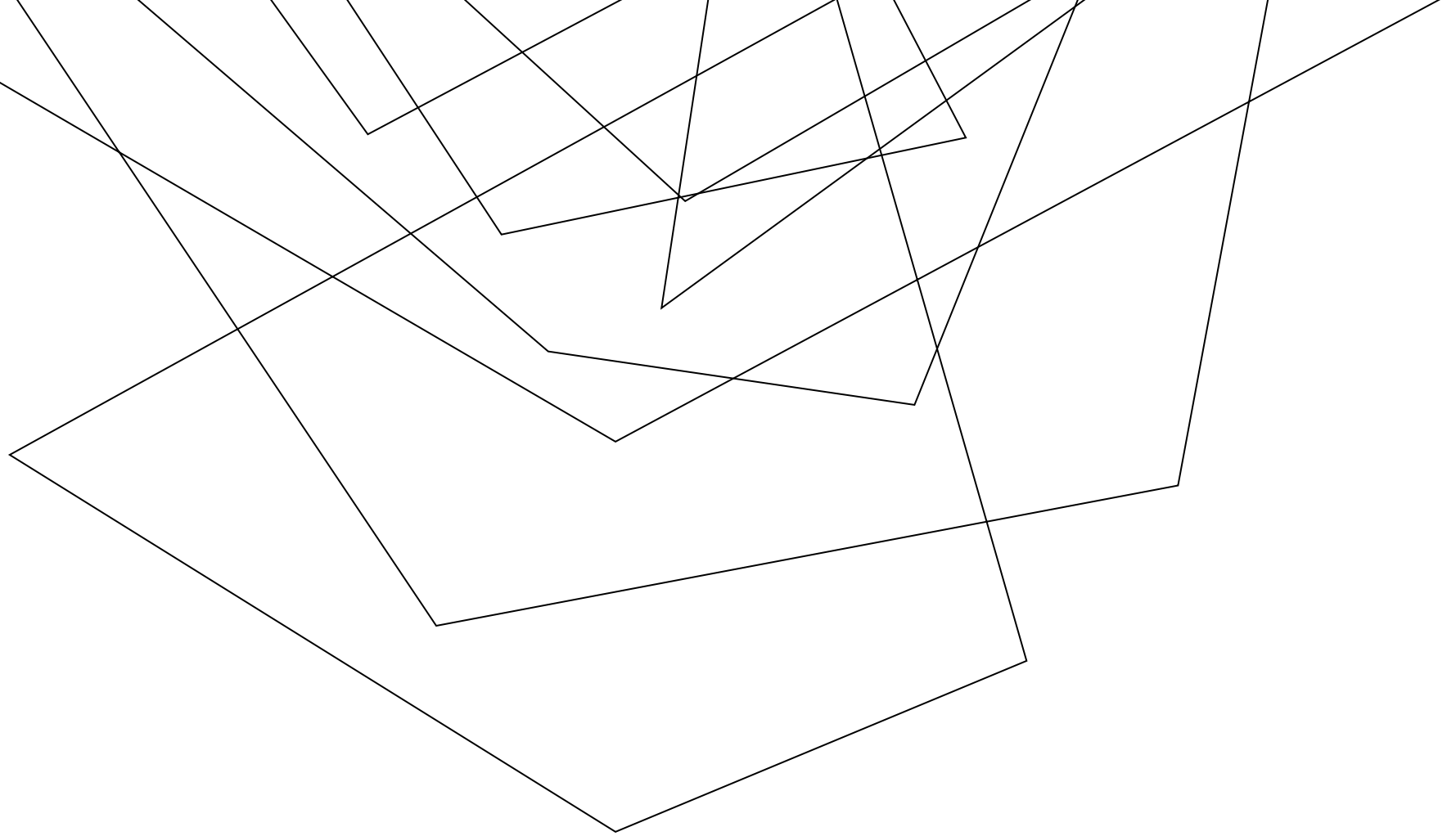


RESPONSIBLE AI

Week 5: AI Regulations

IMPORTANT ANNOUNCEMENTS

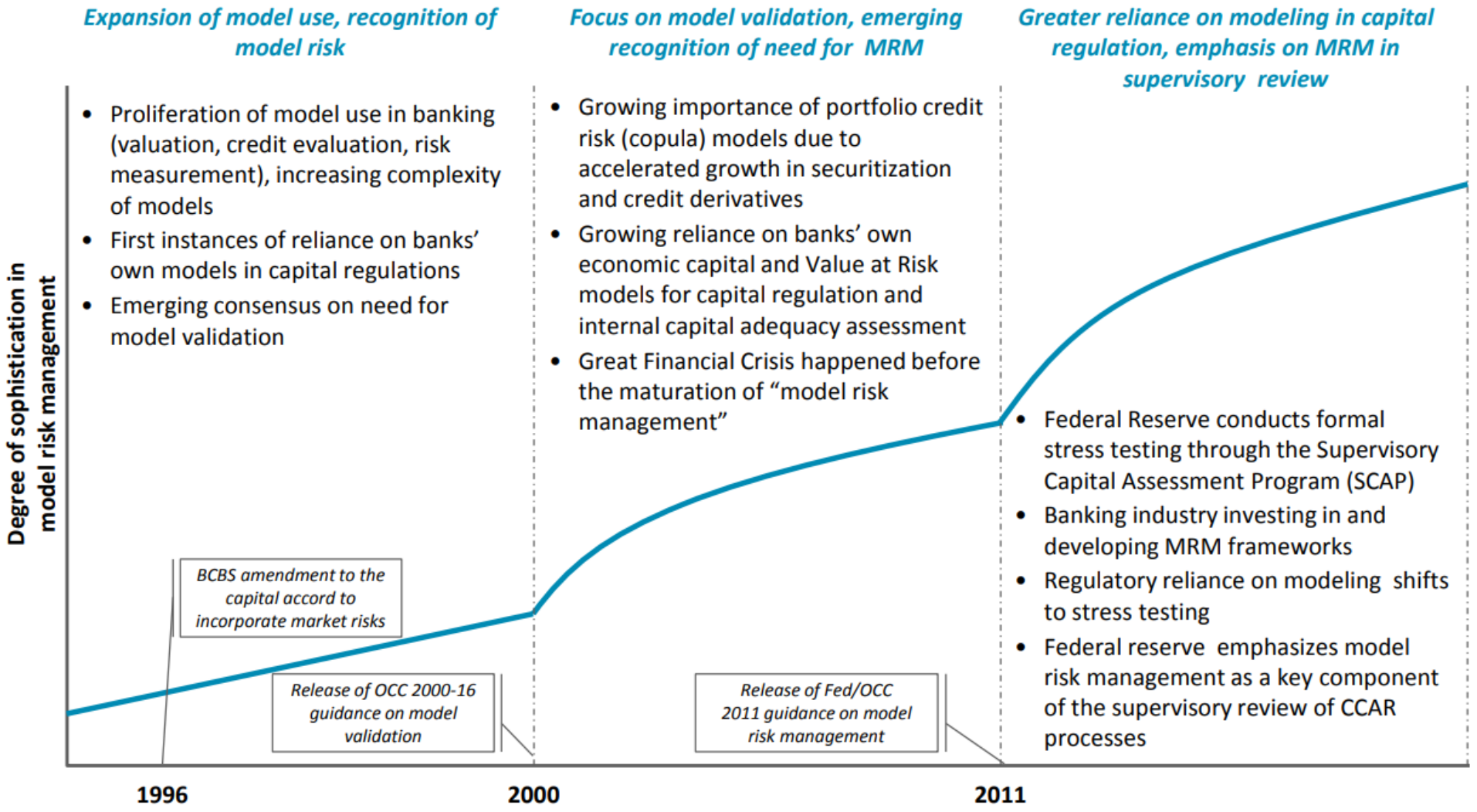
- **Next weeks class will be replaced by Tuesday (NOV 7TH) office hours, 3:30-4:30pm PDT.**

- **3 minutes to fill participation survey!!**

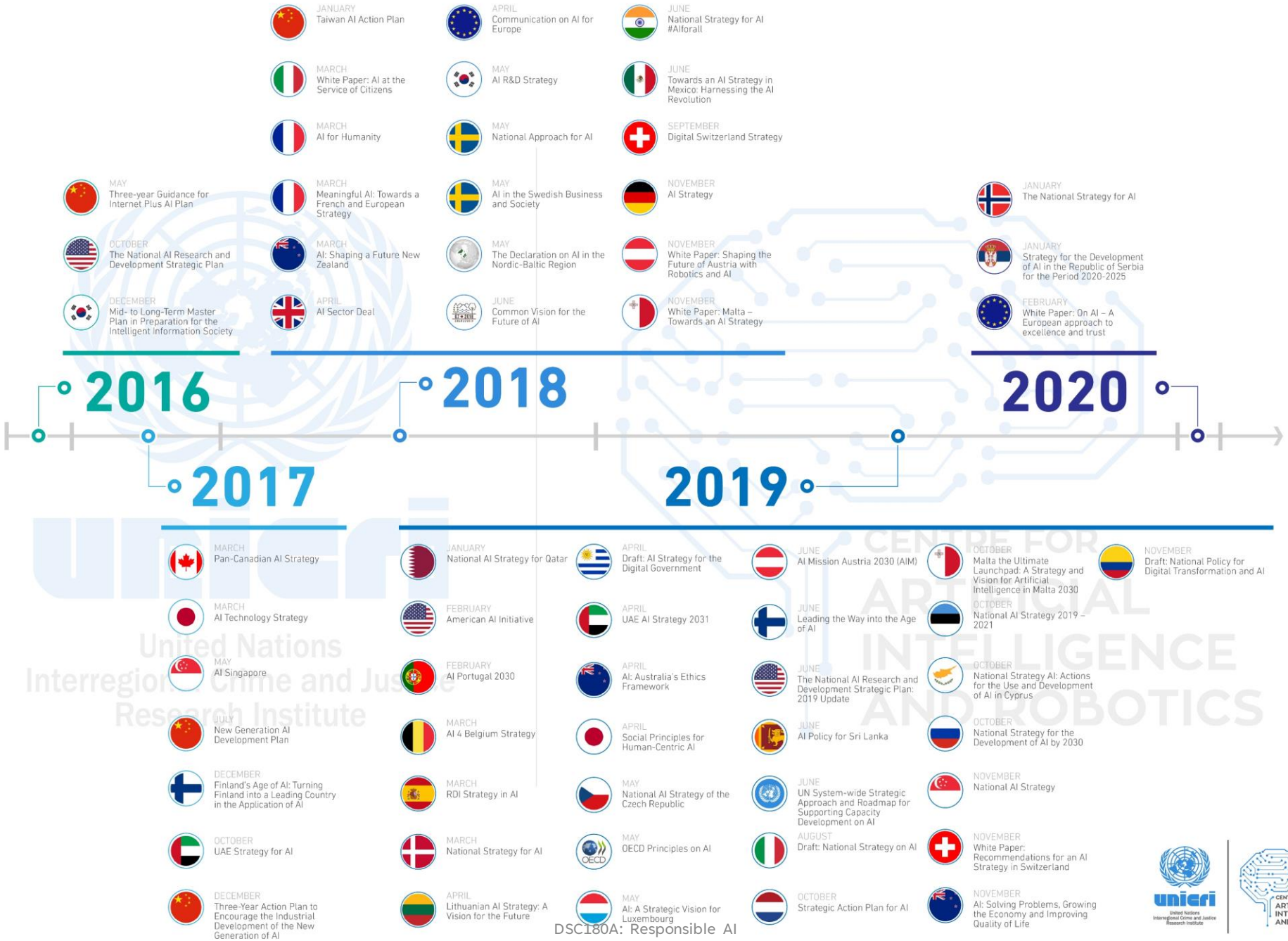
TODAY'S OBJECTIVES

- What precedents are there for AI regulation? Why does AI need special regulation?
- What does the impossibility theorem mean for designing laws around AI? Where is humanity today in terms of AI regulation? What will the regulatory landscape look like as you graduate into data science roles in the workforce?
    - It looks a lot different now than it did a year ago!!
- **How can regulations require AI to be fair and unbiased when those definitions aren't standardized or compatible?**

HISTORY OF REGULATIONS AROUND QUANTITATIVE MODELS



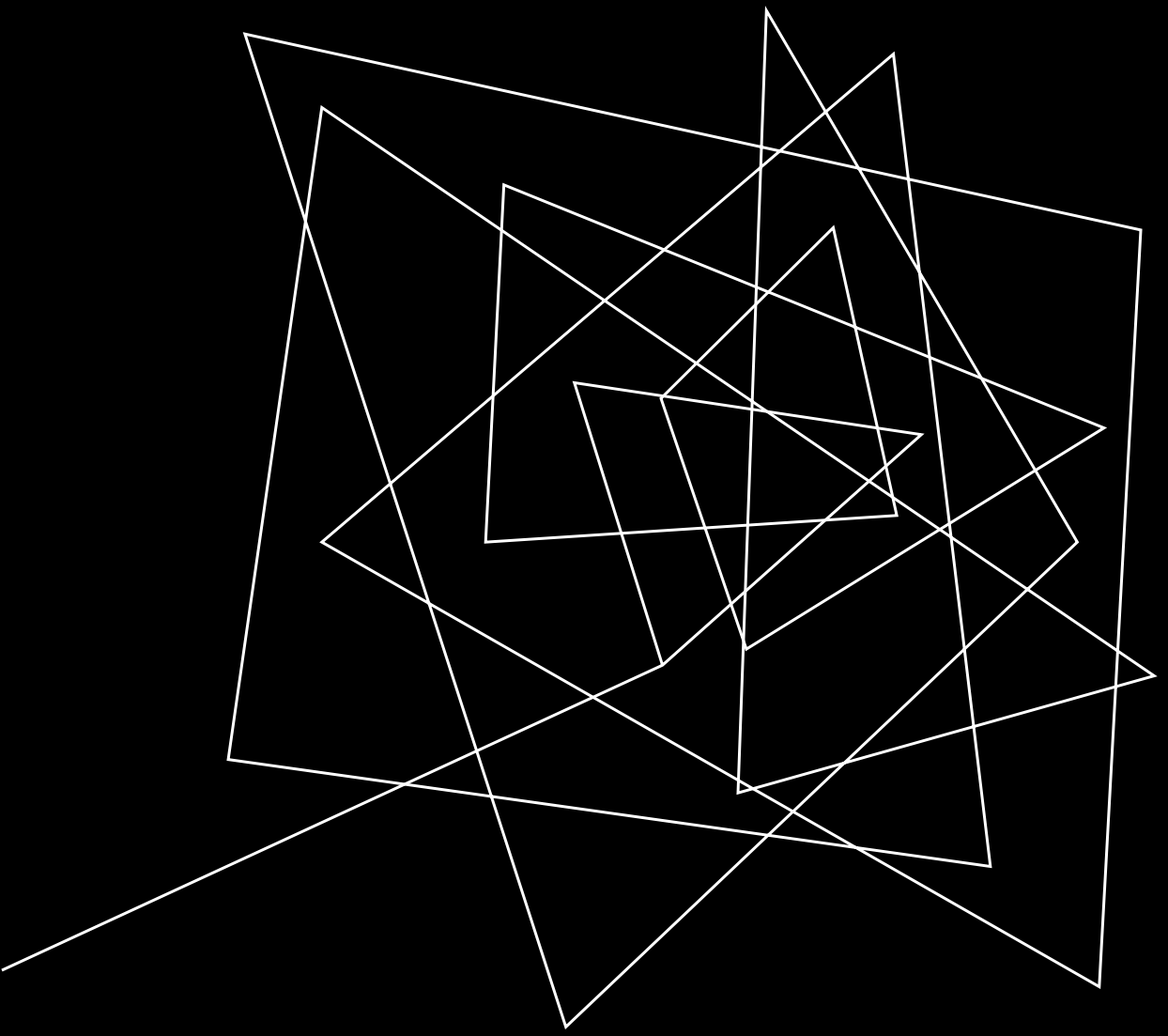
Source: https://wifpr.wharton.upenn.edu/wp-content/uploads/2021/07/15-01.pdf



https://en.wikipedia.org/wiki/Regulation_of_artificial_intelligence#/media/File:AI_Strategic_Documents_Timeline_UNICRI.jpg

WHY DO WE NEED AI-SPECIFIC REGULATION?

- Scale at which these models are being used
- Intersection and application across specialized fields
  - Cybersecurity and privacy
  - Medicine
  - Human capital / employment law
  - Content moderation
  - Housing
  - Ethics
    - Other considerations: the environment?
- Reliance on data
  - Bias within different stages of the AI model lifecycle
- Complexity
- Opacity
  - Lacking transparent-ness about AI / AI processes



READING PRESENTATION #1

Big Data's Disparate Impact (Barocas and Selbst)

Challenges & Limitations to AI Regulations

- **Impossibility Theorem:** How to choose and balance definitions of fairness?
- **(Lack of) Standardization:** There is no "source of truth" for AI Regulations – how will these regulations be consistently assessed and implemented?
- **False Sense of Security:** Will users "let their guard down" if they believe their AI is regulated?
  - Are individuals *actually* secure?
- **Legal Issues:** Bias Mitigation as Affirmative Action?
  - Equal-protection doctrine may threaten algorithmic fairness. Equal protection doctrine prevents government entities from treating similarly situated individuals differently ("anticlassification"). **This approach prohibits making decisions based on protected attributes** (i.e. hiring someone because of their race or gender). However, leading bias mitigation approaches often **purposefully use protected attributes to promote fair or unbiased outcomes in their algorithms**
  - There is a high level of legal scrutiny required to bypass anticlassification, which is why affirmative action cases are so frequently brought to the courts
  - The courts are likely to consider leading bias mitigation techniques a type of affirmative action, which subjects them to strict scrutiny (= harder to get passed) of the law and even potentially renders them illegal

Affirmative Algorithms: The Legal Grounds for Fairness as Awareness – The University of Chicago Law Review Online (uchicago.edu)

AI Policy Efforts in the US

FEBRUARY 11, 2019

Executive Order 13859: Maintaining American Leadership in Artificial Intelligence

Announced the American AI Initiative

*"The United States must foster **public trust and confidence in AI technologies** and protect civil liberties, privacy, and American values in their application in order to fully realize the potential of AI technologies for the American people."*

FEBRUARY 26, 2020

American AI Initiative: Year One Annual Report

Provided a summary of progress and long-term vision for the American AI Initiative, emphasizing the need to "***embrace trustworthy AI** for government services and missions."*

NOVEMBER 17, 2020

Office of Management and Budget (OMB) Memorandum M-21-06: Guidance for Regulation of Artificial Intelligence Applications

Provided guidance to Federal agencies, including considerations for reducing barriers to AI development and adoption

*"The government's regulatory and non-regulatory approaches to AI should contribute to public trust in AI by **promoting reliable, robust, and trustworthy AI applications**."*

DECEMBER 3, 2020

Executive Order 13960: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government

Outlined a set of principles and actions to accelerate trustworthy AI use and development

JANUARY 1, 2021

National AI Initiative Act of 2020

"*The mission of the National AI Initiative is to ensure continued U.S. leadership in AI research and development, **lead the world in the development and use of trustworthy AI** in the public and private sectors…*"

OCTOBER 4, 2022

Blueprint for the AI Bill of Rights

Outlines five protections Americans should have in the AI age:

1. Safe and Effective Systems
2. Algorithmic Discrimination Protection
3. Data Privacy
4. Notice and Explanation, and
5. Human Alternatives, Consideration, and Fallback

What would binding AI regulation look like?

About this Framework

The Blueprint for an AI Bill of Rights is a set of five principles and associated practices to help guide the design, use, and deployment of automated systems to protect the rights of the American public in the age of artificial intel-ligence. Developed through extensive consultation with the American public, these principles are a blueprint for building and deploying automated systems that are aligned with democratic values and protect civil rights, civil liberties, and privacy. The Blueprint for an AI Bill of Rights includes this Foreword, the five principles, notes on Applying the The Blueprint for an AI Bill of Rights, and a Technical Companion that gives concrete steps that can be taken by many kinds of organizations—from governments at all levels to companies of all sizes—to uphold these values. Experts from across the private sector, governments, and international consortia have published principles and frameworks to guide the responsible use of automated systems; this framework provides a national values statement and toolkit that is sector-agnostic to inform building these protections into policy, practice, or the technological design process. Where existing law or policy—such as sector-specific privacy laws and oversight requirements—do not already provide guidance, the Blueprint for an AI Bill of Rights should be used to inform policy decisions.

Safe and Effective Systems | Algorithmic Discrimination Protections | Data Privacy | Notice and Explanation | Human Alternatives, Consideration, and Fallback

https://www.whitehouse.gov/ostp/ai-bill-of-rights/

https://artificialintelligenceact.eu/the-act/

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence OCT 30, 2023

- The Federal Government will seek to promote responsible AI safety and security principles and actions with other nations, including our competitors, while leading key global conversations and collaborations to ensure that AI benefits the whole world, rather than exacerbating inequities, threatening human rights, and causing other harms.

- https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

ADDRESSING RISKS

01

Require that developers of the most powerful AI systems share their safety test results and other critical information with the U.S. government

02

Develop standards, tools, and tests to help ensure that AI systems are safe, secure, and trustworthy.

03

Protect against the risks of using AI to engineer dangerous biological materials

04

Protect Americans from AI-enabled fraud/deception by establishing standards and best practices for … authenticating official content.

05

Establish an advanced cybersecurity program to develop AI tools to find and fix vulnerabilities in critical software

06

Order the development of a National Security Memorandum that directs further actions on AI and security,

Importance of Adopting Trustworthy AI Practices

Sound AI practices can help companies ensure success with AI by protecting against four key risks...

Strategy and Reputation

Loss of public trust and loyalty due to lack of **transparency, fairness, and accountability**

Example:

Many social media users think they have been shadowbanned; lack of transparency over why account blocking or disabling decisions were made reduces trust between users and the service

Cyber and Privacy

Inadequate data protection and improper use of sensitive data can lead to **security and privacy breaches**

Example:

Meta had to pay $725 million to resolve a class-action lawsuit accusing the social media giant of allowing third parties, including Cambridge Analytica, to access users' personal information.

Legal and Regulatory

Unfair practices, compliance violations, or legal action due to **unfair / biased data** or a lack of **explainability**

Example:

Meta sued by Justice Department for discriminating advertising for housing; first case to challenge algorithmic discrimination under the Fair Housing Act

Operations

Without **robust and reliable** AI systems, operational inefficiencies due to inaccurate or inconsistent results

Example:

Inconsistent automated content moderation policies result in confusion among creators and additional work for employees managing platform policy compliance manually

Credit: Meira Gilbert, AI Ethics & Policy, Slide 11

Trustworthy AI | Regulatory Landscape for Human Capital

Federal labor laws and protections have been found in courts to extend to AI-based decisions…

… to protect candidates regardless of age, gender, race, religion, nationality, disabilities, etc. …

… throughout the sourcing, contacting, screening, and interview of potential candidates…

… all the way down to the state and local level.

Title VII of the Civil Rights Act of 1964 (Title VII)

- a federal law that protects employees and applicants against discrimination based on certain specified characteristics such as race, color, national origin, sex, and religion
- prohibits discrimination based on disparate treatment and/or disparate impact
- a court could find that an employer faces the same liability for a program exhibiting the unconscious bias of its programmer as it would if the programmer had made the hiring decision him or herself, based on that bias.

Age Discrimination in Employment Act (ADEA)

- forbids age discrimination against people who are age 40 or older. It does not protect workers under the age of 40, although some states have laws that protect younger workers from age discrimination

Americans with Disabilities Act (ADA)

- An A.I.-hiring practice could also implicate the Americans with Disabilities Act ("ADA") if an algorithm discerns an applicant's physical disability, mental health, or clinical diagnosis, all of which are forbidden inquiries in pre-employment candidate assessments. The ADA Amendments Act of 2008 broadened the statutory definition of "disability," increasing the scope of individuals whom the ADA protects.

Equal Employment Opportunity Commission (EEOC)

- enforces disability discrimination laws with respect to employers in the private sector and the federal government
- has issued guidance qualifying the expanded list of personality disorders identified in the psychiatric literature as protected mental impairments
- has already investigated at least two instances of alleged A.I. bias, and has made clear that employers using A.I. hiring practices could face liability for any unintended discrimination.
- in 2017, found reasonable cause to believe an employer violated the ADEA by advertising on Facebook for a position within its company and "limiting the audience for their advertisement to younger applicants."

Uniformed Services Employment and Reemployment Rights Act (USERRA)

- Maintains reemployment, antidiscrimination, and antiretaliation rights for members of uniformed services

State and Local Legislation

- New York City
  - Local law 144 of 2021 prohibits employers from using an automated employment decision tool unless such tool has been subject to a bias audit within one year of the use of the tool. (Laws goes live in Jan 2023).
- Illinois (Artificial Intelligence Video Interview Act)
  - imposes strict limitations on employers who use A.I. to analyze candidate video interviews. Employers must a) notify applicants that A.I. will be used in their interviews; b) obtain consent to use A.I.; c) explain to applicants how the A.I. works and what characteristics the A.I. will track in relation to the position; d) limit sharing of the video interview; and e) comply with requests to destroy the video within 30 days
- Maryland
  - the prohibition of facial/voice recognition software in interviews

DISCUSSION

Biden recently signed an executive order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence **into effect Oct 30, 2023.**

OCTOBER 30, 2023

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

BRIEFING ROOM • STATEMENTS AND RELEASES

Today, President Biden is issuing a landmark Executive Order to ensure that America leads the way in seizing the promise and managing the risks of artificial intelligence (AI). The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more.

Follow the link provided in chat and discuss these questions in your breakout rooms:

- **What is the goal of this order? Will things change?** Let's consider the issue Ben found in the EDA Code.
- **What are the strengths and weaknesses of this bill**? Does this go far enough? Why or why not?
- **What does challenges could this pose for different stakeholders** (think through the different stakeholders we discussed in previous lectures!)**?**

•"You should not face discrimination by algorithms and systems should be used and designed in an equitable way."

Please share your thoughts with the room when we reconvene!

FOR NEXT WEEK

- Complete next week's **readings**
  - If you signed up to present **Ethical Machine Learning in Health Care (Chen et al. 2021)** come prepared to present next week during office hours.

  - **CLASS IS ON TUESDAY, NOVEMBER 7TH**
    - **If you cannot make it, please let me know. It is an important class – so please try to be there.**

- Submit your answers to next week's participation questions on Gradescope by 10 AM PT, Thursday, November 9th
- Writeup #3 will be sent out and due by Thursday 10am as Participation

- **Replication Part #2:** Notebook and writeup on model development and fairness metric assessments
  - Primary contact for replication project: Emily Ramond (eramond@deloitte.com), Parker Addison (paddison@deloitte.com).
  - Emily's Tuesday 3:30-4:30pm

- First replication project checkpoint was yesterday: submit EDA notebook
  - Error in preprocessing