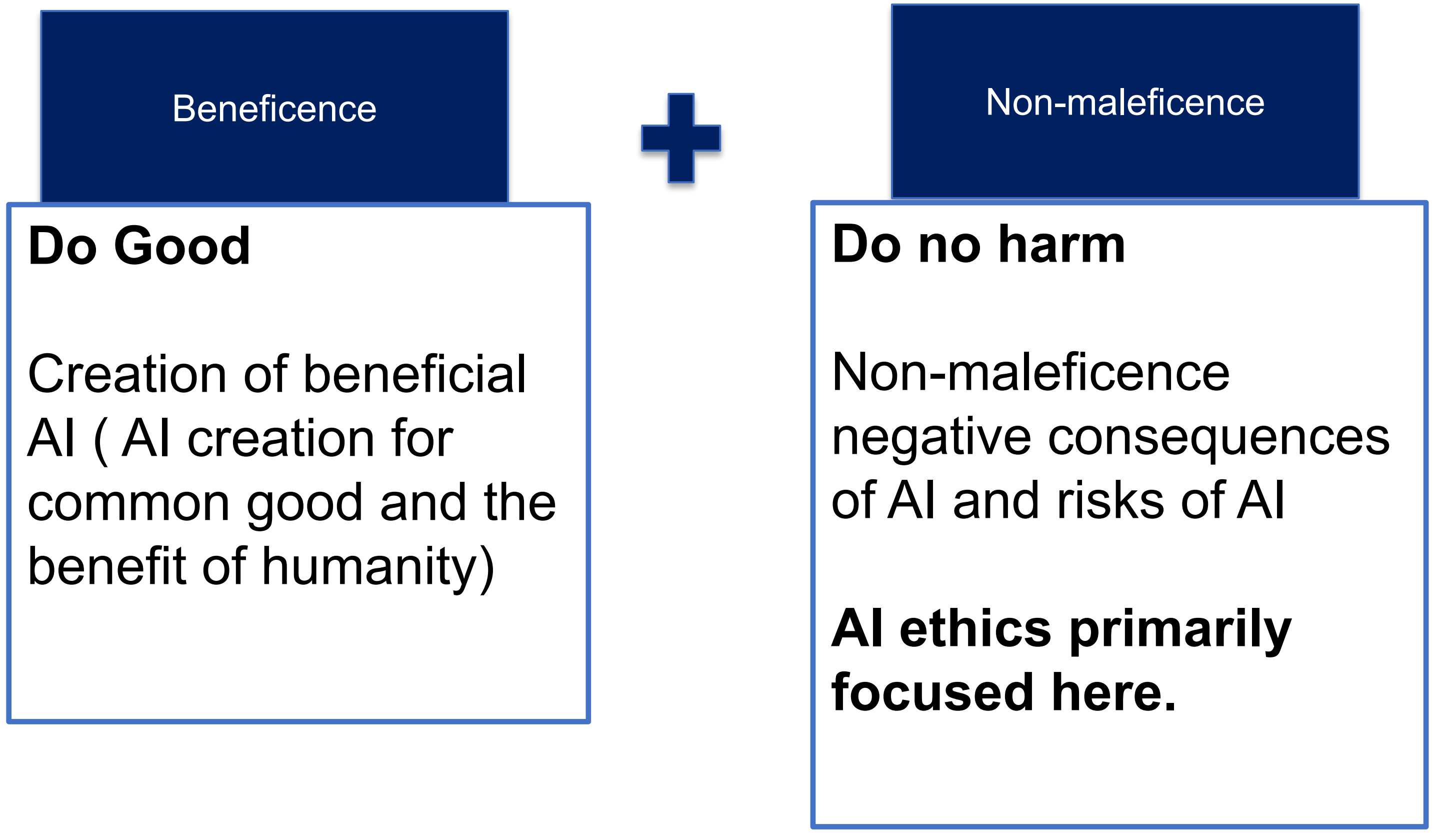
Prof. Dr. Christoph Lütge

Director, TUM Institute for Ethics in Artificial Intelligence

Technical University of Munich

https://www.youtube.com/watch?v=Y8nfvBVB_FM

The principle of Common Good



Beneficence

Do Good

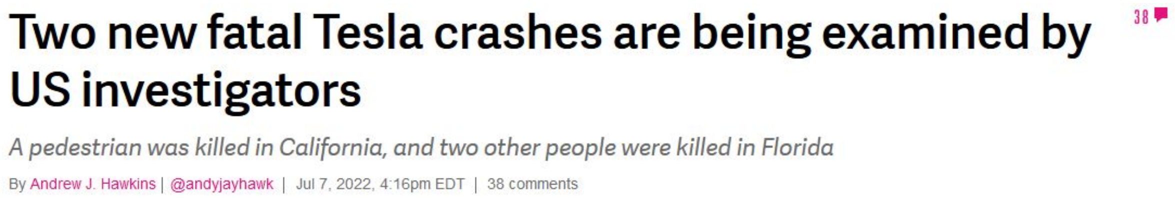
Creation of beneficial AI (AI creation for common good and the benefit of humanity)

Non-maleficence

Do no harm

Non-maleficence negative consequences of AI and risks of AI

AI ethics primarily focused here.



Two new fatal Tesla crashes are being examined by US investigators

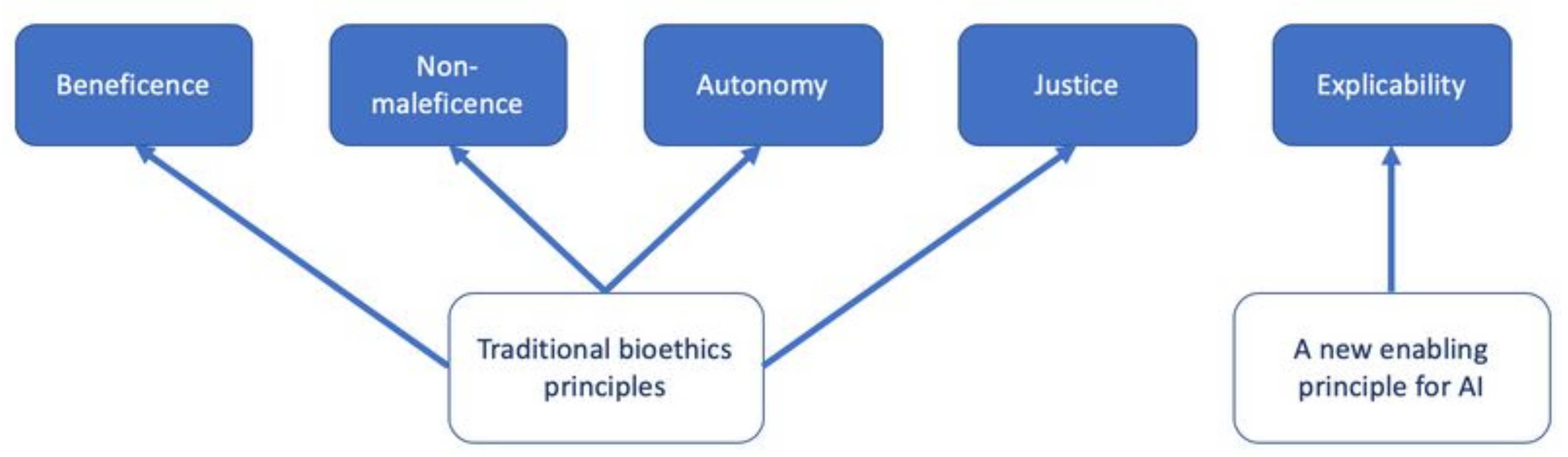
38

A pedestrian was killed in California, and two other people were killed in Florida

By Andrew J. Hawkins | @andyjayhawk | Jul 7, 2022, 4:16pm EDT | 38 comments

https://www.theverge.com/2022/7/7/23198997/tesla-fatal-crashes-california-florida-autopilot-nhtsa

Unified Framework for Five Principles for AI in Society


Beneficence
Non-maleficence
Autonomy
Justice
Explicability
Traditional bioethics principles
A new enabling principle for AI

https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/8

AI Ethics

AI Ethics Concerns

- Security, Privacy, and transparency of AI Systems
- Impact of AI in high-risk scenarios – Military, Medical, Justice, and Education system
- Ethics on developing and implementing AI in society

AI Ethics 5 Core Principles

- Non-Maleficence (Avoid Harm)
- Responsbility or Accountability
- Transparency or Explainability
- Justice or Fairness
- Human rights – Privacy and Security

https://arxiv.org/ftp/arxiv/papers/1906/1906.11668.pdf

AI Ethics 5 Core Principles Questions

1. **Should we use AI for good and not for causing harm**? (the principle of beneficence/ non-maleficence)
2. **Who should be blamed when AI causes harm**? (the principle of accountability)
3. **Should we understand what and why AI does whatever it does**? (the principle of transparency)
4. **Should AI be fair or non-discriminative**? (the principle of fairness)
5. **Should AI respect and promote human rights**? (the principle of respecting fundamental human rights)

https://arxiv.org/ftp/arxiv/papers/1906/1906.11668.pdf

AI Ethics Focus – Non-Maleficence

- Discussion has focused chiefly on how developers, manufacturers, authorities, or other stakeholders should minimize the ethical risks—discrimination, privacy protection, and physical and social harms—that can arise from AI applications.

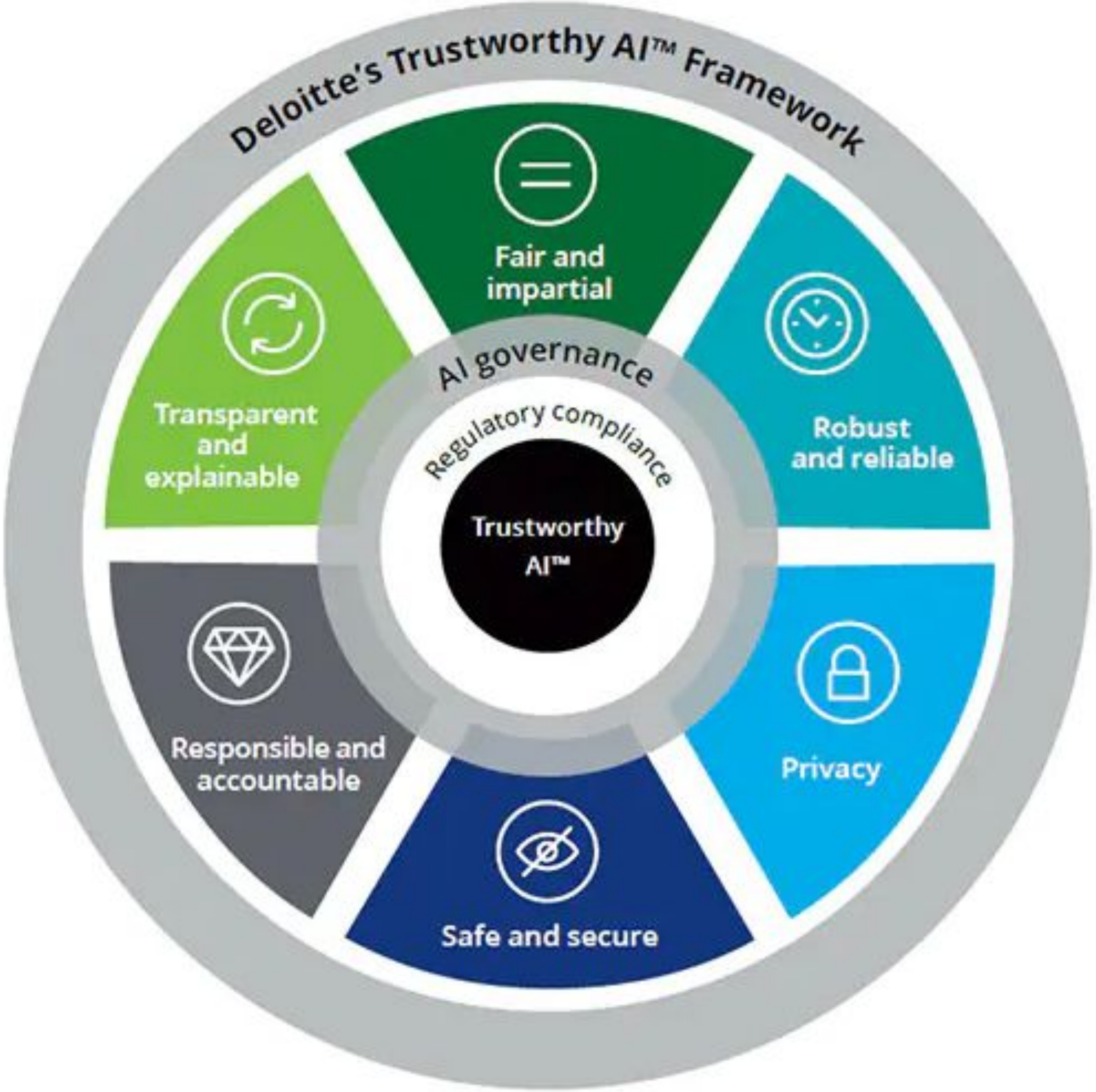
- Intentional misuse, malicious hacking, technical measures, or risk-management strategies.

- Technology focus - *Evgeny Morozov calls this "tech solutionism" – the conviction that problems caused by technology can permanently be fixed by more technology*. – Results from deep ethical problems are oversimplified or unanswered

AI Ethics Frameworks – (Course will cover each of these components in detail)

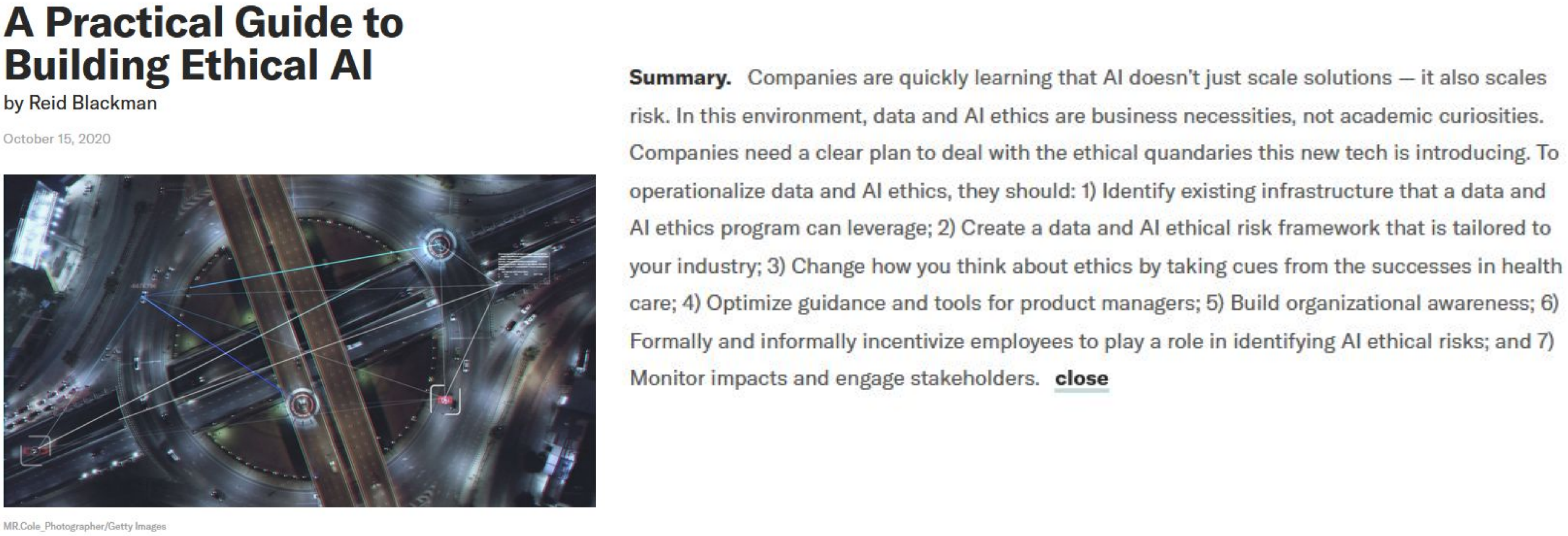
Terms used – Trustworthy, Responsible, Secure, Ethical AI

Deloitte Framework



https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html

Harvard Business Review



A Practical Guide to Building Ethical AI

by Reid Blackman

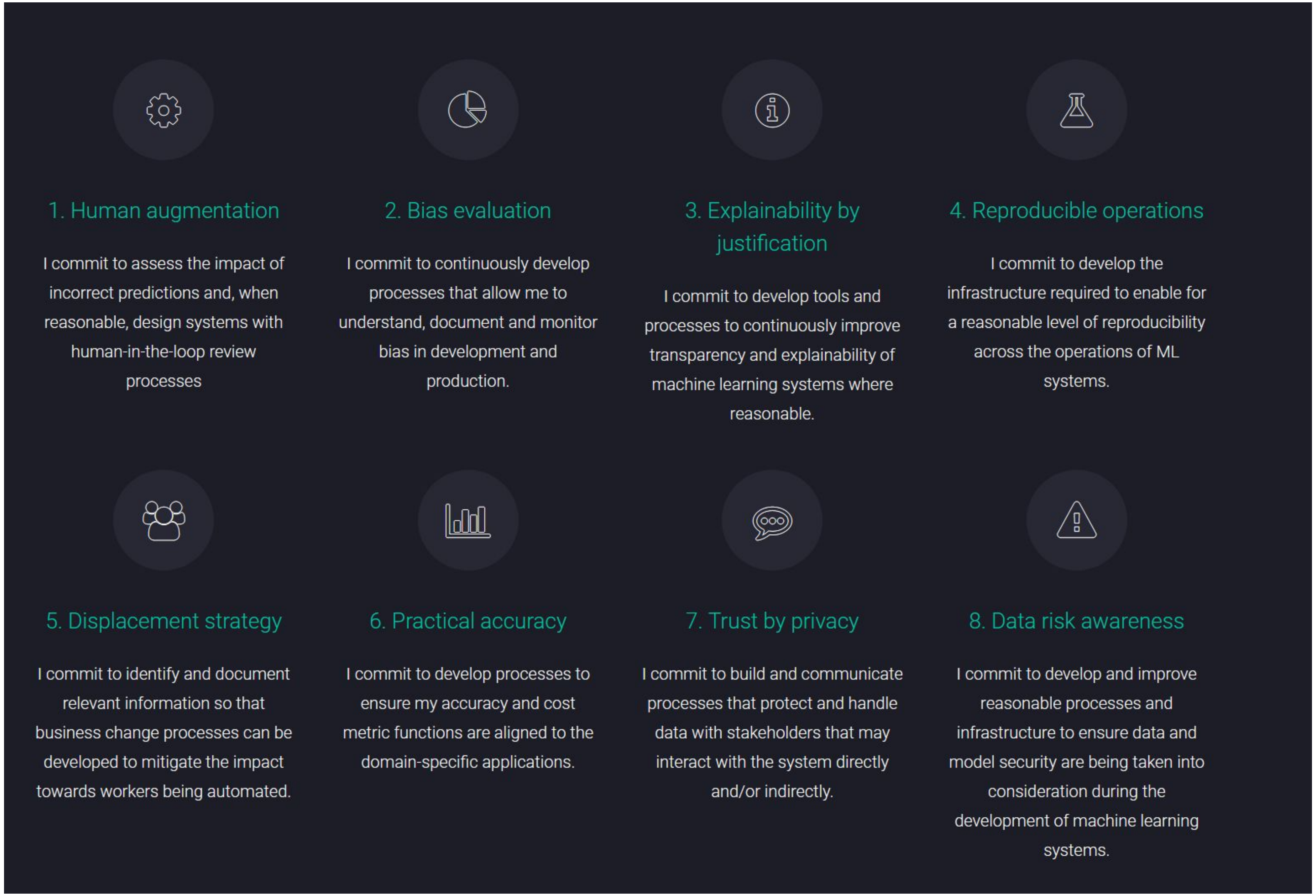
October 15, 2020

MR.Cole_Photographer/Getty Images

Summary. Companies are quickly learning that AI doesn't just scale solutions — it also scales risk. In this environment, data and AI ethics are business necessities, not academic curiosities. Companies need a clear plan to deal with the ethical quandaries this new tech is introducing. To operationalize data and AI ethics, they should: 1) Identify existing infrastructure that a data and AI ethics program can leverage; 2) Create a data and AI ethical risk framework that is tailored to your industry; 3) Change how you think about ethics by taking cues from the successes in health care; 4) Optimize guidance and tools for product managers; 5) Build organizational awareness; 6) Formally and informally incentivize employees to play a role in identifying AI ethical risks; and 7) Monitor impacts and engage stakeholders.

https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai

Responsible Machine Learning Principles



1. Human augmentation

I commit to assess the impact of incorrect predictions and, when reasonable, design systems with human-in-the-loop review processes

2. Bias evaluation

I commit to continuously develop processes that allow me to understand, document and monitor bias in development and production.

3. Explainability by justification

I commit to develop tools and processes to continuously improve transparency and explainability of machine learning systems where reasonable.

4. Reproducible operations

I commit to develop the infrastructure required to enable for a reasonable level of reproducibility across the operations of ML systems.

5. Displacement strategy

I commit to identify and document relevant information so that business change processes can be developed to mitigate the impact towards workers being automated.

6. Practical accuracy

I commit to develop processes to ensure my accuracy and cost metric functions are aligned to the domain-specific applications.

7. Trust by privacy

I commit to build and communicate processes that protect and handle data with stakeholders that may interact with the system directly and/or indirectly.

8. Data risk awareness

I commit to develop and improve reasonable processes and infrastructure to ensure data and model security are being taken into consideration during the development of machine learning systems.

https://ethical.institute/principles.html

This Eyeball Orb Wants to Save Us From AI | Hello World

https://youtu.be/bCOWSfPXuP8?si=ha6pBvzcZCVqo7Ad

Break - 5 min

Human-Centered Design

Understand the fundamentals of AI Ethics and how to design AI systems to ensure they serve the needs of the people it is intended for.

Human Centered Design

- **Human-centered design (HCD)** is an approach to designing systems that serve people’s needs.

HCD involves people in every step of the design process. Your team should adopt an HCD approach to AI as early as possible - ideally, from when you begin to entertain the possibility of building an AI system.

1.Understand People Needs

Understand people’s needs to define the problem – Observing people, conducting interviews, focus groups and reading user feedback . Use a multi-disciplinary team

1. A company wants to address the **problem of dosage errors** for immunosuppressant drugs given to patients after liver transplants.
2. T**he company starts by observing physicians, nurses and other hospital staff throughout the liver transplant process**. It also **interviews** them about the current dosage determination process - which relies on published guidelines and human judgment - and shares video clips from the interviews with the entire development team.
3. The company also **reviews research studies and assembles focus groups of former patients and their families**. All team members participate in a freewheeling brainstorming session for potential solutions.*

**Source: Kaggle*

2. Does AI Add value to the solution ?

Ask if AI adds value to any potential solution
Ask the questions:

a. **Would people generally agree that what you are trying to achieve is a good outcome?**
b. **Would non-AI systems - such as rule-based solutions, which are easier to create, audit and maintain - be significantly less effective than an AI system?**
c. **Is the task that you are using AI for one that people would find boring, repetitive or otherwise difficult to concentrate on?**
d. **Have AI solutions proven to be better than other solutions for similar use cases in the past?**

1. A disaster response agency is working with first responders to reduce the time it takes to rescue people from disasters like floods.
2. The time- and labor-intensive human review of drone and satellite photos to find stranded people increases rescue time.
3. Everybody agrees that speeding up photo review would be a good outcome since faster rescues could save more lives.
4. The agency determined that an AI image recognition system would likely be more effective than a non-AI automated system for this task.
5. It is also aware that AI-based image recognition tools have been applied successfully to review aerial footage in other industries, like agriculture. The agency, therefore, decides to explore the possibility of an AI-based solution further.

3. AI System harm considerations

Consider the potential harms that the AI system could cause

- ❖ Weigh the benefits of using AI against the potential harms throughout the design pipeline, from collecting and labeling data to training a model to deploying the AI system.
- ❖ Consider the impact on users and society.
- ❖ Your privacy team can help uncover hidden privacy issues and determine whether privacy-preserving techniques like differential privacy or federated learning may be appropriate. Take steps to reduce harm, including by embedding people - and therefore, human judgment

1. An online education company wants to use an AI system to ‘read’ and automatically assign scores to student essays while redirecting company staff to double-check random essays and to review essays that the AI system has trouble with.
2. The system would enable the company to get scores back to students quickly.
3. The company created a harms review committee, which recommends that the system not be built.
4. Some of the significant harms flagged by the committee include the potential for the AI system to pick up bias against specific patterns of language from training data and amplify it (harming people in the groups that use those patterns of language), to encourage students to ‘game’ the algorithm rather than improve their essays and to reduce the classroom role of education experts while increasing the role of technology experts.

4.Prototype

Prototype – Start with non-AI Solutions

Develop a non-AI prototype of your AI system quickly to see how people interact with it. This makes prototyping easier, faster and less expensive. It also gives you early information about what users expect from your system and how to make their interactions more rewarding and meaningful.

The people giving feedback should have diverse backgrounds – including along race, gender, expertise and other characteristics. They should also understand and consent to what they are helping with and how.

1. A movie streaming startup wants to use AI to recommend movies to users, based on their stated preferences and viewing history.
2. The team first invites a diverse group of users to share their stated preferences and viewing history with a movie enthusiast, who then recommends movies that the users might like.
3. Based on these conversations and on feedback about which recommended movies users enjoyed, the team changes its approach to how movies are categorized.
4. Getting feedback from a diverse group of users early and iterating often allows the team to improve its product early, rather than making expensive corrections later.

5.Challenge the system

Provide ways for people to challenge the system

People who use your AI system once it is live should be able to challenge its recommendations or easily opt out of using it. Put systems and tools in place to accept, monitor and address challenges.

Talk to users and think from the perspective of a user: if you are curious or dissatisfied with the system's recommendations, would you want to challenge it by:

- Requesting an explanation of how it arrived at its recommendation?
- Requesting a change in the information you input?
- Turning off certain features?
- Reaching out to the product team on social media?
- Taking some other action?

1. An online video conferencing company uses AI to blur the background automatically during video calls.
2. The company has successfully tested its product with diverse people from different ethnicities. Still, it knows that there could be instances in which the video may not properly focus on a person's face.
3. So, it makes the background blurring feature optional and adds a button for customers to report issues. The company also creates a customer service team to monitor social media and other online forums for user complaints.

6. Build Safety Measures

Build in Safety Measures

Safety measures protect users against harm. They seek to limit unintended behavior and accidents, by ensuring that a system reliably delivers high-quality outcomes. This can only be achieved through extensive and continuous evaluation and testing. Design processes around your AI system to continuously monitor performance, delivery of intended benefits, reduction of harms, fairness metrics and any changes in how people are actually using it.

Human oversight of your AI system is crucial:

- **Create a human 'red team' to play the role of a person trying to manipulate your system into unintended behavior**. Then, strengthen your system against any such manipulation.
- Determine how people in your **organization can best monitor the system's safety once it is live**.
- Explore ways for your **AI system to quickly alert a human when it encounters a challenging case**.
- Create ways for users and others to flag potential safety issues.

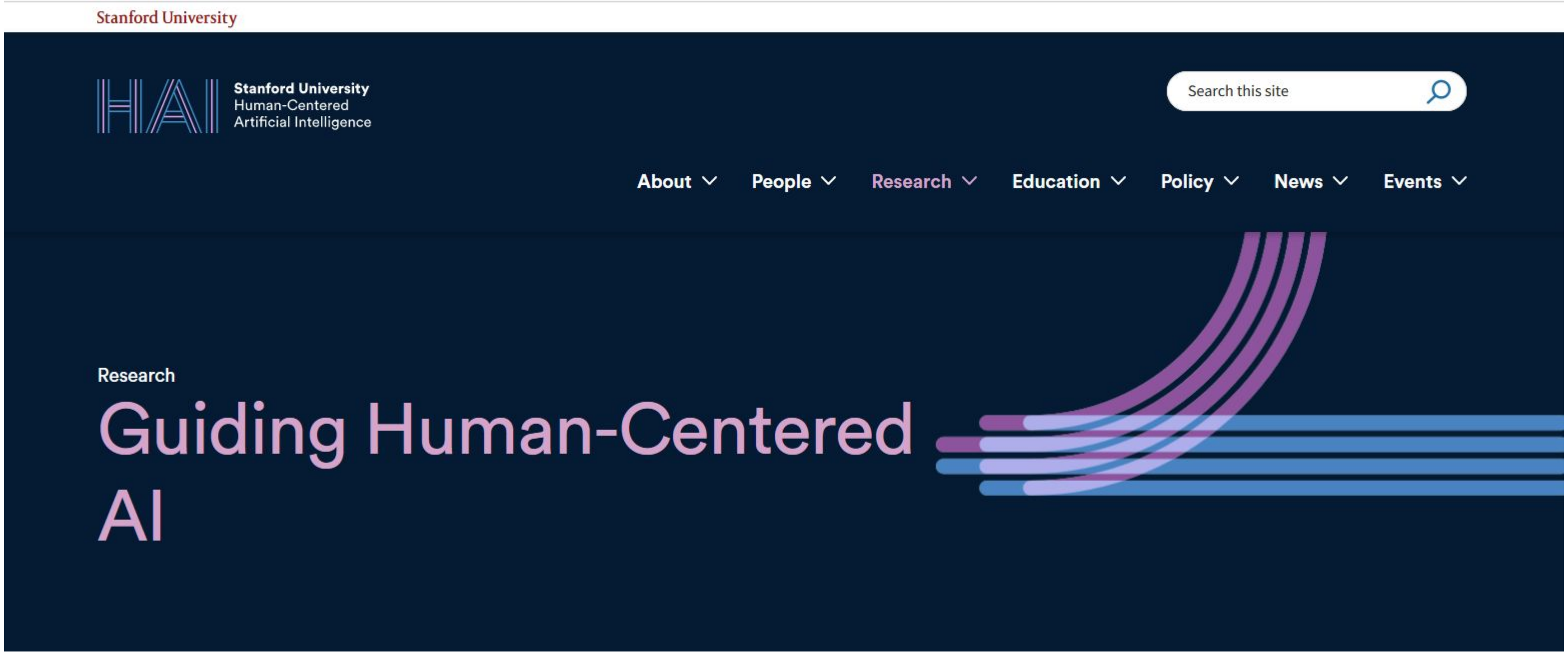
1. To bolster the safety of its product, a company that develops a widely-used AI-enabled voice assistant creates a permanent internal 'red team' to play the role of bad actors that want to manipulate the voice assistant.
2. The red team develops adversarial inputs to fool the voice assistant.
3. The company then uses 'adversarial training' to guard the product against similar adversarial inputs, improving its safety.

https://openai.com/index/red-teaming-network/

MIT 6.S093: Introduction to Human-Centered Artificial Intelligence (AI)

https://www.youtube.com/watch?v=bmjamLZ3v8A

Stanford Human Centered Artificial Intelligence



Research

Guiding Human-Centered AI

https://hai.stanford.edu/research

Appendix